médecine/sciences 2022 ; 38 : 609-12

Chroniques génomiques

Séquence du génome : la fin du commencement

Bertrand Jordan

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

Un génome très utile mais imparfait

Le premier avril 2022 est paru un numéro spécial de la revue *Science* consacré à la séquence du génome humain, et dont l'article principal est intitulé *The complete sequence of a human genome* [1]. S'agit-il d'un poisson d'avril ? Après tout, l'obtention de la séquence de l'ADN humain a été annoncée en 2000, publiée début 2001 [2] *(Figure 1)*, et la version « finie » a, elle, été publiée en 2004 [3]. Pourquoi y revenir près de vingt ans plus tard ? Cette séquence a en réalité fait l'objet d'améliorations continues [4] aboutissant à la trente-huitième version actuellement utilisée et appelée GRCh38.4 [5], mais elle reste entachée d'imperfections majeures. Elle comporte de nombreux « trous » (*gaps*) dont la séquence n'a pas pu être déterminée et dont la longueur totale est estimée à plus de cent cinquante millions de bases (mégabases). De ce fait, elle n'est pas d'un seul tenant : elle est formée de 949 séquences distinctes (*contigs*) alors qu'idéalement chaque chromosome devrait être couvert de manière continue d'un télomère à l'autre. Les régions centromériques de tous les chromosomes, les télomères, les bras courts de chromosomes acrocentriques (chromosomes 13, 14, 15, 21 et 22) ne sont pas représentés. En effet, ils contiennent de multiples séquences répétées dont l'assemblage est quasiment impossible à partir des courts segments (moins de 500 bases) lus par la technique dominante *Illumina*. Pour la même raison, le répertoire des duplications est incomplet. De plus, cette séquence est une sorte de *patchwork* puisque les ADN séquencés proviennent de plusieurs individus ; enfin, le taux d'erreur est estimé à moins d'une erreur par 10 000 bases ce qui est déjà une belle performance mais reste néanmoins trop élevé pour certaines applications comme la détection des nouvelles mutations. Telle qu'elle est, cette séquence a néanmoins révolutionné la biologie, permettant ou accélérant de très nombreuses études : lorsqu'on cherche le gène impliqué dans une maladie génétique, par exemple, la séquence fournit immédiatement la liste des gènes existant au voisinage du point désigné par l'analyse génétique. Autrefois, il fallait établir l'anatomie détaillée de cette région pour espérer y trouver le gène : pour la maladie de Huntington, il a fallu dix ans pour aller de la localisation génétique du gène impliqué (1983) à son isolement effectif (1993) [6]...

Un ensemble d'avancées techniques

Comme nous le verrons plus loin, il ne s'agit pas cette fois d'une nième (trente-neuvième ?) version de la séquence, d'une amélioration incrémentale, mais au contraire d'un réel saut qualitatif aboutissant à la séquence complète de chaque chromosome d'une extrémité à l'autre. Ce résultat n'aurait pas été possible sans l'utilisation de nouvelles approches et de techniques de séquençage récemment mises au point.
La première avancée concerne l'ADN étudié : le consortium T2T (*telomere to telomere*) a séquencé l'ADN d'un tissu très particulier, une « môle hyaditiforme » [7]. Il s'agit d'une anomalie rare de la grossesse qui se manifeste par la croissance d'une masse cellulaire dont les chromosomes maternels sont absents et où tous les chromosomes paternels sont dupliqués. Les cellules de la môle sont donc totalement homozygotes avec deux jeux de chromosomes paternels (et deux X, les cellules YY n'étant pas viables). C'est un avantage considérable pour la lecture du génome puisqu'il n'y a aucune hétérozygotie et aucun problème d'attribution d'une séquence à un chromosome ou à son homologue (ils sont tous identiques). Proposée dès les débuts du programme génome [8], cette approche commence à être appliquée depuis quelques années. On s'assure aussi par ce choix que l'on séquence bien un génome donné et non un *patchwork* composite provenant de plusieurs individus.

Les autres avancées se situent au niveau du séquençage proprement dit, et il s'agit bien sûr des nouvelles techniques de lecture longue (*long-read sequencing*). Deux approches sont actuellement opérationnelles et commencent à concurrencer *Illumina*, le leader du marché. Il s'agit du système *PacBio (Pacific BioSystems)* dont le principe de lecture est similaire à celui d'*Illumina*, mais qui, grâce à un système optique sophistiqué, parvient à séquencer en temps réel une molécule unique d'ADN (SMRT, *single molecule real-time sequencing*), sans étape d'amplification, et à lire ainsi une ou plusieurs dizaines de kilobases d'un seul tenant [9]. Jusqu'à récemment, cette technique était beaucoup plus chère et moins exacte que la méthode *Illumina*, mais elle a fait de grands progrès et est devenue plus abordable. L'autre méthode de séquençage est bien sûr celle des nanopores, commercialisée par l'anglais *Oxford Gene Technology*. Elle consiste à faire passer la molécule d'ADN à travers un nanopore de très petites dimensions, et à déterminer la séquence au passage grâce aux signaux électriques générés par les différentes bases [10, 11] (➜).

(➜) Voir la Chronique génomique de B. Jordan, *m/s* n° 8-9, août-septembre 2017, page 801, et la Synthèse de F. Montel, *m/s* n° 2, février 2018, page 161

Le taux d'erreurs est relativement élevé, mais cette technique est capable de lire une centaine de kilobases d'un seul tenant (on est même allé jusqu'à plusieurs mégabases). Ces lectures ultra-longues sont essentielles car elles permettent de lire des régions contenant de nombreuses séquences répétées (ce qui est le cas des centromères et télomères). Enfin, de nouvelles méthodes informatiques permettent d'intégrer toutes ces informations pour obtenir une séquence que l'on peut vraiment considérer comme « finie ». Elle est désignée par l'acronyme T2T-CHM13 pour *Telomere to Telomere – Complete Hyaditiform Mole* 13, CHM13 en abrégé dans la suite de cette chronique.

« La séquence complète d'un génome humain »

Le titre très sobre de l'article présentant cette nouvelle séquence [1] affirme que cette dernière est « complète », et c'est bien le cas. Les trous (*gaps*) encore présents dans la version GRCh38 (plus de neuf cents) ont été comblés grâce aux lectures longues, et à chaque chromosome correspond maintenant une séquence d'un seul tenant allant d'un télomère à l'autre : le consortium T2T qui a coordonné ce travail a bien mérité son nom. Il n'y a plus aucune séquence non reliée (*unplaced*) alors que celles-ci représentaient plus de onze mégabases dans GRCh38. Et bien sûr, centromères et télomères sont maintenant inclus dans la séquence finale. La *Figure 2* montre, à titre d'exemple, l'étendue et la nature des informations nouvelles apportées par la séquence CHM13 par rapport à l'ancienne référence GRCh38 pour le chromosome 20. On y voit, en noir, les vides qui ont été couverts par la nouvelle séquence : centromère, télomères, mais aussi sept zones dispersées le long du chromosome. Au-dessus du schéma, sont montrés les duplications segmentaires (en bleu)[1] et les satellites centromériques (en jaune)[2]. Enfin, les gènes déjà repérés sur la séquence GRCh38 sont figurés en gris, ceux que révèle la nouvelle séquence sont indiqués en rouge.

Figure 1. ***Couverture de la revue* Nature *(15 février 2001) annonçant la séquence du génome humain.***

On voit que les modifications apportées par CHM13 sont substantielles : il ne s'agit pas d'un « polissage » comme celui qui a marqué les états successifs de la séquence GRCh, mais bien d'un saut qualitatif aboutissant à la représentation exacte et complète d'un génome [12]. Celle-ci ajoute 238 mégabases à la séquence déjà connue, soit 8 % du génome, et 99 nouveaux gènes codant des protéines. Et les bras courts des chromosomes acrocentriques (13, 14, 15, 21 et 22), jusque-là *terra incognita*, sont maintenant intégralement séquencés. Certes, ils sont essentiellement constitués de différents types d'ADN répété, mais ils contiennent aussi les gènes des ARN ribosomiques et ne sont pas sans importance biologique et médicale. Enfin le taux d'erreurs est estimé à moins d'une erreur par dix mégabases – une amélioration d'un facteur mille par rapport aux premières séquences du génome [2, 3]. Notons aussi qu'il s'agit de la séquence d'UN génome, comme le souligne le titre de l'article. C'est le génome d'une cellule totalement homozygote, ce qui lève bien des ambiguïtés, et de cette seule cellule plutôt

[1] Les duplications segmentaires sont des éléments de séquence longs de quelques kilobases, répartis dans le génome et fortement homologues entre eux. Elles sont aussi appelées répétitions à faible copie (*low copy repeats*).

[2] L'ADN répété centromérique est principalement de l'ADN α-satellite dont le motif unitaire mesure 171 bases.

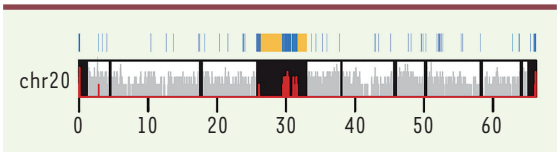

Figure 2. ***Le chromosome 20 avec les nouvelles informations apportées par la séquence complète CHM13.*** Échelle du bas : position le long du chromosome en mégabases ; voir le texte pour la signification des couleurs. Extrait partiel et modifié de la figure 1 de [1].

qu'un assemblage de séquences d'origines diverses comme précédemment. Mais ce n'est pas « le » génome humain, comme on a eu tendance à dire par le passé : nous avons maintenant conscience de la diversité génétique humaine et des multiples différences (ponctuelles ou non) trouvées entre l'ADN de différents individus. Pour bien faire, il faudra répéter cet exercice, nous y reviendrons. Notons enfin que le chromosome Y n'est pas représenté. Comme les môles hyaditiformes YY ne sont pas viables, il faudra pour le séquencer appliquer les mêmes techniques à une cellule XY – mais cela ne devrait pas poser de difficulté majeure puisque le chromosome Y y figure à l'état haploïde.

On n'a pas fini de séquencer

Cet aboutissement (une séquence réellement complète) n'est pas un point final. Comme le disait Churchill dans un tout autre contexte[3] : « *Ceci n'est pas la fin, ni même le commencement de la fin, mais c'est peut-être la fin du commencement* ». Le succès du consortium T2T ne sera pas un tour de force sans lendemain. Selon de bons experts, il montre que nous en sommes maintenant au stade où un laboratoire peut produire en quelques semaines et pour quelques dizaines de milliers de dollars une séquence quasiment sans défauts d'un génome humain [13]. Vision un peu optimiste, car les génomes à séquencer seront généralement diploïdes et hétérozygotes : leur assemblage complet demandera encore un perfectionnement des algorithmes utilisés [14]. Pour que cette séquence CHM13 soit réellement utile, il va aussi falloir se préoccuper d'y reporter l'ensemble des annotations fonctionnelles accumulées au fil des années et repérées sur la séquence GRCh38 – c'est un problème non trivial et qui va lui aussi nécessiter de gros travaux en bioinformatique. Et, très logiquement, il va falloir réévaluer la diversité génétique humaine en comparant des séquences complètes (au sens de CHM13). On savait déjà que l'homogénéité à 99,9 % des génomes humains, annoncée à grand bruit en 2001 [15], était surestimée – en tenant compte des différences non ponctuelles (duplications, délétions, inversions) on arrivait plutôt à 99,5 %. Gageons que de nouvelles évaluations fondées sur la comparaison de séquences complètes feront encore baisser ce chiffre. Plus généralement, ce génome sans « matière noire » va sûrement aider à résoudre des questions de génétique médicale impliquant des zones jusqu'ici non séquencées (ou non alignées). C'est indubitablement une avancée majeure qui va encore décupler l'utilité de cette séquence dont certains doutaient, au départ, qu'elle soit intéressante [16]... (➜). ◊

(➜) Voir la Chronique génomique de B. Jordan, *m/s* n° 10, octobre 1990, page 906

[3] La victoire d'El Alamein, en novembre 1942.

SUMMARY

Human genome: The end of the beginning

Two decades after its original publication, a new human genome sequence has just been published. Far from being an incremental improvement, it is at last really complete, covering each chromosome from one end to the other, with full elucidation of repeated sequences and an extremely low error rate. This is a major advance that tremendously increases our knowledge of our genome and will lead to important scientific and clinical developments. ◊

LIENS D'INTÉRÊT

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

RÉFÉRENCES

1. Nurk S, Koren S, Rhie A, *et al*. The complete sequence of a human genome. *Science* 2022 ; 376 : 44-53.
2. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* 2001 ; 409 : 860–921.
3. International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome. *Nature* 2004 ; 431 : 931-45.
4. Church DM, Schneider VA, Graves T, *et al*. Modernizing reference genome assemblies. *PLOS Biol* 2011 ; 9 : e1001091.
5. GRCh38.p14 2022/02/03 (Genome Reference Consortium) https://www.ncbi.nlm.nih.gov/assembly/GCA_000001405.29.
6. Bates GP. History of genetic disease: the molecular genetics of Huntington disease - a history. *Nat Rev Genet* 2005 ; 6 : 766-73.
7. Kajii T, Ohama K. Androgenetic origin of hydatidiform mole. *Nature* 1977 ; 268 : 633-4.
8. Eichler EE, Surti U, Ophoff R. Proposal for construction a human haploid BAC library from hydatidiform mole source material (2002). www.genome.gov/Pages/Research/Sequencing/BACLibrary/HydatidiformMoleBAC021203.pdf
9. Rhoads A, Au KF. PacBio Sequencing and Its Applications. *Genomics, Proteomics Bioinformatics* 2015 ; 13 : 278-89.
10. Jordan, B. Séquençage d'ADN : l'offensive des nanopores *Med Sci (Paris)* 2017 ; 33 : 801-4.
11. Montel F. Séquençage de l'ADN par nanopores – Résultats et perspectives. *Med Sci (Paris)* 2018 ; 34 : 161–5.
12. Church DM. A next-generation human genome sequence. *Science* 2022 ; 376 : 34-5
13. Robison K. The End of the Beginning of Human Genome Sequencing? *Omics! Omics!* 31 mars 2022. http://omicsomics.blogspot.com/2022/03/the-end-of-beginning-of-human-genome.html
14. Cheng H, Jarvis ED, Fedrigo O, *et al*. Haplotype-resolved assembly of diploid genomes without parental data. *Nat Biotechnol* 2022. doi: 10.1038/s41587-022-01261-x.
15. Collins FS, Mansoura MK. The Human Genome Project. Revealing the shared inheritance of all humankind. *Cancer* 2001 ; 91 : 221-5
16. Jordan B. Feu sur le quartier général : le génome en balance ? *Med Sci (Paris)* 1990 ; 6 : 906–8.

TIRÉS À PART

B. Jordan