médecine/sciences 2000 ; 16 : 988-90

Le nombre de gènes dans le génome humain : les paris sont ouverts

Pourquoi est-il si important de connaître le nombre de gènes dans le génome humain ? Depuis que l'on sait l'influence de l'hérédité sur le déroulement de la vie, il est naturel de vouloir mesurer et quantifier cette information génétique pour mieux comprendre la manière dont cette influence s'exerce. Connaître le nombre de gènes qui contrôlent notre développement et nos fonctions physiologiques revient aussi à circonscrire l'information à la base de la vie, un peu comme connaître la circonférence de la terre était important pour les navigateurs d'autrefois. Les motivations peuvent aussi être d'ordre plus pragmatique et d'intérêt plus immédiat. Environ 85 % du génome humain est aujourd'hui séquencé, et la prochaine étape consiste à annoter cette séquence, c'est-à-dire justement à localiser les gènes. Ce travail est crucial et la compétition est grande. Connaître à l'avance le nombre total de gènes à trouver revient à savoir, au départ d'une course à pied, à quelle distance se trouve l'arrivée. Mais il est également un enjeu commercial à cette connaissance : un certain nombre d'entreprises spécialisées en biotechnologies monnayant l'accès à des bases de séquences de gènes privées, il est évident que la valeur marchande de ces bases de données est influencée par le nombre de gènes qui s'y trouvent.

Il y a encore moins d'un an, le nombre de gènes contenus dans le génome humain était estimé entre 60 000 et 80 000. Cette fourchette était fondée principalement sur deux études. La première repose sur l'observation expérimentale que le génome contiendrait 45 000 îlots CpG non méthylés, que 56 % des gènes sont associés à un îlot CpG et que chaque îlot CpG est associé à un gène. Une simple extrapolation prédit 80 000 gènes dans le génome [1]. La deuxième est fondée sur l'assemblage d'un petit nombre de séquences EST *(expressed sequence tag)* en groupes distincts, puis mesure la fraction qui reconnaît des gènes connus. La différence permet d'estimer la fraction inconnue, donc de reconstituer le nombre total de gènes [2]. Cette étude prédit entre 60 000 et 70 000 gènes. Ces deux mesures sont relativement proches, et ont donc servi de base de référence pendant plusieurs années. De plus, il existe un indicateur des progrès effectués dans la découverte de ces gènes par la communauté scientifique. En effet, une base de données localisée au *National Center for Biotechnology Information* (NCBI) aux États-Unis et appelée *Unigene* assemble les séquences d'EST produites à travers le monde et les regroupe sur la base de leur similitude (http://www.ncbi.nlm.nih.gov/unigene). Il est ainsi possible de créer une liste non redondante à partir de l'ensemble de ces séquences. Il est plus ou moins admis que chaque groupe (ou *cluster*) représente un gène différent, et que le nombre total de groupes permet d'estimer combien de gènes ont été clonés et séquencés jusqu'à ce jour. Cependant, vers la fin de l'année 1999, le nombre de groupes uniques a dépassé 80 000, puis 90 000, et il a donc fallu réévaluer à la hausse le nombre de gènes dans le génome. Un chiffre non officiel a alors couru : 100 000 gènes. Peut-être encouragée par cette tendance, *Incyte*, une société de biotechnologie, a annoncé par communiqué de presse jusqu'à 140 000 gènes différents, une valeur calculée selon la même méthode qu'*Unigene* mais à partir d'une base de données privée. Certes ces chiffres ont surpris, mais sans argument scientifique rigoureux permettant de les contredire, la communauté scientifique a semblé accepter qu'il soit possible de repousser ainsi la limite du nombre de gènes dans le génome, peut-être au regard de l'incroyable complexité de l'organisme humain.

Les choses en étaient là lorsque, au début de l'année 2000, une équipe de chercheurs du Génoscope met la touche finale à un nouvel outil de détection de gènes humains fondé sur des recherches d'homologies avec le génome d'un autre vertébré, le poisson « bulle » *Tetraodon nigroviridis*. Cet outil, baptisé Exofish, a pour but initial de contribuer à annoter la séquence du génome humain alors en cours d'achèvement. Après avoir analysé tout l'ADN humain séquencé à ce jour, les chercheurs font une estimation du nombre de gènes que contiendrait le génome humain. La surprise est de taille et l'on croit d'abord à une erreur : Exofish prédit entre 28 000 et 34 000 gènes. Cependant les résultats semblent robustes et sont annoncés lors d'une conférence à Cold Spring Harbor près de New York. Ces chiffres sont en contradiction avec les estimations admises jusqu'alors, et ils soulèvent un important débat parmi les participants. Il devient rapidement clair que beaucoup ne sont pas convaincus, au point que des paris sont organisés avec une décision finale en 2003. Le gagnant obtiendra une copie dédicacée du livre *La Double hélice* de James Watson (les statistiques de ce pari sont disponibles à http://www.ensembl.org/genesweep.html). Les résultats du Génoscope sont ensuite publiés dans *Nature Genetics* [3] et un autre groupe, celui de Phil Green

de l'Université de Washington à Seattle, publie dans le même numéro une estimation proche (35 000 gènes), mais fondée sur une méthode différente [4]. Cependant, le plus surprenant est qu'un troisième groupe, celui de J. Quackenbush au TIGR *(The Institute for Genomic Research)*, présente aussi dans ce numéro une troisième estimation, avec plus de 120 000 gènes [5]. Cette situation inhabituelle (au moins l'un des articles se trompe largement !) reflète malgré tout assez bien la réalité, d'une part quant à l'importance scientifique que peut avoir la connaissance du nombre de gènes, et d'autre part quant à l'origine de ces estimations pour le moins contrastées.

Le débat scientifique s'articule d'abord autour des différentes méthodes utilisées afin d'obtenir ces récentes estimations et, au-delà, tente de concilier une complexité physiologique inégalée dans le règne animal avec un nombre de gènes de trois à quatre fois plus petit que prévu. La méthode utilisée par l'équipe du Génoscope est originale car, pour la première fois, de très larges fractions de génomes de vertébrés ont été comparées, sans connaître à l'avance les positions des gènes dans l'une ou l'autre des espèces. Le principe repose simplement sur l'observation que les séquences d'ADN codant pour des protéines (les exons) subissent moins de changements au cours de l'évolution que l'ADN des introns ou celui situé entre les gènes. Les séquences qui se ressemblent aujourd'hui entre deux espèces éloignées sont donc très souvent celles qui sont dérivées des gènes de leur ancêtre commun, alors que les autres séquences ont accumulé un très grand nombre de mutations et sont devenues très différentes. Ce principe était connu depuis longtemps et est à la base de la génomique dite «comparative» qui permet d'exploiter la séquence d'ADN d'organismes modèles comme la souris, la drosophile, ou la levure. Le Génoscope a cependant choisi un nouvel organisme modèle, le poisson *Tetraodon nigroviridis*, en raison de son génome compact (huit fois plus petit que le génome humain) et de sa grande distance évolutive qui permet un bon contraste entre les séquences conservées (les gènes) et les séquences non conservées. *Tetraodon* est un cousin d'eau douce de *Fugu rubripes*, une espèce marine déjà proposée par Sydney Brenner comme modèle pour la génomique comparative [6]. Après avoir séquencé environ un tiers du génome de *Tetraodon*, les chercheurs l'ont comparé à un très grand nombre d'ADN complémentaires (ADNc) complets connus (plus de 4 800 gènes) et ont noté que globalement un gène humain possède 3,18 régions conservées (nommées aussi *ecore* pour *evolutionary conserved region*) avec les séquences de *Tetraodon*. Après cette calibration, ils ont comparé la séquence de *Tetraodon* à la séquence du génome humain disponible dans les bases de données publiques (42 % du génome) et ont trouvé plusieurs dizaines de milliers d'*ecores*. La suite est très simple, il suffisait d'extrapoler au génome entier puis de diviser par 3,18 pour trouver le nombre de gènes dans le génome humain : environ 30 000 gènes. La méthode adoptée par l'équipe de Phil Green repose sur des données entièrement différentes. Le principe est fondée sur la construction d'un premier ensemble de 7 662 séquences d'ADNc complets à partir de banques de données. Cet ensemble est considéré comme représentatif de tous les gènes du génome. Ils ont ensuite construit un deuxième ensemble à partir de plus d'un million d'EST, en regroupant les séquences identiques pour créer environ 43 000 groupes non redondants. La comparaison des deux ensembles indépendants identifie 23 % des ADNc complets ; en d'autres termes, ces 7 662 ADNc représentent 23 % des gènes humains et il suffit alors de ramener cette valeur à 100 % pour trouver 33 630 gènes dans le génome. Cette approche, totalement indépendante de celle du Génoscope, arrive à un résultat pratiquement identique, ce qui renforce naturellement leur caractère prédictif. La troisième méthode, adoptée par l'équipe du TIGR, utilise également des séquences d'EST, mais les analyse de façon différente. A partir d'environ 1,6 million d'EST, des groupes de séquences identiques sont créés afin d'obtenir une liste non redondante de 73 655 groupes distincts. La différence clé avec l'analyse précédente est que les auteurs supposent que chaque groupe d'EST représente un gène différent. Comme la comparaison de cet ensemble à un ensemble de gènes connus ne permet l'identification que de 55 % de leurs groupes d'EST, ils estiment alors que 45 % des gènes sont absents de la collection de 73 655 EST. Cela implique que le génome humain contiendrait 134 000 gènes et, après diverses corrections, leur estimation finale est de 110 000 à 118 000 gènes.

Il est important d'essayer de comprendre les raisons pour lesquelles deux études fondées sur les mêmes données initiales parviennent à des résultats entièrement différents. Moins que la méthode de calcul elle-même, la différence provient surtout de la manière de considérer les séquences d'EST. En raison de leur très grande redondance, une étape critique consiste à les assembler en groupes distincts et non redondants sur la base de similitudes de séquences. Ce processus est compliqué par plusieurs facteurs. Les EST sont essentiellement de courtes séquences (environ 400 bases) issues de la région non traduite située en 3' des ADNc (3'UTR). Le processus de regroupement de ces séquences suppose que chaque gène ne possède qu'un seul 3'UTR. Or il est de plus en plus fréquent de trouver des cas de gènes qui peuvent utiliser différents UTR en fonction de leur domaine d'expression, grâce à l'utilisation de sites de polyadénylation alternatifs, combinés ou non à un épissage alternatif. Par conséquent, plusieurs groupes d'EST différents peuvent provenir du même gène. Par ailleurs, certains EST sont assez longs pour contenir une partie des exons codants ou sont séquencés à partir de l'extrémité 5' des ADNc : un épissage alternatif des exons peut là aussi créer plusieurs EST différents, mais qui proviennent tous du même gène. Ces deux phénomènes conduisent les programmes d'assemblage d'EST à créer plusieurs groupes pour un même gène et par conséquent ont tendance à « gonfler » le nombre de gènes apparents contenus dans les banques d'EST. Un troisième facteur

provient de la contamination des banques d'ADNc avec des EST chimériques ou des séquences d'ADN génomique. Malgré le soin apporté au nettoyage des banques de données, il est impossible de les éliminer tous, et chaque séquence contaminante contribuera à la création d'un nouveau gène si elle est incluse dans les assemblages. Ces observations montrent surtout que l'assemblage des séquences d'EST est un exercice très complexe à partir duquel il est hasardeux de dériver des extrapolations. En revanche, les méthodes « d'échantillonnage direct » adoptées par le Genoscope et Phil Green reposent sur des approches plus simples qui réduisent les risques d'erreur.

Mis à part les faiblesses ou les points forts de telle ou telle méthode, le débat scientifique sur le nombre de gènes concerne aussi l'évolution des organismes vivants. Peut-on concevoir que l'être humain ne soit doté que de si peu de gènes au regard d'autres organismes dont le génome est déjà séquencé ? Cela ne semble pas si aberrant si l'on admet que la complexité physiologique n'est pas directement proportionnelle au nombre de gènes. Il suffit par exemple de doubler le nombre de gènes (6 000) de la levure, un eucaryote unicellulaire, pour obtenir celui de la drosophile (13 600), un métazoaire complexe doté d'un système nerveux. Il n'est donc pas impensable de considérer qu'un doublement supplémentaire puisse expliquer la différence entre la drosophile et l'être humain. On sait par ailleurs que la plante *Arabidopsis thaliana* posséderait environ 25 000 gènes, sans que cela explique une complexité semblable à celle des vertébrés. Il est donc clair que les gènes sont utilisés de façon différente selon les organismes. Chez l'être humain, la régulation de l'expression des gènes, l'épissage alternatif, les modifications post-transcriptionnelles des protéines et leurs interactions mutuelles participent sans doute beaucoup à cette complexité qui reste encore un mystère si fascinant.

RÉFÉRENCES

1. Antequera F, Bird A. Number of CpG islands and genes in human and mouse. *Proc Natl Acad Sci USA* 1993 ; 90 : 11995-9.
2. Fields C, Adams MD, White O, Venter JC. How many genes in the human genome ? *Nat Genet* 1994 ; 7 : 345-6.
3. Roest Crollius H, Jaillon O, Bernot A, *et al.* Human gene number estimate provided by genome wide analysis using *Tetraodon nigroviridis* genomic DNA. *Nat Genet* 2000 ; 25 : 235-8.
4. Ewing B, Green P. Analysis of expressed sequence tags indicates 35,000 human genes. *Nat Genet* 2000 ; 25 : 232-4.
5. Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL, Quackenbush J. Gene Index analysis of the human genome estimates approximately 120,000 genes. *Nat Genet* 2000 ; 25 : 239-40.
6. Brenner S, Elgar G, Sandford R, Macrae A, Venkatesh B, Aparicio S. Characterization of the pufferfish (Fugu) genome as a compact model vertebrate genome. *Nature* 1993 ; 366 : 265-8.

Hugues Roest Crollius
Olivier Jaillon

Génoscope, 2, rue Gaston-Crémieux, 91057 Évry Cedex, France.