m/s SYNTHÈSE

médecine/sciences 2001 ; 17 : 299-308

Premiers regards sur la séquence du génome humain

Roland Heilig
Thomas Brüls

Le séquençage du génome humain marque un tournant majeur dans notre approche de la biologie et de la médecine. Fruit d'une collaboration internationale, il a bénéficié des avancées technologiques et informatiques autorisant de hauts débits de séquençage. L'obtention de la séquence préliminaire du génome humain nous offre pour la première fois une vision globale de son organisation et apporte un éclairage sur l'histoire évolutive des vertébrés et de l'homme. La définition du premier catalogue des gènes humains, dont le nombre semble rejoindre définitivement les estimations les plus basses, constitue une ressource essentielle pour aborder l'étude des réseaux d'interactions fonctionnelles qui sous-tendent les processus biologiques normaux et pathologiques.

Cette fin de siècle a vu le franchissement d'une étape importante pour la connaissance de notre biologie : la détermination, sous la forme d'une version encore « préliminaire », de la séquence du patrimoine génétique de notre espèce [1]. Cette réalisation est le fruit d'une collaboration internationale regroupant 20 laboratoires de 6 pays : États-Unis, Grande-Bretagne, Japon, France, Allemagne et Chine. Ce défi paraissait pourtant utopique il y a encore dix ans, et les financements favorisaient alors les programmes axés sur la partie codante (~ 2 %) du génome. Le séquençage du génome humain ne repose cependant pas sur une révolution technologique majeure, mais résulte d'une succession d'améliorations ponctuelles apportées à une méthode de séquençage proposée dès 1977 par Fred Sanger [2]. L'adoption d'approches globales et d'importants efforts d'automatisation ont permis la réduction des coûts d'un facteur 100. Conjointement, les progrès accomplis dans le domaine de l'informatique ont permis le développement d'outils qui se sont révélés déterminants pour la gestion et l'analyse des flux massifs de données. Ces développements ont nourri des projets de séquençage de plus en plus ambitieux dont le bilan s'établit à ce jour à la connaissance de la séquence des génomes de 595 virus, 31 eubactéries, 7 archaebactéries, 1 levure [3], 1 plante [4], 1 nématode [5], 1 insecte [6] et 1 mammifère [1].

Cependant, si l'enjeu était considérable, tant sur le plan de la biologie fondamentale que sur le plan médical, voire économique, l'ampleur de

ADRESSE

R. Heilig, T. Brüls : Genoscope et Cnrs UMR-8030, 2, rue Gaston-Crémieux, 91057 Évry Cedex, France.

la tâche ne l'était pas moins : la taille du génome humain (3,2 milliards de paires de bases) correspond à 25 fois celle du plus grand génome (drosophile), et à 8 fois la totalité des génomes séquencés auparavant. Il s'agissait de plus du premier génome à forte proportion en éléments répétés, accentuant encore le caractère ambitieux de ce projet.

Émergence du projet « génome humain »

Les premières réunions exploratoires envisageant la possibilité du séquençage de grandes régions du génome humain datent de la fin des années 1980. Cependant, les conditions – scientifiques, technologiques et économiques – n'étant pas réunies à l'époque pour envisager une approche systématique, l'accent fut mis sur la construction de cartes pouvant servir de supports à un séquençage futur : cartes génétiques de première génération (par RFLP, *restriction fragment length polymorphism*) et de seconde génération (par microsatellites), cartes physiques par *contigs* de YAC *(yeast artificial chromosome)* et par hybrides d'irradiation (*voir* l'article de D. Le Paslier et A. Bernot, *p. 294 de ce numéro*).

Le projet de séquençage du génome humain acquit une existence réelle à partir de 1995, et les participants se sont répartis la tâche au niveau international. L'objectif était l'obtention d'une séquence « définitive » en 2005 [7, 8]. Une déontologie d'accès libre et immédiat aux données fut adoptée. Le choix stratégique balançait entre une approche totalement aléatoire *(shotgun)* impliquant le fractionnement du génome entier en une multitude de fragments courts, et une approche dirigée nécessitant l'élaboration, assez fastidieuse, de cartes précises de « pré-séquençage » au moyen de clones de grande taille (cosmides, PAC, BAC, *bacterial artificial chromosome* ou YAC). Cette dernière approche permettait cependant de maîtriser et d'optimiser le séquençage en minimisant la redondance. La première méthode présentait quant à elle l'avantage d'une gestion simplifiée [9], mais repoussait l'assemblage des séquences jusqu'à la fin de la phase d'acquisition des données brutes, rendant impossible un traitement prioritaire du séquençage de régions particulières. La taille importante du génome humain et sa grande richesse en éléments répétés plaidaient en faveur de l'approche dirigée [10].

En juillet 1998, Craig Venter et le groupe privé *Celera* annoncèrent leur intention de déchiffrer la totalité du génome humain par une approche aléatoire globale pour la fin de l'année 2000. Malgré le scepticisme que suscita une telle opération, la crainte d'une privatisation des données stimula les acteurs du projet public qui révisèrent à la hausse leurs objectifs en distinguant deux phases : l'obtention d'une version « préliminaire », « de travail », du génome humain pour fin 2000, puis la production d'une version définitive pour fin 2003.

Plusieurs projets pilotes avaient en effet démontré qu'il était possible d'opérer en deux phases. La première, réalisée avec une couverture de 4-5 équivalents (chaque base étant lue 4 à 5 fois en moyenne), permettait d'acquérir rapidement près de 95 % de l'information de séquence. Elle était suivie d'une phase dirigée, pouvant être différée, destinée à consolider les résultats (résolution des lacunes et assurance de la qualité). Cette phase de finition, en général plus coûteuse en moyens comme en temps, nécessitait une couverture préalable de 8-10 équivalents.

Le succès des projets de séquençage des génomes de *Saccharomyces cerevisiae* et *Caenorhabditis elegans*, en valorisant une information inaccessible par les seules études fondées sur l'ADNc, a suscité une forte demande pour des données de séquence génomiques humaines de la part de la communauté scientifique internationale.

Le consortium international

Cette accélération nécessitait une complète réorganisation et une coordination rigoureuse. Une stratégie aléatoire hiérarchisée fut définie. Dans un premier temps, la complexité génomique, essentiellement due aux éléments répétés, fut réduite en fractionnant aléatoirement le génome en segments d'ADN d'une longueur de 100 à 200 kb. Ces fragments furent organisés en collections de clones fortement redondants. L'analyse du chevauchement entre ces clones permit ensuite de définir un sous-ensemble de fragments de manière à réduire la redondance présente dans la collection initiale. Les clones sélectionnés furent alors individuellement soumis à un nouveau cycle de fractionnement, aboutissant à des fragments de quelques kilobases qui servirent finalement de réactifs au séquençage. L'assemblage des séquences obtenues étant ainsi confiné à un segment restreint du génome (typiquement un clone BAC), les difficultés liées aux éléments répétés et aux duplications génomiques furent ainsi considérablement réduites.

Pour mettre en œuvre cette approche, des ressources communes furent créées. Parmi celles-ci figurent des banques génomiques de clones, des collections d'empreintes de restriction et de séquences d'extrémités de clones.

Les deux principales banques génomiques de clones ont été construites dans des vecteurs BAC, qui combinent une forte stabilité à une grande capacité d'insertion (100-200 kb). Elles totalisaient une couverture en clones (ou « profondeur ») de 40 équivalents génomiques (chaque point du génome étant représenté en moyenne 40 fois).

La collection d'empreintes de restriction a été produite à partir des profils engendrés par l'enzyme HindIII, sur plus de 310 000 clones des deux banques génomiques. Elle a été utilisée pour déduire les chevauchements entre clones et reconstituer leur agencement le long de la molécule de départ. Cette analyse permit de regrouper les clones BAC testés en 942 ensembles *(contigs)* de clones chevauchants, couvrant 96 % de la fraction euchromatique du génome [11, 12]. Des clones furent ensuite sélectionnés dans ces *contigs* et distribués aux différents partenaires du consortium.

Enfin, une collection de 750 000 séquences, dérivées des extrémités de clones des deux banques génomiques fut créée [13]. Cette ressource fut largement utilisée par le Genoscope pour le séquençage du chromosome 14 humain [14], selon une stratégie alternative dénommée

«STC» *(sequence tag connectors)* [15]. Cette approche débute par le séquençage complet de clones BAC « de nucléation », régulièrement espacés le long du chromosome. La séquence obtenue est alors utilisée pour interroger la ressource de séquences d'extrémités afin d'identifier des clones BAC permettant d'étendre la séquence de part et d'autre des points d'initiation. Le processus est réitéré, jusqu'à la jonction de *contigs* adjacents. Cette stratégie ne nécessite pas de cartographie fine et permet une meilleure maîtrise de la redondance à chaque étape.

Les efforts d'automatisation réalisés par les différents centres furent considérables : en fin de phase d'acquisition massive des données (juin 2000), le flux du séquençage atteignait 1 000 nucléotides par seconde pour l'ensemble du consortium, soit 1 équivalent génomique en moins de 6 semaines, ce qui représentait une augmentation de la capacité de production d'un facteur 8 en 8 mois.

Le Golden Path

Les assemblages de clones individuels

Le génome humain contient à peu près 3,2 milliards de paires de bases, dont approximativement 88 % ont été séquencés par le consortium international. La version préliminaire de la séquence du génome humain est composée de centaines de milliers de fragments *(contigs)* de séquence de taille variable, provenant de l'assemblage individuel d'environ 30 000 clones BAC. L'ordre et l'orientation de ces fragments ne sont généralement pas déterminés par le processus de séquençage lui-même.

Le super-assemblage

La construction du *Golden Path* (http://genome.ucsc.edu) a un double objectif : (1) assembler entre eux les fragments chevauchants ; (2) ordonner et orienter les fragments non chevauchants sur la base d'informations complémentaires fournies par les ARN messagers, les EST *(expressed sequence tags)* et les séquences d'extrémités de clones plasmidiques et de BAC.

Bien que les relations d'ordre et les orientations qui sont déduites contiennent des erreurs, le *Golden Path* élimine la majorité de la redondance inhérente aux recouvrements, parfois importants, entre les clones génomiques sélectionnés par le consortium. La taille moyenne des *contigs* de séquence est aussi sensiblement plus grande dans le *Golden Path* que dans les assemblages de clones individuels.

Sans entrer dans les détails du processus de construction du *Golden Path*, on peut mentionner la succession de plusieurs étapes : (1) le filtrage des données initiales afin d'éliminer les contaminations par des séquences non humaines ; (2) la construction de *contigs* de séquences, puis de *contigs* de clones chevauchants ; (3) la structuration des données, et l'intégration de l'information provenant des ARN messagers, des EST et des séquences d'extrémités de clones permettant de déterminer l'ordre des fragments de séquence ; (4) enfin, la déduction d'une séquence consensus pour les fragments de séquences, l'enchaînement des séquences consensus au sein des *contigs* de clones, et l'identification des différents types de lacunes.

La construction du *Golden Path* a exploité des données de cartographie provenant essentiellement des empreintes de restriction *(fingerprinting)*, éventuellement complétées par les séquences d'extrémités de clones BAC, afin de restreindre le nombre de comparaisons à effectuer pour dériver l'agencement des clones. Certaines erreurs présentes dans les cartes de *fingerprint* ont donc pu être propagées dans le *Golden Path*. Les cartes d'empreintes de restriction présentent par exemple des faiblesses dans la résolution des régions dupliquées de grande taille.

L'algorithme de construction du *Golden Path* peut également échouer dans l'assemblage de certaines séquences en raison de différences alléliques ou d'erreurs présentes dans les *contigs* de séquence initiaux, provoquant des duplications locales erronées. Inversement, des séquences réellement dupliquées peuvent se retrouver incorrectement fusionnées. Enfin, il faut mentionner l'existence d'un petit nombre de *contigs* de séquences ne contenant pas de marqueur qui permette de les localiser sur le génome.

La hiérarchie des différents types de *contigs* est illustrée dans la *figure 1*. Les *contigs* de séquence initiaux sont intégrés pour constituer des *contigs* de séquence plus grands. Ceux-ci sont alors liés pour former des ensembles de séquence encore plus grands qui sont localisés au sein des *contigs* de clones séquencés. Ces derniers sont finalement ancrés sur des *contigs* de *fingerprint* qui sont positionnés sur les chromosomes grâce à des données de cartographie par hybrides d'irradiation. Des fragments de gènes hypothétiques détectés à partir de la séquence génomique sont également mentionnés.

Les *Tableaux I* et *II* récapitulent l'état actuel du projet.

Le « paysage » génomique

La disponibilité de la séquence de plus de 90 % du génome humain sous la forme d'un continuum globalement ordonné nous en livre pour la première fois une vue d'ensemble. Plusieurs caractéristiques connues à partir de bases expérimentales ou de données fragmentaires peuvent alors être replacées dans un contexte plus général, et réexaminées à différentes échelles.

La composition nucléotidique du génome n'est pas uniforme. Des régions étendues (de l'ordre du mégabase) possèdent une proportion de G+C très éloignée de la moyenne. Les travaux de G. Bernardi sur les propriétés de l'ADN analysé par des gradients de densité avaient révélé l'existence de fractions discrètes, baptisées isochores, correspondant à des contenus en G+C distincts. Cinq classes d'isochores avaient alors été définies [16]. Cependant, une analyse fondée sur la séquence du génome ne permet pas de distinguer des régions discrètes ayant une composition strictement homogène en G+C. Ceci suggère plutôt d'associer aux isochores une notion de compartiments de composition localement hétérogène. Toutefois, devant l'existence de corrélations entre le contenu en G+C et différentes propriétés biologiques, telles que la distribution des séquences répétées et la densité en gènes, il sera important de préciser cette notion dans le futur.

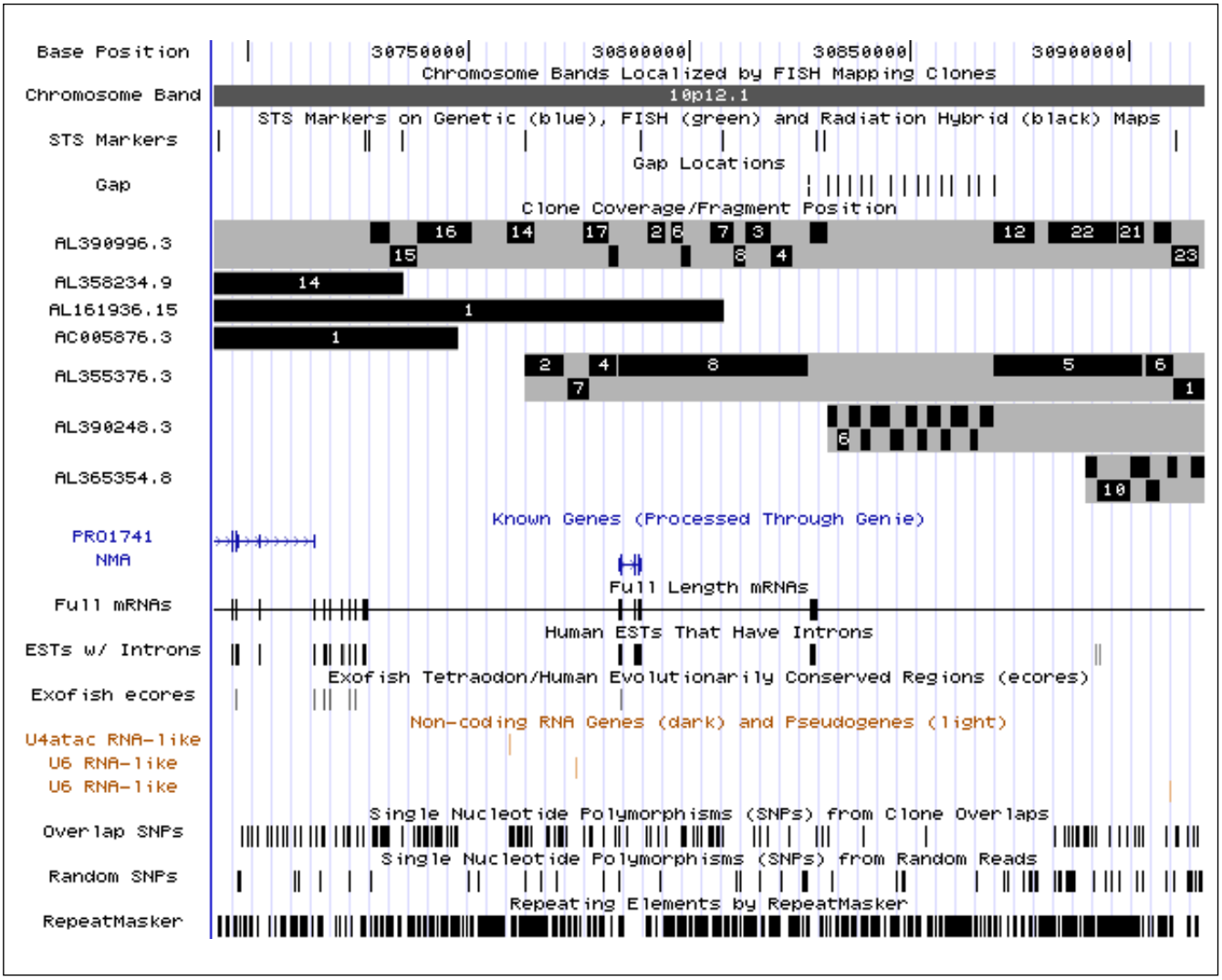

Figure 1. ***Représentation du* Golden Path.** *Les informations données par le* Golden Path *(http://genome.ucsc.edu) sont de haut en bas: la position (en bases) de la région analysée, sa localisation cytogénétique, les marqueurs STS* (sequence tag sites) *cartographiés, puis les lacunes résiduelles* (gap) *séparant les* contigs *de séquence. Ces lacunes résultent de la superposition des assemblages individuels des clones séquencés, qui sont représentés par les barres horizontales épaisses (en noir apparaissent les* contigs *de séquence initiaux) et référencés dans la marge par un identifiant des bases de données. Dans l'exemple présenté on peut voir que presque toutes les lacunes sont localisées au niveau du sixième clone (AL390248). Les 6 autres clones se chevauchent partiellement, ce qui permet d'assembler leurs* contigs *initiaux en des* contigs *de séquence plus grands. Sont ensuite représentés les prédictions de gènes, les segments transcrits (ARNm complets et EST: chaque barre verticale correspond à un exon), les segments conservés chez Tétraodon, les gènes produisant des ARN non codants et les pseudogènes. Les trois dernières lignes représentent les SNP* (single nucleotide polymorphism)*, puis les séquences répétées. Plusieurs fonctions permettent de naviguer dans le* Golden Path*, notamment une fonction zoom permettant d'ajuster à la taille voulue la région analysée, et une fonction permettant de se déplacer le long de la région ciblée.*

La distribution des îlots CpG dans le génome humain a également été examinée. Ce dinucléotide est fortement sous-representé dans le génome des vertébrés en raison de sa fréquence élevée de mutation, liée aux phénomènes de méthylation de l'ADN. On le trouve préférentiellement dans les régions peu soumises à la méthylation, comme les régions transcriptionnellement actives du génome. De fait, dès 1987, A. Bird avait montré que de tels îlots étaient fréquemment associés au voisinage de la région 5' des gènes [17]. L'analyse de la version préliminaire du génome montre que la densité de ces îlots est très variable le long des chromosomes et fortement corrélée à la densité en gènes. Deux extrêmes significatifs illustrent cette tendance: le chromosome 19, très riche en gènes, a une densité de 43,4 îlots/Mb, et le chromosome Y, pauvre en gènes, de 2,9 îlots/Mb.

L'énigme des séquences répétées

La complexité phénotypique d'un organisme n'est pas corrélée à la taille de son génome. Une amibe possède par exemple un génome 200 fois plus grand que celui de l'homme.

Tableau I. État d'avancement du projet génome humain.

Taille du génome humain	3,2 Gb*
nombre de contigs de *fingerprints* utiles	942**
nombre de clones initiaux séquencés	29 298
Somme des données de séquence brutes	23,147 Gb
Somme des contigs des assemblages individuels	4,260 Gb
Total séquence non redondante (super-assemblage)	2,724 Gb
– séquence finie	0,835 Gb
– séquence non finie	1,889 Gb

** Dont 2,95 Gb d'euchromatine.*
*** Soit ~ 96 % de l'euchromatine.*

Tableau II. Statistiques actuelles du super-assemblage *(Golden Path)*.

Chr	Taille estimée (Mb)	Total assemblé (nr) (Mb)	Séquence finie (%)	Contigs de séquence (nb)	Contigs de séquence longueur médiane* (kb)
1	263	214,1	22,9 %	12 406	58,2
2	255	222,9	20,8 %	13 308	57,3
3	214	186,9	12,8 %	15 096	38,0
4	203	169,0	9,5 %	13 183	33,2
5	194	171,0	25,1 %	10 747	72,3
6	183	165,0	50,7 %	5 652	168,0
7	171	149,4	60,8 %	4 666	303,2
8	155	125,2	11,1 %	9 004	38,4
9	145	107,4	15,9 %	6 239	56,0
10	144	127,9	15,0 %	9 142	48,1
11	144	129,2	11,2 %	8 508	40,2
12	143	125,2	25,7 %	8 466	62,5
13	** 98	93,4	22,3 %	5 222	69,9
14	** 93	89,3	81,5 %	829	1 371,0
15	** 89	73,5	3,4 %	5 845	30,3
16	98	74,0	28,3 %	4 923	108,5
17	92	73,4	35,9 %	4 443	88,0
18	85	73,1	7,6 %	4 476	51,3
19	67	56,0	48,2 %	2 531	122,5
20	72	63,3	78,9 %	577	1 229,0
21	** 34	33,8	99,5 %	5	28 515,3
22	** 34	33,7	99,3 %	11	23 048,1
X	164	131,2	50,9 %	4 671	202,6
Y	35	21,8	57,8 %	142	1 045,7
NA	-	14,8	-	613	46,0
Total	**~ 3 200**	**2 724,5**	**30,1 %**	**150 832**	**81,9**

** Cinquante pour cent des bases font partie de* contigs *de taille supérieure ou égale à la longueur médiane.*
*** Bras long uniquement (chromosomes acrocentriques).*
nr : non redondant. NA : non assigné.

Ceci est dû à l'extrême variabilité du contenu en séquences « de remplissage ». Chez l'homme, les régions codantes occupent moins de 2 % du génome, tandis que plus de 50 % sont envahis par différentes familles d'éléments répétés [18, 19]. Parmi celles-ci, on distingue (1) des séquences de faible complexité, constituées par la répétition plus ou moins exacte d'un motif très court (par exemple, les microsatellites de type (CA)n); (2) des copies rétrotransposées d'ARN structuraux ou d'ARN messagers cellulaires (par exemple, les pseudogènes rétro-transcrits); (3) des blocs de séquences répétés en tandem (par exemple, l'ADN satellite des centromères); (4) des séquences dérivées de transposons, dispersées le long du génome.

Beaucoup de ces éléments sont considérés comme de l'ADN inutile, « de remplissage », mais certains jouent un rôle important dans la plasticité des génomes en tant que relais des mécanismes de l'évolution. D'autres, tels certains microsatellites très polymorphes, ont trouvé des applications en cartographie génétique.

Avec leurs 3 millions d'exemplaires, les éléments transposables et leurs dérivés constituent la classe majoritaire. Ils représentent en effet plus de 90 % de l'ensemble des séquences répétées et près de 45 % du génome. Ils se répartissent en 4 groupes principaux: (1) les « LINE » *(long interspersed elements)*, 14 % du génome. D'une taille d'environ 6 kb, ils se propagent par la transcription inverse de leur messager au niveau du nouveau site d'insertion. Sur les trois familles qui existent chez l'homme, une seule est active; (2) les « SINE » *(short interspersed elements)*, 20 % du génome. Leur taille est plus réduite (100 à 400 pb) que celle des « LINE » dont ils utilisent la machinerie pour leur transposition. Seule la famille « Alu » est active chez l'homme; (3) les rétrotransposons à « LTR » *(long terminal repeat)*, 8 % du génome. Ce sont des éléments autonomes qui s'apparentent aux rétrovirus; (4) les transposons-ADN, 3 % du génome. Ils ressemblent aux transposons bactériens. Leur transposase ne faisant pas la différence entre les éléments actifs et inactifs, leur durée de vie est limitée chez leur hôte.

Ces éléments transposables n'étant pas soumis à une contrainte sélective, ils ont pu accumuler des mutations qui permettent de les dater individuellement à partir de la séquence du génome. L'origine de la plupart d'entre eux est antérieure à la radiation des mammifères placentaires. Ceci illustre l'extrême lenteur du processus d'élimination des séquences non fonctionnelles chez les vertébrés. Les transposons-ADN présentent un pic d'activité de part et d'autre de cette radiation. Étant susceptibles de provoquer des réarrangements chromosomiques à grande échelle, ils ont pu jouer un rôle important dans les événements de spéciation. Ils ne semblent plus être actifs chez l'homme

depuis près de 50 millions d'années. Les rétrotransposons à LTR semblent quant à eux proches de l'extinction.

Le déclin de l'activité de transposition dans les lignées qui mènent à l'homme peut s'expliquer à différents niveaux. Tout d'abord, le déficit d'éléments à transmission horizontale chez les mammifères pourrait être lié au développement d'un système immunitaire efficace. Ensuite, une comparaison avec la souris montre que la densité des 4 types d'éléments transposables est sensiblement la même que chez l'homme. Cependant, les différences dans la distribution des âges révèlent que l'activité de transposition des éléments à transmission verticale n'a pas subi le même déclin chez la souris que chez l'homme. Des arguments issus de la génétique des populations pourraient rendre compte de ces différences.

La distribution des éléments répétés dans la séquence du génome est extrêmement variable: de 98 % pour un segment de 200 kb en Xp11, elle chute à moins de 2 % sur les 100 kb entourant chacun des 4 locus des gènes homéobox. Cette distribution dépend notamment de la composition locale en nucléotides. Ainsi, en tant que parasites génomiques, les LINE ont une prédilection marquée pour les régions riches en A+T, pauvres en gènes, où ils risquent moins de provoquer des perturbations. Paradoxalement, les séquences Alu, qui dépendent pourtant de la machinerie LINE pour leur propagation, sont sur-représentées dans les régions de composition G+C plus élevée, riches en gènes. L'examen de leur âge montre que les Alu les plus jeunes sont en fait en excès dans les régions riches en A+T, mais qu'un mécanisme, de sélection positive encore inconnu, doit être responsable de leur accumulation dans les régions riches en G+C. L'existence d'une telle sélection positive suggère que les séquences Alu pourraient avoir eu une fonction bénéfique dans l'histoire évolutive, mais cette dernière reste à découvrir.

L'annotation: vers un catalogue des gènes

Une conséquence immédiate de la détermination de la séquence du génome humain est de pouvoir accéder, dans son environnement génomique propre, à l'ensemble du répertoire des gènes qui le caractérisent. Cette connaissance constitue un préalable à l'étude systématique de la régulation des gènes et du réseau complexe d'interactions de leurs produits dans des conditions physiologiques normales ou pathologiques. Cependant, si l'on a souvent tendance à réduire la notion de gène à ceux qui codent pour des protéines, le génome humain contient aussi des ARN doués de propriétés catalytiques.

Les gènes dont l'ARN est non codant

En marge des gènes codant pour des protéines, il existe chez l'homme plusieurs milliers de gènes dont le produit de transcription n'est pas traduit. Ils se répartissent en plusieurs classes: les ARN de transfert (ARNt), les ARN ribosomiques, et les petits ARN nucléolaires et nucléaires.

En ce qui concerne les ARN de transfert, 535 gènes fonctionnels ont été identifiés, et autant de pseudogènes potentiels. Ils représentent 37 des 39 classes prédites comme étant suffisantes pour assurer le décodage des 61 codons-sens, selon les règles du *wobble* [20]. Ces gènes se trouvent dispersés le long des chromosomes selon une distribution qui semble non aléatoire. Par exemple, une région de 4Mb du chromosome 6 regroupe 148 gènes d'ARNt, presque représentatifs de l'ensemble des classes.

La comparaison du répertoire d'ARNt humain avec ceux du nématode et de la drosophile indique que l'homme possède plus de gènes d'ARNt que la drosophile, mais moins que le nématode. Ce résultat suggère que le nombre de gènes d'ARNt chez les métazoaires n'est pas lié à la complexité phénotypique des organismes, mais plutôt à des idiosyncrasies des niveaux d'ARNt requis dans certains tissus ou stades de développement. Contrairement à la situation observée chez les procaryotes et les eucaryotes inférieurs, le biais des codons (défini comme l'usage préférentiel de codons synonymes) ne semble pas, chez l'homme, corrélé au niveau d'expression des gènes, mais plutôt à leur localisation dans des régions de contenus en G+C distincts. Comme on pouvait s'y attendre dans ce cas, on n'observe qu'une très approximative corrélation entre le nombre de gènes d'ARNt et le biais des codons ou la fréquence des acides aminés.

Quatre ARN ribosomiques (ARNr) distincts, les ARNr 28S, 5,8S, 5S, et 18S, constituent les deux sous-unités des ribosomes. Les gènes des ARNr 28S, 5,8S et 18S sont organisés en unités de 44 kb. On estime qu'au total 150 à 200 copies de ces unités sont distribuées en tandem sur les bras courts des chromosomes acrocentriques (chromosomes 13, 14, 15, 21 et 22). Aucune copie intacte de cet « ADN ribosomique » n'a été retrouvée dans la version actuelle de la séquence du génome: les clones BAC porteurs d'une telle organisation présentaient en effet un profil de restriction de complexité réduite et ont été intentionnellement écartés. Il existe environ 2 000 gènes d'ARNr 5S, qui sont également organisés en tandem en différents locus. Le plus grand regroupement « d'ADNr » 5S, localisé sur le chromosome 1, est dépourvu de sites EcoRI et HindIII, contribuant à la sous-représentation de ces gènes dans les banques de clones génomiques utilisées. Leur séquençage nécessitera donc des efforts ciblés.

Les petits ARN nucléolaires sont impliqués dans les modifications post-transcriptionnelles des ARN ribosomiques, et 87 % d'entre eux ont été retrouvés dans la séquence actuelle.

En ce qui concerne les petits ARN nucléaires, qui font partie des complexes d'épissage, 47 gènes dispersés ont été identifiés pour l'ARN U6, et 15 pour l'ARN U1. Pour des raisons déjà évoquées, certaines familles de gènes organisés en tandem ont pu être manquées à cette étape.

Quant aux autres ARN non codants, qui demeurent encore énigmatiques, l'absence de définition opérationnelle rend leur identification systématique problématique.

Les gènes codants

Bien qu'ils ne représentent qu'une faible proportion du génome, ces gènes suscitent bien évidemment un intérêt biologique et médical majeur.

Leur détection fait appel essentiellement à deux approches.
La première repose sur des prédictions qui combinent des modèles statistiques de gènes, le biais de composition nucléotidique et la présence de signaux dans l'ADN [21]. Si, chez les micro-organismes, la détection des gènes codant revient essentiellement à identifier de longs cadres de lecture ouverts, il n'en est pas de même pour les organismes supérieurs. Chez les vertébrés notamment, les exons sont de petite taille (100-200 pb) et séparés par des introns de taille très variable. Les programmes de prédiction, qui traduisent notre connaissance imparfaite de la structure des gènes et de leurs mécanismes d'expression, se heurtent alors à un problème de bruit de fond [22].
La seconde approche est fondée sur la recherche de similitudes de séquences, utilisant soit des collections de séquences exprimées (ADNc, EST), soit des collections de séquences génomiques qui proviennent le plus souvent d'autres organismes. Des collections de séquences dérivées de transcrits (ADNc, EST) ont été constituées à partir de différents tissus [23, 24]. De telles séquences permettent de reconstituer, au moins en partie, la structure des gènes correspondants. Cependant, ces collections sont nécessairement incomplètes et posent le problème de la représentativité des messagers les moins abondants, ou exprimés dans des types cellulaires minoritaires, ou encore dont l'expression est limitée dans le temps. En outre, les collections d'EST sont inévitablement contaminées par des produits artéfactuels (contaminants génomiques, transcrits immatures…) parfois difficiles à discriminer. La détermination de la structure génomique exacte des gènes est compliquée par le phénomène d'épissage alternatif, responsable d'une diversification des produits d'expression. Il faut mentionner enfin le problème souvent épineux de la distinction entre gènes et pseudogènes.
Quant à la détection systématique des gènes par des comparaisons génomiques entre les espèces, son succès est pour le moment limité par le nombre peu élevé de génomes séquencés à ce jour, et elle constitue une des motivations importantes pour le séquençage du génome de la souris [25] et du tetraodon.

Propriétés des gènes connus

Afin d'étudier les caractéristiques structurales des gènes connus, des séquences d'ADNc pleine longueur (maintenues dans la base de données RefSeq) ont été comparées à la séquence du génome humain. Plus de 85 % de ces séquences purent être alignées sur la quasi-totalité de leur longueur et 92 % sur une partie seulement. Environ 16 % des séquences d'ADNc s'alignaient en plusieurs endroits de la séquence génomique, reflétant probablement la présence de gènes paralogues ou de pseudogènes. Cette analyse a permis de comparer les structures des gènes humains à celles de la drosophile et du nématode. S'il existe certaines différences dans les distributions des tailles d'exons, leur taille typique est semblable dans les trois espèces, ce qui suggère l'existence de composantes exoniques conservées de la machinerie d'épissage. La variabilité de la taille des introns est beaucoup plus importante chez l'homme et résulte en une plus grande variation dans la taille des gènes. Les alignements de séquences d'ADNc contre la séquence génomique donnent une première idée de l'étendue du phénomène d'épissage alternatif. Au moins deux variants d'épissage ont été trouvés pour 59 % des gènes analysés. Cette estimation, probablement en deçà de la réalité, est néanmoins sensiblement plus élevée que la fréquence observée chez le nématode (22 %).

Vers un catalogue des gènes humains

La question du nombre de gènes a été très débattue [26]. Les premières estimations, fondées sur des études de cinétique de réassociation, concluaient à une valeur approximative de 40 000 gènes. Plus récemment, des analyses reposant sur les collections d'EST proposaient des valeurs très disparates, de 34 000 à 120 000 gènes, selon le poids accordé aux contaminations génomiques et la proportion estimée de messagers alternatifs. Une comparaison récente de fractions représentatives des génomes de l'homme et du poisson *Tetraodon nigroviridis* propose une estimation de 30 000 gènes, proche des extrapolations fondées sur la séquence complète des chromosomes 21 [27] et 22 [28].
L'analyse de la séquence du génome humain permet d'estimer le nombre de gènes distincts à environ 32 000. Cette estimation repose sur des hypothèses portant sur le pourcentage de pseudogènes, le taux de fragmentation (lorsqu'un même gène est compté plus d'une fois), le taux de faux positifs et les limites de sensibilité des outils de prédiction. Malgré les incertitudes qui demeurent sur le nombre exact de gènes, et qui sont liées essentiellement à la difficulté d'augmenter la sensibilité des méthodes de détection sans pour autant réduire fortement la spécificité, il semble que l'homme ne possède qu'à peu près deux fois plus de gènes que la drosophile ou le nématode. Toutefois, les gènes humains diffèrent de ceux de ces organismes par plusieurs aspects importants : ils s'étendent sur des régions génomiques beaucoup plus grandes et contiennent plus d'exons qu'ils utilisent pour construire beaucoup plus de transcrits alternatifs, à l'origine d'une diversité fonctionnelle accrue.
Ce premier catalogue des gènes humains est bien sûr imparfait. L'intégration des données futures, provenant tant de l'effort de finition de la séquence génomique que de l'enrichissement des collections de transcrits (notamment d'ADNc complets) en améliorera la qualité. Ce catalogue, maintenu à jour et enrichi par des données fonctionnelles, constitue déjà une ressource essentielle pour la communauté scientifique.
La taille moyenne d'un gène étant estimée à 27 kb chez l'homme et sa longueur codante moyenne à 1 340 pb, environ 28 % du génome est transcrit et 1,4 % correspond à des séquences codantes. Des écarts importants sont cependant observés. Le plus long gène est celui de la dystrophine (dont les mutations sont responsables des myopathies de Duchenne et de Becker), qui s'étend sur 2 400 kb. Le gène de la titine cumule quant à lui les records du messager le plus long (80 780 bases), du nombre d'exons le plus élevé (178), et de l'exon codant le plus long (17 106 pb). La densité

moyenne des gènes dans le génome humain est de 11,1 par Mb, avec des extrêmes allant de 26,8 pour le chromosome 19 à 6,4 pour le chromosome Y. Il est d'ailleurs probable que cette dernière valeur soit surévaluée, le chromosome Y étant particulièrement riche en pseudogènes.

Analyse comparée du protéome

L'obtention de la séquence du génome humain a permis d'effectuer une analyse comparative du répertoire protéique humain avec celui d'autres organismes (plusieurs micro-organismes, *S. cerevisiae*, *C. elegans*, *Drosophila melanogaster*, *Arabidopsis thaliana*) *(Tableau III)*. Elle a notamment conduit à la découverte surprenante de 223 protéines qui n'ont de similitudes qu'avec des protéines bactériennes, et sont probablement le fruit d'événements de transferts horizontaux de gènes. La plupart de ces gènes ont depuis acquis des introns chez l'homme.

Une homologie est retrouvée entre le protéome humain et 61 % du protéome de la drosophile, 43 % de celui du nématode et 46 % de celui de la levure. Un noyau constitué de 1 322 groupes de protéines communs à ces organismes assure les fonctions de base de la cellule (réplication et réparation de l'ADN, transcription, traduction et métabolisme). Une classification des protéines sur la base de motifs fonctionnels et de domaines montre que seules 7 % des familles ainsi définies paraissent spécifiques des vertébrés. Cela montre que les mécanismes d'innovation chez les vertébrés ne sont pas dominés par la création de nouveaux domaines. Il faut noter que ces classifications reposent sur la détection d'homologies de séquence, et que beaucoup de gènes dérivés d'un ancêtre commun ont pu diverger depuis. Ainsi, l'utilisation du programme PSI-BLAST révèle des homologies lointaines entre des gènes d'organismes non vertébrés et environ 45 % des gènes appartenant à l'ensemble spécifique des vertébrés.

Innovations chez les vertébrés

La faible proportion de familles de domaines protéiques spécifiques des vertébrés suggère que peu de motifs nouveaux ont été inventés depuis leur radiation. L'augmentation évidente de la complexité phénotypique entre la levure, le nématode ou la drosophile, et les vertébrés ne s'accompagnant que d'une augmentation relativement modérée du nombre de gènes, l'innovation se place donc à un autre niveau. Un mécanisme important d'innovation chez les vertébrés est la création de nouvelles combinaisons de motifs et domaines, résultant en des architectures nouvelles (arrangements linéaires de domaines à l'intérieur d'une protéine), associées à de nouvelles fonctions. L'expansion de familles de protéines particulières, comme la famille des immunoglobulines et des facteurs de croissance impliqués dans le développement, a constitué un autre mécanisme de diversification fonctionnelle. Environ 60 % des familles de protéines contiennent plus de membres chez l'homme que chez n'importe lequel des organismes dont le génome est à ce jour séquencé. Ceci suggère que la duplication de gènes a été une force évolutive majeure dans l'histoire des vertébrés. Une complexification, même modeste, à différents niveaux, peut engendrer une densification des réseaux fonctionnels par un phénomène d'amplification combinatoire. La prise en considération de l'ensemble des sources de diversité fonctionnelle, agissant aussi bien au niveau pré-traductionnel que post-traductionnel, conduit à une estimation du nombre de protéines distinctes produites par l'organisme humain de plusieurs millions.

Retombées biomédicales

Les données partielles produites par le consortium pendant toute la durée du projet ont été déposées quotidiennement dans des bases de données publiques, et ont été exploitées par de nombreuses équipes, publiques et privées, engagées dans des projets d'identification de gènes responsables de maladies génétiques.

Jusqu'alors, l'approche classique du « clonage positionnel » nécessitait de déployer des ressources considérables en moyens techniques et humains. La disponibilité de la séquence du génome humain élimine une grande partie de ce travail fastidieux en établissant un lien direct entre l'intervalle génétique déterminé et le segment physique correspondant. Elle permet en outre d'affiner cet intervalle en fournissant de nouveaux marqueurs génétiques (microsatellites ou SNP), autorise une identification rapide des gènes candidats et, par la connaissance de leur structure génomique, facilite la recherche de mutations. En conséquence, une entreprise qui nécessitait auparavant des années de labeur, de gros moyens et des compétences techniques pointues, devient souvent réalisable en quelques mois par une équipe de recherche de taille réduite.

La connaissance de la séquence du génome humain permet également l'identification rapide des paralogues d'un gène d'intérêt, au moyen d'une recherche d'homologie à faible stringence. De tels paralogues peuvent par exemple être de bons candidats pour des formes voisines d'une pathologie étudiée : c'est ainsi que la préséniline-2 impliquée dans les formes précoces de la maladie d'Alzheimer a été identifiée à partir du gène de la préséniline-1 [29]. De

Tableau III. Analyse comparée des protéomes.

	Homme	Drosophile	Nématode	Levure	Arabidopsis
Nombre de gènes identifiés	~ 32 000	13 338	18 266	6 144	25 706
Nombre de familles de domaines (InterPro)	1 262	1 035	1 014	851	1 010
Nombre d'architectures de domaines	1 695	1 036	1 018	310	
Fraction homologue au protéome humain		61 %	43 %	46 %	

même, un nouveau récepteur de la sérotonine (de forte conductance) a été identifié à partir du récepteur connu (de faible conductance), alors que les analyses fondées sur les EST et des expériences d'hybridations croisées avaient échoué.
Enfin, plus de 400 000 SNP ont été déterminés dans les zones de recouvrement entre les clones BAC séquencés, et une carte de 1,42 million de SNP a pu être dressée en intégrant toutes les données disponibles. L'espacement moyen entre SNP est de 1,9 kb, ce qui correspond à environ 15 SNP par gène. Cette collection de marqueurs a déjà un impact important en génétique médicale et en génétique des populations.

Conclusions

Le séquençage du génome humain marque une étape importante dans l'entreprise scientifique qui vise à comprendre les mécanismes du développement de l'homme, sa physiologie et son histoire évolutive. Le génome est sans aucun doute un lieu où des informations moléculaires essentielles sont conservées, et nous avons résumé quelques-unes des observations dérivées des premières analyses globales de la séquence du génome humain. Cependant, observer n'est pas comprendre, et l'extraction de connaissances à partir des données de la génomique occupera les scientifiques de disciplines variées dans les décennies à venir.
Le mode de fonctionnement du projet génome humain, notamment sa politique de transparence et de libre accès aux données en temps réel, a joué un rôle déterminant dans l'accomplissement du projet. Il n'y a pas de doute que cette information a aussi été utile aux industries du secteur privé qui ont pu y incorporer leurs propres données. Lorsque l'on considère l'ampleur de la tâche à venir (aussi bien en volume qu'en complexité), les répercussions scientifiques, humaines et économiques, et le nombre croissant d'acteurs dans le domaine (tant dans le secteur public que privé), il est peu probable qu'un tel mode de fonctionnement puisse être conservé intact.
Bien que la motivation initiale du projet génome humain ait été de faciliter l'identification de gènes associés à des maladies, on a vu s'afficher progressivement de nouveaux objectifs liés à une connaissance plus systémique des propriétés du génome. Si une approche génétiquement déterministe sous-tend la réussite des travaux de clonage positionnel, il est probable qu'une approche réductionniste, strictement fondée sur le génome, échouera à rendre compte de certaines propriétés macroscopiques, du type de celles qui sont étudiées en physiologie.
Il est difficile d'imaginer à ce stade comment la connaissance de l'information génétique permettra d'élaborer une connaissance de propriétés systémiques telles que l'homéostasie cellulaire, les propriétés cognitives ou plus simplement la cinétique de formation d'un complexe protéique. L'un des principaux résultats de la connaissance de la séquence du génome est de fournir le répertoire des composants cellulaires qui sont codés génétiquement. Les interactions de ces composants entre eux, et avec des composants non codés génétiquement, nécessitent d'explorer d'autres niveaux d'organisation avec de nouveaux outils technologiques, mais aussi avec de nouveaux concepts. L'amélioration des techniques d'analyse des données d'expression et les techniques de typage de SNP pour l'analyse génétique des maladies complexes sont des exemples d'avancées technologiques étroitement liées aux données génomiques. Les nouveaux centres pluridisciplinaires dédiés à l'analyse des systèmes complexes illustrent la recherche de ces nouveaux concepts, qui seront peut-être les facteurs limitants dans les décennies qui suivront l'obtention de la séquence du génome humain.
Les implications éthiques, légales et sociales des avancées dans le domaine de la génomique sont telles qu'il est absolument indispensable d'impliquer l'ensemble de la société dans les prises de décisions qui orienteront les recherches futures [30]. Bien que nous n'ayons rapporté ici que certains des aspects scientifiques issus du séquençage du génome humain, une réflexion sur les conséquences pour la société de ces découvertes et de leur exploitation est d'une importance au moins égale ■

RÉFÉRENCES

1. International Human Genome Sequencing Consortium. The human genome: initial sequencig and analysis. *Nature* 2001; 409 : 860-921.

2. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* 1977; 74: 5463-7.

3. Goffeau A, Barrell BG, Bussey H, *et al.* Life with 6000 genes. *Science* 1996; 274: 546: 563-7.

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

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

6. Adams MD, Celniker SE, Holt RA, *et al.* The genome sequence of *Drosophila melanogaster*. *Science* 2000; 287: 2185-95.

7. Marshall E. NIH to produce a «working draft» of the genome by 2001. *Science* 1998; 281: 1774-5.

8. Collins FS, Patrinos A, Jordan E, Chakravarti A, Gesteland R, Walters L. New goals for the US Human Genome Project: 1998-2003. *Science* 1998; 282: 682-9.

9. Weber JL, Myers EW. Human whole-genome shotgun sequencing. *Genome Res* 1997; 7: 401-9.

10. Green P. Against a whol-genome shotgun. *Genome Res* 1997; 7: 410-7.

11. The International Human Genome Mapping Consortium. A physical map of the human genome. *Nature* 2001; 409: 934-41.

12. Soderlund C, Humphray S, Dunham A, French L. Contigs built with fingerprints, markers, and FPC V4.7. *Genome Res* 2000; 10 : 1772-87.

13. Zhao S, Malek J, Mahairas G, *et al.* Human BAC ends quality assessment and sequence analysis. *Genomics* 2000; 63: 321-32.

14. Brüls T, Gyapay G, Petit JL, *et al.* A physical map of human chromosome 14. *Nature* 2001; 409: 947-8.

15. Venter JC, Smith HO, Hood L. A new strategy for genome sequencing. *Nature* 1986; 381: 364-6.

16. Bernardi G, Olofsson B, Filipski J, *et al.* The mosaic genome of warm-blooded vertebrates. *Science* 1985; 228: 953-8.

17. Bird AP. CpG islands as gene markers in the vertebrate nucleus. *Trends Genet* 1987; 3: 342-7.

18. Smit A. Interspersed repeat and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev* 1999; 9: 657-63.

RÉFÉRENCES

19. Prak EL, Prak Jr HHK, *et al.* Mobile elements and the human genome. *Nat Rev Genet* 2000 ; 1 : 134-44.

20. Crick FH. Codon-anticodon pairing : the wobble hypothesis. *J Mol Biol* 1966 ; 19 : 548-55.

21. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* 1997 ; 268 : 78-94.

22. Guigo R, Agarwal P, Abril JF, Burset M, Fickett JW. An assessment of gene prediction accuracy in large DNA sequences. *Genome Res* 2000 ; 10 : 1631-42.

23. Adams MD, Kelley JM, Gocayne JD, *et al.* Complementary DNA sequencing : expressed sequence tags and human genome project. *Science* 1991 ; 252 : 1651-6.

24. Hillier LD, Lennon G, Becker M, *et al.* Generation and analysis of 280,000 human expressed sequence tags. *Genome Res* 1996 ; 6 : 807-28.

25. Marshall E. Public-Private project to deliver mouse henome in 6 months. *Science* 2000 ; 290 : 242-3.

26. Aparicio SAJR. How to count… human genes. *Nat Genet* 2000 ; 25 : 129-30.

27. Hattori M, Fujiyama A, Taylor TD, *et al.* The DNA sequence of human chromosome 21. The chromosome 21 mapping and sequencing consortium. *Nature* 2000 ; 405 : 311-9.

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

29. Rogaev EI, Sherrington R, Rogaeva EA, *et al.* Familial Alzheimer's disease in kindreds with missense mutations in a gene in chromosome 1 related to the Alzheimer's disease type 3 gene. *Nature* 1995 ; 376 : 775-8.

30. Kevles DJ, Hood L (eds.). The code of codes : scientific and social issue in the human genome project. Cambridge, Massachusetts and London : Harvard University Press, 1992.

TIRÉS À PART

R. Heilig.

Summary

Initial view of the human genomic sequence

The vast amount of information that has recently become available with the release of the « working draft » sequence, makes it possible to take an initial, global look at the contents of the human genome. The draft genome sequence is the result of an international collaboration and was primarily generated from a physical map of the human chromosomes that covered 96 % of the euchromatin portion of the genome. The sequence itself covers about 90%, and has been made available without restriction to the scientific community ever since the beginning of the effort. This sequencing project is of particular interest because it represents (1) the largest genome analyzed to date, (2) the first vertebrate sequence, and (3) the genome of our own species. We discuss some scientific issues related to the generation and assembly of the data, and on key observations derived from the initial analysis of the genome sequence.