ANTHROPOLOGIE • XLVIII/1 • pp. 41–56 • 2010

ANNA DEGIOANNI, GERAUD GOURJON

LE MÉLANGE DANS LES POPULATIONS HUMAINES: MODÈLES ET MÉTHODES D'ESTIMATION

RÉSUMÉ: Tout au long de l'histoire, les populations humaines se sont déplacées et ont échangé leurs allèles, donnant naissance à des populations mélangées. Ces dernières possèdent des caractéristiques génétiques qui dépendent des apports génétiques des populations parentales. Depuis les premiers travaux de Bernstein, de très nombreux estimateurs de ces contributions (taux ou coefficients de mélange) ont été proposés. Les méthodes d'estimation peuvent être simples, corrélant selon une combinaison linéaire les fréquences alléliques observées dans les populations parentales à celles observées dans la population mélangée. Elles peuvent aussi être plus complexes, proposant une estimation par des approches bayésiennes ou de maximum de vraisemblance. Certaines de ces méthodes sont encore actuellement très utilisées car elles présentent un intérêt au niveau des paramètres pris en compte, ou car elles ont prouvé leur fiabilité. D'autres ont été abandonnées ou oubliées, malgré des avantages intéressants et une fiabilité notable pour certaines. Nous proposons ici une revue d'une quarantaine de méthodes d'estimation des taux de mélange et une réflexion sur leur utilisation.

MOTS-CLÉS: Mélange génétique humain – Méthodes d'estimation – Population mélangée – Hybride – Flux génique

ABSTRACT: Throughout their history, human populations have exchanged alleles, leading to new populations showing varying genetic characteristics depending on parental population contributions. Since the early works of Bernstein, a lot of estimators of parental population contributions (admixture rates or coefficients) have been published. These admixture estimation methods can correlate observed allele frequencies in parental and admixed populations by a simple linear combination, allowing an immediate estimation of admixture rates. They can be more complicated, suggesting an estimation of admixture coefficients through bayesian or full likelihood approaches, which can include some evolutionary forces. Some have been widely used since their models offer a real interest or because of their great reliability. Some others have been phased out or forgotten, despite some interesting characteristics and/or a rather good reliability. We propose here a review of about forty estimation methods of admixture rates, and few comments about their use.

KEY WORDS: Human genetic admixture – Estimation methods – Admixed population – Hybrid – Gene flow

INTRODUCTION

Fait social et culturel tout autant que biologique, le mélange entre les populations a façonné la dynamique évolutive de l'Homme au cours de son histoire. La question anthropologique du "mélange des populations" est complexe. Elle ne renvoie pas seulement à un phénomène de mouvement et d'échange, mais également à la définition même des populations, à leurs différences et à leurs relations socioculturelles. Les populations ne sont pas des entités fixes et immuables, elles sont le produit d'une évolution génétique sur de nombreuses générations. Cette évolution est rythmée par des facteurs et des comportements humains entremêlés tel que le choix marital, la fécondité,

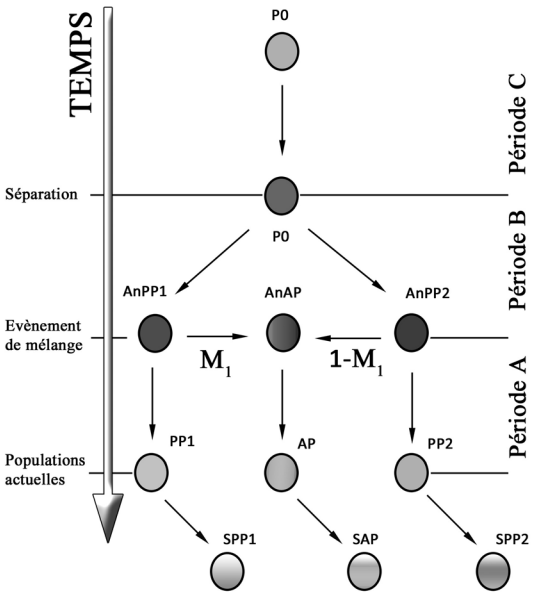

FIGURE 1. Modèle de mélange instantané. Instantaneous admixture model.

l'environnement socioculturel ou bien encore la mobilité géographique (autrement dit la migration).

La migration est un phénomène originel ayant joué dans la préhistoire et la protohistoire un rôle fondamental. Sous l'effet de cataclysmes, d'extensions glaciaires, d'inondations ou de sécheresses, les populations ont dû se déplacer. Morgan (1922) insistait sur le fait que "plus l'homme est primitif, moins il lui est aisé de se soustraire aux lois de la nature".

D'un point de vue génétique, la migration correspond à un mouvement de gamètes entre des populations d'une même espèce. Ces mouvements, induits par le déplacement des individus (flux de gènes), vont modifier les fréquences alléliques de la population qui les reçoit. La migration constitue en cela une pression évolutive majeure. Elle homogénéise les fréquences alléliques au sein des populations (Hartl 1994). Les populations migrantes ou receveuses (les populations parentales ancestrales) vont participer à la constitution d'une population nouvelle mélangée dont le pool génétique présente une distribution des fréquences alléliques intermédiaire entre celles des populations parentales (Bernstein 1931, Chakraborty 1986). Nous appelons ce phénomène le "mélange génétique". Les forces évolutives vont ensuite continuer à modifier les distributions des fréquences alléliques dans les populations, mélangée et parentales. La "population" à une génération donnée apparaît ainsi comme l'état temporaire d'un croisement à une période de l'histoire.

Nous parlons par contre de "métissage culturel", ou de "métissage" quand de nouveaux caractères culturels sont introduits dans une population "hôte". Ce métissage se fait unilatéralement ou bilatéralement, et à des degrés divers, la population intégrant au sein de sa propre culture un nombre plus ou moins importants d'éléments.

Les Comores (archipel de l'Océan Indien peuplé à partir du VII^{e} siècle par des populations bantoues, arabo-persiques, et asiatiques) expriment parfaitement ce désaccord entre les dynamiques de métissage culturel. La langue comorienne, tout en gardant une base et une structure grammaticale bantoue, a incorporé de nombreux éléments de vocabulaire arabe et perse (Vérin 1994), et aucune trace asiatique n'est perceptible (Lafon 1991). Du point de vue religieux, l'islam est devenu la religion officielle et unique, modifiant les institutions et le pouvoir politique sur l'archipel (Hrbek, El Fasi 1997).

Toute réalité historique, culturelle ou linguistique, ne traduit pas invariablement une réalité génétique et *vice-versa*. Les données historiques et archéologiques peuvent ainsi s'avérer fallacieuses lorsqu'il s'agit de mettre en évidence un mélange génétique. Les Comores en sont encore un bon exemple: les hommes d'origine asiatique n'ont laissé que peu de traces génétiques de leur contribution (Chiaroni *et al.* 2004, Msaïdie *et al*. 2010) tandis que leurs apports culturels sont notables: la culture du riz ou l'utilisation de la pirogue à balancier par exemple (Vérin 1994). D'un point de vue génétique, il n'existe pas non plus d'uniformité, chaque type de marqueur (ADN mitochondrial, chromosome Y, et marqueurs autosomaux) étant soumis à une dynamique particulière de mélange (Gourjon *et al.* 2010).

MODÈLES ET ESTIMATION DU MÉLANGE GÉNÉTIQUE

Le protocole d'étude

L'étude du mélange génétique implique l'utilisation de modèles pour le définir. Il existe une réalité implexe au mélange, réalité que ces modèles biomathématiques essaient d'approcher. L'ensemble des paramètres ne peuvent toutefois être pris en compte dans le modèle, nombre d'entre eux n'étant pas connu et/ou estimable. En conséquence, il faut au préalable définir la dynamique de l'évènement de mélange puis identifier les forces évolutives qui vont s'exercer sur les populations et qui vont modifier leurs fréquences alléliques. Il faut rechercher les éventuels facteurs exogènes (comme les facteurs sociétaux) qui peuvent avoir eu une influence sur la dynamique de l'évènement pour essayer de les inclure dans les modèles. Il est peu probable d'arriver à une superposition exacte de la réalité biologique et du modèle statistique. La finalité du protocole d'étude sera d'essayer de superposer le plus d'éléments, spécifiquement ceux ayant le plus d'influence sur les procédures d'estimation. La précision des estimations dépendra de la minimisation de l'écart entre théorie et réalité.

Les modèles classiques

Le terme de mélange peut se référer classiquement à deux processus distincts: le modèle de mélange instantané (MI) et le modèle de flux génique continu (FGC). Ces deux modèles de base se déclinent en de nombreuses versions, et peuvent

présenter dans leur dynamique diverses caractéristiques propres. Il est par exemple possible que les hommes et les femmes d'une population contribuent différemment au mélange (biais sexuel) ou même qu'un seul des deux sexes y participe (mélange sexuellement spécifique). Le modèle de flux génique continu est le plus fréquemment observé dans les populations humaines.

Le modèle de MI est le plus simple et le plus intuitif (*Figure 1*). Il est également nommé "intermixture" par Long (1991) ou "hybrid-isolation" par Pfaff *et al.* (2001). Il considère une population ancestrale P_0 qui évolue pendant un certain temps (Période C) avant de se scinder en deux (ou plus) populations parentales ancestrales, $AnPP_1$ et $AnPP_2$. Les populations parentales ancestrales évoluent indépendamment et se différencient pendant X générations (Période B) avant de se mélanger à la génération G_0 durant une seule génération. Elles contribuent respectivement en proportion de gènes M_1 (contribution de la $AnPP_1$) et $1–M_1$ (contribution de $AnPP_2$) pour former la population mélangée ancestrale, AnAP (avec $M_1 \in [0,1]$). A la génération G_1, les populations $AnPP_1$, $AnPP_2$, et AnAP, se séparent et évoluent indépendamment pendant Y générations (Période A) pour aboutir aux populations parentales actuelles PP_1, PP_2, et à la population mélangée actuelle, AP, dans lesquelles l'échantillonnage est effectué: SPP_1 (échantillon de la population parentale 1), SPP_2 (échantillon de la population parentale 2), et SAP (échantillon de la population mélangée) (d'après Bertorelle, Excoffier 1998).

Le modèle de FGC (*Figure 2*) considère qu'après l'événement de mélange à la génération G_0, au moins une des populations parentales ancestrales continue à être en contact avec la population mélangée ancestrale pendant j générations (Période A), contribuant à chaque génération au pool génétique de la population mélangée ancestrale à un taux M_{11} à la génération 1, M_{12} à la génération 2, …, M_{1j} à la génération j La valeur de M_{1j} n'est pas fixe et peut varier à chaque génération. La contribution finale des populations parentales ancestrales est calculée sur l'ensemble des j générations.

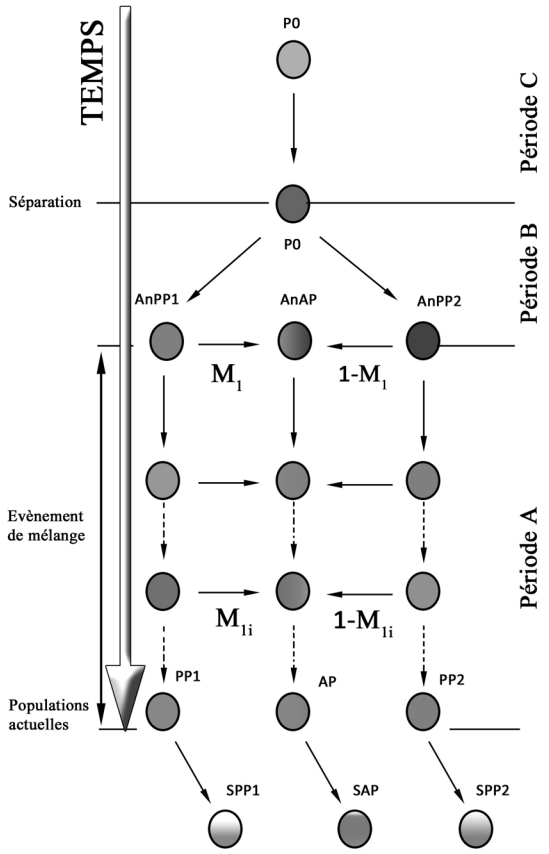

FIGURE 2. Modèle de flux génique continu. Continuous gene flow model.

L'estimation des contributions parentales

Chaque AnPP participe à un certain degré au mélange génétique, degré proportionnel à sa population efficace N_e et aux autres facteurs évolutifs. L'étude de l'origine d'une population mélangée actuelle (AP) passe par l'estimation de manière plus ou moins précise de la contribution de chaque AnPP au pool génétique de la population mélangée ancestrale (AnAP). Depuis un siècle, de nombreuses méthodes théoriques permettant d'évaluer ces contributions relatives (ou taux ou coefficients de mélange) ont été publiées. Nous notons ces taux de mélange M_k pour le taux de contribution des k AnPP ou m_k pour le taux de contribution estimé dans les SPP. L'estimateur/estimation général(e) est noté(e) M*.

Les premières méthodes considéraient les populations parentales actuelles (PP) comme étant similaires aux AnPP. Il est rare qu'un mélange génétique s'effectue sur un laps de temps d'une seule génération (génération G_0 à G_1) et les flux géniques qui suivent cet événement, de même que les forces évolutives, peuvent avoir modifié les fréquences alléliques des différentes populations. En conséquence, plus le nombre de générations G_i séparant l'événement de mélange et l'échantillonnage est important, plus les fréquences alléliques actuelles diffèrent des fréquences ancestrales.

Pour une estimation sans erreur, la situation idéale est celle où aucune des populations parentales n'a d'allèles en commun avec les autres et où l'échantillonnage est effectué à la génération suivant le mélange génétique (G_1). Chaque allèle dans la population mélangée est directement issu de la population parentale le possédant et le coefficient de mélange est obtenu en comptant le nombre d'allèles à assigner à chaque population parentale. Cette idée est reprise dans la méthode des "allèles uniques" par Chakraborty *et al.* (1991), se basant sur la publication de Neel (1973). Ce cas est excessivement rare dans la nature, et même inexistant dans les populations humaines. De plus, l'échantillonnage est non exhaustif. Pour déterminer les taux de mélange, il faut donc considérer plusieurs points.

Tout d'abord, il s'agit de sélectionner des marqueurs suffisamment différenciés dans les PP (en terme de fréquences alléliques et, pour certaines méthodes, en terme de divergence moléculaire) et donc suffisamment informatifs. Ensuite, le deuxième élément essentiel est le

choix de PP les plus représentatives des AnPP (voir Gourjon 2010, pour une discussion au sujet du choix des AnPP selon le modèle de mélange). Ce choix est très souvent restreint par la disponibilité des données. Idéalement, les données devraient porter sur les mêmes marqueurs pour toutes les populations impliquées, et devraient de plus être obtenues suivant le même protocole d'échantillonnage, et d'extraction/séquençage. En théorie, il faudrait échantillonner pour chaque étude les populations parentales et mélangée, de manière à avoir un jeu de données représentatif qui lui est propre et qui correspondrait bien au cadre précis de l'étude. Dans la pratique, l'échantillonnage est presque toujours effectué uniquement dans la AP, et les données sur les PP sont issues de la littérature.

Ces éléments pris en compte, il s'agit d'estimer les contributions des AnPP à la AnAP. Depuis les premiers travaux de Bernstein (1931), de nombreux estimateurs ont été proposés. Parmi toutes les méthodes disponibles, il faut choisir la mieux adaptée au cadre de l'étude et également aux données. Ce choix doit reposer sur les caractéristiques techniques de la méthode: (i) prise en compte de telle ou telle force évolutive et/ou de l'erreur due au biais d'échantillonnage, (ii) utilisation des fréquences alléliques dans les procédures d'estimation ou des fréquences phénotypiques, (iii) adaptabilité aux données moléculaires (comme les méthodes basées sur la théorie de la coalescence), ou bien moins spécifique, (iv) incorporation de modèle de mélanges dihybride, trihybride, ou impliquant un nombre quelconque de AnPP (multihybride), (v) implémentation dans un logiciel ou non, etc. Généralement, les méthodes implémentées dans des logiciels sont les plus utilisées car elles ne nécessitent pas de la part de l'utilisateur des connaissances mathématiques solides et elles réduisent les temps de calcul.

Une revue des méthodes

Les méthodes d'estimation du mélange génétique sont utilisées depuis des décennies, et chaque publication les utilisant en donne une description succincte. Ces descriptions rapides ont pour but de fournir les grandes lignes méthodologiques nécessaires à l'appréhension et l'interprétation des résultats sans avoir à se référer à la publication originale de la méthode, en insistant souvent sur les détails méthodologiques importants pour l'étude considérée. Cependant, ces descriptions sont parfois imprécises. Par exemple dans la publication de Mc Evoy *et al.* (2006), nous retrouvons une description de la méthode de régression linéaire de Chakraborty *et al.* (1992) comme étant une version similaire à celle de Long (1991). La méthode de Chakraborty est en réalité une version simplifiée de cette dernière, puisqu'elle ne prend en compte ni la dérive génétique, ni l'erreur due au biais d'échantillonnage, et n'inclue que deux populations parentales dans le modèle. Les différences entre les deux méthodes sont importantes et ont une influence notable sur les résultats.

Il nous semblait donc essentiel de fournir en premier lieu une revue des méthodes existantes, et, en particulier, des méthodes utilisées dans les publications actuelles, avec une description reprenant les publications originales. Il était également intéressant de regrouper ensemble toutes ces méthodes d'estimation qui sont actuellement dispersées dans les publications et parfois méconnues. Leur nombre est très important (plusieurs dizaines) et certaines méthodes impliquent des formulations mathématiques complexes. Nous avons, pour cette raison, décidé de ne fournir les formulations mathématiques que pour les méthodes "simples". Pour les méthodes plus complexes, nous rappelons seulement les procédures de base d'estimation (régression linéaire multiple, moindres carrés pondérés, coalescence, approche bayésienne, etc.).

LES MÈTHODES D'ESTIMATION DU MÉLANGE POPULATIONNEL

Les méthodes de régression linéaire

Bernstein (1931) part du principe que si les fréquences alléliques d'un locus sont connues pour les deux AnPP, il est possible de prédire les fréquences alléliques dans la AnAP en considérant la proportion de l'apport de chaque AnPP (modèle de MI). Inversement, si les fréquences dans les AnPP et AnAP sont connues, il est possible de déterminer les proportions de l'apport de chacune des AnPP.

Le modèle de base de Bernstein (1931) et de Cavalli-Sforza, Bodmer (1971)

Soit un locus unique et diallélique avec une codominance des allèles pour une simplification algébrique (il est également possible de supposer une dominance, mais des modifications sont alors nécessaires dans la formule). Supposons que P_1, P_2, P_3, …, P_k, soient les fréquences de ces allèles dans k AnPP, P_H la fréquence dans la AnAP. Ces AnPP contribuent au mélange en proportions M_1, M_2, M_3, …, M_k, respectivement, avec $M_i \in [0,1]$ et $\Sigma M_i = 1$, M_i étant le taux de mélange de la $i^{\text{ème}}$ AnPP.

Sous un modèle de MI:

$$P_H = \sum_i^k M_i P_i \,.$$

Pour $k = 2$, la formule de Bernstein se résume par:

$$P_H = M_1 P_1 + (1 - M_1) P_2 \,. \qquad (1)$$

En considérant que les fréquences alléliques sont connues sans erreur, nous obtenons:

$$M_1 = \frac{p_H - p_2}{p_1 - p_2} \,. \qquad (2)$$

Cette relation n'est valable que si les fréquences alléliques des échantillons (SPP et SAP) sont les fréquences exactes dans l'ensemble de la population échantillonnée (PP et AP) qui sont elles-mêmes les fréquences exactes de celles des populations ancestrales (AnPP et AnAP).

Notons p_1, p_2, p_3, …, p_k, les fréquences de ces allèles dans les SPP des k AnPP, m_i la contribution de la $i^{\text{ième}}$ SPP. Nous remplaçons P par p dans l'équation (2):

$$m_1 = \frac{p_H - p_2}{p_1 - p_2}. \quad (3)$$

Dont la variance peut être estimée par:

$$Var(m_1) \approx \frac{Var(p_H) + m_1^2 Var(p_1) + (1-m_1)^2 Var(p_2)}{(p_1 - p_2)^2}. \quad (4)$$

Avec $Var(p_i) = \left(p_i(1-p_i)\right)/n_i$ étant la variance de la fréquence dans l'échantillon de la population i et n_i le nombre d'individus échantillonnés dans la population i. Si plusieurs échantillons de populations sont disponibles, Cavalli-Sforza et Bodmer (1971) proposent d'insérer la variance inter-échantillons dans l'équation (4).

A partir de cette relation linéaire simple, les taux de mélange sont calculés pour chaque locus séparément. Pour obtenir une estimation unique à partir de plusieurs loci diallélіques, Cavalli-Sforza et Bodmer (1971) pondèrent la valeur de m_1 à chaque locus par l'inverse de sa variance. La valeur ainsi obtenue $\bar{m}_1$ peut être utilisée comme estimation de m_1. Considérons q loci, et m_{1i} l'estimation de m_1 au $i^{\text{ème}}$ locus.

$$\bar{m}_1 = \frac{\sum_i^q \left(\frac{m_{1i}}{Var(m_{1i})} \right)}{\sum_i^q \left(\frac{1}{Var(m_{1i})} \right)}.$$

Avec une variance

$$Var(\bar{m}_1) = \frac{1}{\sum_i^q \left(\frac{1}{Var(m_{1i})} \right)}.$$

La méthode de Bernstein et ses variations ont été très utilisées jusque dans les années 50–60 (et par la suite notamment dans les approches basées sur les distances génétiques ou la couleur de la peau). Il faut attendre Roberts et Hiorns (1965) et Elston (1971) pour que des approches multiloci, multialléliques soient présentées.

La méthode de Glass et Li (1953): G&L53

Glass et Li introduisent le premier modèle de mélange sous forme de flux génique (FGC). Pour cela, ils reprennent le principe de migration (une population donneuse et une population receveuse) pour déterminer le taux de mélange à la première génération de mélange, suivie d'un flux unidirectionnel des gènes de la population donneuse d'origine européenne vers la population receveuse d'origine africaine. A la génération G_1, la fréquence p' d'un allèle A dans la population nouvellement mélangée est:

$$p' = (1-m)p + mp_e.$$

Avec p la fréquence de l'allèle A dans la population africaine avant le mélange (G_0), p_s la fréquence du même allèle dans la population européenne, et m la contribution de la population européenne au pool génétique de la AnAP à la génération 1.

La valeur de m est calculée de la même façon:

$$m = \frac{p_j - p_s}{p - p_s}.$$

Glass et Li posent l'hypothèse d'une migration (flux génétique) constante à chaque génération pour la simplification des calculs. Si le processus de mélange dure j générations, à la génération j, la fréquence de l'allèle A dans la population mélangée est:

$$p_j = (1-m)\,p_{j-1} + p_s.$$

Après substitution des termes pour chaque génération, simplification et réarrangement de la formule, l'estimation du taux de migration s'écrit:

$$(1-m)^j = \frac{p_j - p_s}{p - p_s}.$$

La méthode de Roberts et Hiorns (1962, 1965): R&H65

Roberts et Hiorns proposent des modèles similaires et avec une possibilité de mélange bidirectionnel et introduisent une première généralisation du modèle avec un modèle de mélange impliquant plus de deux AnPP.

Modèle de deux AnPP et générations discrètes

Pour ce premier modèle, seules deux AnPP sont impliquées dans le mélange, et les générations sont discrètes (non chevauchantes). Le mélange est panmictique, et aucune force évolutive n'intervient pour modifier les fréquences alléliques.

Soit p_{1j} et p_{21j} la fréquence de l'allèle considéré d'un locus dans la $AnPP_1$ et la $AnPP_2$, respectivement, après j générations. Soit m_{1j} le flux génique de la $AnPP_1$ vers la $AnPP_2$, et m_{21j} le flux génique de la $AnPP_2$ vers la $AnPP_1$.

Roberts et Hiorns posent la même équation et après j générations:

$$p_{1j} = (1-m_{2j})p_{1(j-1)} + m_{2j}p_{2(j-1)},$$

$$p_{2j} = (1-m_{1j})p_{2(j-1)} + m_{1j}p_{1(j-1)}.$$

Avec: $\Delta_j = p_{1j} - p_{2j}$ et: $\Sigma_j = p_{1j} + p_{2j}$.

Ils obtiennent après récurrence et substitutions, et pour des taux constants de flux génique, les fréquences:

$$p_{1j} = \frac{1}{2}\Big[\Sigma_0 + (m_1 - m_2)\sum (1-m_1-m_2)^{k-1}$$
$$\Delta_0 + (1-m_1-m_2)^j \Delta_0\Big],$$

$$p_{2j} = \frac{1}{2}\Big[\Sigma_0 + (m_1 - m_2)\sum (1-m_1-m_2)^{k-1}$$
$$\Delta_0 - (1-m_1-m_2)^j \Delta_0\Big].$$

Avec $k \neq j$ et $k = 1, 2, \ldots, j$.
Si le flux est unidirectionnel, ils obtiennent:

$$p_{1j} = p_{10},$$
$$p_{2j} = p_{10}(1-m)^j + \left[1-(1-m)^j\right] p_{10}.$$

Modèle de deux AnPP et migration continue
Ce deuxième modèle introduit une migration continue et les générations peuvent être chevauchantes. Ils proposent de remplacer les générations par un laps de temps court δt et de déterminer les taux de mélange à un instant t en considérant les taux de migration pendant la période $(t; t+\delta t)$.

Modèle de plus de deux AnPP et migration constante
Utilisant le modèle précédent, Roberts et Hiorns suggèrent l'utilisation des moindres carrés pour résoudre l'estimation des taux de mélange pour plus de deux AnPP. Notons P_{ij} la fréquence du $i^{\text{ème}}$ allèle $(i = 1, 2, 3, \ldots, k)$ d'un locus codominant dans la $j^{\text{ème}}$ population $(j = 1, 2, 3, \ldots, k)$:

$$P = (P_1, P_2, \ldots, P_n) = \left(\left(P_{ij}\right)\right)_{k,n}.$$

Notons $P_{H,i} = \left(P_{H,1}, P_{H,2}, \ldots, P_{H,k}\right)$ la fréquence du $i^{\text{ème}}$ allèle dans l'AnAP, obtenue par le mélange de n AnPP aux taux $M_{i,j} = \left(M_{i,1}, M_{i,2}, \ldots, M_{i,n}\right)$. Si les P_{ij} sont connues sans erreur (aucun biais d'échantillonnage, aucune influence des forces évolutives):

$$E\left(P_{H,i}\right) = P_{i,j} \cdot M_{i,j}.$$

Les MCO (Moindres Carrés Ordinaires) permettent d'obtenir une estimation m de M:

$$m = (X'X)^{-1} X'y.$$

Avec X la matrice de k sur n dimensions: $X = P_{i,j}$; et y un vecteur de dimension k représentant les fréquences alléliques dans la AnAP: $y = P_{H,i}$.

La somme des fréquences à un locus donné pour une AnPP donnée devant être égale à 1, $X'X$ est une matrice singulière (non inversible) et il n'est pas possible de calculer $(X'X)^{-1}$. Roberts et Hiorns proposent pour cela d'éliminer arbitrairement les données pour un allèle afin que $X'X$ devienne non singulière. Le choix de l'allèle peut avoir un impact sur l'estimation.

La méthode des moindres carrés généralisés de Elston (1971): R&H-ELS71
Elston reprend la méthode de Roberts et Hiorns et présente une approche par les Moindres Carrés Généralisés (MCG). Considérant que les fréquences alléliques dans toutes les populations sont connues sans erreur, il pose:

$$m = \left(X'V^{-1}X\right)^{-1} X'^{V^{-1}} y.$$

Avec V étant la matrice de variance-covariance de y. Toutefois, V est toujours singulière si tous les allèles sont inclus, mais le choix de l'allèle à enlever n'a plus d'influence sur les résultats. Il propose une deuxième formule, avec la contrainte que $\sum_1^n m_{i,j} = 1$ et tous les $m_{i,j} > 0$:

$$m^* = (X^{*\prime} X^*)^{-1} X^{*\prime} y^*,$$

$$m_n = 1 - \sum_{j=1}^{n-1} m_j.$$

Avec $y^* = y - P_n$ et X^* une matrice de k sur $n-1$ dimensions pour laquelle la $j^{\text{ème}}$ colonne est $P_j - P_n$ et m^* est un vecteur de dimension $n-1$ représentant les contributions des $n-1$ premières AnPP à la AnAP. m_n est la contribution de la $n^{\text{ème}}$ AnPP.

La méthode d'identité des gènes de (Chakraborty 1975, 1985): CHAK85
La méthode de Chakraborty (1975, 1985) se base sur les coefficients d'identité moyenne (Nei 1972, 1973) à l'intérieur et entre les populations. Notons p_i et q_i les fréquences du $i^{\text{ème}}$ allèle dans les AnPP X et Y respectivement et x_i et y_i les fréquences dans les échantillons correspondants. Les distances génétiques de Nei sont définies par:

$$D = -\ln\left(\frac{G_{XY}}{\sqrt{G_X G_Y}}\right).$$

Avec G_X, G_Y et G_{XY} étant, respectivement, les moyennes de Σp_i^2, Σq_i^2, $\Sigma p_i q_i$ sur l'ensemble des loci. Pour calculer D, on remplace G_X, G_Y et G_{XY} par les identités géniques des échantillons, respectivement J_X, J_Y et J_{XY} qui sont les moyennes de Σx_i^2, Σy_i^2, $\Sigma x_i y_i$ sur les r loci utilisés dans l'estimation.

Chakraborty utilise la même relation pour définir J_{11}, J_{12} et J_{1H} comme étant les moyennes arithmétiques entre les loci des probabilités d'identité des gènes respectivement à l'intérieur de la population AnPP_1, entre les populations AnPP_1 et AnPP_2 et entre les populations AnPP_1 et AnAP.

Pour un mélange instantané, il déduit de la formule de Bernstein:

$$J_{1H} = mJ_{11} + (1-m)J_{12}.$$

Dans sa publication de 1985, il introduit une autre approche pour calculer m pour un nombre quelconque de loci, et d'allèles et un nombre p de AnPP:

$$E\left[J_{iH} - J_{ip}\right] = \sum_{j=1}^{p-1} m_j \left(J_{ij} - J_{ip}\right).$$

La résolution se fait par les moindres carrés pour obtenir la valeur de m correspondant au coefficient de mélange.

Chakraborty (1975) présente également un modèle de mélange continu, avec une migration considérée comme constante à chaque génération.

La méthode des moindres carrés pondérés de Long (1991): LONG91

Cette méthode permet l'estimation simultanée des proportions de mélange et du F_{st} sous un modèle de MI. Elle permet une partition de la variance des taux de mélange en deux composants: le premier mesure l'erreur inhérente à l'échantillonnage dans les AnPP et AnAP (l'erreur due au biais d'échantillonnage) tandis que le deuxième mesure l'erreur qui s'accumule à cause des changements évolutifs dans la AnAP uniquement. En l'absence de sélection naturelle, ce dernier composant est une mesure de la dérive génétique (Long 1991).

Considérons un allèle codominant pour chacun des loci non lié. Si le mélange et la dérive génétique sont les seuls processus évolutifs qui affectent le pool génique de la population mélangée, alors:

$$p_{ih} = mp_{i(1)} + (1-m)\,p_{i(2)} + \varepsilon_{si}\,.$$

Avec au locus $i\left(i=1,2,\ldots,l\right)$, p_{ih}, $p_{i(1)}$, $p_{i(2)}$, respectivement les fréquences de l'allèle *A* dans les AnAP, $AnPP_1$, et $AnPP_2$, *m* la contribution de la $AnPP_1$, $(1-m)$ la contribution de la $AnPP_2$, et ε_{si} l'erreur due à la dérive génétique.

$\tilde{p}_{ih}$ est l'estimation par maximum de vraisemblance de la fréquence allélique de la population mélangée basée sur l'échantillon de la population. L'erreur due au biais d'échantillonnage est incluse et son espérance est égale à 0. Néanmoins, sa variance est différente de 0 et elle est pondérée par la taille des SPP/SAP. En augmentant la taille des échantillons on se rapproche des fréquences dans les AnPP/AnAP, et l'effet de l'échantillonnage a moins d'influence sur les estimations. Les erreurs dues au biais d'échantillonnage et les erreurs dues à la dérive génétique sont ainsi prises en considération et envisagées comme indépendantes. Si $p_{i(1)}$ et $p_{i(2)}$ sont des paramètres connus, la méthode des moindres carrés pondérées est utilisée pour estimer *m* à partir de l'équation reformulée:

$$\left(\tilde{p}_{ih} - p_{i(2)}\right) = m\left(p_{i(1)} - p_{i(2)}\right) + \varepsilon_i\,.$$

Le terme ε_i est égal à la somme des deux composantes d'erreurs. Le taux de mélange *m* est estimé en définissant deux vecteurs et une matrice diagonale de variance-covariance. La solution est obtenue par itération et à chaque étape les espérances des fréquences alléliques de la AnAP sont approximées. L'avantage de l'implantation de cette méthode dans un logiciel est de permettre l'exécution d'un nombre d'itérations important, l'estimation pouvant alors être calculée et recalculée, en l'améliorant à chaque fois. Un intervalle de confiance est ensuite construit en utilisant le EQM (Erreur Quadratique Moyenne = "Mean Squared Error" MSE), qui est la variance standard des fréquences alléliques de la AnAP. Le EQM est aussi utilisé pour estimer le F_{st} qui dans le modèle est une mesure directe de la dérive génétique. Cette quantité est donc très informative sur la structure de la descendance de la population mélangée.

La méthode des moindres carrés pondérés modifiée de Chakraborty et al. (1992): CHAK92

Chakraborty *et al*. (1992) proposent une formulation simplifiée de LONG91, ne prenant en compte ni la dérive, ni le biais d'échantillonnage et n'incluant que deux AnPP dans le modèle de mélange. Soient *L* loci $\left(l=1,2,\ldots,L\right)$ et r_l allèles au locus *l* $\left(i=1,2,\ldots,r_l \text{ pour un locus } l \text{ donné}\right)$ et les fréquences p_{Hil}, p_{1il}, p_{21il}, du $i^{\text{ème}}$ allèle du $l^{\text{ème}}$ locus, pour les AnAP, $AnPP_1$, et $AnPP_2$, respectivement.

Le taux de mélange est:

$$m = \frac{\sum_{l=1}^{L}\sum_{i=1}^{r_l}\frac{\left(p_{2il}-p_{1il}\right)\left(p_{Hil}-p_{1il}\right)}{p_{Hil}}}{\sum_{l=1}^{L}\sum_{i=1}^{r_l}\frac{\left(p_{2il}-p_{1il}\right)^2}{p_{Hil}}}\,.$$

Et la variance de *m*:

$$Var(m) = \frac{EQM}{\sum_{l=1}^{L}\sum_{i=1}^{r_l}\frac{\left(p_{2il}-p_{1il}\right)^2}{p_{Hil}}}\,.$$

$$\text{Avec } EQM = \frac{\sum_{l=1}^{L}\sum_{i=1}^{r_l}\frac{\left[\left(p_{Hil}-p_{1il}\right)\left(mp_{2il}-p_{1il}\right)\right]^2}{p_{Hil}}}{r-L}\,.$$

Avec *r* étant la somme de tous les allèles de tous les loci.

Les méthodes de vraisemblance

La méthode du maximum de vraisemblance de Krieger et al. (1965): KR65

La vraisemblance des génotypes des individus de la SAP est exprimée comme une fonction des fréquences alléliques des AnPP et des taux de mélange. Sous un modèle de MI et en considérant deux allèles *i* et *j* d'un locus donné *k* d'un individu provenant d'un échantillon de la AnAP, la fréquence des allèles dans la AnAP est:

$$p_{ih} = mp_{i(1)} + (1-m)\,p_{i(2)}\,,$$
$$p_{jh} = mp_{j(1)} + (1-m)\,p_{j(2)}\,.$$

En supposant l'équilibre d'Hardy-Weinberg, la probabilité des génotypes est égale à $\left(p_{ih}\right)^2$ et $\left(p_{jh}\right)^2$ si le locus est homozygote respectivement pour l'allèle *i* ou *j*, et elle est égale à $2\left(p_{ih}p_{jh}\right)$ s'il est hétérozygote. En considérant que les loci sont indépendants, la probabilité pour un génotype multiloci est le produit des probabilités à chaque locus et la probabilité d'observer les fréquences de la SAP totale est obtenue en multipliant les probabilités entre les individus. Il est ensuite possible de déterminer la valeur de *m* qui permet de maximiser *L*.

La modification de KR65 par Elston (1971): KR-ELS71

Elston propose une amélioration de la méthode KR65,

toujours sous un modèle de MI et sans biais d'échantillonnage ni forces évolutives. Pour résoudre les équations de l'estimation par maximum de vraisemblance (EMV), il introduit l'utilisation d'une technique itérative basée sur la méthode de Newton-Raphson. Si m_1 est une approximation de m (m étant déjà le taux de mélange estimé depuis l'échantillon, donc une approximation de M, taux de mélange des AnPP), une meilleure approximation m_2 est définie par:

$$m_2 = m_1 + K^{-1}(m_1)s(m_1).$$

Avec $K(m_1)$ étant le (i, j) ème élément de la matrice des informations observées de dimension $(p-1)(p-1)$ avec p étant le nombre de SPP. Et avec $s(m)$ étant la dérivé partielle de $[\delta L/\delta m_1]_{m=M}$, qui est le $i^{\text{ème}}$ élément d'un vecteur de taille $(p-1)$.

Elston suggère d'utiliser des fréquences de génotypes pour cette méthode mais présente une solution pour utiliser des fréquences alléliques (la plupart des éléments de calcul est divisée par $2n$). Il développe de plus une formulation permettant de prendre en compte la dominance de certains allèles.

La méthode du maximum de vraisemblance de Wang (2003): WANG03

Elle se base sur un modèle de MI avec une prise en compte de l'évolution des AnPP avant l'événement de mélange. Elle permet d'estimer conjointement les taux de mélange, la dérive génétique s'exerçant sur les AnPP et AnAP, et la dérive génétique s'étant exercée sur les AnPP à la période B. Elle implique $3d+2$ paramètres (d étant le nombre de AnPP impliquées dans le mélange): $d-1$ taux de mélange, 2 périodes de temps (correspondant à la période A et B du modèle de MI), d tailles efficaces des AnPP (pour chaque population la moyenne de ses N_s sur l'ensemble des générations de la période A), $d+1$ tailles efficaces des AnPP et AnAP (pour chaque population la moyenne de ses N_s sur l'ensemble des générations de la période B). Certaines modifications aux procédures de base de calcul ont été apportées par Wang, de manière à réduire le temps de calcul.

Plusieurs paramètres étant inclus dans le modèle, plusieurs maximums peuvent exister. Wang choisit comme solution la méthode de convergence quadratique de Powell (Press *et al.* 1992) qui nécessite différents points de départ pour déterminer le maximum de vraisemblance. Les valeurs suggérées sont de 10 à 20, et plus le nombre est important, plus le maximum de vraisemblance donné sera correct (mais plus le temps de calcul sera long). De plus, les fréquences alléliques de la population ancestrale P_0 (avant de se diviser en d AnPP) ne sont pas connues. Pour cela, il va falloir déterminer quelles peuvent être les valeurs les plus vraisemblables. Wang considère que toutes les fréquences possibles sont équiprobables pour un allèle donné. Ainsi pour chaque valeur possible, le maximum de vraisemblance va être calculé. Plus on assigne de valeurs discrètes possibles à cette fréquence (nombre de points d'intégration), plus l'estimation sera précise. La méthode prend en compte le biais d'échantillonnage et l'effet de dérive génétique en réduisant le temps par la taille effective de chaque population ($2N$), ce qui réduit le nombre de paramètres à $3d$. La dispersion est donnée sous forme d'intervalle de confiance à 95% pour l'ensemble des paramètres.

Les méthodes basées sur la coalescence et les approches bayésiennes

La méthode basée sur la coalescence des gènes de Bertorelle et Excoffier (1998): B&E98

Bertorelle et Exoffier décident de développer deux estimateurs de mélange qui utilisent des informations moléculaires explicites, le nombre de sauts évolutifs qui séparent des allèles différents pouvant aussi être informatif. Ces nouveaux estimateurs de coefficients de mélange peuvent être appliqués à n'importe quel type de données moléculaires (séquences d'ADN, RFLPs et microsatellites), et par conséquent sont moins adaptés à des données dites "classiques" (comme les fréquences alléliques des groupes sanguins). La méthode originale qui permet de modéliser un mélange entre seulement deux AnPP a été étendue à n'importe quel nombre de AnPP par Dupanloup et Bertorelle (2001). Elle se base sur la coalescence des gènes et incorpore des informations sur la diversité moléculaire présente dans les AnPP et AnAP (différences génétiques entre les allèles). Elle s'appuie sur le modèle de MI.

Le premier estimateur, m_x, est dérivé du temps de coalescence estimé de deux gènes échantillonnés dans la AnAP. Le deuxième estimateur, m_y, est dérivé du temps de coalescence entre un gène échantillonné dans chacune des populations (AnPP et AnAP). Les résultats observés avec m_x ne sont pas aussi précis que ceux obtenus avec m_y, et le développement par Dupanloup et Bertorelle ainsi que toutes les applications de la méthode ont été faits sur l'estimateur m_y.

Pour ce dernier estimateur, il s'agit de considérer le temps de coalescence moyen de deux gènes. Un gène tiré au hasard dans la AnAP peut provenir de la AnPP_1 ou de la AnPP_2 à une probabilité respective de m_1 ou $(1-m_1)$. Pour chaque paire de gènes, deux situations peuvent se présenter. Les gènes peuvent coalescer après l'événement de mélange au niveau de la $i^{\text{ème}}$ AnPP ou bien avant l'événement de mélange dans la population ancestrale unique P_0. Deux modèles mutationnels permettent l'estimation du taux de mélange: le modèle de sites infinis pour les séquences d'ADN et le modèle "*Mutation pas à pas*" pour les microsatellites. En considérant que chaque nouvelle mutation apparaît à un site précédemment monomorphe (il n'existe qu'un type d'allèle à ce locus), la moyenne des temps de coalescence peut être estimée depuis la moyenne du nombre de différences de paires de bases entre les deux allèles. Pour des données multiloci, un estimateur du mélange peut être construit par une estimation du temps de coalescence séparément pour chaque locus et par l'utilisation de leurs valeurs moyennes

calculées sur l'ensemble des loci. Il est nécessaire d'avoir le même taux de mutation à chaque locus. Dans le cas contraire, les coefficients de mélange peuvent être calculés séparément pour des classes de loci partageant les mêmes taux de mutation, et une estimation finale est obtenue par la moyenne entre les classes.

L'approche bayésienne par la vraisemblance par Chikhi et al. (2001): CHIK01

Cette méthode est basée sur la théorie de la coalescence sous un modèle de MI impliquant uniquement deux AnPP. La méthode prend en compte le biais d'échantillonnage et la dérive génétique dans toutes les populations (les AnPP et AnAP peuvent être sujettes à des taux indépendants de dérive génétique). Le temps depuis l'événement de mélange et la contribution d'une des deux AnPP sont estimés simultanément. Bien que ce soit la distribution complète du paramètre "taux de mélange" qui est informative, l'utilisation d'estimateur ponctuel permet d'avoir une idée plus nette de ce paramètre. Les auteurs recommandent l'utilisation de la médiane comme estimateur ponctuel, en particulier lorsque la distribution est à peu près symétrique. La mesure de la dispersion est donnée par l'écart type mais également par l'écart entre les quantiles 5% et 95% (écart interquantile, ou "*equal-tail probability interval"* ETPI). Cette dernière mesure rend mieux compte de la dispersion.

La procédure bayésienne nécessite un *prior* (valeur *a priori*) pour tous les paramètres du modèle. Ils vont être fixés à des valeurs données, sauf les fréquences dans les AnPP qui vont prendre toutes les valeurs possibles de manière équiprobable (distribution uniforme de Dirichlet). Ceci évite d'émettre des hypothèses sur l'histoire évolutive des AnPP pendant la période A. Les temps (en générations) sont pondérés par les tailles efficaces des populations pour l'estimation des temps de coalescence. Pour obtenir la distribution du *posterior*, une chaîne Monte Carlo de Markov permet d'échantillonner aléatoirement parmi les valeurs possibles du jeu de paramètres caractérisant le *prior*. Il est nécessaire de déterminer quand l'équilibre de la chaîne est atteint. Les auteurs utilisent la méthode de Gelman *et al.* (1995), basée sur le lancement de plusieurs chaines MCMC courtes avec des points de départ dispersés dans l'espace des paramètres. Lorsque la variance entre les chaînes est inférieure à 5% de celle observée à l'intérieur des chaînes, l'équilibre est atteint.

Les méthodes ABC (Approximative Bayesian Computation)

Elles offrent une alternative aux méthodes de maximum de vraisemblance et permettent de faire des inférences sur des modèles dynamiques plus complexes. Comme toute approche bayésienne, elles s'appuient sur le choix d'un jeu de paramètres afin d'extraire des informations des données observées.

Introduction de l'approche par Beaumont et al. (2002)

Beaumont *et al.* (2002) introduisent l'approche ABC et la définissent, notamment dans l'optique d'une application en génétique des populations. Les propriétés de la distribution *a posteriori* des paramètres sont approximées sans calcul complet de la vraisemblance. Pour cela, il faut établir une régression linéaire reliant différentes valeurs simulées des paramètres du modèle à différentes valeurs simulées des statistiques de l'échantillon, ("*summary statistics*") puis remplacer les valeurs simulées des paramètres les plus vraisemblables afin de déterminer les valeurs des paramètres secondaires. La méthode combine les avantages des inférences bayésiennes ainsi que la robustesse des méthodes basées sur les données statistiques. Les paramètres secondaires, tels que la variance, vont être directement inclus dans les étapes de calculs et il est ainsi possible de traiter un grand nombre d'entre eux simultanément. La méthode ABC permet d'estimer simultanément tous les paramètres nécessaires à l'étude du mélange: coefficient de mélange, temps de mélange, temps de divergence entre les populations parentales, taille effective des populations. Cette approche s'appuie sur des algorithmes de simulations/rejets massifs et demande beaucoup de temps de calcul.

La méthode de Excoffier et al. (2005): EXC05

La méthode se résume en trois étapes. La première est une étape de simulations. Il s'agit de créer des jeux de données artificiels (environ un million) pour plusieurs loci, avec des caractéristiques similaires aux jeux de données observés (même nombre d'échantillons, même taille pour les SAP/SPP, même nombre de loci), en tirant aléatoirement des valeurs dans les distributions des *prior* (taille des populations efficaces, contribution des populations parentales, temps depuis le mélange, temps de divergence des AnPP avant le mélange, taux de mutation moyen et individuel, paramètres moyen et individuel de longueur de pas du modèle mutationnel). Les distributions de ces paramètres suivent diverses lois de distribution (LogUniforme, Uniforme, Gamma, et Beta). La seconde étape consiste en une comparaison des jeux simulés de données avec les jeux observés de données afin de retenir les simulations les plus proches des observations et en rejetant les autres. La sélection se fait en calculant les distances euclidiennes entre les "*summary statistics*" des données observées et celles des données simulées et en retenant les 1000 simulations ayant les distances les plus faibles (Beaumont *et al.* 2002). La troisième étape consiste à estimer les paramètres du modèle de mélange en réalisant un ensemble de régressions linéaires pondérées entre les statistiques de l'échantillon et les simulations retenues. En plus des paramètres du modèle donnés ci dessus (les *prior*), les autres paramètres sont les temps de divergence et de mélange pondérés par la taille des populations efficaces, les tailles des populations pondérées par les taux de mélange, les temps de divergence et de mélange pondérés par les taux de mutation.

Un synopsis schématique de ces étapes est présenté dans Excoffier *et al.* (2005) et donne une vue générale du processus ABC (voir la figure 2 de la publication originale).

La méthode de Sousa et al. (2009)
Les auteurs présentent une méthode ABC qui utilise la distribution complète des fréquences alléliques. Contrairement aux modèles précédents qui n'incluent que deux AnPP, cette méthode peut être appliquée à un modèle de mélange impliquant trois AnPP. Elle se base sur le modèle de mélange présenté par Thompson (1973) qui prend en compte la dérive génétique. La distance entre les donnés simulées et les données observées est calculée de deux manières: une distance euclidienne standardisée, et une distance génétique, toutes les deux basées sur les fréquences alléliques.

Les autres méthodes

La méthode de Ottensooser (1944): OTT44/OTT62
Ottensooser (1944) présente la même méthode que Bernstein (1931), et certaines études des années 50 et 60 (comme Saldanha 1962) lui en attribuent la paternité à tort. Il introduit ensuite (Ottensooser 1962) une approche permettant de calculer le taux de mélange pour trois AnPP si le taux de mélange est déjà connu pour une des trois. Ce taux peut être connu soit car il provient de la même étude, mais estimé dans un modèle dihybride, soit parce qu´ il est issu d'une autre étude et ne portera donc pas sur le même échantillon. Il est facile de se rendre compte des risques d'erreurs que pose l'utilisation de cette méthode (approche, formulation). Dans le premier cas, on aura une surestimation du taux de mélange si le mélange est en réalité trihybride, et dans le deuxième cas le protocole étant différent, les résultats ne peuvent être fiables.

Les méthodes basées sur les distances (Cavalli-Sforza, Bodmer 1971, Cavalli-Sforza et al. 1994, Pollitzer 1964): CAV94
Pollitzer (1964) est le premier à formaliser l'idée selon laquelle lorsque le flux génique en provenance de deux populations parentales est différent, il y a une relation directe entre les taux de flux génique et les similarités entre la population mélangée et les populations parentales. Cela implique une possibilité d'estimer les taux de mélange depuis les similarités génétiques observées (les distances génétiques). Il considère le carré des distances comme inversement proportionnel au taux de mélange. Sa formulation est applicable à tout type de mélange, impliquant deux populations ou plus. Le taux de mélange de la population parentale *i* sera déterminé par:

$$m_i = \frac{\frac{1}{d_{iH}^2}}{\sum_{j=1}^{n} \frac{1}{d_{jH}^2}}.$$

Avec *n* étant le nombre de populations parentales, d_{jH}^2 la distance au carré entre la population parentale *j* et la population mélangée.

Cavalli-Sforza et Bodmer (1971) introduisent la première méthode réelle basée sur les distances génétiques. Il s'agit d'une méthode de régression linéaire qui part du principe que la distance entre une population mélangée et deux populations parentales peut être utilisée comme estimation du taux de mélange si les distances sont linéaires:

$$d_{12} = d_{1H} + d_{2H}.$$

Avec d_{12} étant la distance entre les deux populations parentales, d_{1H} et d_{21H} celles entre la population mélangée et les $AnPP_1$ et $AnPP_2$ respectivement. Si cette hypothèse est acceptée, l'apport de la $AnPP_2$ au pool génétique de la population mélangée peut être estimé selon la formule:

$$m = \frac{d_{1H}}{d_{12}}.$$

Cette relation est reprise par Lees et Relethford (1978) en considérant les distances entre populations sur la base de réflectivité de la peau.

La méthode de maximum de vraisemblance de Thompson (1973): THOM73
Thompson est la première à présenter une méthode prenant en compte la dérive génétique ainsi que le biais d'échantillonnage dans toutes les populations (AnPP et AnAP). Elle présente un modèle de MI, avec une évolution des populations après l'événement de mélange, les fréquences alléliques fluctuant au cours de la période A dans un espace euclidien de fréquences. Elle transforme donc les fréquences alléliques (p_H, p_1 et p_2) de la AP et des deux PP en vecteur dans un espace à *X* dimensions, *X* étant le nombre d'allèles indépendants. Si il y a *l* loci et que le $i^{\text{ème}}$ locus à *k* allèles, alors:

$$l = \sum_{i=1}^{L} (k_i - 1).$$

Elle assimile ensuite l'effet de la dérive génétique à un mouvement brownien dans l'espace à X dimensions, et dont la variance est $1/8N_e$ (avec N_e étant les tailles efficaces des populations). Plus la taille de la population (ici des échantillons) est importante, plus la variance tend vers 0.

Les fréquences alléliques inconnues des AnPP et AnAP sont les vecteurs P_H, P_1 et P_2 définis par:

$$P_H = MP_1 + (1 - M)P_2.$$

Un maximum de vraisemblance est utilisé pour déterminer le taux de mélange *m* (estimation de *M* à partir de l'échantillon). Sa méthode est valable si la taille des échantillons est identique pour tous les loci au sein d'une même population.

La méthode linéaire multiallèles, multiloci de Korey (1978)
La formule présentée par Cavalli-Sforza et Bodmer (1971) permettant l'estimation pour plusieurs loci bialléliques a été étendue par Korey pour permettre l'estimation pour des loci multialléliques. Il formule $\bar{m}_1$ suivant l'équation:

$$\overline{m}_1 = \frac{\sum_{i=1}^{q} \frac{1}{r_1} \sum_{j=1}^{r_i} m_{1(ij)} Var^{-1}\left(m_{1(ij)}\right)}{\sum_{i=1}^{q} \frac{1}{r_1} \sum_{j=1}^{r_i} Var^{-1}\left(m_{1(ij)}\right)}.$$

Avec $m_{1(ij)}$ étant l'estimation de m_1 au $j^{\text{ème}}$ allèle du $i^{\text{ème}}$ locus, $Var^{-1}\left(m_{1(ij)}\right)$ étant l'inverse de sa variance et r_i le nombre d'allèles au $i^{\text{ème}}$ locus.

La méthode basée sur les allèles uniques ("private alleles") de Chakraborty et al. (1991): CHAK91

La notion de "private alleles" a été introduite par Neel (1973) pour désigner des allèles présents que dans une seule population. L'intérêt des allèles uniques est une simplification de l'équation linéaire de Bernstein: $P_H = M_1 P_1$.

Soit dans l'échantillon: $p_H = m_1 p_1$.

La méthode considère successivement les fréquences alléliques des allèles uniques de SPP1 et de SPP2 et se base sur la méthode de régression de Madansky (1959) pour en extraire les taux de mélange. L'estimateur de Madansky permet de prendre en compte le biais d'échantillonnage. Les estimations issues de chaque locus sont combinées et pondérées par l'inverse de leur variance respective:

$$m = \frac{\frac{m_1}{Var^2\left(m_1\right)} + \frac{1-m_2}{Var^2\left(m_2\right)}}{\frac{1}{Var^2\left(m_1\right)} + \frac{1}{Var^2\left(m_2\right)}}.$$

Avec:

$$Var(m) = \frac{\frac{1}{Var\left(m_1\right)} + \frac{1}{Var\left(m_2\right)}}{\frac{1}{Var^2\left(m_1\right)} + \frac{1}{Var^2\left(m_2\right)}}.$$

Et d'autres encore…

Certaines méthodes d'estimation du mélange génétique n'ont eu qu'un intérêt très limité, et/ou un impact méthodologique faible. D'autres, malgré un potentiel notable, sont passées totalement inaperçues. Nous donnons ici une liste non-exhaustive de quelques unes de ces autres méthodes.

– **Szathmary et Reed** (1978) proposent une méthode visant à donner une estimation du taux de mélange maximum dans le cas où une estimation ponctuelle est impossible (dans le cas de la disparition de certains allèles spécifiques des AnPP et qui ne sont plus présents dans les SPP). Ils se basent sur le mélange de populations d'origine européenne avec des populations amérindiennes, événement historiquement rare, et considèrent que la répartition des allèles rares spécifiques des populations européennes dans les populations amérindiennes suit une loi de Poisson et que la limite supérieure de l'intervalle de confiance de la loi (à 95% ou 99%) peut donc servir à l'estimation du taux maximum de mélange de la population européenne. La limite inférieure doit être considérée comme égale à 0. L'événement de mélange suit un modèle de MI.

– **Lathrop** (1982) propose une méthode d'estimation par maximum de vraisemblance basée sur la création d'un arbre phylogénétique retraçant l'évolution des populations impliquées. La méthode permet de détecter des mélanges génétiques et les temps de divergence des AnPP. Elle se base sur un modèle de MI avec des populations efficaces de tailles constantes. La méthode s'appuie sur celle de THOM73. La dérive y est également modélisée comme un mouvement brownien.

– **Long et Smouse** (1983) étendent la méthode basique d'EMV pour un nombre quelconque de populations parentales ainsi que pour un nombre quelconque de loci multialléliques (avec n'importe quel type de dominance entre eux). Le principe de maximisation de la vraisemblance pour obtenir l'estimation reste similaire. L'événement est un modèle de MI et ni les forces évolutives, ni le biais d'échantillonnage ne sont pris en compte.

– **Wijsman** (1984) propose une technique prenant en compte le biais d'échantillonnage, et la dérive génétique dans toutes les populations, sous un modèle de MI. C'est une méthode de régression linéaire avec la possibilité d'avoir des tailles d'échantillons variables en fonction des différents loci. Des poids sont assignés à chaque population (à ses fréquences alléliques) de manière inversement proportionnelle à l'erreur (variance), à son éloignement de la droite de régression (voir les figures 1 et 2 dans Wijsman 1984).

– **Blangero** (1986) présente une généralisation du modèle par maximisation de la vraisemblance pour une estimation du mélange basée sur des traits quantitatifs multivariés, avec l'hypothèse de neutralité des traits polygéniques utilisés. Sa méthode prend en compte la dérive génétique et le biais d'échantillonnage dans toutes les populations (AnPP et AnAP). Sa généralisation permet également d'impliquer n'importe quel nombre de AnPP, avec un flux génique continu (avec la nécessité de connaître le nombre de générations de mélange). Pour chaque population, les données utilisées regroupent: la taille efficace (pour la dérive génétique), la valeur moyenne des phénotypes ("*mean phenotype vector*"), la matrice de covariance génétique ("*additive genetic covariance matrix*").

– **Manderscheid et Rogers** (1996) s'intéressent à l'expansion de l'Homme anatomiquement moderne et à son mélange potentiel avec les populations archaïques. Ce mélange potentiel induit qu'une fraction q de mitochondries dans les populations actuelles dérive des populations archaïques remplacées (complètement ou partiellement) par les populations modernes juste après leur expansion. Par conséquent, q va mesurer d'une part le niveau de mélange (taux de mélange), et d'autre part la fréquence initiale de mitochondries dites "divergentes[1]".

[1] Les mitochondries divergentes sont les formes mitochondriales archaïques qui sont présentes dans la population d'hommes modernes juste après l'expansion et donc après les premiers mélanges.

Sous un modèle de MI, la méthode prend en compte la possibilité d'absence de ces mitochondries "divergentes" sous l'effet du biais d'échantillonnage ainsi que de la dérive génétique.

– **Estoup *et al.*** (1999) proposent une méthode suivant le modèle de MI et utilisant les JMS (Juxtaposed Microsatellite Systems). Leur méthode s'appuie sur cinq hypothèses pouvant être résumées par le fait que les JMS doivent être spécifiques des AnPP et qu' il n'y a eu ni mutation ni recombinaison au niveau des JMS. La méthode d'estimation du taux de mélange est similaire à l'approche par les "private alleles", avec une dispersion mesurée par la méthode de bootstrap.
– **Helgason *et al.*** (2000) proposent une nouvelle méthode permettant de gérer le problème inhérent au nombre "infini" d'allèles possibles pour le chromosome Y (les allèles étant les haplotypes, il peut exister autant d'haplotypes que de combinaisons possibles de mutation sur une séquence d'ADN). Ils proposent tout d'abord une approche itérative heuristique pour définir la distribution des taux de mélange définissant le mieux les données. Elle incorpore les informations des "private haplotypes" de la AnAP.
– **Di Benedetto *et al.*** (2001) considèrent un flux unidirectionnel instantané depuis une population donneuse vers une population receveuse, le taux de migration étant considéré comme le taux de mélange. Ils proposent également un mélange sous modèle de FGC avec un taux de migration/taux de mélange m_C sur t générations obtenu par résolution de l'équation:

$$p' = p_s + (1 - m_C)^t (p - p_s).$$

La dispersion est mesurée avec l'écart type.

– **Collins-Schramm *et al.*** (2002) présentent deux approches sous un modèle de MI. D'une part, ils considèrent les fréquences alléliques attendues en fonction de différents taux de mélange qu'ils assignent à chacune des AnPP. Ils comparent ensuite les fréquences alléliques attendues aux fréquences alléliques observées de manière à déterminer quel taux de mélange est le plus susceptible de donner les fréquences observées. C'est une approche graphique ayant des similarités avec les approches par maximum de vraisemblance. Dans leur deuxième approche, ils estiment les taux de mélange selon une méthode de régression linéaire. Les taux de mélange sont obtenus en minimisant l'équation:

$$\sum_{j=1}^{l} \sum_{i=1}^{k} \left[mP_1 + (1-m) P_2 - P_H \right]^2 .$$

Avec l étant le nombre de loci et k le nombre d'allèles. Un intervalle de confiance est construit par la méthode de bootstrap pour la mesure de la dispersion.

– **Salas *et al.*** (2004) présentent une approche bayésienne pour déterminer les taux de contribution de diverses régions sources d'Afrique à une population mélangée. Ils assument que le nombre d'ADNmt dans chaque cluster (groupe) d'un échantillon de la AnAP ($n_i : 1 \le i \le C$, C étant le nombre de clusters) est tiré au hasard suivant une distribution multinomiale de paramètre "la taille de la SAP" ($N = \sum_{i=1}^{C} n_i$), et:

$$p_i = \sum_{j=1}^{R} m_i f_{ji} .$$

Avec R étant le nombre de sources d'AnPP africaines, f_{ji} la fréquence du $i^{\text{ème}}$ cluster dans la $j^{\text{ème}}$ région source, et m_i étant le taux de mélange. Le modèle de mélange est un modèle de MI.

– **Wang** (2006) présente une méthode très similaire à celle de Wang (2003). A la génération T_A après le mélange, un échantillon d'individus est prélevé de la AP et des PP (deux PP impliquées). Les séquences ADN des échantillons sont ensuite analysées au niveau de plusieurs loci et utilisées pour déterminer les taux de mélange, le temps de divergence des AnPP, le temps depuis l'événement de mélange, et les tailles efficaces des populations. Ces paramètres sont ensuite pondérés par le taux de mutation. Lorsque plusieurs loci sont utilisés pour l'estimation, les taux de mutation (constants sur toutes les générations) sont pris comme la moyenne des taux de mutation sur l'ensemble des loci. Pour prendre en compte les informations moléculaires et la mutation dans l'estimation des taux de mélange, un modèle mutationnel de site infini est défini. Le nombre de mutations est égal au nombre de sites polymorphiques sur la séquence d'ADN. L'ensemble de ces paramètres permet de définir une généalogie des séquences d'ADN jusqu'au plus récent ancêtre commun. La dispersion est mesurée par l'intervalle de confiance à 95% ainsi que par l'écart type par une procédure de bootstrap.

Les méthodes basées sur la couleur de la peau (CdP)

Pour conclure sur les différents types de méthodes qui ont été présentées depuis presque un siècle, il faut citer celles permettant l'estimation des taux de mélange en se basant sur la couleur de la peau. Ces méthodes sont très controversées. Trevor (1953) est le premier à s'intéresser aux caractères quantitatifs pour l'estimation des contributions au mélange, et Harrison et Owen (1964) sont les premiers à proposer une formulation mathématique, modifiée par Lees et Relethford (1978). Ces derniers reprennent la relation exploitée par Harrison *et al.* (1967) mais proposent d'utiliser la moyenne obtenue avec les différents filtres du spectrophotomètre et ajoutent une transformation possible des valeurs de réflectivité. Korey (1980) introduit la variance suivant Reed (1969) et Cavalli-Sforza et Bodmer (1971).

INTERETS ET INTERPRETATIONS DE L'ESTIMATION DU MELANGE

L'estimation des taux de mélange s'avère essentielle dans l'étude des structures génétiques des populations et de leur histoire. Les populations humaines étant le résultat d'un

mélange plus ou moins récent entre diverses populations, il en résulte que le pool génétique de celles-ci peut être défini à partir de ses populations parentales. La connaissance des contributions des populations parentales au pool génétique de la population mélangée a toujours joué un rôle significatif dans l'analyse de ces populations hybrides. Plusieurs études sont conçues dans un but historique (Abe-Sandes *et al.* 2004, Alves-Silva *et al.* 2000, Arpini-Sampaio *et al.* 1999, Avena *et al.* 2001, 2006, Bertoni *et al*. 2003, Bonilla *et al.* 2004, 2005, Buentello-Malo *et al.* 2008, Carvajal-Carmona *et al*. 2000, Carvalho-Silva *et al.* 2001, Castrì *et al.* 2007, Cerda-Flores *et al.* 2002, Loyo *et al.* 2004, Madrigal *et al.* 2001, Martínez Marignac *et al*. 2004, Martínez-Cortés *et al*. 2010, Mendizabal *et al*. 2008, Parra *et al.* 2001, Salas *et al.* 2004, Sans 2000, Sans *et al.* 2002, Seldin *et al.* 2007), en Amérique du Nord ou Latine et dans les populations hispano-américaines, pour fournir un appui génétique important à l'histoire de la colonisation des Amériques. Les estimations des taux de mélange permettent également de faire ressortir l'existence de flux migratoires différentiels homme/femme ou bien de mettre en évidence des différences dans la structure génétique d'une même population mélangée ou des populations parentales selon le lieu d'échantillonnage (Gourjon *et al.* 2010). Dans le cas de mélange supposé, les résultats sont toutefois à prendre avec plus de recul.

Les estimations sont donc un soutien à des données historiques, linguistiques ou archéologiques, mais ne sont pas des certitudes et il est fallacieux de s'appuyer sur une seule estimation pour certifier un fait. La qualité d'une estimation dépend de trois choix principaux: le choix des marqueurs, le choix des populations parentales et le choix de la méthode. Les résultats obtenus varient en fonction de ces choix (Gourjon *et al.* 2010). Nous avons mené une réflexion sur les deux premiers types de choix (Gourjon 2010). Lorsque toutes les hypothèses sur lesquelles s'appuient les méthodes sont remplies, au moins partiellement, ces méthodes peuvent fournir des résultats en adéquation avec les données extérieures (historiques, archéologiques, linguistiques, culturelles), et alors être considérées comme exploitables et fiables. Mais l'utilité pratique de ces méthodes devient alors ambigüe. Si les estimations sont concordantes avec des taux de mélange attendus, nous pouvons évidemment nous questionner sur l'intérêt d'une estimation génétique qui ne fait que donner le reflet d'autres types de données ; si les résultats obtenus ne sont pas concordants, ils peuvent susciter des critiques. On cherche, alors à expliquer les différences observées entre les estimations attendues et les estimations observées. Ceci, en s'appuyant en grande partie sur les discordances possibles entre les modèles théoriques (sur lesquels reposent les méthodes employées) et le mélange réel survenu entre les populations étudiées. Dans tous les cas, la volatilité des estimations sur une simple modification d'un paramètre, selon les choix des populations parentales et de l'échantillonnage, ainsi que selon le type de marqueurs employés (Gourjon *et al.* 2010, Merriwether *et al.* 1997, Mona *et al*. 2009), nous conduit à suggérer une autre évaluation du mélange génétique. Non plus des taux, mais une valeur indicative, un intervalle dont les modalités de calcul permettent de prendre en compte les estimations obtenues à partir de plusieurs estimateurs, ainsi que plusieurs combinaisons de populations parentales. Chaque estimation est pondérée d'une part par la taille des échantillons et d'autre part par des paramètres tels que l'écart-type de chaque estimateur, évitant ainsi de donner trop de poids aux estimations "aberrantes". Un intervalle est obtenu pour chaque composante parentale: l'intervalle indicatif de mélange (IIM), donnant les tendances de contribution parentale, les tendances d'ancestralité.

CONCLUSION

Cette revue des méthodes d'estimation du mélange génétique populationnel souligne la complexité de l'apport des composantes parentales au pool génétique d'une population mélangée, et ses enjeux anthropologiques. Pour conclure nous gardons en mémoire ce qu'écrivait De Quatrefages, donnant une vision holistique du mélange entre les populations humaines, et qui, un siècle plus tard, semble être fondamentalement visionnaire. "Sans doute, le métissage, favorisé, multiplié par la facilité croissante des communications, me semble préparer une ère nouvelle. Les races de l'avenir moins différentes de sang, rapprochées par les chemins de fer et les steamers, auront bien plus de penchants, de besoins, d'intérêts communs. […] En dépit des croisements, la variété, l'inégalité persisteront sur la terre. Mais dans son ensemble l'humanité se sera complétée ; elle aura grandi ; et les civilisations de l'avenir, sans faire oublier celles du présent, les dépasseront dans quelque direction encore inconnue, comme les nôtres ont dépassé leurs devancières" (De Quatrefages 1888).

RÉFÉRENCES

ABE-SANDES K., SILVA W. J., ZAGO M., 2004: Heterogeneity of the Y chromosome in Afro-Brazilian populations. *Hum. Biol.* 76, 1: 77–86.

ALVES-SILVA J., DA SILVA SANTOS M., GUIMARÃES P., FERREIRA A., BANDELT H., PENA S., PRADO V., 2000: The ancestry of Brazilian mtDNA lineages. *Amer. J. Hum. Genet.* 67, 2: 444–461.

ARPINI-SAMPAIO Z., COSTA M., MELO A., CARVALHO M., DEUS M., SIMÕES A., 1999: Genetic polymorphisms and ethnic admixture in African-derived black communities of northeastern Brazil. *Hum. Biol.* 71, 1: 69–85.

AVENA S., GOICOECHEA A., DUGOUJON J., SLEPOY M., SLEPOY A., CARNESE F., 2001: Análisis antropogenético de los aportes indigena y Africano en Muestras Hospitalarias de la ciudad de Buenos Aires. *Rev. Arg. Antr. Biol.* 3: 21.

AVENA S., GOICOECHEA A., REY J., DUGOUJON J., DEJEAN C., CARNESE F., 2006: Gene mixture in a population sample from Buenos Aires City. *Medicina (B Aires)* 66, 2: 113–118.

BEAUMONT M., ZHANG W., BALDING D., 2002: Approximate Bayesian computation in population genetics. *Genetics* 162, 4: 2025–2035.

BERNSTEIN F. 1931: Die geographische verteilung der blutgruppen und ihre anthropologische bedeutung. In: C. Ginni (Ed): *Comitato Italiano per lo Studio dei Problemi della Populazione*. Pp. 227–243. Istituto poligrafico dello Stato, Roma.

BERTONI B., BUDOWLE B., SANS M., BARTON S., CHAKRABORTY R., 2003: Admixture in Hispanics: distribution of ancestral population contributions in the Continental United States. *Hum. Biol.* 75, 1: 1–11.

BERTORELLE G., EXCOFFIER L., 1998: Inferring admixture proportions from molecular data. *Mol. Biol. Evol.* 15, 10: 1298–1311.

BLANGERO J., 1986: Admixture estimation using multivariate quantitative traits: A maximum likelihood approach. *Amer. J. Phys. Anthrop.* 69: 177.

BONILLA C., BERTONI B., GONZÁLEZ S., CARDOSO H., BRUM-ZORRILLA N., SANS M., 2004: Substantial Native American female contribution to the population of Tacuarembó, Uruguay, reveals past episodes of sex-biased gene flow. *Amer. J. Hum. Biol.* 16, 3: 289–297.

BONILLA C., GUTIÉRREZ G., PARRA E., KLINE C., SHRIVER M., 2005: Admixture analysis of a rural population of the state of Guerrero, Mexico. *Amer. J. Phys. Anthrop.* 128, 4: 861–869.

BUENTELLO-MALO L., PEÑALOZA-ESPINOSA R., SALAMANCA-GÓMEZ F., CERDA-FLORES R., 2008: Genetic admixture of eight Mexican indigenous populations: based on five polymarker, HLA-DQA1, ABO, and RH loci. *Amer. J. Hum. Biol*. 20, 6: 647–650.

CARVAJAL-CARMONA L., SOTO I., PINEDA N., ORTÍZ-BARRIENTOS D., DUQUE C., OSPINA-DUQUE J., MCCARTHY M., MONTOYA P., ALVAREZ V., BEDOYA G., RUIZ-LINARES A., 2000: Strong Amerind/white sex bias and a possible Sephardic contribution among the founders of a population in northwest Colombia. *Amer. J. Hum. Genet.* 67, 5: 1287–1295.

CARVALHO-SILVA D., SANTOS F., ROCHA J., PENA S., 2001: The phylogeography of Brazilian Y-chromosome lineages. *Amer. J. of Hum. Genet.* 68, 1: 281–286.

CASTRÌ L., OTÁROLA F., BLELL M., RUIZ E., BARRANTES R., LUISELLI D., PETTENER D., MADRIGAL L., 2007: Indentured migration and differential gender gene flow: the origin and evolution of the East-Indian community of Limón, Costa Rica. *Amer. J. Phys. Anthrop.* 134, 2: 175–189.

CAVALLI-SFORZA L. L., BODMER W., 1971: *The genetics of human populations*. W.H. Freeman and Company, San Francisco. 943 pp.

CAVALLI-SFORZA L. L., MENOZZI P., PIAZZA A., 1994: *The History and Geography of Human Genes*. Princeton University Press, Princeton. 1088 pp.

CERDA-FLORES R., VILLALOBOS-TORRES M., BARRERA-SALDAÑA H., CORTÉS-PRIETO L., BARAJAS L., RIVAS F., CARRACEDO A., ZHONG Y., BARTON S., CHAKRABORTY R., 2002: Genetic admixture in three Mexican Mestizo populations based on D1S80 and HLA-DQA1 loci. *Amer. J. Hum. Biol.* 14, 2: 257–263.

CHAKRABORTY R., 1975: Estimation of race admixture- a new method. *Amer. J. Phys. Anthrop.* 42, 3: 507–511.

CHAKRABORTY R., 1985: Gene identity in racial hybrids and estimation of admixture rates. In: Y. R. Ahuja and J. V. Neel (Ed): *Genetic differentiation in human and other animal populations*. Pp. 171–180. Indian Anthropological Association, Delhi.

CHAKRABORTY R., 1986: Gene admixture in human-populations – Models and predictions. *Yearb. Phys. Anthropol.* 29: 1–43.

CHAKRABORTY R., KAMBOH M., FERRELL R., 1991: "Unique" alleles in admixed populations: a strategy for determining "hereditary" population differences of disease frequencies. *Ethn. Dis.* 1, 3: 245–256.

CHAKRABORTY R., KAMBOH M., NWANKWO M., FERRELL R., 1992: Caucasian genes in American blacks: new data. *Amer. J. of Hum. Genet.* 50, 1: 145–155.

CHIARONI J., TOUINSSI M., FRASSATI C., DEGIOANNI A., GIBERT M., REVIRON D., MERCIER P., BOËTSCH G., 2004: Genetic characterization of the population of Grande Comore Island (Njazidja) according to major blood groups. *Hum. Biol.* 76, 4: 527–541.

CHIKHI L., BRUFORD M., BEAUMONT M., 2001: Estimation of admixture proportions: a likelihood-based approach using Markov chain Monte Carlo. *Genetics* 158, 3: 1347–1362.

COLLINS-SCHRAMM H., PHILLIPS C., OPERARIO D., LEE J., WEBER J., HANSON R., KNOWLER W., COOPER R., LI H., SELDIN M., 2002: Ethnic-difference markers for use in mapping by admixture linkage disequilibrium. *Amer. J. Hum. Genet.* 70, 3: 737–750.

DE QUATREFAGES A., 1888: *L'espèce humaine*. Félix Alcan Editeur, Paris. 366 pp.

DI BENEDETTO G., ERGÜVEN A., STENICO M., CASTRÌ L., BERTORELLE G., TOGAN I., BARBUJANI G., 2001: DNA diversity and population admixture in Anatolia. *Amer. J. Phys. Anthrop.* 115, 2: 144–156.

DUPANLOUP I., BERTORELLE G., 2001: Inferring admixture proportions from molecular data: extension to any number of parental populations. *Mol. Biol. Evol.* 18, 4: 672–675.

ELSTON R. C., 1971: The estimation of admixture in racial hybrids. *Ann. Hum. Genet.* 35, 1: 9–17.

ESTOUP A., CORNUET J.-M., ROUSSET F., GUYOMARD R., 1999: Juxtaposed Microsatellite Systems as Diagnostic Markers for Admixture: Theoretical Aspects. *Mol. Biol. Evol.* 16, 7: 898–908.

EXCOFFIER L., ESTOUP A., CORNUET J., 2005: Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. *Genetics* 169, 3: 1727–1738.

GELMAN A., CARLIN J., STERN H., RUBIN D., 1995: *Bayesian Data Analysis*. Chapman and Hall, London. 552 pp.

GLASS B., LI C., 1953: The dynamics of racial intermixture; an analysis based on the American Negro. *Amer. J. Hum. Genet.* 5, 1: 1–20.

GOURJON G., 2010: *L'estimation du mélange génétique dans les populations humaines*. PhD Thesis. Université de la Méditerranée, Aix-Marseille II, Marseille. 297 pp.

GOURJON G., BOËTSCH G., DEGIOANNI A., 2010: Gender and population history: sex bias revealed by studying genetic admixture of Ngazidja population (Comoro Archipelago). *Amer. J. Phys. Anthrop.* In Press.

HARRISON G., OWEN J., 1964: Studies on the inheritance of human skin colour. *Ann. of Hum. Genet*. 28: 27–37.

HARRISON A., OWEN J., DA ROCHA F., SALZANO F., 1967: Skin colour in southern Brazilian populations. *Hum. Biol.* 39, 1: 21–31.

HARTL D. J., 1994: *Génétique des populations*. Flammarion, Paris. 305 pp.

HELGASON A., SIGURÐARDÓTTIR S., NICHOLSON J., SYKES B., HILL E., BRADLEY D., BOSNES V., GULCHER J., WARD R., STEFÁNSSON K., 2000: Estimating Scandinavian and Gaelic ancestry in the male settlers of Iceland. *Amer. J. Hum. Genet.* 67, 3: 697–717.

HRBEK I., EL FASI M., 1997: *Histoire générale de l'Afrique:*

L'Afrique du VIIème au XIème siècle. Udicef Unesco, Evreux. 559 pp.

KOREY K., 1978: A critical appraisal of methods for measuring admixture. *Hum. Biol.* 50, 3: 343–360.

KOREY K., 1980: Skin colorimetry and admixture measurement: some further considerations. *Amer. J. Phys. Anthrop.* 53, 1: 123–128.

KRIEGER H., MORTON N., MI M., AZEVÊDO E., FREIRE-MAIA A., YASUDA N., 1965: Racial admixture in north-eastern Brazil. *Ann. Hum. Genet.* 29, 2: 113–125.

LAFON M., 1991: *Lexique français-comorien (shingazidja)*. L'Harmattan, Le-Poiré-sur-Vie. 239 pp.

LATHROP G., 1982: Evolutionary trees and admixture: phylogenetic inference when some populations are hybridized. *Ann. of Hum. Genet.* 46, 3: 245–255.

LEES F., RELETHFORD J., 1978: Admixture estimation using skin reflectance data. *Amer. J. of Phys. Anthrop.* 49, 4: 505–509.

LONG J., 1991: The genetic structure of admixed populations. *Genetics* 127, 2: 417–428.

LONG J., SMOUSE P., 1983: Intertribal gene flow between the Ye'cuana and Yanomama: genetic analysis of an admixed village. *Amer. J. Phys. Anthrop.* 61, 4: 411–422.

LOYO M., DE GUERRA D., IZAGUIRRE M., RODRIGUEZ-LARRALDE A., 2004: Admixture estimates for Churuguara, a Venezuelan town in the State of Falcón. *Ann. of Hum. Biol*. 31, 6: 669–680.

MADANSKY A., 1959: The fitting of straight lines when both variables are subject to error. *J. Am. Stat. Assoc.* 54: 173–205.

MADRIGAL L., WARE B., MILLER R., SAENZ G., CHAVEZ M., DYKES D., 2001: Ethnicity, gene flow, and population subdivision in Limón, Costa Rica. *Amer. J. Phys. Anthrop*. 114, 2: 99–108.

MANDERSCHEID E., ROGERS A., 1996: Genetic admixture in the late Pleistocene. *Amer. J. Phys. Anthrop.* 100, 1: 1–5.

MARTÍNEZ-MARIGNAC V., BERTONI B., PARRA E., BIANCHI N., 2004: Characterization of admixture in an urban sample from Buenos Aires, Argentina, using uniparentally and biparentally inherited genetic markers. *Hum. Biol.* 76, 4: 543–557.

MARTÍNEZ-CORTÉS G., NUÑO-ARANA I., RUBI-CASTELLANOS R., VILCHIS-DORANTES G., LUNA-VÁZQUEZ A., CORAL-VÁZQUEZ R., CANTO-CETINA T., SALAZAR-FLORES J., MUÑOZ-VALLE J., SANDOVAL-MENDOZA K., LÓPEZ, Z., GAMERO-LUCAS, J. J., RANGEL-VILLALOBOS, H., 2010: Origin and genetic differentiation of three Native Mexican groups (Purépechas, Triquis and Mayas): Contribution of CODIS-STRs to the history of human populations of Mesoamerica. *Ann. Hum. Biol.* 37, 6: 801–819.

McEVOY B., BRADY C., MOORE L., BRADLEY D., 2006: The scale and nature of Viking settlement in Ireland from Y-chromosome admixture analysis. *Eur. J. Hum. Genet.* 14, 12: 1288–1294.

MENDIZABAL I., SANDOVAL K., BERNIELL-LEE G., CALAFELL F., SALAS A., MARTÍNEZ-FUENTES A., COMAS D., 2008: Genetic origin, admixture, and asymmetry in maternal and paternal human lineages in Cuba. *BMC Evol. Biol.* 8: 213.

MERRIWETHER D., HUSTON S., IYENGAR S., HAMMAN R., NORRIS J., SHETTERLY S., KAMBOH M., FERRELL R., 1997: Mitochondrial versus nuclear admixture estimates demonstrate a past history of directional mating. *Amer. J. Phys. Anthrop*. 102, 2: 153–159.

MONA S., GRUNZ K., BRAUER S., PAKENDORF B., CASTRÌ L., SUDOYO H., MARZUKI S., BARNES R., SCHMIDTKE J., STONEKING M., KAYSER M., 2009: Genetic admixture history of Eastern Indonesia as revealed by Y-chromosome and mitochondrial DNA analysis. *Mol. Biol. Evol.* 26, 8: 1865–1877.

MORGAN J., 1922: Des origines des Sémites et de celles des Indo-Européens. *Rev. de Synth. Hist*. 34.

MSAIDIE S., DUCOURNEAU A., BOETSCH G., LONGEPIED G., PAPA K., ALLIBERT C., YAHAYA A., CHIARONI J., MITCHELL M., 2010: Genetic diversity on the Comoros Islands shows early seafaring as major determinant of human biocultural evolution in the Western Indian Ocean. *Eur. J. Hum. Genet*. doi:10.1038/ejhg. 2010.128.

NEEL J., 1973: "Private" genetic variants and the frequency of mutation among South American Indians. *Proc. Natl. Acad. Sci. USA* 70, 12: 3311–3315.

NEI M., 1972: Genetic distance between populations. *Am. Naturalist*. 106: 283–292.

NEI M., 1973: The theory and estimation of genetic distance. In: N. Morton (Ed): *Genetic Structure of Populations*. Pp. 45–54. University of Hawaii Press, Honolulu.

OTTENSOOSER F., 1944: Calculo do grau de mistura racial atraves dos grupos sanguineos. *Rev. Bras. Biol.* 4: 531–537.

OTTENSOOSER F., 1962: Analysis of trihybrid populations. *Amer. J. Hum. Genet.* 14: 278–280.

PARRA E., KITTLES R., ARGYROPOULOS G., PFAFF C., HIESTER K., BONILLA C., SYLVESTER N., PARRISH-GAUSE D., GARVEY W., JIN L., McKEIGUE P., KAMBOH M., FERRELL R., POLLITZER W., SHRIVER, M., 2001: Ancestral proportions and admixture dynamics in geographically defined African Americans living in South Carolina. *Amer. J. Phys. Anthrop*. 114, 1: 18–29.

PFAFF C., PARRA E., BONILLA C., HIESTER K., McKEIGUE P., KAMBOH M., HUTCHINSON R., FERRELL R., BOERWINKLE E., SHRIVER M., 2001: Population structure in admixed populations: effect of admixture dynamics on the pattern of linkage disequilibrium. *Amer. J. Hum. Genet.* 68, 1: 198–207.

POLLITZER W., 1964: Analysis of a triracial hybrid. *Hum. Biol.* 36: 362–373.

PRESS W., TEUKOLSKY W., VETTERLING W., FLANNERY B., 1992: *Numerical Recipes in Fortran 77*. Cambridge University Press, Cambridge. 992 pp.

REED T., 1969: Caucasian genes in American Negroes. *Science* 165, 895: 762–768.

ROBERTS D., HIORNS R., 1962: The dynamics of racial intermixture. *Amer. J. Hum. Genet.* 14: 261–277.

ROBERTS D., HIORNS R., 1965: Methods of analysis of the genetic composition of a hybrid population. *Hum. Biol.* 37: 38–43.

SALAS A., RICHARDS M., LAREU M., SCOZZARI R., COPPA A., TORRONI A., MACAULAY V., CARRACEDO A., 2004: The African diaspora: mitochondrial DNA and the Atlantic slave trade. *Amer. J. Hum. Genet*. 74, 3: 454–465.

SALDANHA P., 1962: Racial admixture among Northeastern Brazilian populations. *Am. Anthropol.* 64: 751–759.

SANS M., 2000: Admixture studies in Latin America: from the 20th to the 21st century. *Hum. Biol.* 72, 1: 155–177.

SANS M., WEIMER T., FRANCO M., SALZANO F., BENTANCOR N., ALVAREZ I., BIANCHI N., CHAKRABORTY R., 2002: Unequal contributions of male and female gene pools from parental populations in the African descendants of the city of Melo, Uruguay. *Amer. J. Phys. Anthrop.* 118, 1: 33–44.

SELDIN M., TIAN C., SHIGETA R., SCHERBARTH H., SILVA G., BELMONT J., KITTLES R., GAMRON S., ALLEVI A., PALATNIK S., ALVARELLOS A., PAIRA S., CAPRARULO C., GUILLERÓN C., CATOGGIO L., PRIGIONE C., BERBOTTO G., GARCÍA M.,

PERANDONES C., PONS-ESTEL B., ALARCON-RIQUELME, M., 2007: Argentine population genetic structure: large variance in Amerindian contribution. *Amer. J. Phys. Anthrop.* 132, 3: 455–462.

SOUSA V., FRITZ M., BEAUMONT M., CHIKHI L., 2009: Approximate bayesian computation without summary statistics: the case of admixture. *Genetics* 181, 4: 1507–1519.

SZATHMARY E., REED T., 1978: Calculation of the maximum amount of gene admixture in a hybrid population. *Amer. J. Phys. Anthrop.* 48, 1: 29–33.

THOMPSON E., 1973: The Icelandic admixture problem. *Ann. Hum. Genet.* 37, 1: 69–80.

TREVOR J., 1953: Race crossing in man: The analysis of metrical characters. *Eugenics Lab. Memoirs* 36: 1–45.

VERIN P., 1994: *Les Comores*. Karthala, Paris. 264 pp.

WANG J., 2003: Maximum-likelihood estimation of admixture proportions from genetic data. *Genetics* 164, 2: 747–765.

WANG J., 2006: A coalescent-based estimator of admixture from DNA sequences. *Genetics* 173, 3: 1679–1692.

WIJSMAN E., 1984: Techniques for estimating genetic admixture and applications to the problem of the origin of the Icelanders and the Ashkenazi Jews. *Hum. Genet.* 67, 4: 441–448.

Anna Degioanni
Unité d'Anthropologie Bioculturelle
UMR 6578, Université d'Aix Marseille II – CNRS – EFS
13916 Faculté de Médecine Secteur Nord
Bâtiment A
CS80011, Bd Pierre Dramard
13344 Marseille, France
E-mail: Anna.Degioanni@univmed.fr

Géraud Gourjon
Unité d'Anthropologie Bioculturelle
UMR 6578, Université d'Aix Marseille II – CNRS – EFS
13916 Faculté de Médecine Secteur Nord
Bâtiment A
CS80011, Bd Pierre Dramard
13344 Marseille, France
E-mail: ggourjon@hotmail.fr