m/s

médecine/sciences 2001 ; 17 : 309-11

Le génome humain : une séquence pour le prix de deux ?

La publication fort médiatisée de l'analyse de la séquence du génome humain par le consortium public Human Genome Project [1] et par la société *Celera* [2] est le point final d'une histoire de frères ennemis. Au cours de ces trois dernières années, les effets d'annonces de *Celera*, habilement amplifiés par son président Craig Venter, n'ont eu de cesse d'agacer les leaders du projet HGP. En l'absence de faits objectifs pouvant servir de bases aux discussions, la répartie n'était pas aisée et les dialogues ont souvent tourné en prise de becs par médias interposés. La dernière étape du processus de publication des articles résume à elle seule les enjeux et les tensions des mois passés. Il était convenu depuis l'été 2000 que le projet HGP et *Celera* publieraient conjointement deux articles séparés dans un même numéro de la revue *Science*. Au mois de novembre, on apprenait que *Celera* envisageait de revenir sur l'une des fameuses annonces de Craig Venter : l'accès à sa séquence ne serait plus libre, *via* une base de données publique, mais serait restreint à son propre site web, après authentification, formulaire à remplir et acceptation de conditions d'utilisation restrictives. La quantité maximale de séquence qu'il serait possible de télécharger serait limitée à 1 mégabase (Mb) par semaine et par personne. Cette attitude, acceptée par la revue *Science*, était particulièrement critiquable en raison des termes implacables généralement imposés aux chercheurs lors de la publication d'un article, qui incluent la soumission obligatoire des séquences à des bases de données d'accès public. Les responsables du projet HGP ont immédiatement réagi en soumettant leur article à la revue *Nature*, en lieu et place de *Science*. Aujourd'hui les articles sont publiés, les résultats sont là et peuvent être comparés.

On s'aperçoit rapidement, à la lecture de l'article de *Celera*, que réunir les chiffres nécessaires à la comparaison est une entreprise difficile. Dans un style qui ne laisse planer aucune ambiguïté sur le succès de l'entreprise (à commencer par le titre de l'article), le texte omet un certain nombre de chiffres clés qu'il est néanmoins possible de retrouver par recoupement. La première surprise provient du fait que *Celera* ne décrit pas de tentative d'analyser ses propres données seules. En effet, les deux assemblages de la séquence du génome humain décrits dans l'article incorporent la séquence du projet HGP. Pourquoi deux assemblages ? La stratégie initialement annoncée était pourtant relativement claire dans son principe et ne parlait pas de deux méthodes [3]. Il s'agissait de fragmenter le génome humain en morceaux aléatoires de petite taille, cloner et séquencer massivement ces fragments afin d'atteindre une redondance suffisante, et finalement assembler le tout en une seule étape sur la base des chevauchements de séquence. Cette méthode dite de Séquencage Global Aléatoire (SGA) était annoncée comme devant rendre obsolète les méthodes utilisées par le projet HGP, car beaucoup plus rapide et efficace. En effet, la stratégie du projet HGP comprend une étape supplémentaire qui consiste d'abord à cartographier environ 150 000 clones BAC (fragment d'ADN clonés de 100 à 200 kilobases) puis en sélectionner un sous-ensemble (environ 20 000) qui sera séquencé. Cette stratégie cible des régions individuelles qui sont traitées séparément par les différents groupes du consortium et finalement ré-assemblées grâce aux données de cartographie.

La première version de la séquence de *Celera* est effectivement une tentative d'appliquer directement la stratégie SGA, comme ceci avait été annoncé. Pour cela, deux ensembles de données ont été utilisés : environ 5,1 équivalents génome* de séquences brutes produites par la société à partir d'un séquençage aléatoire du génome, et les 7,5 équivalents génome produits par le projet HGP à partir de séquençage de clones BAC *(Tableau I)*. L'ensemble représente donc 12,5 équivalents génome. Pour utiliser les séquences du consortium HGP, *Celera* est parti des BAC assemblés dans *Genbank*, qui représentent 1,45 fois la taille totale du génome en raison d'une certaine redondance entre clones voisins. *Celera* a ensuite « décomposé » chaque clone en fragments de 550 bases (baptisés du nom évocateur de *« faux-reads »*) afin de mimer une stratégie de séquençage aléatoire de ses séquences. Cependant, pour garder une information de chevauchement, cette opération a été réalisée deux fois, en décalant les points de coupure de la deuxième copie de 225 bases. En réalité donc, les séquences consécutives obtenues par ce découpage se recouvrent parfaitement sur la moitié de leur longueur. Cet agencement idéal de séquences retient bien toute l'information de l'assemblage de la séquence HGP, tout en donnant l'impression de « repartir à zéro ». La séquence HGP ainsi décomposée représente 2,9 équivalents du génome (2 fois 1,45), un chiffre qui est repris dans l'article de *Celera*, mais en se gardant bien de mentionner les 7,5 équivalents génome initiaux qui ont servis à les constituer. Cette mystification a

** Équivalent génome : nombre moyen de fois qu'une base a été lue au cours d'une phase de séquençage aléatoire ou est présente dans une banque de clones.*

Tableau I. Comparaison entre les deux assemblages produits par *Celera* et celui du projet HGP public. La méthode SGA (Séquençage Global Aléatoire) tente d'assembler toutes les données de séquence en une seule fois, tandis que la méthode de compartimentation subdivise d'abord la séquence en sous-ensembles sur la base d'informations issues du HGP. Les pourcentages cités ici tiennent compte de la fraction euchromatique du génome humain, qui est estimée à 2,93 milliards de bases.

	Celera 1 SGA	*Celera 2* compartimentation	HGP
Équivalents génome Données brutes utilisées dans l'assemblage	5,1 équiv. *(Celera)* + 7,5 équiv. (HGP) 12,6 équiv. au total	5,1 équiv. *(Celera)* + 7,5 équiv. (HGP) 12,6 équiv. au total	7,5 équiv. (HGP)
Couverture totale de l'assemblage, trous inclus	2,85 milliards de bases	2,91 milliards de bases	3,07 milliards de bases
Génome séquencé sans compter les trous (% du total)	2,57 milliards de bases (88 %)	2,65 milliards de bases (90 %)	2,69 milliards de bases (92 %)
Fraction du génome			
– Séquencée (brut)	99,9 %	> 99 %	94 %
– Assemblée	88 %	90 %	92 %
– Non assemblée	~ 12 %	~ 10 %	~ 2 %
Nombre de contigs (et aussi de trous)	221 036	170 033	149 821
Nombre d'ossatures	118 968	53 591	87 757
Plus longue séquence	1,2 million de bases	2,0 millions de bases	23,0 millions de bases

clairement pour but de minorer l'impact de la séquence du HGP sur l'assemblage de *Celera*.

Au vu de ces chiffres, on pourrait s'attendre à ce que la séquence de *Celera* soit un net progrès par rapport à la séquence du HGP. En fait, la stratégie SGA permet un assemblage qui couvre environ 88 % du génome en 221 000 morceaux de séquences (donc également 221 000 trous), dont le plus grand ne fait que 1,2 million de bases (0,04 % du génome). Certains fragments ont pu être positionnés et orientés les uns par rapport aux autres, et produisent des édifices appelés ossatures. Il apparaît que seules 119 000 ossatures ont pu être construites à partir des 221 000 morceaux de séquences. Cette version ne constitue donc pas un assemblage réussi car il reste très fragmenté, et laisse de côté environ 12 % de lectures brutes sous forme de centaines de milliers de «bouts» de séquence qui n'ont pu être assemblés. Pourquoi la stratégie SGA ne fonctionne-t-elle pas sur le génome humain ? Probablement parce que ce dernier contient plus de 50 % de séquences répétées, un chiffre qui contraste avec les 10 % trouvés chez la drosophile. En fait, l'assemblage par *Celera* de la séquence de drosophile [4], qui était censé démontrer la puissance de l'approche SGA, ne contenait que 2 % de séquences répétées car les régions hétérochromatiques avaient été écartées dès le départ. Malgré cela, l'assemblage final avait nécessité une profondeur de 14 équivalents génome en séquences brutes. Comment croire aujourd'hui que seul 12,5 équivalents suffiraient à assembler la séquence du génome humain ?

Devant ce résultat décevant, *Celera* a donc décidé de construire une deuxième version. Cette fois, la stratégie consiste à s'aligner sur la position des clones utilisés par le projet HGP afin de compartimenter les données en régions plus petites et moins complexes. Ici aussi, la séquence assemblée du HGP est décomposée en petits fragments de 550 bases additionnées aux lectures de *Celera*. Le tout est ensuite assemblé mais, cette fois-ci, localement, en utilisant les clones BAC du HGP et donc sa cartographie préalable. La séquence globale est finalement reconstituée en assemblant les régions ainsi pré-assemblées. Le résultat de cette deuxième stratégie est légèrement meilleur : 90 % du génome y est inclus, en 170 000 fragments eux-mêmes regroupés en 53 500 ossatures, et la plus longue séquence continue mesure 2 millions de bases. Malgré tout, environ 10 % des séquences brutes n'en font toujours pas partie. Par comparaison, la séquence du consortium HGP couvre 92 % des bases du génome en 150 000 fragments de séquence qui permettent d'assembler 88 000 ossatures. La plus longue séquence continue mesure ici 23 millions de bases.

La séquence assemblée par *Celera* selon la stratégie de «compartimentation» est donc très semblable à la séquence du HGP du point de vue de la qualité de l'assemblage. Cette

observation est confirmée par une étude informatique fondée sur la comparaison des deux séquences [5]. Ce résultat est une surprise, car on aurait pu croire que le supplément de données brutes, ajouté aux séquences du consortium public comblerait un grand nombre de trous présents dans la séquence de ce dernier. Or la différence reste très modeste. Ceci est probablement dû aux difficultés rencontrées par *Celera* pour assembler ses propres séquences dans les régions qui n'étaient pas déjà couvertes par les séquences HGP. Il apparaît ainsi clairement que la séquence produite par *Celera* repose entièrement sur l'assemblage du projet public. Cette conclusion n'est certainement pas celle que l'on tire à la lecture de l'article de *Science* qui brosse un tableau idéalisé des résultats, et encore moins à l'écoute des divers communiqués de presse distillés par la société. Cela eût été de bonne guerre s'il ne s'était pas agi d'une imposture qui vise à diminuer l'impact du travail de centaines de chercheurs et techniciens sur plus de 10 ans.

Même si une partie n'a pu être ni assemblée ni positionnée sur les chromosomes, les données de séquence brute de *Celera* couvrent plus de 99% du génome car elles sont issues d'un clonage direct de tout l'ADN nucléaire. En raison de la similarité entre l'assemblage du HGP et celui de *Celera*, quel avantage y aurait-il à se soumettre aux conditions d'accès imposées par la société ? En fait la séquence de *Celera* représente une version « affinée » de la séquence HGP en plusieurs régions où le supplément de données brutes permet de réorienter certains fragments, de corriger certaines erreurs d'assemblage, voire de colmater certains trous. Il est difficile d'estimer la qualité réelle de la séquence de *Celera*, mais environ 35 % de la séquence HGP est finie tandis que 65 % est encore sous forme d'ébauche. Il reste donc un minutieux travail de finition à faire, qui consiste à boucher les trous et à réduire le taux d'erreur. Seul le HGP envisage de terminer sa séquence (d'ici 2003) mais il ne fait nul doute que *Celera*, là aussi, saura profiter de la situation.

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* 2001 ; 409 : 860-921.
2. Venter JC, Adams MD, Myers EW, *et al.* The Sequence of the Human Genome. *Science* 2001 ; 291 : 1304-51.
3. Weissenbach J, Salanoubat M. Séquences des génomes : le feu d'artifice. *Médecine/Science* 2001 ; 16 : 10-6.
4. Myers EW, Sutton GG, Delcher AL, *et al.* A whole-genome assembly of Drosophila. *Science* 2000 ; 287 : 2196-204.
5. Aach J, Bulyk ML, Church GM, *et al.* Computational comparison of two draft sequences of the human genome. *Nature* 2001 ; 409 : 856-9.

Hugues Roest Crollius

Genoscope et Cnrs FRE2231, 2, rue Gaston-Crémieux, 91057 Évry Cedex, France.

TYPON