Interpréter la diversité humaine

Entretien avec Bertrand Jordan

Claude-Olivier DORON et Jean-Paul LALLEMAND-STEMPAK

L'étude de marqueurs polymorphiques dans l'ADN humain a ouvert la voie à de nouveaux modes d'interprétation de la diversité humaine aux applications très diverses. Mais comment ces interprétations sont-elles construites ? La nouveauté technique ne cache-t-elle le vieux concept de race ?

La vie des idées : Grand vulgarisateur des recherches actuelles sur les polymorphismes humains sur lesquelles vous avez notamment publié *L'humanité au pluriel : la génétique et la question des races* (Seuil, 2008), vous n'êtes pas vous-même praticien de ce champ. Comment en êtes-vous venu à travailler sur la question de la diversité humaine d'un point de vue génétique en partant d'une thèse de physique nucléaire ?

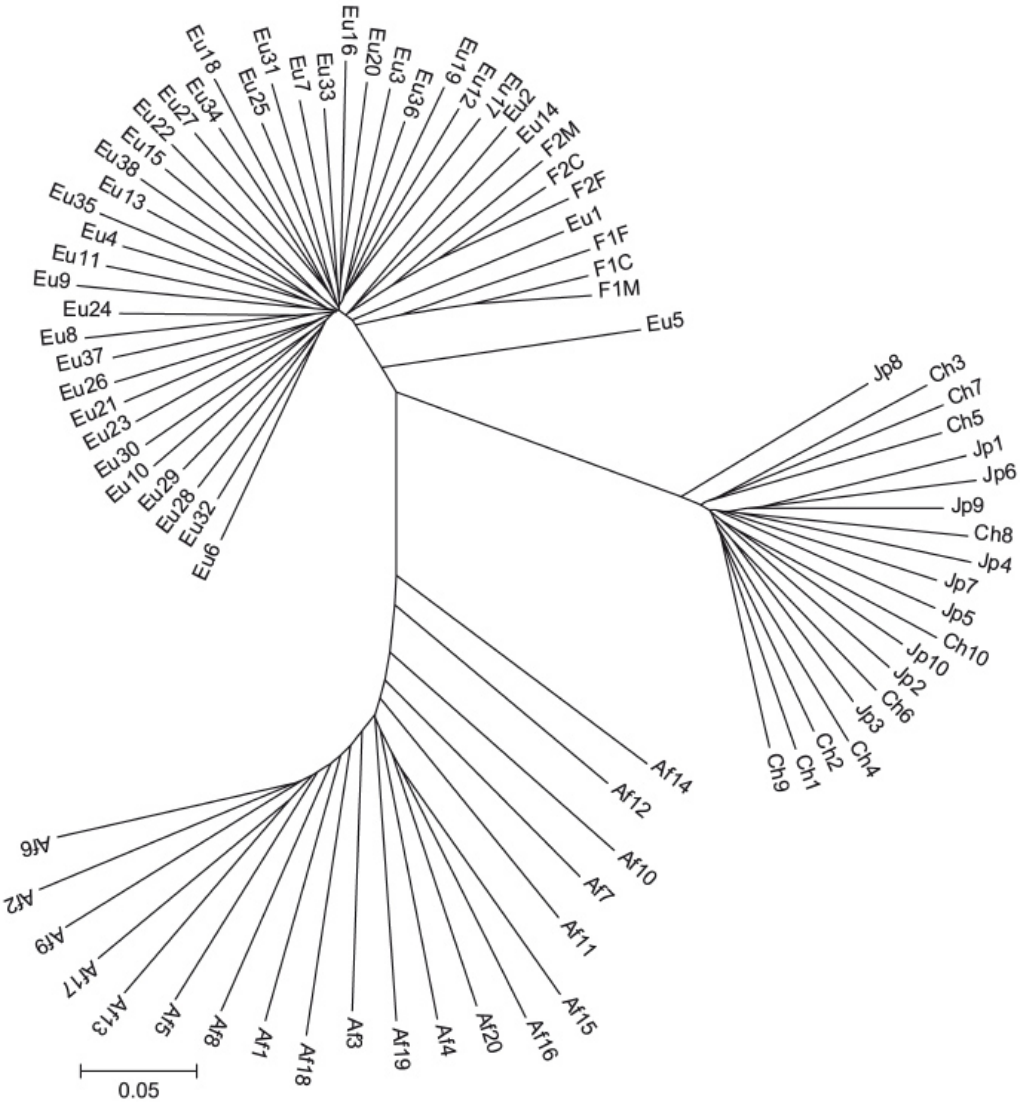
Bertrand Jordan : J'ai en effet commencé par faire une thèse en physique des particules, mais je me suis tourné vers la biologie moléculaire dès le milieu des années 1960, en raison des types très différents de recherches que ces deux domaines impliquaient. Je me suis rendu compte que la physique des particules – à l'époque déjà, et encore plus maintenant – était vraiment de la « *big science* », pratiquée par de très grands groupes, séparant la théorie et l'expérience, et dans laquelle le rôle de l'individu n'était pas évident. On y faisait une expérience par an et ce n'était pas comme ça que j'avais envie de faire de la recherche.

À l'époque, c'était le début du développement de la biologie moléculaire. La double hélice date de 1953 et j'ai soutenu ma thèse de Physique en 1965, au moment où l'on venait juste de déchiffrer le code génétique – la correspondance entre les triplets dans l'ADN et les acides aminés dans les protéines. Il était clair que c'était un domaine qui allait bien se développer et qui était intéressant. De plus, il y avait à l'époque la DGRST (Délégation générale de la recherche scientifique et technique) qui a joué un rôle important dans le développement de la recherche moléculaire en France. Elle cherchait en particulier à attirer vers la biologie moléculaire des gens ayant une formation de physique ou de chimie et non pas de naturalistes. Ça a donc été très facile pour moi de changer de voie. J'ai eu une bourse de reconversion, puis je suis entré au CNRS et j'ai fait une carrière de biologiste moléculaire.

La question du racisme et de la diversité humaine m'a, quant à elle, toujours intéressé. Je me suis éveillé à la politique avec la guerre d'Algérie – qui renvoyait forcément à des problèmes de racisme – et j'ai aussi vécu mai 68 : c'était donc un sujet qui me concernait.

En fait, j'ai découvert ce qui s'était passé en recherches génétiques sur la diversité humaine à l'occasion d'un congrès de cancérologie à Charleston, en 2006. Un des grands

représentants de ce domaine, Shriver,[1] proposait des méthodes pour rendre les essais cliniques plus rationnels en évitant d'avoir une population de malades et une population de témoins qui aient des origines différentes.[2] Il y présentait des schémas (illustration 1)[3] – qui m'ont alors beaucoup frappé.



Jusque-là, je n'imaginais pas qu'à partir d'une analyse d'ADN, on pouvait remonter à l'ascendance d'une personne. À l'époque, on parlait uniquement des empreintes génétiques. On a dû me demander, en 2004 ou 2005, si par une analyse de l'ADN on pouvait connaître la « race », la « population d'origine » et j'ai répondu « non ». C'était ce que pensaient les gens à l'époque, y compris les chercheurs. En fait, les possibilités d'analyses sont relativement récentes. Ce n'est qu'à partir du moment où on a eu les moyens de regarder un grand nombre

[1] Mark Shriver est un généticien, chercheur au laboratoire d'anthropologie de la Penn State University, réputé en particulier pour avoir développé des techniques, fondées sur un ensemble de marqueurs informatifs d'ascendance, permettant d'estimer le mixte d'ascendances biogéographiques d'un individu donné. Ces techniques furent utilisées en particulier en médecine légale et dans les entreprises de généalogies génétiques commerciales. Il s'intéresse par ailleurs aux déterminants génétiques de la pigmentation de la peau et de l'iris.

[2] On s'efforce, dans un tel cas, de préciser la structure et la composition génétique des populations cas et témoin, à travers des techniques comparables à celles évoquées précédemment (*admixture mapping*).

[3] L'article dont est tirée cette illustration est disponible à l'adresse :
http://www.humgenomics.com/content/pdf/1479-7364-1-4-274.pdf

de marqueurs polymorphiques dans l'ADN d'un grand nombre de personnes qu'on a pu dire quelque chose sur leur origine[4]. C'était quelque chose de relativement nouveau mais je pense que c'était bien plus nouveau pour moi que ça ne l'était pour les chercheurs américains.[5]

100 euros pour observer un million de points dans l'ADN d'une personne

La vie des idées : Comment utilise-t-on ces marqueurs polymorphiques pour établir l'ascendance d'un individu ?

Bertrand Jordan : Au sein de la population humaine, qui est particulièrement homogène (par rapport à la population de chimpanzés ou la population de gorilles, par exemple) mais qui présente quand même une certaine diversité, si vous séquencez intégralement mon ADN et le vôtre, on va trouver un certain nombre de différences (quelques millions). Si on se limite aux différences ponctuelles[6]– car il existe aussi des différences sur des petits segments qui sont dupliqués ou qui sont absents chez l'un ou chez l'autre mais qui fonctionnent à peu près de la même façon[7] –, il y a quelques millions de points dans votre ADN où il y a une base différente de celle qu'on va trouver, au même endroit, dans le mien. La très grande majorité de ces variations n'a aucune influence parce qu'elles se situent en dehors des gènes, dans des régions qui ne sont pas critiques – les gènes n'occupant que quelques pourcents de notre ADN. La plupart du reste de notre ADN n'a pas de fonction précise, même si certaines séquences jouent un rôle de régulation, etc. La majorité de ces millions de différences n'a donc aucune conséquence sur notre phénotype. Et puis il y en a quelques-unes, on pense maintenant que ça tourne autour d'une ou deux centaines, qui touchent effectivement des gènes. Mais, il ne suffit pas qu'elles les touchent, il faut encore qu'elles modifient de façon significative la protéine qui est codée par le gène. Parce que, là encore, dans une protéine, tout n'est pas critique. Il y a des endroits où vous pouvez changer un acide aminé pour un autre sans altérer le fonctionnement de la protéine. Il y a d'autres endroits où ça va, au contraire, modifier ses propriétés. La centaine de différences génétiques réellement significatives qui peut exister entre deux personnes, est ce qui est responsable des différences d'aspect, de taille, de couleur de peau et de fonctionnement métabolique.

Qu'est-ce qu'un single-nucleotide polymorphism (SNP) ou polymorphisme nucléotidique?

Notre ADN est constitué de plus de 3,3 milliards de paires de bases azotées (guanine, cytosine, adénine, thymine). Un SNP ou polymorphisme nucléotidique désigne une variation ponctuelle sur une base. Si, par exemple, nous avons sur le chromosome 1 d'un individu, en un emplacement donné, la série AACCTT et pour un autre, au même emplacement, la série AAC**T**TT, on parlera d'un polymorphisme. Les polymorphismes nucléotidiques sont relativement fréquents dans le génome (environ 15 millions par génome, soit plus 1 fois/300 nucléotides), et affectent dans leur majorité les séquences non-codantes de l'ADN, ce qui signifie qu'elles n'ont généralement aucun impact sur le phénotype. Par contre, certains polymorphismes (ou groupes de polymorphismes) ont des fréquences qui varient de manière significative selon les populations humaines. Lorsqu'un

[4] Ces évolutions sont intimement liées aux progrès techniques considérables en matière de séquençage du génome qui ont eu lieu depuis les années 1990 et au développement de logiciels informatiques permettant de traiter une masse très importante de données.

[5] Les premiers travaux de Shriver sur le sujet datent par exemple de 1997 (Shriver, M. D., Smith, M. W., Jin, L., Marcini, A., Akey, J. M., Deka, R., & Ferrell, R. E. (1997). "Ethnic-affiliation estimation by use of population-specific DNA markers." *American Journal of Human Genetics*, *60* (4), 957).

[6] Ou Single-Nucleotide Polymorphism (SNP).

[7] Ce qu'on appelle des Short Tandem Repeats ou microsatellites, c'est-à-dire des séquences de 2 à 10 nucléotides qui peuvent se répéter un nombre de fois variable.

polymorphisme apparaît avec, en moyenne, une fréquence au moins 30-50% plus importante dans une population que dans une autre, il est décrit comme un marqueur informatif de l'ascendance (AIM). Une sélection d'AIMs peut ainsi permettre d'assigner avec une forte probabilité un individu à un groupe d'ascendance.

Pour estimer les ascendances d'un individu, on ne se focalise pas du tout sur les gènes : on regarde l'ensemble. Aujourd'hui, on peut techniquement observer très facilement un million de points dans l'ADN d'une personne pour une centaine d'euros. Là où c'est plus compliqué, c'est que ces différences ne se trouvent pas n'importe où. Il y a des points « polymorphes » où l'on observe plus souvent une différence d'une personne à l'autre dans une population. Dans une population donnée, vous allez trouver 20% de personnes qui ont à cet endroit un G et 80% de personnes qui ont à cet endroit un A (la séquence ADN est composée de quatre bases puriques, la guanine (G), l'adénine (A), la cytosine (C) et la thymine (T). Ce sont ces points, souvent hautement variables, que l'on regarde. Plutôt que de lire intégralement votre ADN et le mien, ce qu'on commence d'ailleurs à pouvoir faire à un prix accessible, on regarde les endroits où on a le plus de chances de trouver des variations.

Par rapport à la question des ascendances ou des « races », on peut premièrement se poser la question : « est-ce qu'il y a des endroits où il y a *toujours* une certaine base dans l'ADN d'un Européen et une autre dans celui de quelqu'un d'origine africaine ? ». La réponse est « *non* ». Il n'y a pas d'allèle qui soit *spécifique* d'une population donnée à 100% (On parle « d'allèle » car il y a un point dans l'ADN qui peut exister sous au moins deux formes : deux « allèles »). Cela, bien sûr, si on regarde un nombre d'individus suffisant, car si on regarde dix personnes, il y a certains allèles qui ne seront pas représentés dans une population et seront représentés dans une autre. Mais si on regarde un nombre suffisant de personnes, on retrouve presque toute la diversité humaine dans n'importe quelle population. Par contre, ce qu'on voit, c'est que les *fréquences* des deux allèles varient parfois selon les populations.

Si vous considérez un point donné dans l'ADN qui est polymorphique, et que vous regardez une population 'bretonne', vous allez par exemple trouver à peu près aussi souvent un allèle que l'autre : soit 50% A et 50% G. Si vous regardez une tribu papoue de Papouasie orientale, vous trouverez encore les deux allèles mais vous trouverez peut-être 20% de l'un et 80% de l'autre. Ceci est vrai pour une partie des points variables, pour à peu près 10% d'entre eux – enfin, c'est une question de degré. En tout cas, il y a une petite partie des points variables pour lesquels la répartition des deux variants diffère selon la population. Bien entendu, regarder un variant ne vous suffit pas pour rattacher une personne à une population. Mais si vous trouvez un moyen de choisir ceux qui sont les plus variables dans une population et d'en regarder 1000 ou 2000 à la fois, vous allez pouvoir à ce moment-là – avec une assez bonne probabilité – dire : « cet ADN là provient de quelqu'un qui a des ancêtres européens » ou bien « cet ADN là provient de quelqu'un qui a des ancêtres africains ».

Le problème des groupes de références

La vie des idées : Mais comment sont déterminés au préalable ces *groupes de référence* auxquels on rapporte les individus ? Il y a plusieurs questions en lien avec ça. La première est immédiate : « combien d'individus prend-on pour définir une population de référence ? ». Vous disiez qu'il faut observer un nombre suffisant de génomes individuels pour identifier les allèles caractéristiques d'une population donnée, mais à combien faut-il évaluer ce nombre suffisant ? En effet, souvent, c'est sur la base du génome de quelques individus seulement qu'on va définir « une population ». Qu'est-ce que ça peut poser comme type de problèmes ?

Bertrand Jordan : Tout commence par un problème insoluble : c'est qu'on n'a pas accès aux « populations ancestrales ». Par définition, ce qu'on souhaiterait avoir comme référence, ce sont les populations humaines telles qu'elles existaient il y a 500 ou 1000 ans, avant qu'il n'y ait le mélange particulièrement important lié à la colonisation, aux voyages, etc. Ces populations n'étant plus là, on ne dispose pas de leur ADN. On prend donc comme référence des populations dont on a des raisons de penser qu'elles sont proches des populations ancestrales – parce qu'elles ont peu bougé, qu'il n'y a pas eu d'immigration – par exemple des Yorubas du Nigeria.[8]

La vie des idées : Quand on dit qu'on dispose de l'ADN des Yorubas du Nigeria, ne se base-t-on pas sur quelques individus seulement ? Cela ne pose-t-il pas problème en termes d'*inférence statistique* ?

Bertrand Jordan : Oui, je dirais que le nombre d'individus observés est de l'ordre de la centaine. La base de données officielle du *Human Genome Diversity Panel* contient ainsi les échantillons tirés de 1050 individus issus de 52 populations.[9] Maintenant, les entreprises et les chercheurs essayent d'élargir ce domaine et d'avoir des populations de référence pour permettre de faire des attributions plus fines. Mais, on ne sait pas très bien ce qu'il y a dans les panels constitués par les entreprises.

La vie des idées : Par ailleurs, par rapport à ces populations de référence, une autre question vient nécessairement à l'esprit : comment sont-elles *nommées et délimitées* ? On se déclare Yoruba quand on nous fait un prélèvement génétique ou bien s'agit-il d'anthropologues qui ont déterminé, au préalable, ce qu'était une ethnie Yoruba, une ascendance biogéographique Yoruba (ou bretonne, auvergnate, etc.) ?

Bertrand Jordan : Ça n'est certainement pas auto-déclaré, dans ce cas du moins. Dès les débuts du programme « génome », il y a eu une tentative de lancer un programme visant à cartographier la diversité humaine au niveau génétique. C'était Cavalli-Sforza[10] qui s'en occupait à cette époque et le projet a eu beaucoup de difficultés à se mettre en place – justement parce que ça posait un certain nombre de problèmes éthiques, politiques, etc.[11] On craignait par exemple que le recueil de données génétiques puisse servir de base à une discrimination ou éventuellement à une exploitation. Il y avait aussi la question de la propriété des données génétiques : il y a eu un certain nombre de cas – celui des Pima, par exemple, une tribu indienne aux États-Unis – où les gens étaient d'accord pour donner leur sang pour faire des études sur le diabète mais où les échantillons auraient été utilisés ensuite pour des études anthropologiques pour lesquelles les gens n'avaient pas donné un accord préalable.

---

[8] Ce que l'anthropologie et la génétique des populations ont longtemps qualifié d'isolats, c'est-à-dire des populations relativement fermées (du fait d'obstacles géographiques et sociaux) présentant une composition génétique relativement plus homogène.

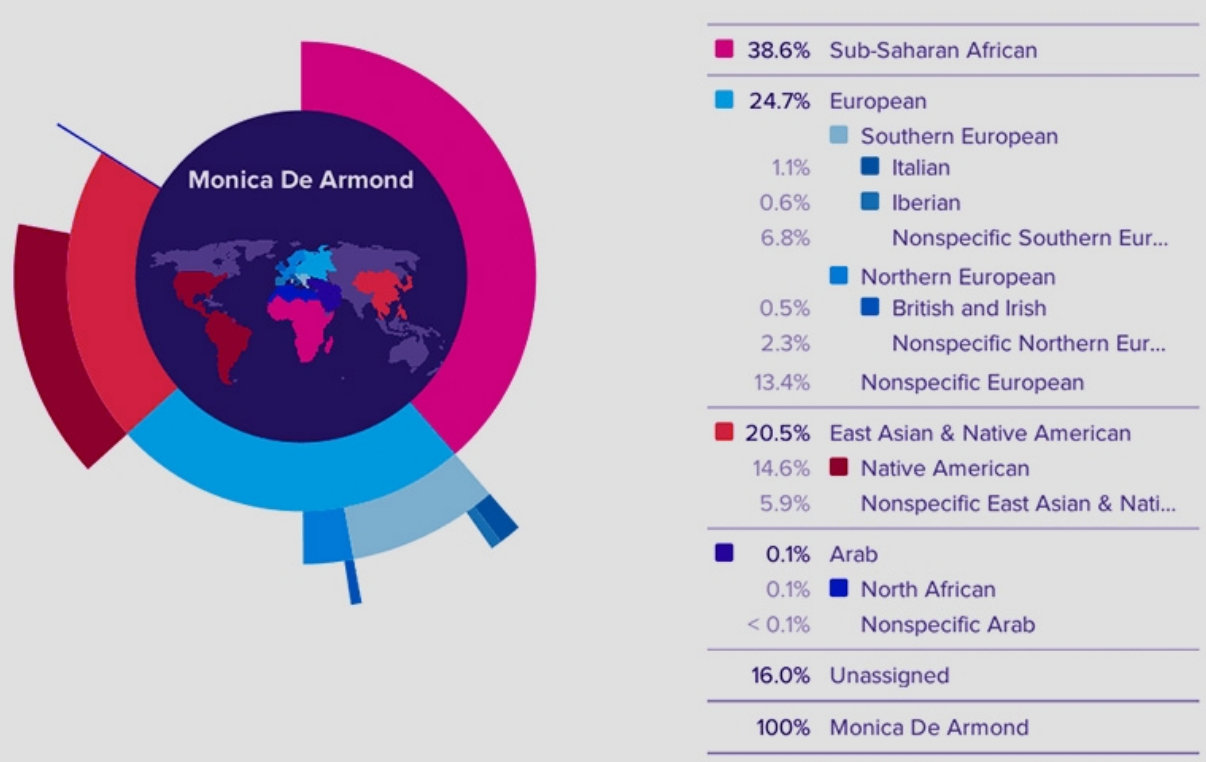
[9] À titre d'exemple, la « population Yoruba » est représentée par 25 individus.
Voir http://www.cephb.fr/en/hgdp/table.php

[10] Luca Cavalli-Sforza est l'un des pères de la génétique des populations humaines. Il a joué un rôle essentiel dans le développement des recherches visant à cartographier la distribution statistique et la stratification des variants génétiques au niveau géographique et à établir des corrélations entre ces distributions et des mouvements migratoires, des transferts de technologies etc. Il a par ailleurs développé, en particulier avec Anthony Edwards, des arbres phylogénétiques permettant de décrire l'évolution des populations humaines.

[11] Voir Reardon, Jenny, *Race to the finish : identity and governance in an age of genomics*, Princeton University Press, 2004 et M'Charek, Amade, *The human genome diversity project: an ethnography of scientific practice*, Cambridge University Press, 2005

La vie des idées : Il s'agit donc d'assigner un individu à des populations auxquelles on donne des noms. Or, ces populations sont elles-mêmes déterminées par des individus dont on a séquencé le génome. Mais ces individus, que représentent-ils ? Il y a eu un débat de cet ordre il y a quelques temps à propos de Tokyoïtes dont on avait prélevé les données génétiques : que représentent-ils ? Les Tokyoïtes ? Les Japonais ? Les habitants du Sud-est asiatique ?

Bertrand Jordan : Si vous regardez l'illustration (2) créée par la compagnie 23&me[12], vous trouvez des dénominations « Italien » ou « Européens du sud ». Ces indications sont là surtout pour faire joli, en réalité. En effet, si vous regardez les « Européens du sud » – à supposer que ce soit réellement valable – ils représentent tout le secteur en bleu clair. Le secteur bleu est la partie des Européens et parmi eux, le bleu clair est ce qu'ils ont pu attribuer aux Européens du sud, et le bleu plus sombre aux Européens du nord. C'est-à-dire que le reste, ils peuvent dire que c'est « européen » mais ils ne peuvent pas vraiment distinguer. Si, ensuite, vous regardez parmi les Européens du sud, il y a une part du génome qui est attribuée aux Italiens, une autre part qui est attribuée aux Ibériques et le reste qui n'est pas attribué. Cela revient donc à utiliser une population relativement restreinte pour pouvoir dire quelque chose – si on trouve des marqueurs qui permettent de la désigner – mais sans être exhaustif.



[12] 23andMe, créée en 2006, est l'une des plus importantes entreprises de généalogie génétique en ligne. Elle proposait par ailleurs des services de diagnostic génétique en ligne qui ont été suspendus en novembre 2013, après le refus de la Food and Drug Administration de leur accorder une autorisation. Voir https://www.23andme.com/. On peut consulter un grand nombre de vidéos de personnes recevant leurs tests ADN sur youtube : voir par exemple http://www.youtube.com/watch?v=oV-t0nwQ3uo

La vie des idées : Et sans s'interroger sur les effets de nomination qui vont avec ? Il y a pourtant de très lourds enjeux d'identification derrière – notamment dans le cas d'entreprises comme celle-là qui sont des entreprises de généalogie génétique personnelle.[13]

Bertrand Jordan : Ici, ils ont pris un exemple particulièrement mélangé puisque vous avez des ascendances africaines, européennes et amérindiennes pour le même individu. Mais pour ce genre d'étude, la motivation des gens qui se considèrent comme Européens aux Etats-Unis, c'est essayer de voir de quelle région d'Europe ils proviennent.[14] De la même manière, l'intérêt de ceux qui se considèrent comme Afro-Américains, c'est de voir de quelle région d'Afrique ils proviennent. À ce moment-là, effectivement, le rattachement aux Yoruba ne leur apprendra pas grande chose parce que, ce qu'il faudrait, c'est avoir au moins une dizaine de populations bien caractérisées et bien stables d'Afrique sub-saharienne.[15]

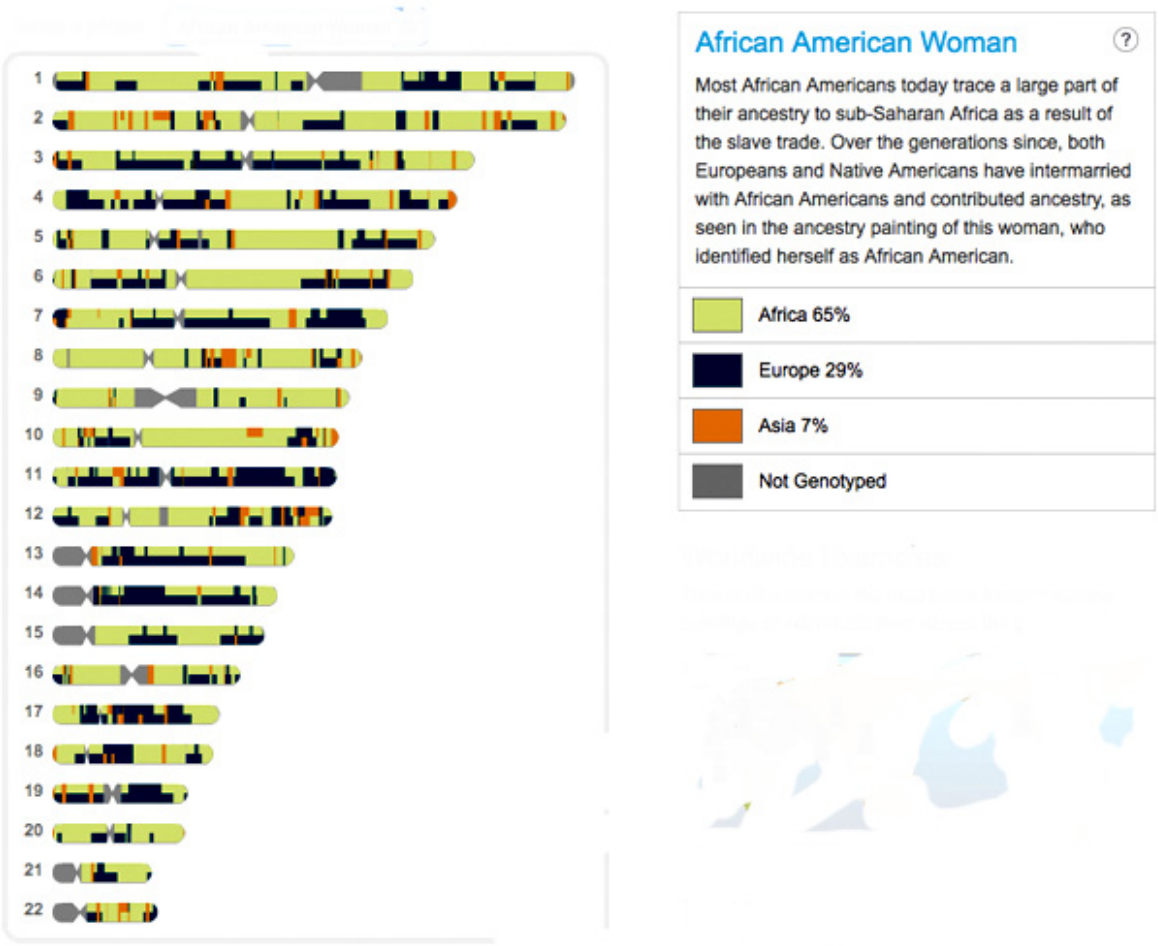
Les pourcentages indiqués sont censés signifier la part du génome qui peut être rattachée à telle ou telle ascendance. Mais il s'agit en fait de marqueurs d'assignation statistique à des populations de référence. Si vous prenez l'illustration 3 – qui provient elle aussi de 23andMe, vous voyez la représentation des chromosomes d'une personne. Sur chaque chromosome, il y a une couleur qui indique la population à laquelle chaque fraction de chromosome a été rattachée. C'est divisé en deux puisque chaque chromosome est en double (un paternel, un maternel). Par exemple, on voit que la région au début du chromosome 1 est rattachée à une population européenne sur l'un des deux chromosomes et à une population africaine sur l'autre. Cette figure est intéressante car elle montre le mélange à l'échelle des chromosomes. Quand on dira, sur cette personne, qu'il y a 29% d'européen, ça signifie que sur l'ensemble du génome, vous avez pu rattacher 29% de la longueur de l'ADN à une population européenne.

---

[13] Sur ces enjeux identitaires, voir les sources évoquées précédemment ainsi que Doron, Claude-Olivier, « L'ascendance biogéographique : génétique des populations et généalogie des individus », à paraître in Luciani, Isabelle & Piétri, Valérie (dir.), *L'incorporation des ancêtres*, Presses Universitaires Aix-Marseille, 2014.

[14] Nash, Catherine, *Of Irish Descent: Origin Stories, Genealogy and the Politics of Belonging*, Syracuse Univ. Press, 2008 pour la diaspora irlandaise.

[15] Les entreprises de généalogies génétiques se distinguent précisément les unes des autres en mettant en avant leurs bases de données plus ou moins fournies sur telle ou telle région et la précision à laquelle elles peuvent arriver à partir de là. Par exemple, des entreprises spécialisées comme AfricanDNA et African Ancestry disposent de bases de données beaucoup plus conséquentes concernant les populations du continent africain.



Mais vous pouvez avoir 29% d'européen avec des profils de chromosomes très différents. Un profil chromosomique peut être marqueur d'une ascendance très ancienne ou bien d'un mixte très récent. Vraisemblablement, il s'agirait ici d'une ascendance relativement ancienne parce que c'est bien morcelé et réparti sur l'ensemble du génome. Mais, si cette figure concernait l'enfant d'un couple européen-africain « pur », à ce moment-là, vous pourriez avoir la même répartition en termes de pourcentage mais avec un chromosome qui serait pour moitié européen et pour moitié africain, ou bien ¾ l'un et ¼ l'autre, ou même un chromosome entier qui serait d'une origine ou de l'autre. Ça aboutirait au même pourcentage avec une distribution différente de gènes.

Concrètement, cela signifie qu'on peut évaluer – plus ou moins – le moment temporel où s'est fait le mélange, sachant qu'il se produit en moyenne une ou deux recombinaisons par chromosome à chaque génération. Donc, tout se morcelle davantage avec le temps. Et cette remarque qui vaut au niveau individuel vaut aussi au niveau d'une population. On peut aussi évaluer – avec les mêmes techniques – le moment où s'est fait le mélange de populations différentes dans une population actuelle.[16]

Autres tests, autres raisonnements

La vie des idées : Il y a deux manières différentes d'évaluer l'ascendance génétique. D'un côté, ces études que nous venons de voir et qui sont faites sur l'ensemble des chromosomes.

[16] C'est d'ailleurs l'un des principes de la technique dite « *admixture mapping* ». Elle fait néanmoins intervenir toute une série d'hypothèses. Voir, sur ce point, notre essai bibliographique.

D'un autre, des techniques d'analyse qui ne sont pas faites sur l'ensemble des chromosomes mais deux types de matériel génétique bien spécifiques : l'ADN mitochondrial et le chromosome Y. Pouvez-vous préciser le type d'informations que leur analyse permet de faire apparaître et quelles sont leurs limites ?

Bertrand Jordan : Ce dernier type d'études est plus facile techniquement et donc pratiqué depuis plus longtemps. C'est ce que faisait – il y a encore peu de temps – la majorité des entreprises proposant un profil d'ascendance. Elle consiste à regarder soit l'ADN mitochondrial, soit le chromosome Y. L'ADN mitochondrial est un petit ADN qui est contenu dans les mitochondries, des bactéries symbiotiques assimilées il y a bien longtemps dans l'organisme humain. Elles ont leur propre génome – un petit ADN ayant des régions hautement variables – et proviennent uniquement de la mère. En effet les mitochondries sont présentes dans le cytoplasme de la cellule, pas dans le noyau, et l'ovule en contient contrairement aux spermatozoïdes : les mitochondries de l'embryon sont donc apportées par la mère. Mes mitochondries proviennent de ma mère, qui les tenait de sa mère, qui les tenait de sa mère à elle, et ainsi de suite. En regardant l'ADN des mitochondries, on peut donc avoir des informations sur l'ascendance mais uniquement à travers la lignée maternelle directe.[17] C'est relativement facile à faire car l'ADN mitochondrial est petit. Il mesure 16 500 nucléotides, la région variable en couvre quelques centaines, et ça fait longtemps qu'on peut les étudier. C'est ainsi qu'on a pu en venir à l'histoire de l'Eve mitochondriale, « notre mère à tous »[18]. Ça a été l'une des premières indications permettant de dire qu'on venait tous d'Afrique. Depuis, cette idée a été renforcée par un tas d'autres données. Mais c'est une façon de faire qui ne nous donne des informations que sur la lignée maternelle.

Dans notre ADN, il y a le chromosome Y, qui vient du père, les mitochondries qui viennent de la mère et tout le reste qui vient des deux parents et qui se mélange – plus ou moins – à chaque génération. Le chromosome Y lui, est transmis uniquement de père en fils. Les mutations qui se sont produites dans le chromosome Y (qui sont transmises et reconnaissables) informent sur la lignée paternelle et peuvent permettre d'avancer une ascendance uniquement au niveau de cette lignée[19]. Il peut donc y avoir des gens qui ont une lignée paternelle qui va d'un côté, et une lignée maternelle qui va complètement d'un autre côté. Si on regarde l'ensemble de l'ADN (tous les chromosomes), on va avoir une version plus globale et plus juste. Cela dit, à partir d'un certain nombre de points variables de mon chromosome Y, en le comparant au chromosome Y d'un certain nombre de populations, on va aussi pouvoir dire que ma lignée paternelle vient d'Asie mineure, par exemple. Dans ce cas, la dimension statistique reste prédominante. La Réunion est un endroit intéressant à cet

[17] Ce qui implique de très fortes *limitations*, régulièrement soulignées dans la littérature. Ainsi, à 10 générations, chacun d'entre nous possède environ 1000 lignes d'ancêtres directs : ces analyses ne nous fournissent d'informations que sur *une* seule de ces lignes (il en est de même pour celles du chromosome Y). Elles sont néanmoins extrêmement populaires, permettant de caractériser des « clans » ou des groupes internet qui se définissent par leur haplotype, c'est-à-dire parce qu'ils partagent une même signature génétique sur leur chromosome Y ou leur ADN mitochondrial. On assimile par ailleurs régulièrement ces haplogroupes aux « clans » réputés traditionnels dans certaines régions, comme par exemple en Irlande ou en Ecosse. Voir notre essai bibliographique.

[18] L'article de référence sur ce point est Cann RL, Stoneking M, Wilson AC (1987), "Mitochondrial DNA and human evolution", *Nature* **325** (6099): 31–36.

[19] Des marqueurs spécifiques sur le chromosome Y ont ainsi permis de caractériser une certaine signature qui serait propre à la lignée des Cohanim, nom qui, selon la tradition juive, est réservé à la lignée qui, de père en fils depuis Aaron, se transmettrait les fonctions sacerdotales. Sur l'histoire de cette association, voir par ex. Abu-El Haj, *The genealogical science*, Univ. of Chicago Press, 2012. De manière plus générale, les informations génétiques transmises par le chromosome Y sont souvent corrélées à la transmission des noms propres qui, dans beaucoup de cultures, obéissent aux mêmes règles de transmission patrilinéaire.

égard car c'est une île qui était inhabitée jusqu'au XVIII^{e} siècle. Les Français s'y sont établis pour des raisons stratégiques et y ont fait des plantations de canne à sucre pour lesquelles ils ont fait venir des esclaves de Madagascar. Ensuite, lorsque la traite a été interdite, ils ont fait venir des travailleurs – soi-disant libres – d'Inde et de Chine. Aujourd'hui, vous y trouvez, en gros, quatre populations : les blancs qui s'appellent les « z'oreilles », les noirs qui s'appellent les « cafres », les Indiens qui s'appellent « les Malbars » et les Chinois. Au niveau paternel, on trouve une ascendance à 90% européenne et au niveau maternel, on y retrouve les quatre groupes : c'est la trace de relations sexuelles très asymétriques entre les colons et leur main d'œuvre.

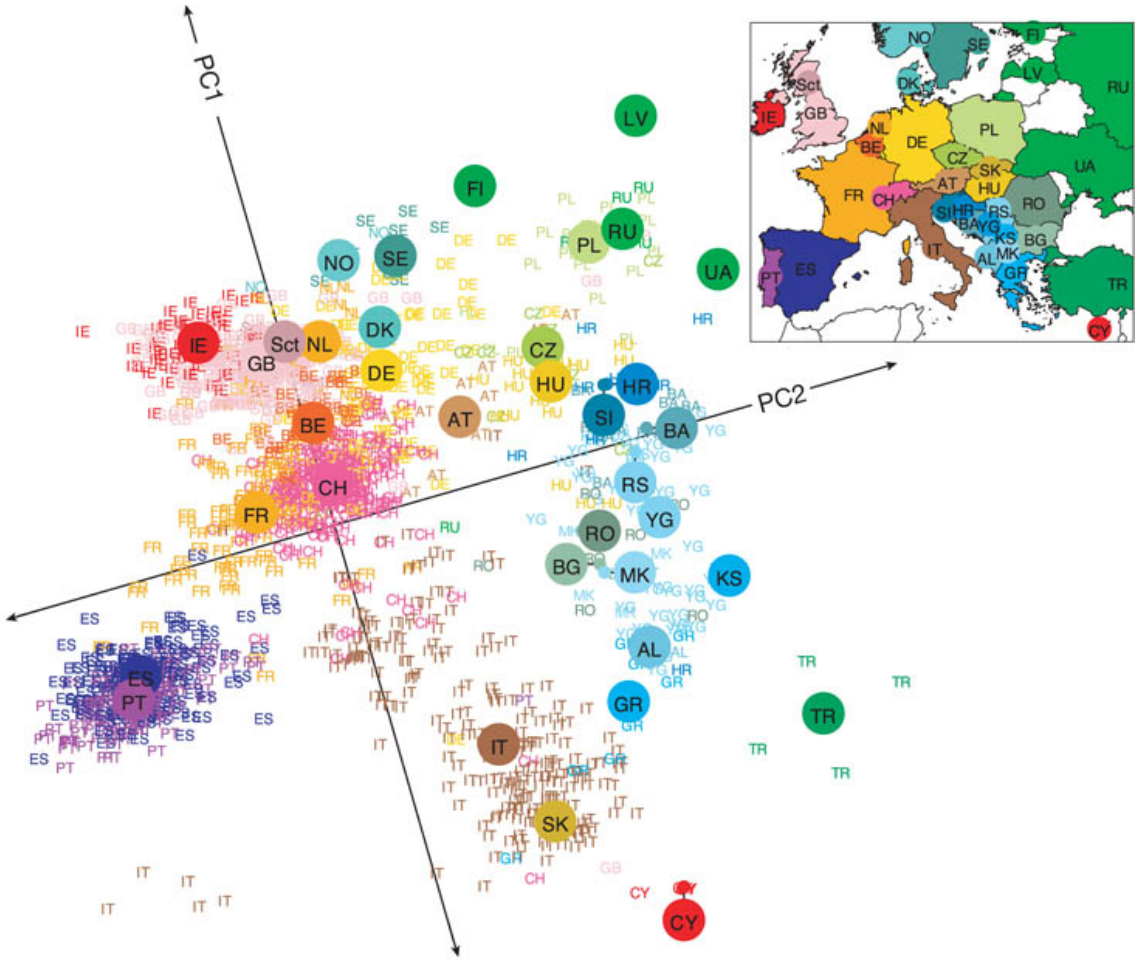
La vie des idées : Mais les logiciels utilisés pour assigner les individus ou les populations à des groupes d'ascendance biogéographique, ne font-ils pas intervenir des hypothèses de base très fortes – notamment le nombre de groupes pour lequel on fait un test ?

Bertrand Jordan : Les modèles qui sont à la base de ces logiciels sont extrêmement difficiles à étudier : il s'agit vraiment de statistique pure et dure. Dans ces papiers, on peut comprendre l'introduction et la conclusion ; on peut plus ou moins comprendre les figures qu'ils donnent mais ce sont des simulations. Ils développent en fait un programme pour retrouver la parenté avec des populations de référence, et ils le testent en générant des génomes artificiels qui ont 10% de telle ascendance, 20% de telle autre, et le reste d'une troisième. Ils voient jusqu'à quel point leur logiciel retrouve les choses correctement, et ils donnent des résultats dans lesquels l'assignation est correcte pour 90% des segments. De toute façon, dans toutes ces méthodes, ils découpent le génome en petits morceaux et ils essayent de voir la parenté de chaque petit morceau avec les populations de référence.

Il faut reconnaître que la marge d'erreur est souvent loin d'être négligeable et je n'ai pas réussi à trouver quelque chose qui me donne une vraie précision. Par exemple, quand on dit « 20% *Native American* » (Indien d'Amérique), c'est 20 plus ou moins combien ? Je vois que les approches les plus récentes et les plus sophistiquées, ne présentent sans doute pas les pires résultats. Ils disent que c'est correct dans 90% des cas. Ça veut aussi dire que les 0,5% de telle ou telle ascendance qu'on retrouve souvent dans les tests ne signifient rien du tout[20].

Cela dit, d'autres approches sont développées qui sont plus objectives, dans la mesure où elles ne se réfèrent pas à des populations de référence. Ainsi de l'analyse en composantes principales (APC). Regardez par exemple cette étude sur les populations européennes, l'illustration 4 ?

[20] Pour un article répertoriant l'ensemble des techniques informatiques disponibles actuellement, voir Liu, Yushi, et al. "Softwares and methods for estimating genetic ancestry in human populations." *Hum. Genomics* 7.1 (2013). Les hypothèses fondant ces logiciels et leurs implications ont été décrites par exemple par Bolnick, D., « Individual ancestry inference and the reification of race as a biological phenomenon » in Koenig, B., Soo-Jin Lee, S. & Richardson, S., (dir.), *Revisiting race in a genomic age*, Rutgers University Press, 2008, p. 70-89.



L'analyse en composantes principales

L'analyse en composantes principales (APC) est une technique d'analyse des données qui permet de réduire un nuage de données complexes, qui mettent en jeu une multiplicité de variables corrélées (par ex., un tableau à n entrées), de sorte à le projeter sur un plan à deux dimensions indépendantes en le déformant le moins possible, c'est-à-dire en conservant le maximum de la variabilité des données. Les différents axes indépendants qui expliquent le mieux la variabilité des données sont appelés des « composantes principales ». L'analyse commence donc généralement par retenir les deux premières composantes, qui sont celles qui résument le maximum de variabilité des données.

Cette figure illustre bien comment marche l'APC et elle est très utile parce qu'elle permet de relativiser les conclusions qu'on aurait tendance à tirer de l'image précédente où l'on séparait trois groupes correspondant aux trois blocs continentaux. Elle permet de montrer la puissance de l'analyse génétique qui permet de faire des distinctions bien au delà ce qu'on pourrait considérer raisonnablement comme « race ».

Il s'agit d'une étude portant sur les populations européennes. On y a analysé un certain nombre de personnes dont les grands parents habitaient dans le même pays, de façon à éliminer les gens qui ont migré récemment (et donc les mélanges que ça peut avoir induit). On a ensuite, regardé environ 500 000 points variables, dans l'ADN de chacune de ces personnes (environ 1 000 dans l'étude). Pour chaque personne, on dispose donc de son profil en 500 000 points du génome. Ensuite, à partir de là, on peut calculer la distance génétique entre toutes ces personnes. On se trouve alors dans quelque chose qui est comme dans un espace à 500

000 dimensions. Et, comme on a un peu de mal à appréhender un tel espace, on va essayer de trouver, parmi ces 500 000 dimensions, celles qui séparent le mieux différents sous-groupes ; tout cela se fait bien sûr de façon informatique. On choisit alors les deux dimensions qui sont les plus discriminantes et on projette les résultats sur le plan qu'elles définissent. Le résultat est celui-ci : la projection de 1000 points répartis dans un espace à 500 000 dimensions selon les dimensions qui différencient le mieux des groupes. Il n'y a aucune référence à des populations d'origine ou quoi que ce soit. Ce n'est établi qu'à partir des données qu'on a obtenues des 500 000 points dans l'ADN de chacune de ces personnes (chaque personne est représentée par une petite lettre). Et on voit que, dans la plupart des cas, les Espagnols et les Portugais se retrouvent d'un côté, les Français d'un autre, les Italiens d'un autre encore, les Irlandais d'un autre, et ainsi de suite.

La seule chose arbitraire dans cette image, c'est qu'on a choisi les échelles et le sens de la projection de façon à retrouver la carte de l'Europe. On pourrait aussi bien regarder l'image en miroir ou la tourner dans un sens ou dans l'autre. Par contre, le fait que cela permette de distinguer, uniquement par l'analyse d'ADN, les Français ayant des grands-parents français des Irlandais ayant des grands-parents irlandais et des Espagnols ayant des grands-parents espagnols, c'est objectif. Il n'y a pas d'hypothèse *a priori* là-dedans. C'est une image que j'utilise pour montrer que – si on y met le prix – l'analyse génétique permet de retomber sur quelque chose qui est relativement évident certes (si vous êtes Français, il est plus probable que vous ayez des liens proches avec un autre Français qu'avec un Irlandais ou un Espagnol), mais ça arrive à un niveau de précision qui va bien au delà de ce qu'on pourrait imaginer appeler « race ».

La vie des idées : Il y a quand même un arbitraire non négligeable, c'est le choix même des composantes principales. On choisit les composantes qui vont dans le sens souhaité, c'est-à-dire celui du maximum de différenciation entre groupes. Vous l'avez dit, il y a en fait 500 000 dimensions, et l'on ne retient que celles qui sont les plus discriminantes. Dans un certain sens, peut-on s'étonner de voir émerger des groupes si clairement différenciés si on s'y prend de cette manière-là ?

Bertrand Jordan : C'est objectif dans le sens où on n'impose pas l'homogénéité de chaque groupe. En clinique, on fait souvent des analyses par composante principale. On mesure un certain nombre de bio-marqueurs sur des malades et des témoins et on fait ensuite une analyse par composante principale, pour voir si ça sépare les malades des témoins. Il arrive que ça sépare deux groupes mais que, dans chaque groupe, il y ait des malades et des témoins. Dans ce cas, on se dit que ce ne sont pas les bons bio-marqueurs et qu'il faut en chercher d'autres.

Dans l'exemple précédent, l'analyse aurait par exemple pu mélanger Irlandais et Espagnols. Or, ce n'est pas le cas, et l'on constate que les groupes qui sont objectivement trouvés par l'analyse en composante principale recoupent des proximités géographiques.

Usages sociaux de l'ascendance biogéographique

La vie des idées : Quand ces procédures qui permettent d'assigner – par le biais des marqueurs polymorphiques – un individu à un groupe d'ascendance, ont été mises en place, qui s'est en emparé ? Et dans quel but ?

Bertrand Jordan : Il y a eu deux types d'usages de ces informations, par le grand public et par les scientifiques. Les usages scientifiques sont, par exemple, reliés à la question de la

prévision/prédiction/détermination des vulnérabilités à différentes maladies. Ces marqueurs ont été utilisés par les généticiens des populations à la fois pour contrôler les groupes dans les essais cliniques, pour s'assurer qu'on n'était pas en train de mélanger des effets de différences de populations avec des effets réellement liés à la pathologie étudiée.[21] Éventuellement aussi, pour étudier des populations mélangées afin d'avoir des pistes de marqueurs de vulnérabilité à des maladies polygéniques[22] comme le diabète, la maladie de Crohn, le cancer de la prostate, etc.[23]

Il y a en effet des maladies qui sont plus fréquentes dans une population donnée que dans une autre. Il existe alors plusieurs cas de figure. Certaines maladies correspondent probablement à des mutations relativement récentes et qui n'ont pas eu le temps de se répandre dans l'ensemble de la population humaine. L'exemple typique est celui de la maladie de Tay-Sachs chez les juifs Ashkénazes. C'est une mutation probablement assez récente (datant de dix mille ans et peut-être moins), apparue dans cette population et qui y est largement restée car il s'agit d'une population relativement fermée, qui s'est peu mélangée avec d'autres. Sinon, quand vous avez la possibilité de faire des statistiques raciales – ce qui existe aux États-Unis mais pas en France –, vous trouvez des différences de prévalence pour un grand nombre de maladies polygéniques entre Européens et Afro-Américains. Ensuite, bien-sûr, la question est de savoir si ces différences sont dues aux conditions de vie, de niveau social, d'accès aux soins, d'alimentation, etc., ou s'il y a une part génétique.

On a vu aussi utilisée la notion d'ascendance biogéographique dans un contexte peu connu en France au niveau médicolégal. En France, la question médicolégale concerne d'abord la définition des empreintes génétiques. On a voulu faire en sorte qu'elles ne puissent pas identifier directement les personnes. Il fallait que l'on puisse dire de tel ADN qu'il provient d'une personne plutôt que d'une autre, mais sans qu'il permette d'avoir des informations sur les caractéristiques de la personne. Dans la loi qui définit l'usage des empreintes génétiques, il est dit que c'est pris en dehors des séquences codantes, de façon à ce

[21] C'est particulièrement vrai en ce qui concerne les *Genome Wide Association Studies* (GWAS) où les différences de fréquence de telle ou telle maladie entre un groupe de cas et un groupe-contrôle peuvent en réalité être liées à des différences dans la composition génétique des groupes. Pour résorber ce risque de créer des « faux positifs », il convient de s'assurer autant que possible que les groupes de cas et de contrôle représentent le même « mixte » de groupes ethniques et connaître donc la manière dont ils sont composés. Voir pour plus de détails, Brookfield, J.F, « Promise and pitfalls of genome-wide association studies », *BMC Biology*, 2010, 8:41.

[22] On distingue les maladies monogéniques, dues à une mutation localisée sur un seul gène, et les maladies polygéniques, qui mettent en jeu des altérations plus ou moins légères sur de nombreux gènes, qui ont de petits effets additifs favorisant (ou non) la maladie. Les maladies polygéniques obéissent donc à des lois héréditaires beaucoup plus complexes que les lois de Mendel, et mettent en jeu des modèles probabilistes plus complexes.

[23]En ce sens, les « populations hybrides » deviennent des objets expérimentaux de premier ordre. Ainsi, plusieurs équipes de recherche se sont-elles efforcées d'utiliser les Chicanos et les Portoricains (les premiers réputés être composés majoritairement d'ascendance européenne et *native american* ; les seconds d'ascendance européenne et africaine) pour tester l'hypothèse d'une association significative entre des fréquences d'asthme plus élevées et une ascendance africaine. Ces études permettent d'identifier de possibles facteurs de susceptibilité qui sont ensuite testés dans des populations réputées plus homogènes (cf. Fullwiley, Duana, « The biologistical construction of race : « Admixture » technology and the new genetic medicine », *Social Studies of Science*, 38(5) : 697-737). La même logique se retrouve à propos de la détermination des facteurs (génétiques et/ou environnementaux) qui expliquent la plus forte fréquence de l'hypertension chez les populations africaines-américaines. On a ainsi comparé les populations « afro-américaines » proprement dites avec d'autres populations – notamment la population jamaïcaine – qui devraient (en tout cas en termes de mélange de populations) posséder le même profil. Or, il s'est avéré que le taux d'hypertension chez les Jamaïcains était beaucoup plus faible que chez les Africains-Américains. Toutes ces études reposent sur un nombre de présupposés considérables, bien résumés in Rajagopalan, R. & Fujimora, J., « Making history via DNA, making DNA from history », in Wailoo, K., Nelson, A. & Lee, C. (dir.), *Genetics and the unsettled past*, Rutgers University Press, 2012, p. 143-163

que cela ne soit pas révélateur de la personne. En fait, compte tenu de tout ce que nous avons vu jusqu'ici, ça n'a pas besoin d'être dans les séquences codantes, puisque l'ascendance se reflète dans tout l'ADN et que les points par lesquels vous rattachez quelqu'un à une ascendance donnée n'ont absolument pas besoin d'être dans des gènes pour être significatifs. Il y a une tendance actuellement qui se développe pour essayer de tirer le maximum d'informations des prélèvements d'ADN qui ont été faits (par exemple sur le lieu d'un crime) pour avoir des indications sur la personne à rechercher. Ce qui existe actuellement, et qui fonctionne relativement bien, c'est l'estimation de la couleur des yeux. Il y a une entreprise, effectivement, qui commercialise un test qui permet d'avoir, en regardant une petite dizaine de marqueurs polymorphiques, une bonne approximation de la couleur des yeux de la personne de qui provient cet ADN.[24] La couleur des yeux, comme celle de la peau, est un caractère très polygénique. Traditionnellement on dit « gène bleu récessif, marron dominant », mais il y a en fait des dizaines de gènes qui interviennent pour définir la couleur des yeux. On arrive à la prédire en regardant les sept ou huit plus significatifs. Pour la couleur de la peau, il y a probablement un très grand nombre de gènes qui interviennent. Si on étudie ça sur l'ensemble des populations humaines, on va de nouveau avoir beaucoup de mal à démêler ce qui est lié à la couleur de la peau, et ce qui est lié à des ascendances globales différentes. Et, si on l'étudie à l'intérieur d'un groupe de populations, on a une gamme de variations de la couleur de la peau qui est relativement limitée. On n'a donc pas de bon prédicteur de la couleur de la peau actuellement. Si l'analyse de l'ADN vous dit que l'ascendance de la personne est à 95% africaine, il y a des chances que sa peau soit relativement foncée. Par contre, si c'est 50%, vous ne pouvez pas savoir. Il y a aussi des gens qui travaillent pour essayer de définir la forme du visage à partir de l'ADN, parce que c'est aussi génétiquement déterminé.[25]

Et le concept de race ?

La vie des idées : En Louisiane, en 2003-2004, dans l'affaire Derrick Todd Lee, on cherchait un *serial killer* dont tout laissait penser qu'il était blanc. En désespoir de cause, les enquêteurs ont envoyé son ADN à *DNAPrint Genomics*. L'entreprise a déclaré que l'individu recherché avait 85% d'ascendance africaine et environ 15% d'ascendance *Native American*, qu'il était donc ou « afro-caribéen ou africain-américain mais en aucun cas caucasien » et a conseillé d'ouvrir le panel des suspects à des gens qui présentaient un phénotype plus probablement associé à de telles origines. Ils ont trouvé le suspect de cette manière.[26] On a pu

---

[24] Il s'agissait de DNAPrint Genomics, dont l'un des fondateurs était Tony Frudakis, particulièrement engagé – avec Mark Shriver – dans le développement d'applications « forensic » des techniques d'estimations des ascendances bio-géographiques, ainsi que dans le « *molecular photofitting* », c'est-à-dire la tentative d'inférer le phénotype d'une personne (la pigmentation de ses yeux ou de sa peau, la forme de son visage) à partir de son génotype. Voir Frudakis, T., *Molecular photofitting : predicting ancestry and phenotype using DNA*, Academic Press, 2007. DNAPrint Genomics s'est arrêtée en 2009, après avoir néanmoins fourni diverses estimations de l'ascendance biogéographique de criminels et permis ainsi à la police américaine ou anglaise de cibler tel groupe de « suspects ».

[25] Sur la base de ce genre de travaux, une artiste new yorkaise, Heather Dewey-Hagborg a essayé de faire des portraits à partir de des mégots de cigarettes et des morceaux de chewing-gum,. Elle a fait des visages, littéralement, à partir de ces objets ramassés dans la rue et utilise ces travaux sur l'ADN pour élaborer ces visages. Voir http://www.npr.org/2013/05/12/183363361/litterbugs-beware-turning-found-dna-into-portraits et http://deweyhagborg.com/strangervisions/samples.html Mais, en vérité, son travail est fondé sur quelques très faibles informations et beaucoup d'imagination.

[26] Pour cette affaire et d'autres similaires, voir par exemple Shankar, Pamela, « Forensic DNA Phenotyping : reinforcing race in law enforcement » in Whitmarsh, I. & Jones, D. (dir.) *What's the use of race ?*, MIT Press, 2010, p. 49 et sqq. et "Forensic DNA Phenotyping: continuity and change in the history of race, genetics and policing" in Wailoo & al., *Genetics and the unsettled past*, op. cit., p. 104 et sqq.

en fin de compte se demander si parler d' « ascendance biogéographique » n'était pas un mot pudique pour dire "race" ? », comme l'explicitent d'ailleurs certains chercheurs.[27] Les deux termes sont-ils en fait équivalents ?

Bertrand Jordan : La différence, pour moi, c'est que l'idée de race représente des sous-ensembles – avec des limites nettes – à l'intérieur de l'espèce humaine. L'appartenance à une race détermine alors votre apparence mais aussi vos capacités, votre caractère etc. Il y a tout un ensemble de caractéristiques variées et relativement homogènes au sein de ce sous-ensemble bien délimité, défini par le fait d'être noir, blanc ou asiatique. Au delà, l'idée traditionnelle de race implique aussi qu'il y aurait une hiérarchie entre les races. Son origine est en fait de justifier la domination d'un groupe de populations sur un autre. Dans la notion de 'race' mais plus largement dans le racisme, il y a une forme de globalisation conduisant à considérer qu'un groupe définit les personnes qui en font partie – et les mettant toutes dans le même sac, d'une certaine façon. Mais je pense aussi que certaines personnes entendent bien « groupes d'ascendance » derrière le mot « race ». Or, si la génétique peut définir des groupes sur la base d'un certain nombre de marqueurs, on voit bien que ces groupes ont des limites floues, qu'il y a presque autant de diversité à l'intérieur d'un groupe, qu'entre la moyenne de deux groupes, etc. Disons qu'on peut trouver des différences entre la moyenne de deux groupes qui, opérationnellement, permettront de distinguer un groupe d'un autre.

J'utilise souvent les chiens comme exemple d'une espèce dans laquelle il y a clairement des races (créées par l'Homme), avec des comportements différents, des aspects et des caractères différents et bien marqués. Chez les chiens, vous déterminez la race extrêmement facilement, à partir de quatre ou cinq marqueurs. On peut donc dire qu'il y a réellement des races à l'intérieur de cette espèce. Il y a une différenciation génétique en groupes qui est vraiment beaucoup plus marquée que dans l'espère humaine. C'est une question de degré. Dans l'espèce humaine, il faut regarder quelques centaines de marqueurs bien choisis ou 500 000 marqueurs tout venant, et faire travailler des programmes sophistiqués pour arriver à retrouver la trace de vos ancêtres dans votre ADN. Dans le cas du chien, vous faites cinq sites polymorphiques et vous savez si c'est l'ADN qui vient d'un chiwawa ou d'un labrador. Ceci est dû au fait que les races canines sont très homogènes et très différenciées les unes des autres. Même si c'est relativement récent, avec un temps de génération d'une année, en deux ou trois siècles vous avez le temps de séparer des groupes jusqu'à ce qu'ils deviennent réellement des races. Le terme de « race » n'est, de toute façon, biologiquement pas précis. « Espèce » a un sens précis mais « race » est une subdivision qu'on met à l'intérieur d' « espèce » ; on met le curseur un peu où on veut.

Illustrations

Fig. 1: "arbre de proximité génétique pour 84 personnes d'origine européenne, africaine, chinoise et japonaise, construit à partir des 1000 SNP's d'indices de fixation (Fst) les plus élevés (sur 8525 étudiés), c'est-à-dire dont la variabilité est maximale entre les groupes". (Shriver et al. "The genomic distribution of population substructure in four populations using 8,525 autosomal Snps", Human Genomics, 2004, vol. 1, p. 274-286)

[27] les demandes de brevet qui avaient été formulées par Mark Shriver et Tony Frudakis sur les les *AIMs* (les marqueurs informatifs de l'ascendance), on qualifiait la « biogeographical ancestry » de « heritable component of "race" » Voir http://www.google.com/patents/WO2004016768A2?cl=en Cette ambiguïté est systématique et se retrouve chez la plupart des scientifiques utilisateurs de la notion d'ascendance biogégraphique, comme en témoignent les entretiens menés par Catherine Bliss in *Race Decoded*, Stanford University Press, 2012, p. 100-133. La notion d'ascendance biogéographique est sans cesse présentée comme la dimension « biologique » ou « génétique » de la « race ».

Fig. 2: "Ancestry painting": composition des différentes ascendances de M. de Armond (https://www.23andme.com/)

Fig. 3: "Chromosome map": détail, chromosome par chromosome, de la composition ancestrale d'un individu (https://www.23andme.com/)

Fig. 4: Analyse en composantes principales sur les données génétiques tirées de 1 387 Européens, fondées sur les deux composantes principales (PC1 et PC2). Les petites lettres colorées représentent chaque individu (IT pour un Italien, FR pour un Français etc.) et sa position sur les deux axes. Les lettres encerclées représentent la médiane des valeurs de tous les individus d'un pays donné pour les deux composantes principales. La carte en haut à droite permet juste de se remémorer la localisation des différents pays. Tiré de November & al. Genes mirror geography within Europe, John Novembre et al., *Nature*. 2008 Nov 6;456(7218):98-101 doi: 10.1038/nature07331. Epub 2008 Aug 31.

Publié dans laviedesidees.fr, le 25 février 2014