médecine/sciences 2024 ; 40 : 563-5

Chroniques génomiques

L'ADN ancien parle de plus en plus

Bertrand Jordan

Biologiste, généticien et immunologiste, Président d'Aprogène (Association pour la promotion de la Génomique), 13007 Marseille, France.
brjordan@orange.fr

IBD et parenté

En génétique des populations, l'exploration des relations entre individus ou entre groupes humains peut maintenant s'appuyer sur la connaissance de nombreuses séquences génomiques complètes ou, à défaut, sur des profils génétiques répertoriant les allèles de SNP (*single nucleotide polymorphism*) présents chez chaque individu. On va alors rechercher les similitudes globales entre ces profils et évaluer ainsi la proximité génétique des individus concernés. Un type de comparaison apparu à la fin des années 2000 [1, 2] et qui s'est avéré très efficace, est la recherche dans les ADN comparés de segments « identiques par la descendance » – on utilise généralement l'expression anglaise *Identical by Descent* et son acronyme IBD. Il s'agit de retrouver dans ces deux ADN de grands segments strictement identiques, s'étendant sur 10 centiMorgans (cM)[1] ou plus. De telles coïncidences ne peuvent être dues au hasard en raison du polymorphisme des génomes humains (plusieurs millions de différences entre deux personnes prises au hasard) et doivent avoir pour origine un ancêtre commun, signant ainsi une relation de parenté entre les deux individus concernés. Compte tenu des recombinaisons qui se produisent à chaque génération (en moyenne une recombinaison par chromosome) et qui vont morceler les IBD, il s'agit nécessairement d'une parenté relativement proche, d'autant plus proche que le segment IBD est long (voir la *Figure 1* pour des exemples). Ces analyses ont été largement utilisées en génétique des populations en analysant l'ADN des populations actuelles et ont permis de préciser de nombreux points concernant leur origine et leur démographie ; de nombreuses approches méthodologiques ont été développées pour améliorer la détection des segments IBD et limiter les temps de calcul impliqués [3].

Le Graal de l'ADN ancien

Depuis une dizaine d'années, il est devenu possible d'obtenir des séquences d'ADN humain à partir d'échantillons anciens et même de fossiles (le record actuel est de 45 000 ans), grâce à de grands progrès techniques dans l'obtention et l'analyse de ces séquences. La recherche de segments IBD dans ces génomes est susceptible d'apporter de précieuses informations sur ces lointains ancêtres, mais elle était jusqu'ici presque impossible en raison de la mauvaise qualité de ces données : alors que la règle pour un génome humain de qualité standard est qu'il soit séquencé avec une redondance de 30 fois (30X), la plupart de ces génomes anciens ont été lus avec une redondance de l'ordre de 1X, parfois moins : ils sont *a priori* truffés d'erreurs et de trous, or il suffit d'une seule erreur dans un segment IBD de 10 cM pour le fractionner. Les centaines ou milliers d'erreurs que contiennent deux génomes anciens comparés rendent ainsi impossible la détection de segments IBD. Il faut donc recourir à des méthodes informatiques et statistiques capables de tolérer de telles imprécisions et incertitudes, et c'est ce que fait avec un certain succès un système appelé ancIBD, présenté dans un article récemment paru [4].

Sous le capot d'ancIBD

Impossible naturellement de décrire en détail le système ancDNA, ensemble très complexe et qui fait

[1] Le Morgan est l'unité de distance utilisée en génétique, correspondant à un segment de chromosome dont la longueur ne permet, en moyenne, qu'une recombinaison par méiose. La longueur totale du génome humain est d'environ 33 Morgan. Un centiMorgan correspond en moyenne à un million de bases d'ADN (une mégabase).

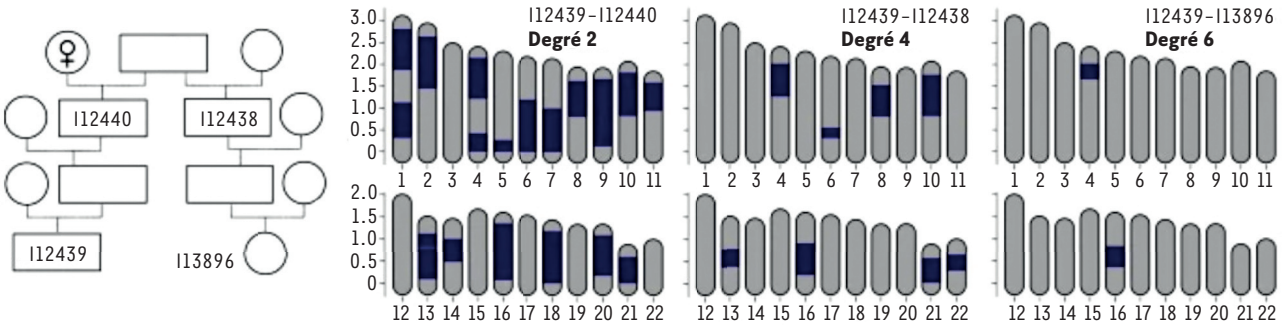

Figure 1. Analyse par ancIBD de quatre individus (numéros) appartenant au groupe indiqué à gauche. À droite, résultats pour les personnes apparentées au deuxième, quatrième et sixième degré. Les 22 autosomes sont figurés de haut en bas, les segments IBD en bleu. L'échelle verticale est en Morgans. Le nombre et l'étendue des segments IBD décroissent quand la parenté devient plus lointaine, mais la relation reste détectable au sixième degré (extrait partiel et modifié de la figure 3 de [4]).

appel à des techniques statistiques sophistiquées. Dans les grandes lignes, il comporte deux phases principales : une étape d'*imputation*, qui complète les données de séquence fragmentaires, et, ensuite, la recherche de segments IBD par un modèle de Markov caché (MMC ou *Hidden Markov Model*, HMM). L'imputation consiste à « remplir » les séquences incomplètes en s'appuyant sur les données actuelles du *1 000 genomes project* ; on pourrait dire que l'on triche, mais cette méthode est déjà largement employée et donne de bons résultats, à condition que les données d'imputation et la séquence « imputée » proviennent de populations comparables. Quant au modèle de Markov caché, il s'agit d'une approche statistique permettant de faire des prévisions dans une situation où on dispose de données très incomplètes et où les « règles » qui les relient sont en partie inconnues. Les MMC sont apparus dans les années 1960 et ont joué un grand rôle dans la mise au point des systèmes de reconnaissance de la parole ; ils ont ensuite été appliqués à l'analyse des séquences d'ADN[2]. Bien entendu, le module d'imputation tout comme le modèle de Markov sont soigneusement paramétrés pour l'usage qui en est fait ici ; et, en aval, un ensemble d'astuces informatiques permet de réduire le temps de calcul à moins d'une seconde pour la recherche de segments IBD entre deux génomes.

Performances et applications d'ancIBD

La qualité des résultats fournis par le système a été évaluée de différentes manières. Une des plus informatives a utilisé quatre séquences d'ADN ancien de qualité exceptionnelle (redondance 5X, ce qui est très élevé pour de tels échantillons) et a testé les performances d'ancIBD sur des jeux de ces séquences volontairement dégradés (*downsampled*) pour correspondre à des redondances bien plus faibles. Ces analyses ont montré que ancIBD appliqué à des données de séquence donnait des résultats fiables jusqu'à une redondance de 0,1X. Les performances sont meilleures pour les séquences que pour les profils de SNP, et la détection de longs segments (plus de 10 cM) est plus fiable que celle de régions plus courtes. Ces résultats sur des données artificiellement dégradées donnent confiance en l'emploi d'ancIBD pour des séquences « typiques » d'ADN ancien.

Les auteurs ont ensuite appliqué leur système à l'analyse de 4 248 séquences d'ADN ancien et ont montré, en comparant leurs résultats avec différentes simulations, que l'on pouvait en déduire des degrés de parenté allant au moins jusqu'au troisième degré. L'analyse de quatre ADN provenant d'individus, dont la généalogie avait été établie par des données archéologiques, montre que ancDNA peut détecter des segments IBD jusqu'au sixième degré de parenté (*Figure 1*), ce qui en fait un instrument très puissant pour évaluer les relations entre populations humaines.

Un autre exemple d'utilisation d'ancIBD est fourni par la comparaison des ADN de deux individus appartenant, d'après les données archéologiques, à la culture Afanasievo (pasteurs des steppes orientales, âge du cuivre, vers – 3 000 ans) mais dont les tumulus funéraires se situent l'un en Mongolie centrale, l'autre en Russie méridionale, à une distance de presque mille cinq cents kilomètres. L'analyse par ancIBD révèle plusieurs segments IBD dont quatre dépassent la taille de 20 cM (*Figure 2*), ce qui suggère une parenté au cinquième degré.

Cela signifie donc qu'au moins une personne dans la chaîne de parenté reliant ces deux individus doit avoir parcouru des centaines de kilomètres au cours de son existence, ce qui va dans le sens de la réévaluation actuelle de degré de mobilité des populations ancestrales, alimentée en grande partie par les analyses d'ADN anciens.

[2] Voir https://fr.wikipedia.org/wiki/Mod%C3%A8le_de_Markov_cach%C3%A9 pour une description complète mais assez aride.

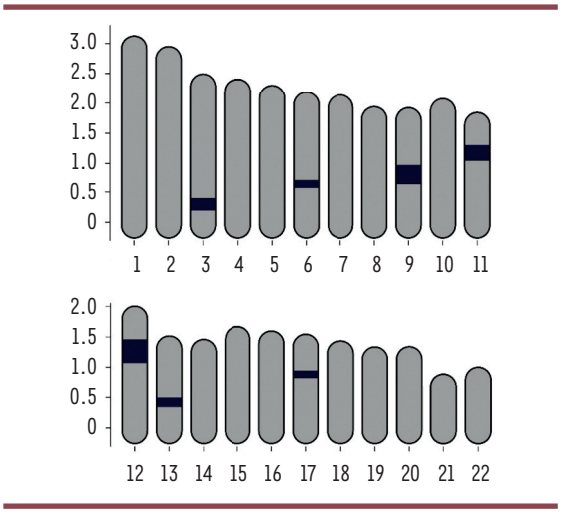

Figure 2. Segments IBD détectés par ancIBD dans les ADN de deux membres de la culture Afanasievo provenant de sépultures situées à une distance de 1 410 km. Les segments IBD détectés sont indiqués en bleu, le plus long (chromosome 12) mesure près de 40 cM. L'échelle verticale est en Morgans (extrait partiel et modifié de la figure 5 de [4]).

L'ADN ancien n'a pas fini de nous étonner

Il y a moins de vingt ans commençaient les tentatives de séquençage de l'ADN de Néandertal, qui se heurtaient à de difficiles problèmes techniques (ADN très dégradé, contaminations omniprésentes) [5] (➜). Ces problèmes ont été progressivement résolus grâce à l'entrée en lice des nouvelles techniques de séquençage (NGS, *new generation sequencing*), à des précautions draconiennes dans l'obtention et le traitement des échantillons, et à des méthodes informatiques qui permettent d'interpréter ces séquences, qui restent de mauvaise qualité comparées à celles que l'on peut obtenir sur des prélèvements « normaux ». On répertorie aujourd'hui plus de 10 000 séquences d'ADN ancien (de 1 000 à presque 50 000 ans d'âge) [6], et ces données ont produit une véritable révolution en anthropologie, révélant des parentés que l'on ignorait ou soupçonnait à peine, documentant les migrations de populations anciennes dont la mobilité était insoupçonnée, et fournissant des estimations fiables de paramètres essentiels comme la taille effective de telle ou telle population[3]. Cette entrée en force de données génétiques dans le champ de l'archéologie a suscité quelques controverses, mais il est clair que les deux approches sont complémentaires et que l'on peut encore en attendre de grandes avancées dans la compréhension de notre lointain passé [7] (➜). ◊

(➜) Voir le Repères de S. Gilgenkrantz, *m/s* n° 1, janvier 2007, page 95

(➜) Voir le Forum de C. Bon, page 556 de ce numéro.

[3] La taille effective d'une population est le nombre de reproducteurs(-trices) qu'elle comporte, elle est par nature inférieure au nombre total d'individus. L'analyse de la diversité génétique (grâce à l'ADN ancien) permet de l'estimer ; on considère par exemple que la taille effective de la population de Néandertaliens était d'environ 10 000 individus (seulement).

SUMMARY

Ancient DNA speaks

Many human DNA sequences have been obtained from ancient remains dating back from several millennia. However, these have low coverage and may contain many errors; this has limited their usefulness for many analyses, in particular the search for Identical By Descent (IBD) segments that is very powerful for detection of kinship. A new method, using imputation from database data and sophisticated statistical analysis, proves able to detect IBD segments (and thus parenthood) in low-quality DNA sequences from individuals linked only by sixth degree parenthood, opening a whole new field of investigation using ancient DNA. ◊

LIENS D'INTÉRÊT

L'auteur déclare n'avoir aucun lien d'intérêt concernant les données publiées dans cet article.

RÉFÉRENCES

1. Browning SR. Estimation of pairwise identity by descent from dense genetic marker data in a population sample of haplotypes. *Genetics* 2008 ; 178 : 2123-32.
2. Browning SR, Browning BL. Identity by descent between distant relatives: detection and applications. *Annu Rev Genet* 2012 ; 46 : 617-33.
3. Sticca EL, Belbin GM, Gignoux CR. Current Developments in Detection of Identity-by-Descent Methods and Applications. *Front Genet* 2021 ; 12 : 722602.
4. Ringbauer H, Huang Y, Akbari A, Mallick S, Olalde I, Patterson N, Reich D. Accurate detection of identity-by-descent segments in human ancient DNA. *Nat Genet* 2024 ; 56 :143-51.
5. Gilgenkrantz S. Les prémices du génome de Néandertal. *Med Sci (Paris)* 2007 ; 23 : 95-8.
6. Callaway E. 'Truly gobsmacked': Ancient-human genome count surpasses 10,000. *Nature* 2023 ; 617 : 20.
7. Bon C. La paléogénétique ou de l'intérêt de l'exploration génétique du passé. *Med Sci (Paris)* 2024 ; 40 : 556.

TIRÉS À PART

B. Jordan