

Chroniques génomiques

Le déclin de l'empire des GWAS

Bertrand Jordan



Marseille-Nice Génopole,
case 901,
Parc Scientifique de Luminy,
13288 Marseille Cedex 9, France.
brjordan@club-internet.fr

Un peu d'histoire pour commencer

Après les succès initiaux de la Génétique inverse dans les années 1980 et 1990 (myopathie, mucoviscidose, chorée de Huntington, pour ne citer que quelques-unes des maladies monogéniques dont le gène fut découvert durant cette période), la même démarche fut appliquée à des affections complexes, multigéniques, notamment à divers troubles psychiatriques. Et là, ce fut la Bérézina, résumée par ce titre d'un article de Risch et Botstein publié en 1996 et faisant le point des travaux sur la psychose maniaco-dépressive : « *A manic depressive history* » [1]. Résultats contradictoires, multiples localisations aussi peu fondées les unes que les autres, conclusions divergentes d'analyses portant sur les mêmes échantillons... Visiblement, la technologie génétique n'était pas assez puissante pour s'attaquer à ces problèmes compliqués, et les méthodes statistiques employées manquaient cruellement de rigueur.

Avec le séquençage du génome humain, après la construction d'un répertoire systématique des variations ponctuelles, les Snip (SNP, *single nucleotide polymorphisms*), et avec les travaux du consortium HapMap démontrant que l'on pouvait caractériser un génome grâce à l'étude de 500 000 Snip « seulement », les perspectives s'amélioreraient nettement. Et la mise au point vers 2004 par les deux grands industriels du secteur des puces à ADN, Affymetrix et Illumina, de « puces »

permettant l'analyse de 500 000 Snip dans l'ADN d'une personne en une seule expérience d'hybridation et à un coût acceptable (quelques centaines d'euros) permit le démarrage effectif des études d'association génétique « génome entier », les fameuses GWAS. Le schéma général de ces travaux est toujours le même [2] : on constitue un groupe de personnes affectées par la maladie étudiée et un groupe témoin indemne mais aussi proche que possible au niveau de l'ascendance et du mode de vie. L'ADN de chacune de ces personnes est caractérisé par l'analyse de 500 000 Snip, c'est-à-dire par la détermination pour chacun d'eux des allèles présents sur les deux chromosomes du sujet. Une fois obtenu l'ensemble des résultats, on examine si certains allèles sont significativement sur- ou sous-représentés chez les malades par rapport aux sujets témoins. Cela désigne ainsi des régions du génome dans lesquelles doivent se trouver des gènes (ou des séquences régulatrices) influençant la vulnérabilité des personnes à l'affection en cause ; reste alors à les identifier par une étude approfondie de ces régions, qui est grandement facilitée par la connaissance de la séquence du génome humain.

Des études statistiques rigoureuses ont montré qu'il faut analyser l'ADN de plusieurs milliers de personnes pour arriver à une « puissance » permettant de déceler des influences génétiques d'intensité modérée. Il s'agit donc de travaux de grande ampleur, nécessitant des financements significatifs (de l'ordre du million d'euros) et la participation de plusieurs équipes, tant pour le recrutement des sujets que pour la réalisation des expériences et pour leur interprétation. Mais leur intérêt est tel qu'il s'en est déjà pratiqué plus d'une

GWAS : l'acronyme barbare de ce titre renvoie aux études d'association génétique à grande échelle (*Genome-Wide Association Studies*), largement pratiquées depuis deux ou trois ans mais dont on commence à percevoir les limites - qui sont le sujet de cette chronique génomique.



centaine, portant sur des dizaines de maladies ou de traits polygéniques... Et ces travaux ont été couronnés de succès : ils ont abouti à l'identification de plus de deux cent cinquante locus génétiques. Identifications qui sont cette fois solides, généralement confirmées par les études ultérieures, et qui ont parfois abouti à la découverte de gènes dont on commence à comprendre le rôle dans le déclenchement de l'affection [3].

Où est donc passée l'héritabilité ?

Tout est-il donc pour le mieux dans le meilleur des mondes ? Non, car ces résultats indiscutables s'accompagnent d'une surprise de taille : les effets phénotypiques associés à ces variations sont faibles, et ne rendent compte que d'une petite partie de l'héritabilité connue pour la maladie ou le caractère phénotypique considérés. Je prendrai l'exemple de la taille à l'âge adulte : c'est un caractère largement héréditaire, dont l'héritabilité est évaluée à 0,8¹. Une étude GWAS récente a identifié 20 locus influençant cette caractéristique [4]. Mais l'effet de chacun d'eux est minime (quoique rigoureusement prouvé), et à eux tous ils ne rendent compte que de 3 % de l'héritabilité ! D'une manière générale, les variants identifiés correspondent à des risques relatifs faibles, de l'ordre de 1,1 à 1,5 au maximum, et leur effet total dépasse rarement 10 % de l'héritabilité mesurée lors, par exemple, de travaux portant sur les jumeaux. Comme le soulignait un récent article de *Nature* dont le titre, *The case of the missing heritability*, rappelait à dessein un roman policier [5], il y a là une énigme qu'il est urgent de résoudre. La faiblesse des effets constatés signifie en effet que la valeur prédictive des variations génétiques observée est quasiment nulle, contrairement à ce que voudraient nous faire croire les entreprises (*Navigenics*, *23andMe*, *Decode genetics*...) qui prétendent révéler votre destin à partir d'une analyse d'ADN vendue de 500 à 2 000 dollars [2]. Les identifications réalisées dans le cadre des études GWAS sont certes scientifiquement valables et utiles pour la compréhension du mécanisme pathogène (donc porteuses d'espoirs thérapeutiques), mais, rendant compte de moins d'un dixième des héritabilités constatées, elles passent visiblement à côté d'un phénomène important...

Une amorce de réponse

Comment résoudre ce paradoxe ? Il faut pour cela revenir sur ce qu'examinent réellement les GWAS. Elles se limitent aux Snip, faisant (pour le moment du moins) l'impasse sur les *copy number variations* (CNV), ces délétions, duplications ou inversions dont on a découvert récemment plusieurs centaines de milliers dans notre génome. Et même pour les Snip, elles ne donnent pas une image complète des variations génétiques entre individus. Par la force des choses, les 500 000 Snip représentés sur les puces d'*Affymetrix* ou d'*Illumina* (et qui ont préalablement été étudiés par le consortium HapMap)

correspondent à des polymorphismes assez facilement repérables dans un échantillon de population : la règle adoptée a été de ne retenir que les Snip pour lesquels la fréquence de l'allèle mineur est au moins égale à 5 %. Cet usage était nécessaire pour limiter les difficultés dans le positionnement des Snip lors de l'établissement des cartes d'haplotypes ; mais il a pour conséquences que les GWAS n'examinent que les variants fréquents... Selon une hypothèse largement répandue, les maladies multigéniques fréquentes (diabète, hypertension, schizophrénie...) seraient dues à la conjonction de plusieurs allèles eux aussi fréquents : c'est la règle « *common disease, common variant* » souvent évoquée depuis une dizaine d'années [6]. Les résultats de la centaine d'études d'association pangénomiques pratiquées à ce jour indiquent que cette hypothèse est très probablement fautive : les variants communs ne rendant compte que d'une faible partie de l'héritabilité, le reste est vraisemblablement dû à des variants rares (ponctuels ou non) dont ces études ne tiennent pas compte puisque les puces utilisées ne les voient pas².

Des puces à la séquence ?

Il va donc falloir reprendre ces études d'association en se mettant en mesure, cette fois, de détecter des variants rares, présents à des fréquences inférieures à un pour cent ou même un pour mille. Ceux-ci, par définition, n'ont pas été répertoriés en tant que Snip : la seule manière de les détecter est de procéder au séquençage intégral de l'ADN de chacun des sujets de l'étude... C'est un projet apparemment chimérique, une approche qui aurait été totalement impossible il y a seulement deux ou trois ans. Mais les techniques de séquençage rapide, apparues depuis deux ou trois ans et commercialisées par *Roche* (454), *Illumina* (*Solexa*) et *Applied Biosystems* (*SOLiD*) sont bien établies, et le coût du séquençage de l'ADN d'une personne se rapproche actuellement de 10 000 dollars [7]. Les perfectionnements apportés de manière continue à ces machines comme aux réactifs, dans un contexte d'intense compétition et d'adoption rapide de ces techniques (près de 1 000 systèmes déjà installés), contribuent à faire encore baisser ce chiffre. De plus différentes techniques de « troisième génération » sont proches de la commercialisation (ou en sont déjà à ce stade, comme *Helicos*) et promettent de gagner encore un ou deux ordres de grandeur... Déjà des entreprises comme le

¹ C'est-à-dire que 80 % de la variation observée au sein d'une population au mode de vie homogène est due à l'hérédité.

² Cette hypothèse des variants rares n'est pas la seule possible, on pourrait éventuellement imaginer une interaction entre allèles communs multipliant fortement leur influence en cas de présence simultanée, mais ceci paraît peu probable et aucune donnée actuelle ne va dans ce sens.



« 1 000 genomes project »³ s'attaquent à ce type de programme [8], et l'on peut envisager que lorsque le « génome à 1 000 dollars » sera une réalité, les études d'association génétique basculeront vers la séquence. Il ne faut pas en sous-estimer la difficulté : les quantités de données à traiter vont être gigantesques (des dizaines de téraoctets, de quoi saturer les capacités de stockage et de transmission de l'information), la précision requise est élevée afin d'éliminer les erreurs de séquence, et le traitement des résultats sera différent de celui mis au point pour des Snip dont la position est connue *a priori*. Le séquençage à très grande échelle n'en représente pas moins l'avenir de ce secteur...

La fin des Snip et des puces ?

Faut-il donc abandonner les études d'association à grande échelle telles qu'elles ont été pratiquées jusqu'ici ? Cette question a fait l'objet de trois « perspectives » publiées fin avril dans le *New England Journal of Medicine*. Pour l'un des auteurs, David Goldstein [9], les variants communs dont l'effet est notable ont déjà été trouvés, et de coûteuses études ultérieures (avec un nombre de sujets plus important) ne pourront aboutir qu'à la mise en évidence de variants dont l'effet est encore plus faible et l'intérêt discutable. D'autres, comme Joel Hirschhorn [10] soulignent l'intérêt potentiel d'études GWAS étendues à des Snip et des variants moins fréquents (descendant jusqu'à des fréquences de 1 %), tout au moins tant que le coût du séquençage n'est pas effectivement arrivé à un niveau comparable à celui d'une analyse par puce à ADN. Peter Kraft et David Hunter [11], eux, insistent sur les incertitudes dans l'évaluation du risque génétique sur la base de données incomplètes. Ces débats sont importants en termes de politique scientifique, puisqu'il s'agit de programmes lourds dans lesquels chaque étude consomme des millions de dollars en appareils, consommables et réactifs... En tout cas, les fabricants de *microarrays* s'inquiètent ; *Affymetrix*, qui est en position financière difficile et perd une vingtaine de millions de dollars par trimestre, tout comme *Illumina*, en meilleure posture grâce notamment à son département de séquençage, prévoient une baisse des ventes de puces au cours de cette année et s'ingénient à trouver de nouvelles applications pour leurs produits. Le déclin de cette approche est indiscutablement amorcé ; sa rapidité dépendra du progrès effectif des projets fondés sur le séquençage, qui vont sans doute mettre quelques temps à résoudre les problèmes évoqués plus haut. Mais la tendance est nette. D'ailleurs, les firmes qui se sont lancées dans le « génome personnel » évoquent elles aussi la perspective du séquençage...

³ <http://www.1000genomes.org/page.php>

Le difficile passage du génotype au phénotype

Tout ceci nous fait toucher du doigt à quel point l'espoir un peu naïf des années 1990, l'idée que la connaissance de « la » séquence humaine allait – presque par magie – nous permettre de comprendre notre fonctionnement biologique et de réparer ses imperfections⁴, à quel point cet espoir faisait l'impasse sur la complexité du vivant. Nos connaissances ne cessent de s'accroître, à un rythme qui continue à s'accélérer ; les moyens techniques dont nous disposons ont fait des progrès inimaginables en quelques années ; mais nous sommes encore bien loin de comprendre comment une séquence d'ADN dirige la construction d'un être vivant... ♦

The decline of genome-wide association studies

RÉFÉRENCES

1. Risch N, Botstein D. A manic depressive history. *Nat Genet* 1996 ; 12 : 351-3.
2. Jordan B. « Génome personnel », gadget ou révolution ? *Med Sci (Paris)* 2008 ; 24 : 91-4.
3. Hardy J, Singleton A. Genomewide association studies and human disease. *N Engl J Med* 2009 ; 360 : 1759-68.
4. Weedon MN, Lango H, Lindgren CM, et al. Genome-wide association analysis identifies 20 loci that influence adult height. *Nat Genet* 2008 ; 40 : 575-83.
5. Maher B. Personal genomes: the case of the missing heritability. *Nature* 2008 ; 456 : 18-21.
6. Reich DE, Lander ES. On the allelic spectrum of human disease. *Trends Genet* 2001 ; 17 : 502-10.
7. Jordan B. Une révolution longuement attendue. *Med Sci (Paris)* 2008 ; 24 : 869-73.
8. Jordan, B. Un, deux, trois... mille génomes ? *Med Sci (Paris)* 2008 ; 24 : 237-8.
9. Goldstein DB. Common genetic variation and human traits. *N Engl J Med* 2009 ; 360 : 1696-8.
10. Hirschhorn JN. Genomewide association studies: illuminating biologic pathways. *N Engl J Med* 2009 ; 360 : 1699-701.
11. Kraft, P, Hunter DJ. Genetic risk prediction: are we there yet? *N Engl J Med* 2009 ; 360 : 1701-3.

⁴ « Des gènes pour guérir », slogan d'un Téléthon pas si ancien (1993).

TIRÉS À PART

B. Jordan