m/s

Chroniques génomiques

par Bertrand JORDAN

médecine/sciences 1995 ; 11 : 1717-9

Génome humain : l'annuaire nouveau est arrivé

Événement scientifique, ou coup médiatique ?

La publication par la revue *Nature* d'un numéro spécial de près de quatre cents pages, le 28 septembre dernier [1], a fait l'objet d'une préparation médiatique approfondie : dossier à l'usage des journalistes développant le contenu des articles et rappelant les biographies des auteurs, conférence de presse sur invitation à Washington, et même distribution sur le réseau *Internet* par « ftp anonyme » des photographies de Craig Venter et de Daniel Cohen... Simple « coup de pub » visant à marquer un point dans l'intense compétition qui oppose *Nature* et *Science*, ou résultats déterminants pour l'avancée des programmes Génome, c'est ce que nous allons tenter de démêler.

On trouve, dans ce *Genome Directory*, une nouvelle version de la carte physique générale, accompagnée de cartes plus détaillées portant respectivement sur les chromosomes 3, 12, 16 et 22 et, enfin, un article sur les ADNc. Dû au groupe de Craig Venter, ce dernier [2] occupe à lui seul 172 pages, décrit l'analyse de « 84 millions de nucléotides de séquence d'ADNc », et constitue, à mon sens, la nouveauté principale de ce supplément. Événement politique autant que scientifique : après des controverses acerbes qui ont défrayé la chronique depuis deux années, voici que Venter publie enfin une bonne partie de ses résultats et révèle 174 472 EST *(expressed sequence tags*, séquences partielles d'ADNc) jusque-là conservés à l'abri des regards indiscrets dans une base de données « privée »...

Les EST enfin révélés

Le dossier de presse indique que près de 90 % des données de séquence de TIGR* sont maintenant disponibles – il est en fait difficile d'évaluer exactement combien d'EST restent confidentiels. Il s'agit naturellement des plus « juteux », ceux dont la séquence risque de mener à de nouveaux inhibiteurs de coagulation ou à des cytokines inédites, donc à des produits commerciaux brevetables. Par ailleurs, la contribution du projet français Genexpress (qui jusqu'à fin 1994 a été le plus important producteur d'EST « publics ») n'est nulle part mentionnée dans ces commentaires... Pourtant, ses séquences font partie de l'analyse présentée par Venter, de même que celles de « l'initiative Merck », vaste programme de séquençage d'ADNc conduit par Robert Waterston (Saint-Louis, États-Unis) avec le soutien de cette firme. Avec, fin août, 172 388 séquences disponibles, cette entreprise ôtait beaucoup de sa valeur au « trésor de guerre » de TIGR/HGS. Son succès a sans doute joué un grand rôle dans la décision de Venter : il était temps de publier ces résultats avant qu'ils ne soient par trop dévalorisés...

Reste que l'étape est décisive : le groupe de Venter a effectué son analyse sur 174 472 EST « maison », auxquels il a ajouté 118 406 séquences supplémentaires tirées de la base spécialisée dbEST, librement accessible sur le réseau. C'est la première fois qu'une étude porte sur un ensemble aussi vaste : elle devrait permettre d'affiner l'estimation du nombre total de gènes contenus dans notre génome, tout en indiquant combien d'entre eux sont déjà étiquetés. Le calcul n'est pas simple, puisque les clones à séquencer ont été pris au hasard et que les gènes fortement exprimés sont sur-représentés dans les banques d'ADNc. Et, malheureusement, il ne suffit pas de comparer les séquences pour repérer celles qui proviennent du même gène. Le déchiffrage effectué par les différentes équipes porte tantôt sur l'extrémité 3', tantôt sur l'extrémité 5' de l'ADNc ; les séquences 5' peuvent elles-mêmes se situer en différents points selon la banque et le clone étudiés. Les erreurs de lecture compliquent encore l'analyse.

En comparant deux à deux ces presque trois cent mille petites séquences qui représentent au total 83 millions de nucléotides, Venter obtient 29 599 groupes de deux ou plusieurs EST se recouvrant totalement ou partiellement. Baptisées THC, pour *tentative human consensus sequences*, ces entités définissent en principe autant de gènes. En raison des recouvrements non détectés, le nombre réel est très certainement plus faible. Les séquences « solitaires », quant à elles, sont au

* *TIGR :* The Institute for Genome Research, *le laboratoire de Craig Venter, est très lié à* Human Genome Science, *lui-même sous contrat avec la firme Smith Kline Beecham. Voir Jordan B. La valse des étiquettes.* médecine/sciences *1995 ; 11 : 273-6.*

nombre de 58 384, et représentent un effectif de gènes difficile à estimer pour les raisons évoquées ci-dessus: 10000?, 30 000 ? La fourchette est large. Enfin, certains gènes très peu exprimés, ou mis en œuvre seulement dans un organe spécifique à un moment précis du développement, ont pu échapper à l'étiquetage (car absents des banques utilisées) et ne sont donc pas comptabilisés. Bref, malgré leur ampleur, ces travaux ne permettent pas encore de fixer le contenu de notre patrimoine génétique. Le précédent de la levure ou du nématode, où le séquençage intégral a révélé trois ou quatre fois plus de gènes qu'attendu, incite d'ailleurs à la prudence. Quoi qu'il en soit, 10 214 seulement des séquences étudiées correspondent à des gènes connus, ce qui montre une nouvelle fois l'étendue de notre ignorance... et l'impact des informations apportées par les EST.

Venter tente aussi d'interpréter les données en termes d'expression : les EST ayant été obtenus à partir d'un grand nombre de banques d'ADNc provenant de divers tissus, leur fréquence dans chacune d'elles donne une idée du niveau d'expression du gène correspondant. C'est une démarche similaire à celle suivie au Japon par Kosaku Okubo et Kenichi Matsubara, quoique moins rigoureuse dans la mesure où l'approche expérimentale (et les banques d'ADNc) n'a pas été conçue dans cette optique ; les résultats ne peuvent donc être qu'indicatifs.

L'accès à ces résultats suppose la consultation de la base de données de TIGR, la *Human cDNA Database* (HCD). L'accès au niveau 1, qui concerne d'après Venter 85 % des données, est libre pour les chercheurs du secteur public sous réserve de la signature d'une décharge de responsabilité. L'accès au niveau 2, contenant les données confidentielles, quant à lui, reste soumis à la signature d'un accord donnant à *Human Genome Sciences* un droit de premier regard sur la valorisation des résultats obtenus. Moins satisfaisante que la disponibilité complète avec archivage dans les bases de données publiques (EMBL, GenBank), cette formule permet aux mécènes privés du projet de garder un certain contrôle sur la diffusion des résultats (vis-à-vis de firmes concurrentes, par exemple) tout en offrant aux chercheurs des possibilités d'analyse intéressantes grâce aux logiciels développés par les informaticiens de TIGR et incorporés dans HCD.

L'article de Venter, et la mise à disposition de plus de cent mille EST nouveaux, représentent un acquis important. La cartographie de ces séquences fait l'objet d'un programme international qui devrait placer 20 000 EST sur le génome dès la mi-1996 : le clonage de gènes de maladies par l'approche des « candidats positionnels » rendra alors quasiment caduc le clonage positionnel classique. C'est ce qu'annonçait d'ailleurs, il y a près d'un an, un connaisseur en la matière nommé Francis Collins... [3].

Carte générale : affinement et consolidation

Passons maintenant aux cartes, et pour commencer à la plus globale (et la plus médiatisée), celle qui émane du CEPH (Daniel Cohen). Elle fait suite à la publication, en 1992, d'une méthode d'alignement de YAC [4] qui était censée donner très rapidement une carte couvrant 90 % du génome. C'est en réalité fin 1993 [5] qu'était publiée une « carte physique de première génération » s'appuyant largement sur les marqueurs de la carte génétique réalisée au Généthon par le groupe de Jean Weissenbach. Les « niveaux » successifs (nombre de YAC connectant deux points « sûrs ») de cette carte présentaient un taux de couverture croissant... et une fiabilité décroissante. Le présent article intitulé quant à lui *« A YAC contig map of the human genome »* [6], indique que les « niveaux » élevés ont été abandonnés, et qu'une combinaison de techniques a été mise en œuvre pour obtenir les données les plus solides possible.

Le résultat, selon les auteurs, est un ensemble de 225 *contigs* de YAC, d'une taille moyenne de dix mégabases, couvrant 75 % du génome. Tout porte à penser que la fiabilité est effectivement au rendez-vous, que les incertitudes ont été correctement évaluées et que ces *contigs* seront une base sérieuse pour les travaux ultérieurs. Le retard enregistré par rapport à des prévisions initiales très optimistes correspond aux difficultés bien réelles rencontrées par tous les laboratoires dans la construction de cartes de grande étendue avec ces réactifs indispensables mais très imparfaits que sont les YAC. La publication de cette carte est une étape majeure dans le balisage physique de notre génome, et va encore accélérer la découverte de gènes impliqués dans des maladies en rendant immédiatement accessibles les YAC « couvrant » toute région désignée par l'analyse génétique. Elle ne constitue cependant pas un saut qualitatif comparable à la première carte génétique parue en 1987 [7] ou à la première carte physique générale publiée en 1993 [5]. De plus, une nouvelle carte physique du génome humain a été établie au *Whitehead Institute* (le *Genome Center* américain dirigé par Eric Lander) et présentée pour la première fois lors du récent « Colloque sur les séquences transcrites » tenu à l'île des Embiez début novembre. Construite selon une technologie analogue, utilisant largement la carte génétique de Généthon et les YAC du CEPH, elle apparaît plus fine (incorporant plus de 11000 STS au lieu de 3000), plus complète (couvrant environ 90 % du génome) et très fiable puisque chaque YAC est relié à son voisin par au moins deux STS. Déjà partiellement disponible sur *Internet*, cette nouvelle avancée relativise donc l'importance du résultat rapporté ici.

Les cartes par chromosome restent indispensables

Les quatre autres articles décrivent des cartes physiques portant chacune sur un chromosome. Ils s'appuient sur la carte générale et sur les YAC du CEPH, mais y ajoutent des marqueurs et des données provenant de multiples équipes souvent regroupées en consortiums. Ces groupes travaillent parfois depuis des années sur différentes régions d'un chromosome, y ont cherché et souvent trouvé

les gènes impliqués dans les maladies qui les intéressent, et ont rassemblé des collections d'hybrides, YAC et autres réactifs. Ils savent quelles sont les régions problématiques, celles dont les séquences répétées brouillent l'analyse, celles où de fréquentes délétions compliquent l'analyse génétique... Les débats contradictoires auxquels se livrent ces experts garantissent une qualité de résultat nettement supérieure à celle atteinte par les grands centres qui, attaquant l'ensemble du génome, ne peuvent pas trop s'attarder sur les zones litigieuses.

Deux des cartes publiées se disent « de deuxième génération », mais cette appellation, comme le numéro de version des logiciels ou la notion de nouveau modèle automobile, repose sur des bases hautement subjectives. A mon avis, seule mérite vraiment ce nom la carte du chromosome 16 [8], coordonnée par le laboratoire de Robert Moysis (Los Alamos, États-Unis) et bénéficiant de la collaboration active du groupe de Grant Sutherland (Adelaide, Australie). Elle s'appuie, comme les trois autres, sur la construction d'un *contig* de YAC couvrant sans trop de lacunes l'ensemble du chromosome ; mais elle offre une résolution beaucoup plus fine grâce à un deuxième niveau reposant sur l'alignement de deux mille cosmides. Rien de très étonnant puisque le groupe de Los Alamos projetait à l'origine (avant la découverte des YAC) d'assembler un *contig* de cosmides sur ce chromosome... L'entreprise s'est avérée impossible avec ces seuls clones, trop petits (30 à 40 kilobases) par rapport aux dizaines de mégabases à baliser ; les YAC, arrivés à la rescousse, ont comblé les trous et affermi une carte qui serait, sinon, restée incomplète. Mais le travail préalable effectué permet, une fois la carte de YAC établie, de passer rapidement à ces réactifs (actuellement) irremplaçables que sont les cosmides, seuls clones de taille raisonnable que l'on puisse facilement préparer, analyser en détail ou séquencer.

Les trois autres articles [9-11] correspondent, quant à eux, à l'affinement de cartes physiques reposant sur des YAC et à l'obtention de *contigs* fiables couvrant la quasi-totalité des chromosomes étudiés (le 3, le 2 et le 22). Cartes fiables et détaillées, directement utilisables pour la recherche de séquences transcrites par *exon-trapping* ou *cDNA capture*; mais pour ces trois chromosomes, on est encore loin du « prêt à séquencer » *(sequence-ready map)* dont se rapproche beaucoup la carte actuelle du 16.

Des stratégies concurrentes mais interdépendantes

L'avancée des travaux sur les EST ou sur les cartes physiques présente une grande continuité, et le choix de la date de publication est assez arbitraire. La sortie, fin septembre 1995, de ce *Genome Directory* doit donc beaucoup à des impératifs politiques et médiatiques ; elle autorise, à tout le moins, un tour d'horizon fort opportun. Bien qu'absente de cet annuaire, la carte génétique ne peut être oubliée : sa version à 5 000 marqueurs, due au projet mené par Jean Weissenbach, paraîtra bientôt. Elle a sans doute atteint le niveau de finesse nécessaire et raisonnable : ce volet est en quelque sorte clos. Le projet Génome se recentre ainsi autour de l'amélioration des cartes physiques, du séquençage et du positionnement des EST, et de l'amorce d'un effort sérieux pour déchiffrer l'ensemble de notre ADN. Ces différents aspects ne sont pas indépendants, et leur poids respectif reste incertain. Comment équilibrer les travaux entre cartes par chromosome, qui font intervenir de nombreux groupes experts en des domaines très divers, et carte générale, qui est plutôt l'affaire de structures importantes comme le CEPH ? La richesse du catalogue des EST, et l'obtention sans doute prochaine d'informations sur leur position dans le génome rendent-ils moins urgent le séquençage exhaustif ? Autant d'interrogations auxquelles les décideurs qui financent ces recherches devront apporter une première réponse dans les mois qui viennent – en attendant que le succès ou l'échec des travaux en cours ne change, peut-être, les données de la question ■

RÉFÉRENCES

1. The Genome Directory. *Nature* 1995 ; 377 (suppl).

2. Adams MD, *et al.* Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* 1995 ; 377 (suppl) : 3-174.

3. Collins F. Positional cloning moves from traditional to perditional... *Nature Genet* 1995 ; 9 : 347-50.

4. Bellanne-Chantelot C, Lacroix B, Ougen P, Billault A, Beaufils S, *et al.* Mapping the whole human genome by fingerprinting yeast artificial chromosomes. *Cell* 1992 ; 70 : 1059-68.

5. Cohen D, Chumakov I, Weissenbach J. A first-generation physical map of the human genome. *Nature* 1993 ; 366 : 698-701.

6. Chumakov IM, *et al.* A YAC contig map of the human genome. *Nature* 1995 ; 377 (suppl) : 175-298.

7. Donis-Keller H, *et al.* A genetic linkage map of the Human Genome. *Cell* 1987 ; 51 : 319-37.

8. Doggett NA, *et al.* An integrated physical map of human chromosome 16. *Nature* 1995 ; 377 (suppl) : 335-66.

9. Gemmill RM, *et al.* A second-generation YAC contig map of human chromosome 3. *Nature* 1995 ; 377 (suppl) : 299-320.

10. Krauter K, *et al.* A second-generation YAC contig map of human chromosome 12. *Nature* 1995 ; 377 (suppl) : 321-34.

11. Collins JE, *et al.* A high-density YAC contig map of human chromosome 22. *Nature* 1995 ; 377 (suppl) : 367-96.