médecine/sciences 2014 ; 30 : 463-6

Chroniques génomiques

Variants rares et explosion démographique

Bertrand Jordan

UMR 7268 ADÉS, Aix-Marseille Université/EFS/CNRS,
Espace éthique méditerranéen,
Hôpital adultes La Timone,
264, rue Saint-Pierre,
13385 Marseille Cedex 05, France.
CoReBio PACA, case 901,
parc scientifique de Luminy,
13288 Marseille Cedex 09, France.
bertrand.jordan@univ-amu.fr
brjordan@orange.fr

Nous retrouvons dans cette chronique les variants rares qui avaient été présentés en 2012 [1] : les premiers séquençages d'exomes réalisés sur des milliers de personnes mettaient en évidence un grand nombre de variants trouvés à basse fréquence (moins de 0,5 %), parfois chez une seule personne sur les milliers séquencées. Rappelons que, dans ce contexte, l'on parle aujourd'hui de SNV (*single nucleotide variant*) en réservant le terme de SNP (*single nucleotide polymorphism*) ou snip à ceux des SNV dont l'allèle mineur est présent à plus de 5 % dans la population - les « SNV communs » en quelque sorte selon cette terminologie qui s'est récemment imposée. Je m'étais surtout intéressé aux conséquences fonctionnelles de ces SNV : beaucoup d'entre eux étant non synonymes, inactivant la protéine correspondante ou modifiant fortement ses propriétés, ils pouvaient apporter une solution (partielle) au problème de l'hérédité manquante [2]. Je mentionnais aussi que leur abondance était liée à l'histoire démographique de notre espèce, à l'expansion très forte et très récente de la population humaine. L'article sur lequel est centrée la présente chronique [3] s'intéresse spécifiquement à cet aspect démographique et, grâce à un *design* expérimental soigné, apporte des données nouvelles sur ce point.

Séquence et anthropologie, l'apport des variants rares

La possibilité d'obtenir rapidement et sans frais prohibitifs des données de séquence sur de nombreux individus ou, parfois, sur des fossiles vieux de plusieurs dizaines de milliers d'années révolutionne l'anthropologie. *Interbreeding* (limité) avec l'homme de Neandertal, découverte des Denisoviens... Sans remettre en cause le schéma général *Out of Africa* selon lequel notre espèce est apparue en Afrique puis a essaimé sur tout le globe à partir de 100 000 ans BCE (*before current era*), l'étude détaillée de l'ADN fossile et des populations actuelles apporte de précieuses informations sur notre histoire ancienne [4]. La population humaine, estimée à quelques millions d'individus en 8 000 BCE (soit il y a 10 000 ans, ou encore 400 générations) a cru de manière explosive pour atteindre sept milliards de personnes aujourd'hui *(Figure 1)*, avec un taux de croissance estimé à 0,5 % par an environ sur ces 400 générations [5-7].

Globalement, c'est cette croissance rapide (trois ordres de grandeur) qui explique la présence de très nombreux variants rares dans notre ADN[1] : ce sont des mutations apparues récemment (fréquence de mutation estimée à 10^{-8} par nucléotide et par génération, soit 3 000 mutations *de novo* par génome haploïde), et que la sélection naturelle n'a pas eu le temps d'éliminer. On peut montrer que la fréquence de ces mutations, et plus précisément le spectre de fréquence par nucléotide (*site frequency spectrum*, SFS) dépend de l'histoire démographique de la population. La *Figure 2* montre, lorsqu'on analyse 500 génomes, que la proportion de mutations retrouvées seulement chez un ou deux individus augmente considérablement si la population a cru d'un facteur 1 000 au cours des 400 dernières générations. Il s'agit là d'une simulation, mais elle montre que l'étude du spectre de fréquences

[1] Variants nombreux lorsqu'on fait le bilan sur quelques centaines ou milliers de personnes, mais dont l'allèle mineur est rare dans la population et aussi pour chaque individu, dans l'ADN duquel dominent les SNP, voir figure 2 de [1].

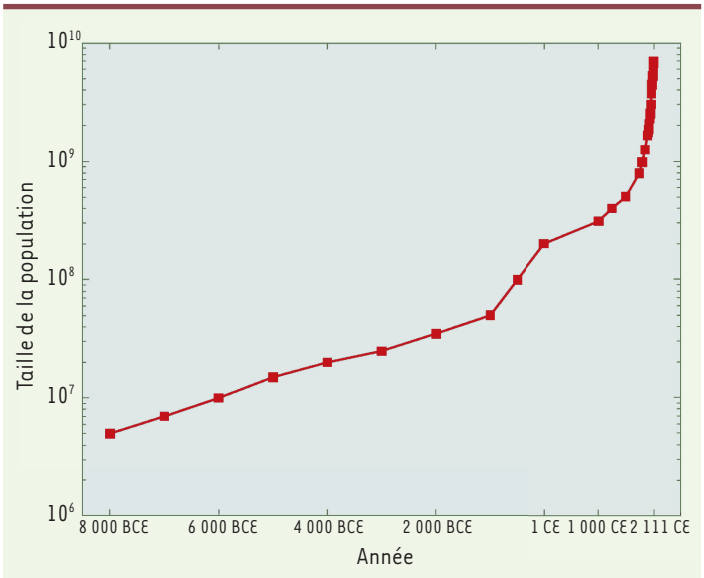

Figure 1. ***Évolution démographique de l'espèce humaine au cours des dix mille dernières années selon Keinan et Clark*** [6] ***(BCE = before current era).*** Taille de la population en ordonnée (échelle logarithmique).

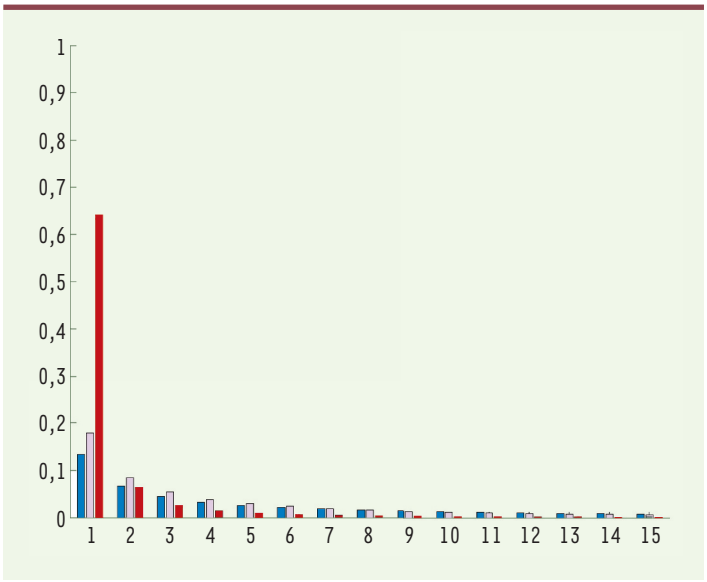

Figure 2. ***Simulation indiquant la probabilité (en ordonnée) de retrouver l'allèle mineur d'un SNV une, deux, trois ou N fois parmi 500 génomes séquencés (abscisse).*** Bleu : population constante ; rose : population ayant subi un goulet d'étranglement puis restée constante ; rouge : population ayant cru d'un facteur 1 000 au cours des 400 dernières générations (soit 10 000 ans) (données extraites de la figure 2 de [6]).

dans la population actuelle peut renseigner assez précisément sur son histoire démographique. Encore faut-il, naturellement, disposer de données fiables et qui ne soient pas brouillées par la sélection naturelle… C'est ce à quoi se sont employés les auteurs de l'article qui est le sujet de cette chronique [3].

Un plan expérimental bien étudié

Pour cela, les auteurs ont fait porter leurs efforts dans trois directions [3]. Tout d'abord, et contrairement aux travaux précédents qui portaient sur des exomes, ils ont cherché à analyser des régions neutres du point de vue évolutif, afin d'éviter l'effet de la sélection qui tend à éliminer du génome les variants délétères et fausse ainsi la distribution des SNV. Pour cela ils ont choisi des régions éloignées des gènes connus, ne comportant pas de duplications ou délétions, et le moins possible d'éléments conservés ou répétitifs - l'algorithme mis au point à cet effet est publié [8]. Les auteurs définissent ainsi quinze locus couvrant chacun de 5 à 20 kilobases, soit au total 216 kilobases, qui seront séquencées chez chaque individu de leur échantillon. Naturellement, ils ne peuvent pas garantir absolument qu'il s'agit de zones neutres du point de vue évolutif, mais au moins ont-ils pris toutes les précautions possibles pour s'en assurer.

L'effort a également porté sur la population à étudier. Il est souhaitable que celle-ci soit homogène du point de vue de son ascendance, afin d'éviter une influence de sa structure sur la distribution des fréquences. Ils ont pour cela étudié une cohorte (cohorte ARIC, *atherosclerosis risk in communities*) de près de 10 000 personnes d'origine européenne, déjà caractérisées par leur profil pour 500 000 snip dans le cadre d'une étude GWAS (*genome-wide association study*). Les données ont été ré-analysées en termes de composants principaux, et les résultats *(Figure 3)* ont permis de définir au sein de cette cohorte un sous-ensemble très homogène de 500 personnes sur lesquelles a porté ensuite l'analyse par séquençage.

Enfin le séquençage proprement dit, effectué sur un système Illumina HiSeq, a été poussé jusqu'à une redondance très élevée : près de 300 fois en moyenne (295x exactement). Une série de vérifications, grâce à des données disponibles par ailleurs sur certains membres de la cohorte, permet au total d'assurer que, pour l'essentiel, les variants observés sont bien réels (et ne sont pas des erreurs de séquençage), même lorsqu'ils ne sont retrouvés que dans un seul ADN sur les 500 étudiés (*singletons*)

Les résultats

Après tous ces contrôles, les auteurs trouvent un nombre élevé de *singletons* : 192 variants ne sont observés qu'une seule fois dans les 500 séquences obtenues, alors que la prévision pour une population dont la taille n'aurait pas varié est de 68. On retrouve

donc bien ces nombreux variants rares qui signent une expansion récente – mais peut-on en dire plus sur les paramètres de cette explosion démographique ? Il faut alors modéliser l'évolution de la population, en déduire le spectre de fréquence par nucléotide (SFS, *site frequency spectrum*) attendu, et le comparer à la courbe observée expérimentalement sur les 500 séquences obtenues. Ici apparaît une complication, le fait que l'évolution ancienne de la population après le dernier goulet d'étranglement subi il y a environ 18 000 ans (ou 720 générations) [9][2] a une forte influence sur les résultats du modèle pour l'expansion récente (depuis 10 000 ans ou 400 générations), ce qui laisse un large intervalle de confiance sur le résultat final. Un modèle plus détaillé qui inclut comme paramètre la taille de la population effective[3] avant le début de l'expansion - fixée à 10 000 dans le modèle précédent - donne un meilleur ajustement à la courbe expérimentale et surtout une sensibilité bien moindre aux conditions initiales *(Figure 4)*.

Le résultat auquel on arrive alors est celui d'une expansion récente, dont le début se situe il y a environ 140 générations (3 500 années) avec un taux de 3,4 % environ par génération *(Figure 5)*, ce qui est très élevé, bien plus que les estimations précédentes qui se situaient, rappelons-le, autour de 0,5 % mais sur une durée plus longue.

Des limites… et des enseignements

Soulignons d'abord que ce travail, portant sur un échantillon homogène de 500 personnes originaires d'Europe du Nord *(Figure 3)* nous renseigne uniquement sur les ancêtres de cette population et ne nous dit rien sur l'histoire démographique des populations Asiatiques ou, bien sûr, Africaines. Par ailleurs, malgré le soin apporté à la définition des régions séquencées, au choix de la population et à la qualité de la lecture de l'ADN, l'approche par modélisation a ses limites et la précision des chiffres annoncés (population effective de 5 633 individus au début de l'expansion, il y a 140,8 générations) ne doit pas faire illusion : le message essentiel est que (pour cette population), l'expansion ne s'est pas répartie également sur les 10 000 ans passés depuis le début de l'agriculture, elle a été nettement plus récente et plus rapide. On voit

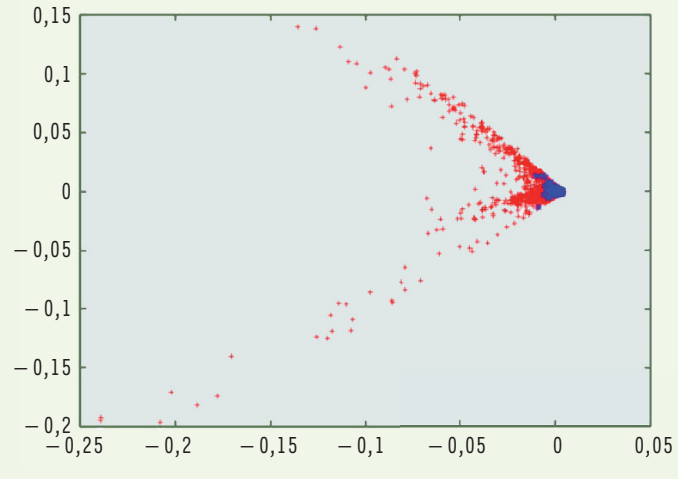

Figure 3. ***Analyse en composants principaux des données de snip sur les 9 716 personnes de la cohorte ARIC.*** Les 500 personnes choisies pour l'étude de Gazavea *et al.* [3] sont figurées en bleu. L'axe principal 1 (en abscisse) correspond principalement à la variation géographique selon la direction Sud/Nord, l'axe 2 (en ordonnée) à la variation Ouest/Est. La population choisie est donc Nord-Européenne.

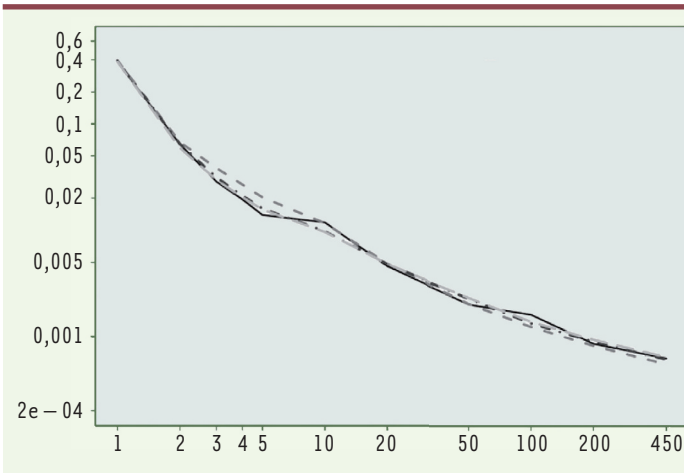

Figure 4. ***Spectre de fréquences observé et prédit par différents modèles.*** En abscisse : nombre d'observations de l'allèle mineur dans l'une des 500 séquences effectuées ; en ordonnée : fréquence de ce résultat (par exemple, la valeur 1 est observée 192 fois sur 500 séquences, soit une fréquence de 38,4 %). Ligne continue : résultat expérimental, lignes pointillées : prévisions des différents modèles, celui qui a été retenu étant le meilleur (pointillés gris longs). Les trois modèles fournissent des estimations assez proches, et des taux d'expansion comparables ; c'est surtout la marge d'erreur qui les différencie.

d'ailleurs sur la *Figure 4* que la différence entre les modèles n'est pas flagrante – et ils aboutissent tous à un taux de croissance entre 3 et 5 %.

Au vu de la *Figure 5*, on peut trouver étonnant que la population soit restée stable durant plusieurs milliers d'années après la dernière crise (goulet d'étranglement il y a 720 générations) pour se mettre

[2] Selon les données historiques et les analyses de Keinan *et al.* [9], la population européenne a subi deux importants goulets d'étranglement, d'une part lors de la sortie d'Afrique (vers 80 000 ans BCE) et, plus récemment, lors du dernier maximum glaciaire (vers 18 000 ans BCE), voir *Figure 5*.

[3] On appelle population effective le nombre d'individus dans une population qui ont une descendance dans la génération suivante (qui participent à la reproduction).

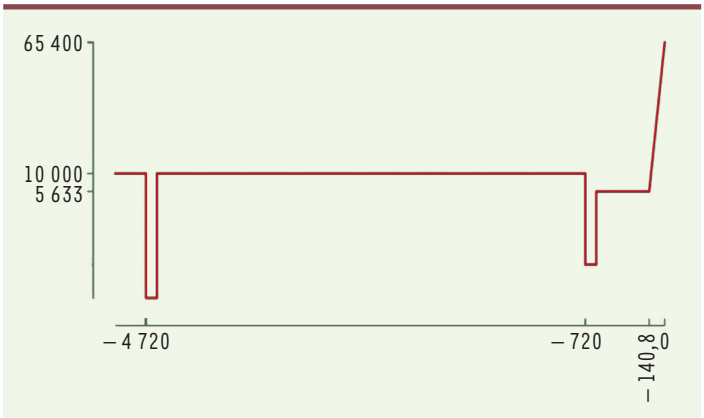

Figure 5. ***Évolution de la population au cours des 120 000 dernières années.*** En abscisse, le nombre de générations (25 ans), en ordonnée la population effective (échelle logarithmique). Les deux creux correspondent aux deux goulets d'étranglement proposés par Keinan *et al.* [9]. Cette représentation est bien sûr très schématique et la précision apparente des chiffres est illusoire.

ensuite à augmenter très rapidement ; pourtant, d'après les auteurs, leurs données excluent une croissance même modérée avant le début de l'accélération, il y a 140 générations. Mais n'oublions pas que ce graphique présente la population *effective*, celle qui participe à la reproduction, et non le nombre total d'individus. Les auteurs suggèrent que les changements sociétaux associés au début de la révolution néolithique et de l'agriculture (à partir de 8 000 ans BCE) pourraient avoir joué un rôle, en accroissant l'espérance de vie et donc le temps de génération, et surtout en favorisant une inégalité sociale qui induit une hiérarchie dans l'accès aux femmes (*differential access to females*), excluant nombre de mâles de la reproduction et donc diminuant la population effective. La population réelle aurait donc pu augmenter au cours de cette période sans changement de la population effective.

En tout cas cet article, qui précise notre histoire démographique récente, souligne, une fois de plus, le caractère exceptionnel de notre espèce par rapport à toutes les autres espèces animales, non seulement par sa taille actuelle et son ubiquité sur notre planète, mais aussi par son rythme d'accroissement. Et, du point de vue méthodologique, on ne peut que rester songeur en voyant jusqu'où peut mener le comptage des variants rares dans l'ADN de cinq cents personnes... Cela s'inscrit en tous cas dans la révolution actuelle de l'anthropologie, qui en deux ou trois ans a vu mettre en évidence la contribution d'un peu d'ADN de Neandertal au génome des populations non Africaines, l'existence des Denisoviens, et la complexité des interactions entre groupes humains tout au long de la préhistoire [4]. ◊

LIENS D'INTÉRÊT

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

SUMMARY

Rare variants and demographic explosion

The abundance of rare variants in human DNA is the consequence of tremendous recent expansion of our population. Careful measurement of neutral variants in a European population points to more recent and more rapid expansion than previously believed. ◊

RÉFÉRENCES

1. Jordan B. Rare is frequent. *Med Sci (Paris)* 2012, ; 28 : 893-6.
2. Jordan B. A la recherché de l'héritabilité perdue. *Med Sci (Paris)* 2010 ; 26 : 541-3.
3. Gazavea E, Maa L, Changa D, *et al.* Neutral genomic regions refine models of recent rapid human population growth. *Proc Natl Acad Sci USA* 2014 ; 111 : 757-62
4. Disotell TR. Archaic human genomics. *Am J Phys Anthropol.* 2012 ; 149 Suppl 55 : 24-39.
5. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet* 2009 : 5(10):e1000695.
6. Keinan A, Clark AG. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* 2012 ; 336 : 740-3.
7. Gravel S, Henn BM, Gutenkunst RN, *et al.* Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci USA* 2011 ; 108 : 11983-8.
8. Arbiza L, Zhong E, Keinan A NRE: a tool for exploring neutral loci in the human genome. *Bmc Bioinformatics* 2012 ; 13 : 301.
9. Keinan A, Mullikin JC, Patterson N, Reich D. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nat Genet* 2007 ; 39 : 1251-5.

TIRÉS À PART

B. Jordan