Chapitre 4
Génétique des populations

E. Génin

La génétique des populations est une discipline née de la synthèse entre la théorie mendélienne de l'hérédité et la théorie darwinienne de l'évolution. Elle étudie la diversité génétique dans une population dans le temps (au cours des générations) et dans l'espace, ainsi que les mécanismes qui ont façonné cette diversité génétique. Pour caractériser la diversité génétique, on étudie des variants génétiques, c'est-à-dire des changements dans la séquence de l'ADN qui sont dus à des mutations. On distingue différents types de variants génétiques :

- les *single nucleotide polymorphisms* (SNIP or SNIP) sont des variations d'une seule base de l'ADN (par exemple une adénine A qui est remplacée par une guanine G). Ces variations sont fréquentes sur le génome humain (en moyenne 4 à 5 millions dans le génome d'un individu). Elles peuvent être uniques (*single nucleotide variant* ou SNV) ou partagées entre plusieurs individus. Elles peuvent se situer dans des gènes ou dans le génome non codant, ce qui est la situation la plus fréquente. Au total, 12,8 millions de SNIP sont répertoriés (https://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi) ;
- les *copy number variations* (CNV) sont des variations du nombre d'exemplaires de fragments du génome mesurant de quelques kilobases à plusieurs mégabases (perte ou gain de fragments). Ces variations structurales couvrent environ 15 % du génome et, comme les SNIP, certaines sont uniques et d'autres sont partagées entre individus. Un catalogue des différents CNV identifiés à ce jour en population générale est disponible (http://dgv.tcag.ca/).

D'autres types de variations existent également, comme par exemple les VNTR (*variable number tandem repeats*), qui sont des petites séquences nucléotidiques répétées en tandem. Le nombre de répétitions varie d'un individu à l'autre et permet donc de définir différents allèles. Ce type de variations a constitué une source importante de marqueurs génétiques et a permis la mise au point des premières cartes de marqueurs pour réaliser les analyses de liaison génétique [1].

À partir des observations sur les distributions de ces variants génétiques dans des échantillons représentatifs des populations étudiées, on pourra caractériser la structure génétique des populations, tester des hypothèses et construire des modèles pour comprendre l'histoire des populations et étudier les mécanismes évolutifs impliqués. Comprendre la structure génétique des populations est un préalable nécessaire dans l'étude du déterminisme génétique des maladies et la mise en évidence des gènes de susceptibilité.

Dans ce chapitre, nous aborderons quelques notions de base de génétique des populations pour caractériser la structure génétique des populations et comprendre les mécanismes qui entrent en jeu pour façonner cette structure génétique et créer la diversité génétique qu'on observe dans les populations. Nous discuterons ensuite de l'impact de cette diversité génétique sur la santé humaine.

Le modèle de Hardy-Weinberg

Présentation du modèle

Ce modèle formulé en 1908 [2, 3] permet de décrire les relations entre les fréquences génotypiques et alléliques au cours des générations. Il fait les hypothèses suivantes concernant la population étudiée : la population est de taille infinie, les unions se font au hasard (panmixie et pangamie), et il n'y a ni migration, ni sélection, ni mutation.

Dans ces conditions, les fréquences alléliques et génotypiques restent les mêmes au cours des générations et il n'y a donc pas de perte de diversité de la population. On peut le démontrer en considérant que, partant d'une génération t, les individus de la génération suivante $t+1$ sont issus d'un tirage aléatoire des allèles dans les urnes gamétiques mâles et femelles. On parle d'ailleurs souvent de l'**équilibre de Hardy-Weinberg**.

On peut résumer le modèle de Hardy-Weinberg comme suit : dans une population idéale (taille infinie, unions au hasard, ni migration, ni sélection, ni mutation), les fréquences alléliques restent les mêmes d'une génération à l'autre et la diversité génétique se maintient. Les fréquences des génotypes se déduisent des fréquences alléliques et sont p^2, $2pq$, q^2 (proportions de Hardy-Weinberg).

Écart aux conditions du modèle de Hardy-Weinberg

Écart à l'hypothèse de panmixie

Dans la plupart des populations humaines, même si le choix du conjoint ne se fait pas complètement au hasard, l'hypothèse de panmixie reste assez réaliste et on observe que les distributions génotypiques sont généralement conformes aux proportions de Hardy-Weinberg. Dans certaines populations cependant, les unions avec des individus apparentés, par exemple des cousins germains, sont favorisées ce qui engendre de la consanguinité dans la population. À l'échelle d'une famille, cela va se traduire par des boucles de consanguinité dans la famille comme illustré sur la figure 4.1 pour le cas d'un enfant issu d'une union entre des parents cousins germains. Sur le plan génétique, la conséquence de la consanguinité est la possibilité qu'un allèle soit reçu identique par descendance (on parle d'IBD pour *identical by descent*) par l'enfant consanguin. La probabilité pour que l'enfant consanguin G reçoive un allèle IBD à un locus dépend du nombre d'ascendants dans la boucle de consanguinité. Cette probabilité est le *taux de consanguinité* (on note en général *F* ce taux de consanguinité). Dans l'exemple de la figure 4.1, le taux de consanguinité *F* est de 1/16, qui est le taux de consanguinité d'un enfant issu de l'union entre deux cousins germains.

À l'échelle de la population, on peut définir un coefficient moyen de consanguinité qui est la moyenne des coefficients de consanguinité des différents types d'unions qui existent dans la population pondérée par la fréquence de ces unions.

En l'absence de panmixie (toutes les autres conditions de la population idéale étant réunies par ailleurs), les fréquences

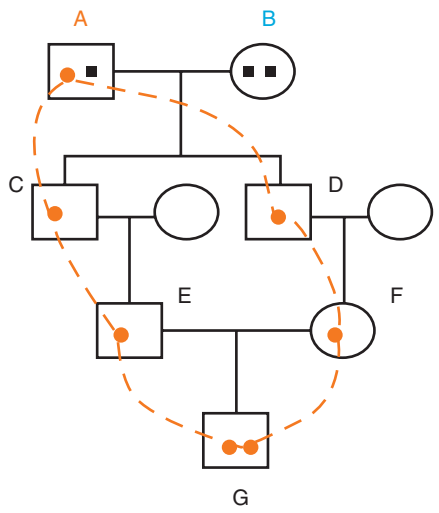

Fig. 4.1. Famille dans laquelle un enfant G est issu d'une union entre des parents cousins germains (E et F).
On observe une boucle de consanguinité et la conséquence génétique est la possibilité qu'un allèle (représenté ici par le point orange) présent chez l'arrière-grand-père A soit transmis en double exemplaire au petit-fils G. Cet allèle est alors identique par descendance ou IBD chez G.

alléliques restent inchangées au cours des générations (on a toujours un équilibre) mais par contre, les proportions génotypiques sont différentes avec un excès d'homozygotes par rapport au modèle panmictique et un déficit d'hétérozygotes. La consanguinité à elle seule ne conduit donc pas à une perte de la diversité allélique.

Écart à la taille infinie – Le phénomène de dérive génétique

Nous allons maintenant considérer la situation où la taille de la population est réduite, ce qui peut par exemple se produire lorsqu'un petit nombre d'individus provenant d'une grande population quitte cette population pour coloniser un nouvel espace, par exemple une île. Dans cette situation, on parle *d'effet fondateur*. La constitution de la nouvelle population ainsi formée va dépendre des allèles portés par les fondateurs. Si tous les fondateurs qui quittent le continent pour conquérir ce nouvel espace ont le même génotype AA, alors dans le nouvel espace, la fréquence de l'allèle A sera de 1 et il y aura eu une perte de la diversité génétique avec fixation de l'allèle A. De manière générale, ce phénomène de taille réduite de la population va conduire à une fluctuation aléatoire de la fréquence des allèles. C'est un phénomène qu'on appelle *dérive génétique*. Cette dérive génétique va aboutir plus ou moins vite dans le temps à une perte du polymorphisme avec la fixation d'un des allèles. Le temps mis pour aboutir à la fixation dépend de la taille de la population : plus la population est petite et plus la fixation d'un allèle se produira rapidement. L'allèle fixé est en général le plus fréquent, mais il est possible aussi que ce soit l'allèle minoritaire qui se fixe par le jeu du hasard (ou encore plus fréquemment si cet allèle est favorisé dans le nouvel environnement). Lorsque l'isolement des populations persiste au cours du temps pendant de nombreuses générations (par exemple du fait de l'existence de barrières géographiques qui empêchent les échanges entre populations), ces différences vont même pouvoir aboutir à la spéciation, et cela en l'absence de toute sélection. La dérive génétique pourrait donc représenter une force évolutive importante comme le postule la théorie neutraliste de l'évolution.

En résumé, une taille réduite de population engendre un phénomène de fluctuation des fréquences alléliques au cours des générations qu'on appelle dérive génétique. Il n'y a donc plus d'équilibre, mais si les unions restent panmictiques, les proportions d'Hardy-Weinberg restent respectées à chaque génération. Si plusieurs populations se séparent, la dérive génétique va engendrer des différences de fréquences alléliques entre ces populations filles.

Écart à l'hypothèse d'absence de sélection

On parle de sélection lorsque la probabilité de transmission des allèles à la descendance varie en fonction du génotype des individus, certains génotypes ayant plus de chances de produire des descendants. La fréquence des

allèles constituant ces génotypes favorisés va donc augmenter au cours des générations. On distingue différentes situations. Si le génotype favorisé est l'un des génotypes homozygotes, alors l'allèle concerné va se fixer et cela d'autant plus vite qu'il est favorisé. C'est ainsi qu'un variant apparu par mutation et initialement présent chez un seul individu peut augmenter en fréquence dans la population jusqu'à envahir toute la population et aboutir à sa fixation et une perte du polymorphisme. Si le génotype favorisé est le génotype hétérozygote (on parle d'un avantage de l'hétérozygote), alors un équilibre pourra s'établir dans lequel les deux allèles seront présents. Un exemple d'avantage de l'hétérozygote chez l'homme est celui qui s'exerce au locus du gène de la bêta-globine qui code l'une des chaînes peptidiques de l'hémoglobine. Un variant S de ce gène est responsable, chez les homozygotes qui le portent, de la drépanocytose ou anémie falciforme. Les porteurs hétérozygotes de cet allèle S sont quant à eux protégés du paludisme. En conséquence, on observe à travers le monde une corrélation entre la présence du paludisme et la fréquence de l'allèle S qui est plus forte dans les zones où le paludisme est endémique. De la même manière, un avantage de l'hétérozygote est soupçonné pour expliquer la fréquence relativement élevée du variant p.Phe508del du gène *CFTR* responsable de la mucoviscidose à l'état homozygote, mais on ne connaît pas la nature de cet avantage même si plusieurs pistes ont été suggérées comme une protection des hétérozygotes vis-à-vis du choléra [4]. Une étude plus récente a permis de dater l'origine et l'expansion du variant p.Phe508del qui semble être associé à la culture campaniforme [5].

Écart à l'hypothèse d'absence de migration ou de mutation

La survenue de mutations comme celle de migrations conduisent à un apport de nouveaux allèles dans la population d'origine et a donc comme conséquence une augmentation de la diversité génétique dans la population.

Les migrations vont par ailleurs contribuer à une certaine homogénéisation des populations séparées par une barrière géographique en créant des flux d'allèles. Elles vont également limiter la consanguinité.

Corrélations entre allèles à différents loci

Après avoir étudié l'évolution des fréquences alléliques à un locus, les généticiens des populations se sont intéressés à décrire ce qui se passait pour les allèles situés à deux loci.

Notion de déséquilibre de liaison

Pour étudier deux loci simultanément, il est nécessaire d'introduire un paramètre noté *D* qui est le déséquilibre de liaison (ou déséquilibre gamétique). On définit le déséquilibre de liaison *D* comme la différence entre la fréquence observée des combinaisons d'allèles et la fréquence attendue si les allèles étaient distribués aléatoirement (qui est le produit des fréquences alléliques).

Évolution du déséquilibre de liaison

Le déséquilibre de liaison évolue d'une génération à l'autre sous l'effet des recombinaisons (*crossing-over*). On s'attend donc à trouver du déséquilibre de liaison surtout entre des loci situés à proximité sur le même chromosome ce qui va alors permettre de définir des *haplotypes*, c'est-à-dire des groupes d'allèles à différents loci situés sur un même chromosome qui vont être transmis ensemble aux descendants (sauf si une recombinaison se produit qui va alors « casser » l'haplotype).

Déséquilibre de liaison et puissance des tests d'association génétique

D'une manière générale, plus le déséquilibre est fort entre les deux loci et plus la puissance pour détecter l'effet de l'un en étudiant l'association d'une maladie avec l'autre sera forte. Avec le séquençage du génome humain et la découverte de nombreux SNIP, la possibilité est apparue de réaliser ces études d'association à l'échelle de tout le génome dans une démarche plus agnostique que l'approche « gène candidat ». Encore fallait-il caractériser les déséquilibres de liaison qui existaient entre les allèles des SNIP dans les populations humaines. C'est ce à quoi s'est attaché le projet HapMap.

Notion de tag SNIP et projet HapMap

Au début des années 2000, le projet international HapMap a cherché à caractériser les SNIP retrouvés le plus fréquemment dans les populations humaines en mesurant les fréquences de leurs allèles dans des échantillons représentatifs des continents européen, africain et asiatique (https://www.genome.gov/10001688/international-hapmap-project). Le projet a également mesuré le déséquilibre de liaison entre les allèles de ces SNIP, et montré qu'il existait des blocs de déséquilibre de liaison ou *blocs haplotypiques* dans lesquels les SNIP présentaient des déséquilibres de liaison forts. À l'intérieur de ces blocs, peu de recombinaisons se sont produites au cours du temps et le nombre d'haplotypes différents est réduit. Un exemple de blocs haplotypiques est donné dans la figure 4.2. La connaissance des allèles présents sur un des SNIP dans un bloc haplotypique permet d'obtenir des informations sur les allèles probablement présents aux autres SNIP du bloc (ainsi, sur la figure 4.2, dans la population africaine YRI, on voit que si dans le premier bloc, un individu porte un allèle G au premier SNIP, il portera un allèle C aux deux autres SNIP du bloc).

À partir des résultats du projet HapMap, il a donc été possible de déterminer le nombre minimum de SNIP nécessaire pour représenter tous les variants communs du génome et pouvoir réaliser des tests d'association pangénomique (*Genome-Wide Association Study* ou GWAS). Ceci a permis la réalisation de puces de SNIP (ou « SNIP-chip »)

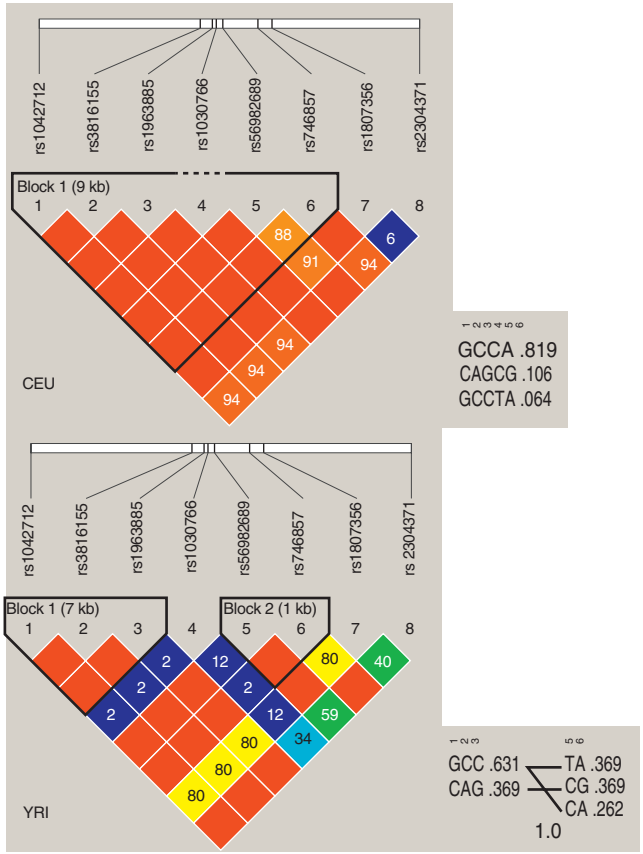

Fig. 4.2. Exemple de blocs haplotypiques dans le gène LCT (chromosome 2) qui code la lactase.
Les couleurs indiquent la force du déséquilibre de liaison (le rouge correspondant aux déséquilibres de liaison les plus forts). Ces déséquilibres sont donnés ici en valeur de D', qui est une mesure du rapport entre le déséquilibre de liaison observé et le déséquilibre de liaison maximum possible étant donné les fréquences alléliques aux loci considérés. Les valeurs sont indiquées dans les différentes cases en pourcentage (%) (lorsque aucun chiffre ne figure dans la case, D' est de 100 %). Dans la partie droite de la figure, on voit les haplotypes observés et leurs fréquences. Les résultats sont donnés pour deux populations du projet 1 000 Génomes, la population européenne (CEU) et la population africaine yoruba (YRI). Les analyses et les représentations graphiques ont été réalisées à l'aide du logiciel Haploview (https://www.broadinstitute.org/haploview/haploview). On observe deux blocs haplotypiques dans la population YRI et un seul dans la population CEU.

permettant de génotyper simultanément tous ces SNIP chez des individus.

Les blocs haplotypiques peuvent être différents d'une population à l'autre, surtout en termes d'étendue puisque les blocs sont généralement plus courts dans les populations africaines plus anciennes que dans les populations européennes, ou encore que dans les populations asiatiques où les blocs sont souvent les plus longs. Ces différences n'ont cependant pas été prises en compte initialement dans la construction des puces de SNIP qui ont été optimisées pour les populations européennes dans lesquelles la plupart des études GWAS ont été réalisées. Plus récemment, différentes initiatives ont été lancées, comme le projet H3Africa, pour mieux caractériser les populations africaines et construire des puces de SNIP mieux adaptées à ces populations (https://h3africa.org/).

Diversité génétique des populations humaines

Après le projet HapMap qui avait déjà permis d'approcher la diversité génétique des populations, différents projets ont été lancés pour décrire la structure des populations et étudier cette diversité dans les populations naturelles. On a pu ainsi mieux comprendre le rôle des différentes forces évolutives que sont la dérive, les migrations, les mutations et la sélection dans la mise en place de cette diversité. Les études de génétique des populations ont montré en particulier l'importance des migrations et du métissage et ont banni l'idée des races. Ces études apportent également un éclairage sur les maladies et leur distribution géographique.

Différences de fréquences alléliques entre populations humaines

Différences pour les gènes du système HLA

Les gènes du complexe majeur d'histocompatibilité humain ou *human leucocyte antigen* (HLA) situés sur le chromosome 6 présentent des niveaux de polymorphismes très importants, et d'importantes différences de fréquences alléliques sont observées ente populations humaines. Ainsi, certains allèles peuvent être absents ou très rares dans certaines populations, et retrouvés avec des fréquences non négligeables dans d'autres populations. C'est par exemple le cas des allèles HLA-B*1502 et HLA-B*5801 qui sont quasiment absents de la plupart des populations mais ont des fréquences de l'ordre de 10 % dans les populations d'Asie de l'Est. Ces allèles sont fortement associés à des réactions cutanées très sévères (syndrome de Lyell et syndrome de Stevens-Johnson) à des médicaments comme la carbamazépine pour le premier et l'allopurinol pour le second. Dans la population chinoise Han, tous les individus qui présentent des réactions à ces médicaments sont porteurs de ces allèles HLA-B, mais on retrouve aussi ces réactions dans les populations européennes chez des individus qui ne portent pas ces allèles [6].

Ces différences de distribution des allèles HLA sont le produit de l'histoire des populations et de phénomènes de sélection qui ont pu, à certains moments, favoriser les individus porteurs de certains allèles dans certaines régions géographiques. Il est également établi qu'une proportion importante des variants des gènes HLA pourrait venir des hommes de *Néandertal*, qui on le sait aujourd'hui se sont mélangés avec nos ancêtres *sapiens*.

Différences sur d'autres loci et notion de marqueurs d'ancestralité

Ces différences géographiques qu'on observe sur des gènes très polymorphes comme ceux du système HLA se retrouvent aussi sur d'autres marqueurs du génome. En analysant des données génétiques obtenues par le génotypage de puces de SNIP, on a pu montrer que les distances génétiques entre populations étaient corrélées aux distances géographiques

selon le principe de l'isolement par la distance. Ainsi, à partir des données du *Human Genome Diversity Project* (HGDP) qui a échantillonné des individus issus de 52 populations à travers le monde [7], on peut mettre en évidence, lorsqu'on étudie simultanément de nombreux SNIP, que les populations se regroupent en fonction de leur continent d'origine (fig. 4.3) [8], mais forment un continuum montrant bien une origine commune. Certains marqueurs peuvent présenter des différences de fréquences alléliques plus marquées entre populations et, en génotypant ces marqueurs, on pourra donc obtenir des informations sur l'origine géographique la plus probable des individus. Ces marqueurs sont appelés AIM pour *ancestry informative markers*. Des panels de SNIP pouvant servir d'AIM ont été définis pour distinguer les origines continentales des individus [9]. Ainsi, avec une trentaine de marqueurs, on pourra distinguer les populations africaines, européennes et asiatiques. Des panels d'AIM ont également été proposés pour déterminer les régions d'origine en Europe des individus, mais ces inférences sont plus difficiles à l'échelle d'un continent car, même si en réalisant une analyse en composantes principales des données génétiques on observe que les individus originaires d'un même pays ont tendance à se regrouper, les différences sont plus subtiles [10]. De plus, comme nous l'avons vu, le métissage a joué un rôle important dans les populations humaines, et vouloir assigner un individu à un groupe n'a pas de sens. Aujourd'hui, les compagnies qui proposent d'utiliser l'information génétique pour retrouver les origines des individus donnent plutôt les résultats en pourcentage du génome de différentes origines. Cependant, pour réaliser ces inférences d'origine, il faut comparer les données individuelles à des panels de référence qui peuvent ne pas être représentatifs de toutes les populations humaines et conduire à des résultats biaisés. Leur interprétation peut donc être sujette à caution pour un public non averti [11].

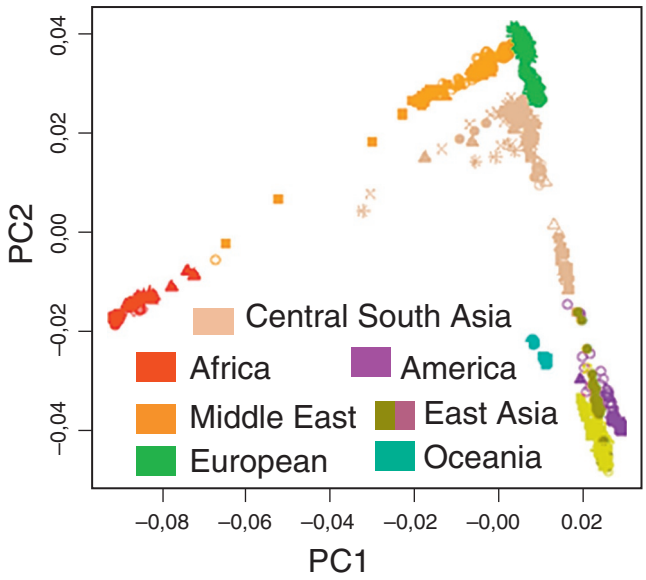

Fig. 4.3. Analyse en composantes principales (ACP) des données génotypiques des 940 individus du panel HGDP.
Les valeurs des deux premières composantes principales sont données pour chacun des individus. Ces deux premières composantes principales (PC1 et PC2) capturent respectivement 27,7 % et 20,6 % de la variance des génotypes. Les couleurs correspondent aux régions géographiques d'origine des individus. L'ACP présentée a été réalisée à partir des données génotypiques sur 647 137 SNIP répartis sur les 22 autosomes en utilisant le logiciel EIGENSOFT/smartpca (https://www.hsph.harvard.edu/alkes-price/software/) avec l'option de prise en compte du déséquilibre de liaison.
Source : modifié à partir de Li JZ, Abshe, D, Tang H, Southwick AM, Casto AM, Ramachandran S et al. Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. Science 2008;319:1100–4 [8].

Références

[1] Nakamura Y. DNA variations in human and medical genetics: 25 years of my experience. J Hum Genet 2009;54:1–8.
[2] Hardy GH. Mendelian Proportions in a Mixed Population. Science 1908;28:49–50.
[3] Weinberg W. Über den Nachweis der Vererbung beim Menschen. Jahresh Ver Vaterl Naturkd Württemb 1908;64:369–82.
[4] Rodman DM, Zamudio S. The cystic fibrosis heterozygote – Advantage in surviving cholera? Med Hypotheses 1991;36:253–8.
[5] Farrell P, Férec C, Macek M, Frischer T, Renner S, Riss K, et al. Estimating the age of p.(Phe508del) with family studies of geographically distinct European populations and the early spread of cystic fibrosis. Eur J Human Genet 2018;26:1832–9.
[6] Génin E, Schumacher M, Roujeau JC, Naldi L, Liss Y, Kazma R, et al. Genome-wide association study of Stevens-Johnson Syndrome and Toxic Epidermal Necrolysis in Europe. Orphanet J Rare Dis 2011;6:52.
[7] Cann H, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, et al. A Human Genome Diversity Cell Line Panel. Science 2002;296:261–2.
[8] Li JZ, Abshe D, Tang H, Southwick AM, Casto AM, Ramachandran S, et al. Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. Science 2008;319:1100–4.
[9] Kosoy R, Nassir R, Tian C, White PA, Butler LM, Silva G, et al. Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. Hum Mutat 2009;30:69–78.
[10] Heath SC, Gut IG, Brennan P, McKay JD, Bencko V, Fabianova E, et al. Investigation of the fine structure of European populations with applications to disease association studies. Eur J Hum Genet 2008;16:1413–29.
[11] Bonniol JL, Darlu P. L'ADN au service d'une nouvelle quête des ancêtres? Civilisations 2014;63:201–19.