m/s SYNTHÈSE
médecine/sciences 2001 ; 17 : 294-8

Le Projet Génome Humain : quinze ans d'efforts

Denis Le Paslier
Alain Bernot

Le Projet Génome Humain a été discuté dès 1985, et le premier programme conçu en 1990. Les programmes suivants ont tenu compte des progressions rapides réalisées dans ce domaine, des évolutions technologiques, ainsi que l'entrée en course du secteur privé. La cartographie génétique et physique, le séquençage de messagers, et de la totalité du génome de l'homme ont été successivement entrepris. Par ailleurs, plusieurs génomes d'espèces modèles ont été entièrement séquencés. Juin 2000 a représenté la date à laquelle 90 % du génome humain étaient obtenus, et la séquence complète est espérée d'ici 2003. Cela devrait conduire à l'identification de tous les gènes humains, et des variations génétiques d'intérêt.

Le séquençage du génome humain constitue une tâche gigantesque, qui a mobilisé – et mobilise toujours – de nombreuses équipes. Il permettra d'identifier tous les gènes présents dans le génome humain. Les chercheurs y voient l'espoir de déceler et de soigner les 6000 maladies génétiques, de comprendre les prédispositions génétiques aux maladies communes, ainsi qu'une multitude de phénomènes biologiques. Les investisseurs pensent à une mine d'or avec la mise au point de tests diagnostiques et de médicaments nouveaux.

La genèse

Cette histoire a débuté il y a un peu plus de 15 ans aux États-Unis. Le projet d'établir la séquence complète du génome humain a été discuté la première fois à Alta (Utah) en 1984, lors d'une réunion organisée par le *Department of Energy* (DOE) [1]. Le but de cette réunion était d'évaluer l'utilité des méthodes d'analyse de l'ADN pour la détection des mutations, et l'éventuelle augmentation de leur fréquence parmi les survivants d'Hiroshima et de Nagasaki. Mais la principale conclusion fut qu'aucune méthode ne permettait l'identification des mutations, à moins qu'un projet de séquençage énorme, complexe et très coûteux, ne soit entrepris. Et la meilleure sensibilité ne serait obtenue que s'il était possible de comparer la séquence complète de parents avec celles de leurs enfants !

Les premières discussions sérieuses ont lieu lors de *Workshops* à Santa Cruz (Californie) en 1985, puis à Santa Fe (Nouveau-Mexique) en 1986 *(figure 1)*. L'établissement de la séquence complète du génome humain, et donc de ses gènes (dont le nombre était alors estimé à 50 000-100 000) serait un atout majeur pour la localisation puis la caractérisation de gènes impliqués dans de nom-

ADRESSE

D. Le Paslier, A. Bernot : Genoscope et Cnrs UMR-8030, 2, rue Gaston-Crémieux, 91057 Évry Cedex, France.

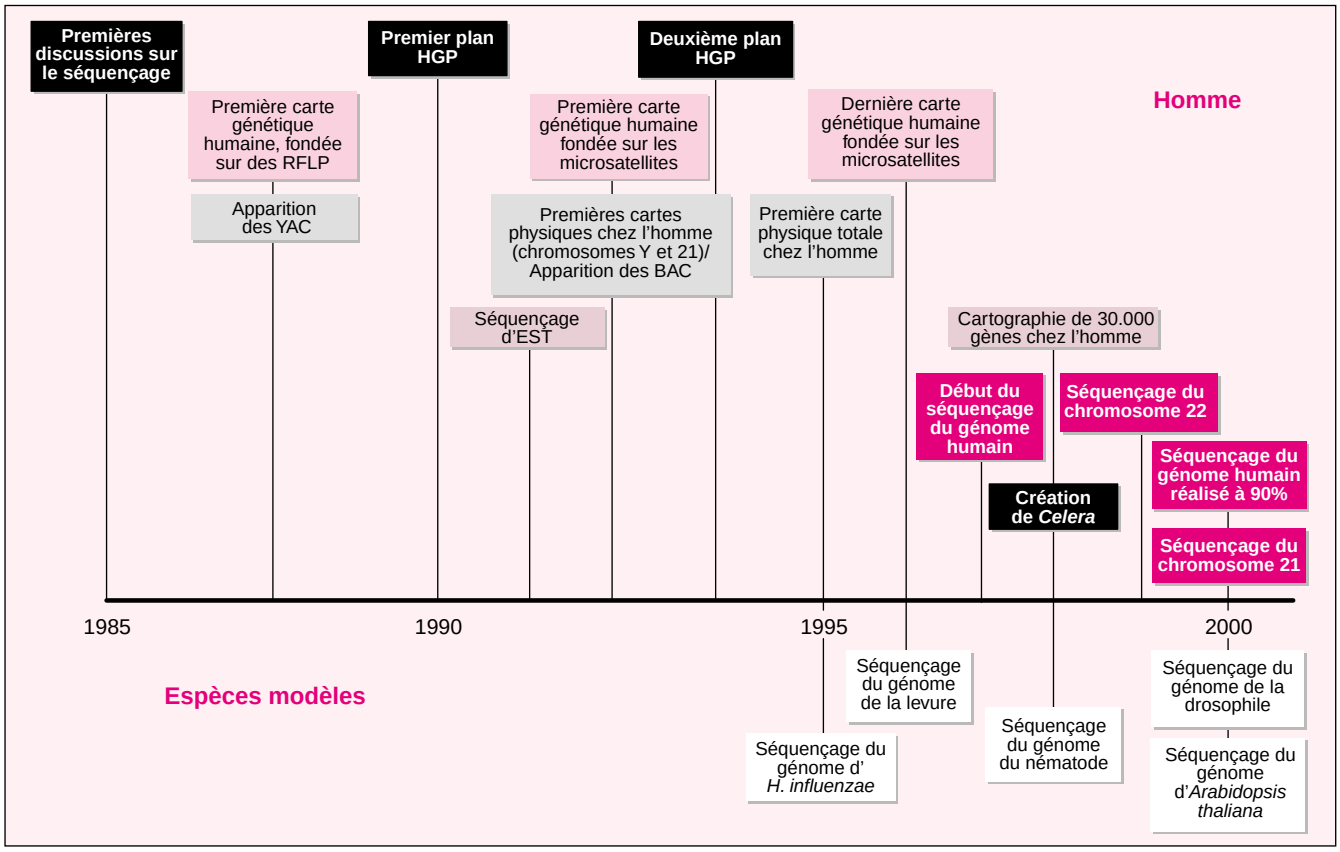

Figure 1. ***Principales étapes du projet génome humain, et des projets des espèces modèles.*** *HGP:* human genome project; *YAC:* yeast artificial chromosome; *BAC:* bacterial artificial chromosome; *EST:* expressed sequence tags.

breuses maladies, communes ou complexes, telles que les cancers, les maladies cardio-vasculaires ou la maladie d'Alzheimer [2, 3]. A cette époque, toutefois, l'établissement d'une séquence nucléique de plusieurs centaines de bases prenait une journée et coûtait 10 dollars la base. Aussi, la suggestion de séquencer le génome humain restait-elle imaginative et ambitieuse, mais d'un coût encore prohibitif.

Après de nombreuses réunions organisées par le DOE, le *National Institutes of Health* (NIH), le *Congressional Office of Technology Assessment* et la *National Academy of Science*, un *Genome Office* a été créé en 1988 par le NIH (qui deviendra en 1989 le *National Center for Human Genome Research* dirigé par J. Watson, puis le *National Human Genome Research Institute* dirigé par F. Collins). Cette même année eut lieu à Cold Spring Harbor le premier congrès annuel consacré à la cartographie et au séquençage du génome. Afin de coordonner les efforts de cartographie et de séquençage, et d'éviter la redondance, la *Human Genome Organisation* (HUGO) fut créée en 1988. HUGO n'a toutefois pas obtenu les moyens à la hauteur de ses ambitions: elle ne peut plus être perçue que comme une société savante qui situe aujourd'hui sa réflexion, ses propositions et ses actions en aval du projet proprement dit.

En 1990, un premier programme de 5 ans a été développé. Le *Human Genome Project* (HGP) représentait la première entreprise d'envergure internationale dans le domaine biomédical, ayant comme but ultime la détermination de la séquence complète de notre génome. Cette formidable entreprise a été alternativement présentée comme l'équivalent du projet Apollo, de la quête du Saint Graal, ou de l'établissement du tableau périodique de la biologie. Ce projet a eu aussi de nombreux détracteurs et engendré de nombreuses luttes d'influence peu favorables à la mise en place de programmes nationaux et internationaux.

Le programme

Ce programme comprenait les points suivants: établir une carte génétique, dont les marqueurs soient espacés de 2 à 5 centimorgans; établir une carte physique, dont les marqueurs soient espacés d'environ 100 kb; améliorer les performances du séquençage automatique (le séquençage complet du génome humain était considéré comme un objectif réservé à une date plus lointaine); séquencer des génomes d'organismes modèles, et élaborer les outils informatiques permettant de traiter, d'archiver, et de communiquer l'ensemble des données ainsi produites. En 1993, un second plan de cinq ans fut défini, comprenant un nouvel objectif: identifier le plus grand nombre possible de gènes par séquençage intensif d'ADNc. En 1997, enfin, fut décidé le séquençage complet du génome humain. Le dernier plan date de 1999: il proposait une première étape de séquençage du génome humain, sous la forme d'une

version de travail *(working draft)*, séquence incomplète, mais recouvrant 90 % du génome, avec une profondeur supérieure à 5 fois (nombre moyen de lectures par région séquencée).

Les premières cartes

Une carte génétique utilisant les RFLP *(restriction fragment length polymorphism)* comme marqueurs a été proposée par D. Botstein et ses collègues en 1980 [4]. L'élaboration d'une telle carte a été rendue possible par la mise à disposition de l'ensemble de la communauté scientifique d'une ressource commune, l'ADN de familles nombreuses, par le Centre d'étude du polymorphisme humain (CEPH) [5, 6]. Ainsi, toutes les données produites à partir de l'ADN de ces familles sont cumulables. La première carte a été publiée en 1987 par un groupe privé américain, *Collaborative Research Inc.* [7]. Ces cartes, utilisant comme marqueurs génétiques des RFLP, ont été supplantées par des cartes utilisant les microsatellites, développées au Généthon [8, 9]. Ces marqueurs, très polymorphes et d'utilisation aisée, ont permis de localiser plusieurs centaines de gènes responsables de maladies.

La façon de construire la carte physique du génome humain a longtemps été débattue. Une approche spécifique de chromosome avait été entreprise par deux laboratoires du DOE : le *Los Alamos National Laboratory* et le *Lawrence Livermore National Laboratory*, pour les chromosomes 16 et 19. Pour cela étaient utilisées des banques de cosmides, dont la capacité de clonage est faible (de l'ordre de 40 kb). L'apport des chromosomes artificiels de levures (YAC : *yeast artificial chromosome*), avec la possibilité de cloner de très grands fragments d'ADN (> 1 000 kb) a été déterminant [10], et les YAC ont permis très rapidement d'établir des cartes pour les chromosomes 21 et Y [11, 12]. Les marqueurs génétiques avaient été indispensables à cette réalisation. Par la suite, l'utilisation des YAC a permis aux équipes du CEPH / Généthon [13] puis du *Whitehead Institute* [14] d'obtenir une carte physique couvrant 95 % du génome humain.

Cependant, les défauts inhérents aux YAC (trop souvent chimériques, réarrangés ou instables) ont finalement rendu leur utilisation impossible pour le séquençage. De nouveaux vecteurs de clonage bactériens (BAC, *bacterial artificial chromosome* ou PAC [15, 16]) ont été développés, permettant de cloner de façon stable des fragments de l'ordre de 200 kb : ils sont devenus le matériel de choix pour le séquençage d'ADNc. Ces nouvelles banques d'ADN génomique ont été construites à partir de donneurs dont l'anonymat est certifié, et qui ont donné leur accord pour une utilisation dans le cadre du séquençage. Ces clones ont été physiquement cartographiés, par criblage avec des marqueurs préalablement localisés, par empreintes de restriction, par séquençage systématique de leurs extrémités…

Le séquençage d'ADNc

A contre-courant de ces projets génomiques, d'importants programmes de séquençage d'ADNc ont été lancés, dans le but de décrypter la partie codante des gènes [17]. L'accumulation de ces séquences partielles (et parfois très redondantes) a été extraordinaire : 14 500 séquences d'EST *(expressed sequence tags)* humains étaient disponibles en 1993, plus de 3 millions en février 2001 ! Cette approche a été menée à grande échelle à la fois dans le secteur public et privé. L'Université de Washington, Merck, TIGR (*The Institute of Genomic Research*, fondé et dirigé par C. Venter), *Incyte*, *Human Genome Science*, *Millennium*... ont constitué des bases de données de plusieurs millions de séquences. Cette approche a fait pour la première fois surgir le problème de la propriété industrielle et intellectuelle des données biologiques produites par les recherches sur le génome humain. Combien de gènes différents ces millions d'EST représentent-ils demeure une question non résolue : les estimations varient de 35 000 [18] à 120 000 [19]. Ces EST ont aussi représenté une source importante pour la cartographie et la localisation de gènes sur les chromosomes.

La cartographie devait connaître un nouveau développement, avec l'apparition des hybrides d'irradiation : des cellules humaines irradiées sont fusionnées avec des cellules de hamster, qui intègrent de façon aléatoire des fragments d'ADN humain [20, 21]. L'ossature de telles cartes est constituée des marqueurs provenant de la carte génétique. Plusieurs cartes d'hybrides d'irradiation ont été successivement construites, ce qui a en particulier permis de localiser 30 000 fragments de gènes [22-24].

Le grand projet

Parallèlement, et après de nombreuses discussions, la stratégie retenue pour le séquençage a été de partager les différents chromosomes entre les divers groupes du consortium international, et d'utiliser les données des cartes obtenues antérieurement. Un nombre minimal de clones recouvrant le génome a été déterminé et partagé entre les membres du consortium. La France, représentée par le Genoscope, a rejoint le HGP en 1998 et séquence le chromosome 14, selon l'approche du séquençage dirigé (*voir* [25] pour le détail de cette stratégie). Cette approche avait déjà été utilisée pour le séquençage d'un certain nombre de génomes de petite taille, mais aussi pour ceux du nématode (100 Mb) [26] et d'*Arabidopsis thaliana* (125 Mb) [27].

La naissance de Celera : ou la menace d'un monopole

La réalisation du HGP a été stimulée par l'entrée dans la course de la société *Celera Genomics Inc.*, créée en 1998 par C. Venter, et qui avait annoncé qu'elle allait séquencer le génome humain en 3 ans, c'est-à-dire bien avant que ne soit réalisé le projet du consortium public. La stratégie mise en œuvre est celle du séquençage aléatoire global, similaire à celle utilisée par le TIGR pour le séquençage de génomes bactériens. La première séquence complète avait été obtenue de cette façon en 1995 (*H. influenzae* : 1 830 137 pb et 1 743 gènes) [28]. Cette stratégie avait été testée pour la première fois avec succès sur un génome complexe, celui de la drosophile (120 Mb d'euchromatine et 13 600 gènes), en collaboration avec des scientifiques du domaine

public. Elle a cependant nécessité un recouvrement de 14 fois le génome [29] ! Pour l'assemblage de ses séquences humaines, *Celera* a pu bénéficier des données du consortium international, déposées dans les 48 heures dans les bases de données publiques.

Suite à l'annonce de *Celera*, le *Wellcome Trust* britannique (fondation finançant le *Sanger Centre*) a annoncé qu'elle doublait sa contribution au projet. La guerre des communiqués entre le HGP et *Celera* a fait rage et le projet s'est accéléré. Chez l'homme, une séquence quasi finie de deux chromosomes a été obtenue dès 1999 (chromosome 22) [30] et 2000 (chromosome 21) [31]. Les séquences ne recouvrent pas les régions centromériques, ni les bras courts de ces chromosomes acrocentriques, et quelques lacunes subsistent. Le contenu en gènes détectés ou prédits a été étonnamment faible : seulement 545 pour le chromosome 22, et 225 gènes pour le chromosome 21 ! Sur la totalité du génome humain, la version de travail – recouvrant 90 % de la partie euchromatique – a été officiellement annoncée le 26 juin 2000, mais le nombre exact de gènes reste à préciser. Il est maintenant estimé à 30 000-40 000. Cette estimation est à rapprocher de celle qui avait été obtenue par l'analyse du génome compact du poisson *Tetraodon nigroviridis* [32], ou de celle obtenue par P. Green [18], qui avaient prédit l'existence de 28 000 à 35 000 gènes. Même avec la séquence finie du génome humain, la caractérisation de tous les gènes ne sera pas chose aisée.

L'ensemble du séquençage a déjà permis d'identifier de nombreux gènes, de caractériser des gènes impliqués dans des maladies et de stimuler de nombreuses recherches fondamentales ou appliquées. Vingt-deux gènes responsables de maladies ont été identifiés à partir de la version de travail, tels que les gènes responsables de la maladie de Usher de type 1C, de la LGMD 2G *(limb-girdle muscular dystrophy)*, de l'ataxie spino-cérébelleuse de type 10, du cancer du sein *(BRCA2)*...

Par ailleurs, l'établissement de la séquence finie a déjà permis d'identifier des mutations ponctuelles de type SNP (*single nucleotide polymorphism* : polymorphisme de séquence biallélique), et plus de 1,5 million de ces marqueurs ont été déposés dans les bases de données. Plus de 12 000 SNP ont par exemple été identifiés et caractérisés à partir des séquences du chromosome 22 [33]. Ce type de marqueur génétique est très utile pour l'identification de gènes responsables de maladies héréditaires multifactorielles. Les entreprises de génomique investissent également dans ce domaine, car le polymorphisme qu'ils peuvent détecter pourrait expliquer pourquoi un individu répond plus ou moins bien à un traitement et adapter ce dernier en conséquence pour faire de la pharmacogénétique.

La fin de l'histoire

La prochaine étape sera de produire une séquence « finie » et exacte à 99,99 %. Il s'agira de sélectionner et de séquencer de nouveaux clones pour les régions non encore couvertes et d'augmenter la couverture de séquençage jusqu'à 10 fois. Cette étape devrait progresser rapidement et s'achever en 2003. La réalité a toutefois toujours été en avance sur les prévisions... ■

RÉFÉRENCES

1. Cook-Deegan R. The Alta summit, December 1984. *Genomics* 1989 ; 5 : 661-3.

2. Dulbecco R. A turning point in cancer research : sequencing the human genome. *Science* 1986 ; 231 : 1055-6.

3. Koshland, D. Sequences and consequences of the human genome. *Science* 1989 ; 246 : 189.

4. Botstein D, White W, Skolnick M, Davis R. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* 1980 ; 32 : 314-31.

5. Dausset J, Cann H, Cohen D, Lathrop M, Lalouel JM, White R. Centre d'étude du polymorphisme humain (CEPH) : collaborative genetic mapping of the human genome. *Genomics* 1990 ; 6 : 575-7.

6. NIH / CEPH Collaborative Mapping Group. A comprehensive genetic linkage map of the human genome. *Science* 1992 ; 258 : 67-86.

7. Donis-Keller H, Green P, Helms C, *et al.* A genetic linkage map of the human genome. *Cell* 1987 ; 51 : 319-37.

8. Weissenbach J, Gyapay G, Dib C, *et al.* A second-generation linkage map of the human genome. *Nature* 1992 ; 359 : 794-801.

9. Dib C, Faure S, Fizames C, *et al.* A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* 1996 ; 380 : 152-4.

10. Burke D, Carle G, Olson M. Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science* 1987 ; 236 : 806-12.

11. Chumakov I, Rigault P, Guillou S, *et al.* Continuum of overlapping clones spanning the entire human chromosome 21q. *Nature* 1992 ; 359 : 380-7.

12. Foote S, Vollrath D, Hilton A, Page D. The human Y chromosome : overlapping DNA clones spanning the euchromatic region. *Science* 1992 ; 258 : 60-6.

13. Chumakov I, Rigault P, Le Gall IS, *et al.* A YAC contig map of the human genome. *Nature* 1995 ; 377 (suppl) : 175-297.

14. Hudson T, Stein D, Gerety S, *et al.* An STS-based map of the human genome. *Science* 1995 ; 270 : 1945-54.

15. Shizuya H, Birren B, Kim U, *et al.* Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc Natl Acad Sci USA* 1992 ; 89 : 8794-7.

16. Ioannou P, Amemiya C, Garnes J, *et al.* A new bacteriophage P1-derived vector for the propagation of large human DNA fragments. *Nat Genet* 1994 ; 6 : 84-9.

17. Adams M, Kelley J, Gocayne J, *et al.* Complementary DNA sequencing : expressed sequence tags and human genome project. *Science* 1991 ; 252 : 1651-6.

18. Ewing B, Green P. Analysis of expressed sequence tags indicates 35,000 human genes. *Nat Genet* 2000 ; 25 : 232-4.

19. Liang F, Holt I, Pertea G, Karamycheva S, Salzberg S, Quackenbush J. Gene index analysis of the human genome estimates approximately 120,000 genes. *Nat Genet 2000* ; 25 : 239-40.

20. Benham F, Hart K, Crolla J, Bobrow M, Francavilla M, Goodfellow P. A method for generating hybrids containing nonselected fragments of human chromosomes. *Genomics* 1989 ; 4 : 509-17.

21. Cox D, Burmeister M, Price E, Kim S, Myers R. Radiation hybrid mapping : a somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. *Science* 1990 ; 250 : 245-50.

22. Walter M, Spillett D, Thomas P, Weissenbach J, Goodfellow P. A method for constructing radiation hybrid maps of whole genomes. *Nat Genet* 1994 ; 7 : 22-8.

23. Schuler G, Boguski M, Stewart E, *et al.* A gene map of the human genome. *Science* 1996 ; 274 : 540-6.

24. Deloukas P, Schuler G, Gyapay G, *et al.* A physical map of 30,000 human genes. *Science* 1998 ; 282 : 744-6.

RÉFÉRENCES

25. Weissenbach J, Salanoubat M. Séquence des génomes : le feu d'artifice. *Med Sci* 2000 ; 16 : 10-6.

26. *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans* : a platform for investigating biology. *Science* 1998 ; 282 : 2012-8.

27. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 2000 ; 408 : 796-815.

28. Fleischmann R, Adams M, White O, *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae*. *Science* 1995 ; 269 : 496-512.

29. Adams M, Celniker S, Holt R, *et al.* The genome sequence of *Drosophila melanogaster*. *Science* 2000 ; 287 : 2185-95.

30. Dunham I, Shimizu N, Roe B, *et al.* The DNA sequence of human chromosome 22. *Nature* 1999 ; 402 : 489-95.

31. Hattori M, Fujiyama A, Taylor T, *et al.* The DNA sequence of human chromosome 21. *Nature* 2000 ; 405 : 311-9.

32. Crollius H, Jaillon O, Bernot A, *et al.* Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nat Genet* 2000 ; 25 : 235-8.

33. Dawson E, Chen Y, Hunt S, *et al.* A SNP Resource for Human Chromosome 22 : Extracting dense clusters of SNPs from the genomic sequence. *Genome Res* 2001 ; 11 : 170-8.

TIRÉS À PART

A. Bernot.

Summary

Human Genome Project : after fifteen years of effort

The Human Genome Project was discussed as early as 1985, and the initial program began in 1990. The subsequent sequencing programmes have taken into account the rapid progress in the field, the technological advances, and the appearance of companies from the private sector. The genetic and physical cartography, sequencing of messengers and of the entireity of the human genome were begun successively. Furthermore, a number of genomes of model species have been fully sequenced. Ninety percent of the human genome was sequenced as of June 2000, and it is hoped that it will be finished in it's entirety by 2003. This should lead to the identification of all of the human genes, and of genetic variations of interest.

:

€