

**M/S : médecine sciences**



## **Séquençage du génome humain: le chromosome 14, quatrième chromosome terminé**

## **Human genome sequencing: chromosome 14, the fourth finished chromosome**

Roland Heilig and Núria Fonknechten

Volume 19, Number 5, mai 2003

Neurosciences

URI: <https://id.erudit.org/iderudit/006614ar>

[See table of contents](#)

Publisher(s)

SRMS: Société de la revue médecine/sciences  
Éditions EDK

ISSN

0767-0974 (print)  
1958-5381 (digital)

[Explore this journal](#)

Cite this article

Heilig, R. & Fonknechten, N. (2003). Séquençage du génome humain: le chromosome 14, quatrième chromosome terminé. *M/S : médecine sciences*, 19(5), 522–525.

Tous droits réservés © M/S : médecine sciences, 2003

This document is protected by copyright law. Use of the services of Érudit (including reproduction) is subject to its terms and conditions, which can be viewed online.

<https://apropos.erudit.org/en/users/policy-on-use/>

**é**rudit

This article is disseminated and preserved by Érudit.

Érudit is a non-profit inter-university consortium of the Université de Montréal, Université Laval, and the Université du Québec à Montréal. Its mission is to promote and disseminate research.

<https://www.erudit.org/en/>

8. Hilvert D. Critical analysis of antibody catalysis. *Annu Rev Biochem* 2000; 69: 751.
9. Jerne NK. Toward a network theory of the immune system. *Ann Immunol (Paris)* 1974; 125C: 373-89.
10. Izadyar L, Friboulet A, Remy MH, et al. Monoclonal anti-idiotypic antibodies as functional internal images of enzyme active sites: production of a catalytic antibody with a cholinesterase activity. *Proc Natl Acad Sci USA* 1993; 90: 8876-80.
11. Avallé B, Thomas D, Friboulet A. Functional mimicry: elicitation of a monoclonal anti-idiotypic antibody hydrolyzing beta-lactams. *FASEB J* 1998; 12: 1055-60.
12. Pillet D, Paon M, Vorobiev II, et al. Idiotypic network mimicry and antibody catalysis: lessons from the elicitation of efficient anti-idiotypic protease antibodies. *J Immunol Meth* 2003; 269: 5-12.
13. Friboulet A, Avallé B, Debat H, Thomas D. A possible role of catalytic antibodies in metabolism. *Immunol Today* 1999; 20: 474-5.
14. Kit YY, Semenov DV, Nevinsky GA. Phosphorylation of different human milk proteins by human catalytic secretory immunoglobulin A. *Biochem Mol Biol Int* 1996; 39: 521-7.
15. Kanyshkova TG, Semenov DV, Khlimankov D, et al. DNA-hydrolyzing activity of the light chain of IgG antibodies from milk of healthy human mothers. *FEBS Lett* 1997; 416: 23-6.
16. Paul S, Volle DJ, Beach CM, et al. Catalytic hydrolysis of vasoactive intestinal peptide by human autoantibody. *Science* 1989; 244: 1158-62.
17. Li L, Paul S, Tyutyulkova S, et al. Catalytic activity of anti-thyroglobulin antibodies. *J Immunol* 1995; 154: 3328-32.
18. Thiagarajan P, Dannenbring R, Matssura K, et al. Monoclonal antibody light chain with prothrombinase activity. *Biochemistry* 2000; 39: 6459-65.
19. Lacroix-Desmazes S, Moreau A, Sooryanarayana S, et al. Catalytic activity of antibodies against factor VIII in patients with hemophilia A. *Nat Med* 1999; 5: 1044-7.
20. Shuster AM, Gololobov GV, Kvaschuk OA, et al. DNA hydrolyzing autoantibodies. *Science* 1992; 256: 665-7.
21. Vlassov A, Florentz C, Helm M, et al. Characterization and selectivity of catalytic antibodies from human serum with RNase activity. *Nucleic Acids Res* 1998; 26: 5243-50.
22. Baranovskii AG, Ershova NA, Buneva VN, et al. Catalytic heterogeneity of polyclonal DNA-hydrolyzing antibodies from the sera of patients with multiple sclerosis. *Immunol Lett* 2001; 76: 163-7.
23. Lacroix-Desmazes S, Bayry J, Misra N, et al. The prevalence of proteolytic antibodies against factor VIII in hemophilia A. *N Engl J Med* 2002; 346: 662-7.
24. Wentworth JP, McDunn JE, Wentworth AD, et al. Evidence of antibody-catalyzed ozone formation in bacterial killing and inflammation. *Science* 2002; 298: 2195-9.
25. Wentworth A, Jones L, Wentworth P, et al. Antibodies have the intrinsic capacity to destroy antigens. *Proc Natl Acad Sci USA* 2000; 97: 10930-5.
26. Nardi M, Tomlinson S, Greco MA, Karpatkin S. Complement-independent, peroxide-induced antibody lysis of platelets in HIV-1-related immune thrombocytopenia. *Cell* 2001; 106: 551-61.
27. Tawfik DS, Chap R, Green BS, et al. Unexpectedly high occurrence of catalytic antibodies in MLR/lpr and SJL mice immunized with a transition-state analog: is there a linkage to autoimmunity. *Proc Natl Acad Sci USA* 1995; 92: 2145-9.
28. Thomas NR. Hapten design for the generation of catalytic antibodies. *Appl Biochem Biotechnol* 1994; 47: 345-72.

## NOUVELLE

### Séquençage du génome humain : le chromosome 14, quatrième chromosome terminé

Roland Heilig, Núria Fonknechten

> En juin 2000, un consortium public regroupant 20 laboratoires de six pays annonça le séquençage de 90 % du génome humain\* [1] sous forme d'un arrangement encore imparfait (*working draft*) de plus de 150 000

\* Il s'agit en fait de la partie euchromatique (soit environ 90 % du génome). L'hétérochromatine, représentée surtout au niveau des centromères et des télomères et quasiment dépourvue de gènes, est difficilement accessible aux techniques usuelles de séquençage, en raison d'une structure répétitive particulière et d'une homologie souvent élevée entre les centromères des différents chromosomes.

segments (*contigs*) d'une longueur médiane\*\* de près de 90 000 paires de bases (pb) et positionnés entre eux, par chromosome, sur la base d'informations de proximité [2] (→). Cette première esquisse constituait une avancée considérable, notamment en servant de support à la recherche des gènes responsables de mala-

\*\* La moitié de la séquence est assemblée en *contigs* d'une taille égale ou supérieure à la longueur médiane.

Genoscope, Centre National de Séquençage, 2, rue Gaston Crémieux, CP 5706, 91057 Évry Cedex, France.  
[heilig@genoscope.cns.fr](mailto:heilig@genoscope.cns.fr)  
[fonk@genoscope.cns.fr](mailto:fonk@genoscope.cns.fr)

dies génétiques. Restait à accomplir une tâche de « finition », tout aussi capitale, visant à déterminer la séquence manquante entre les *contigs*, puis à assurer en tout point un niveau strict de qualité. Cette étape, déjà accom-

plie pour les trois plus petits chromosomes, 22, 21 puis 20 [3-5], devrait être complétée pour les suivants au cours de cette année. Disposer d'une séquence finie et fiable reste essentiel pour nombre d'analyses fines du génome, et en particulier pour une description détaillée des gènes, de leurs mécanismes d'action et de régulation.

Le Genoscope (Centre national de séquençage) a

(→) m/s  
 2001, n° 3,  
 p. 309



apporté sa contribution à cet effort international en publiant récemment [6] la séquence complète et l'analyse du chromosome 14\*, de taille intermédiaire. Pour établir la séquence des chromosomes, une procédure hiérarchisée a été utilisée, qui a nécessité une fragmentation préalable du génome en segments chevauchants, d'une taille manipulable (150-200 kb), clonés dans un vecteur BAC (*bacterial artificial*

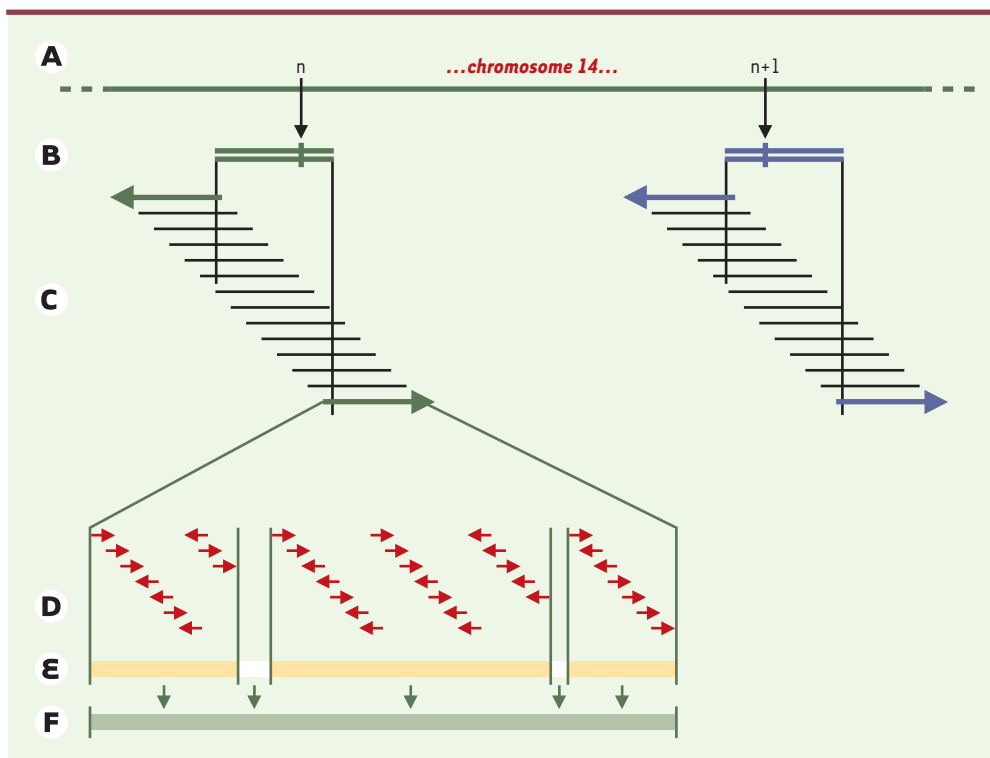
*chromosome*). Les BAC d'intérêt sont ensuite sélectionnés dans cette collection, riche de plus de 900 000 clones. Chaque clone est alors soumis individuellement à une nouvelle fragmentation visant à déterminer la séquence du segment qu'il représente (Figure 1 D, E, F), laquelle prendra place au sein d'une reconstitution de l'ensemble. Pour le chromosome 14, cette procédure a été répétée sur 681 clones. La stratégie utilisée pour leur sélection [7] se distingue de celle qui avait été mise en œuvre pour les autres chromosomes [8]. Plutôt que de nous engager dans un important effort préalable de cartographie, nous avons préféré débiter le processus d'acquisition de la séquence à partir de « balises » réparties le long du

chromosome, puis d'utiliser ces données comme autant de points de départ d'une progression locale, bidirectionnelle et contrôlée (Figure 1 A, B, C), jusqu'à la fusion entre les groupes d'expansion adjacents. Un effort particulier, accompli pour résoudre les dernières lacunes récalcitrantes, a permis d'établir la séquence finale sous forme d'un continuum ininterrompu de 87410661 pb, s'étendant du centromère au télomère du bras long, et qui constitue, outre la plus longue séquence assemblée à ce jour, la première séquence complète d'un chromosome (Tableau 1). Elle a nécessité plus de 1 600 000 réactions de séquençage. La cohérence de son assemblage a été vérifiée par comparaison avec la carte génétique de Généthon, puis avec d'autres collections de marqueurs. Pour atteindre un niveau plus fin de résolution, une procédure expérimentale automatisée a été développée, destinée à s'assurer de l'intégrité des clones sélectionnés et, à la fois, de la conformité de leur séquençage.

La composition nucléotidique moyenne du chromosome 14 est proche de celle de l'ensemble du génome, avec un contenu moyen en (G+C) de 40,9 % [2], mais présente d'importantes variations locales de 32,6 à 61,2 % (pour des fenêtres de 50 kb). Les gènes s'étendent sur 43,6 % du chromosome 14, mais la somme de leurs exons n'en représente que 2,3 %, pour un potentiel codant protéique de 1,1 %. Les séquences répétées, réparties essentiellement en SINE (type court, exemple « Alu ») et LINE (type long), représentent

Figure 1. Stratégie de séquençage du chromosome 14. Ce schéma représente les différentes étapes du processus utilisé pour la sélection des clones à séquencer. **A**. Étape d'initiation : 162 « balises » (courts fragments de séquence connue) sont sélectionnées le long du chromosome 14. **B**. Les BAC (*bacterial artificial chromosome*) d'initiation correspondants à ces balises sont identifiés, par hybridation, à partir de la banque de clones construite pour l'ensemble du génome. **C**. La séquence des BAC d'initiation (obtenue selon D, E, F) sert alors à identifier dans la banque, pour chacun, un ensemble de BAC dont deux seront retenus pour étendre le séquençage dans les deux sens. **D**. Le séquençage de tous les clones BAC est réalisé par fractionnement en segments de quelques kilobases, sous-clonés en vecteur plasmide et séquencés aux extrémités. La combinaison de 1000 à 1500 lectures permet pour chacun de construire un assemblage préliminaire (**E**), servant de base aux opérations de finition (**F**). Le processus C, D, E, F est répété jusqu'à la fusion des groupes adjacents.

\* Il s'agit de la partie euchromatique du chromosome 14, restreinte à son bras long. Le bras court des chromosomes acrocentriques humains (13, 14, 15, 21 et 22), essentiellement hétérochromatique, contient les gènes codant pour des ARN ribosomiques mais, semble-t-il, aucun gène « protéique ». La partie séquençage résulte d'un effort conjoint du Genoscope (86 %), du Multimegabase Sequencing Center, Seattle (USA) (8,5 %) du Genome Sequencing Center, St Louis (USA) (4,5 %) et de la Kyoto University Graduate School of Medicine (Japon) (1 %).



(→) m/s 2001, n° 1, p. 103; n° 5, p. 628; n° 8-9, p. 911 46,2 % (→). La distribution de ces divers éléments apparaît cependant inhomogène, les régions à fort contenu en (G+C) montrant une corrélation étroite avec une forte densité en SINE, une faible densité en LINE et une concentration accrue en exons et en gènes. Ces régions côtoient souvent sans transition des « déserts de gènes », le plus grand atteignant 6 Mb.

L'identification des gènes s'est déroulée en deux étapes. La première, entièrement automatique, a consisté à identifier, en consultant les différentes bases de données, les séquences exprimées (ADNc et peptides) présentant une homologie avec la fraction non répétée de la séquence du chromosome. Après une série de traitements par des programmes spécifiques permettant d'affiner la structure des résultats, des modèles primaires de gènes ont été construits. Indépendamment, les modèles prédits par des programmes spécialisés à partir des seules informations de séquence ont été répertoriés. Enfin, les segments conservés (*ecores*) entre le génome de l'homme et ceux de la souris ou du poisson à génome compact *Tetraodon nigroviridis* [2, 9, 10] (→) ont été recherchés à ce stade.

La seconde étape fait appel à l'expertise humaine pour valider chaque modèle de gène dans son contexte génomique, préciser sa structure et son étendue et identifier d'éventuelles formes d'épissage alternatif.

Nous avons caractérisé 850 gènes sur le chromosome 14, qui se répartissent en 506 gènes déjà connus, 121 gènes nouveaux, 212 gènes présomptifs (sans cadre de lecture significatif) et 11 gènes prédits, ainsi que 296 pseudogènes. Il faut ajouter 200 « segments de gènes » et 97 « pseudo-segments » provenant des gènes codant pour des récepteurs de cellules T (TCR) et les chaînes

lourdes d'immunoglobulines (IGH), ces gènes étant caractérisés par une structure génomique particulière. Si l'on exclut ces deux dernières catégories, le chromosome 14 présente une densité moyenne de 10 gènes par mégabase (Mb), en accord avec l'estimation moyenne définie pour l'ensemble du génome (9,3 à 10,8 gènes/Mb), mais assez différente cependant des valeurs observées pour les autres chromosomes finis (Tableau I). La taille moyenne des gènes connus et nouveaux du chromosome 14 est de 58,7 kb, six d'entre eux dépassent 500 kb et le plus étendu (1691847 pb) code pour la neurexine. Quant aux transcrits, le plus grand (115 exons totalisant 21794 pb) code pour SYNE-2 (ou nesprine-2), une protéine de 6649 acides aminés associée à l'enveloppe nucléaire des cellules musculaires. À noter la forte proportion des gènes présomptifs (25 %), qui combinent plusieurs origines: gènes incomplètement caractérisés (ressources trop fragmentaires), gènes codant pour de très courts polypeptides, ou utilisant une déviation du code génétique (incorporation de sélénocystéine au niveau de codons stop, par exemple), ou encore gènes assurant leur fonction directement par leur transcrit ARN. Il est probable aussi qu'une fraction des gènes de cette catégorie résulte d'un « bruit de fond » contaminant les bases de données d'expression.

Ces résultats sont en accord avec l'estimation de 28 à 32000 gènes pour l'ensemble du génome humain [1, 10]. La comparaison avec les nombres de gènes identifiés chez la drosophile (13379), le nématode (19099) ou une plante, l'ara-

bette (25498), ne reflète évidemment pas les différences phénotypiques. Des mécanismes produisant un niveau supplémentaire de complexité sont à rechercher, tels que l'épissage différentiel. Ce phénomène, observé pour 54 % des gènes du chromosome 14, peut revêtir une importance considérable. Ainsi, dans le cas du gène codant pour la neurexine [11], il offre une potentialité combinatoire susceptible d'engendrer 1728 transcrits distincts.

L'annotation de la séquence du chromosome 14 (accessible sur <http://www.genoscope.fr/chr14/>) représente l'état des connaissances actuelles et devra être maintenue à jour régulièrement, selon l'enrichissement des différentes ressources qui lui servent de base, et devra notamment incorporer les résultats des analyses expérimentales en cours.

Le séquençage du chromosome 14, dans le contexte global du séquençage du génome humain, et sa contribution à la constitution du répertoire des gènes humains, constituent un pas important vers la compréhension des mécanismes biologiques, physiologiques ou physiopathologiques, et des phénomènes moteurs de l'évolution. À ce titre, ils fourniront sans doute matière à de nombreuses études. Les bénéfices les plus importants apportés par la génomique humaine sont attendus dans le domaine de la santé. Plus de 60 maladies génétiques, pour lesquelles plus de 30 gènes morbides restent à identifier, sont associées au chromosome 14, et de nombreux déficits ne sont pas encore assignés. ♦

#### Human genome sequencing: chromosome 14, the fourth finished chromosome

	Chromosomes				Génome [1] (euchromatine)
	22q [3]	21q [4]	20 [5]	14q [6]	
Taille (Mb)	33,4	33,5	59,2	87,4	~ 2900
Nombre de lacunes	11	3	4	0	~ 150000
Contenu en (G+C) (%)	47,8	40,9	44,1	40,9	41,0
Taille moyenne des gènes connus (kb)	-	57,0	51,3	58,7	27,0
Densité génique (gènes/Mb)	16,3	6,7	12,2	10,0	9,3 - 10,8

Tableau I. Caractéristiques comparées entre les chromosomes finis et le génome entier (working draft).



## RÉFÉRENCES

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* 2001; 409: 860-921.
2. Heilig R, Brüls T. Premiers regards sur la séquence du génome humain. *Med Sci* 2001; 17: 299-308.
3. Dunham I, Shimizu N, Roe Ba, et al. The DNA sequence of human chromosome 22. *Nature* 1999; 402: 489-95.
4. Hattori M, Fujiyama A, Taylor TD, et al. The DNA sequence of human chromosome 21. *Nature* 2000; 405: 311-9.
5. Deloukas P, Matthews LH, Ashurst J, et al. The DNA sequence and comparative analysis of human chromosome 20. *Nature* 2001; 414: 865-71.
6. Heilig R, Eckenberg R, Petit JL, et al. The DNA sequence and analysis of human chromosome 14. *Nature* 2003; 421: 601-7.
7. Brüls T, Gyapay G, Petit JL, et al. A physical map of human chromosome 14. *Nature* 2001; 409: 947-8.
8. The International Human Mapping Consortium. A physical map of the human genome. *Nature* 2001; 409: 934-41.
9. Roest 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.
10. Roest Crollius H, Jaillon O. Le nombre des gènes dans le génome humain: les paris sont ouverts. *Med Sci* 2000; 16: 988-90.
11. Rowen L, Young J, Birditt B, et al. Analysis of the human neurexin genes: alternative splicing and the generation of protein diversity. *Genomics* 2002; 79: 587-97.

## NOUVELLE

### Mutations de la neurotrypsine et retard mental

Florence Molinari, Marlène Rio, Arnold Munnich, Laurence Colleaux

Inserm U.393,  
Hôpital Necker-Enfants  
Malades, 149, rue de Sèvres,  
75015 Paris, France.  
[colleaux@necker.fr](mailto:colleaux@necker.fr)

> Le terme retard mental regroupe différentes affections caractérisées selon l'OMS (Organisation mondiale de la santé) par un quotient intellectuel inférieur à 70 (QI < 70) [1]. Le retard mental est le handicap majeur le plus fréquent puisque sa prévalence est de 1 % si on ne considère que le retard mental sévère (QI < 50), et s'élève à 2-3 % si l'on inclut le retard mental léger (QI entre 50 et 70) [2, 3]. Malgré les récents progrès de la génétique médicale, dans 40 % des cas, la cause exacte du retard mental demeure cependant encore totalement inexplicée.

On distingue classiquement deux groupes de maladies: (1) les retards mentaux syndromiques, qui associent un retard mental et des anomalies neurologiques, morphologiques, viscérales ou biochimiques; et (2) les retards mentaux non syndromiques, définis par un retard mental isolé sans autres anomalies cliniques. Les gènes impliqués dans des retards mentaux syndromiques sont très divers et participent à de multiples processus cellulaires de sorte qu'il est souvent difficile de comprendre le lien exact entre le

défaut moléculaire et la déficience cognitive observée. À l'inverse, les gènes responsables de retards mentaux non syndromiques participent directement aux processus liés aux fonctions cognitives: mémorisation, apprentissage, comportement, etc.

Le nombre des gènes responsables de retards mentaux est estimé à environ 300 et seuls 90 d'entre eux ont été identifiés. En effet, l'extrême hétérogénéité clinique et génétique de ces maladies rend difficile l'utilisation des techniques classiques d'analyse de liaison génétique, en particulier dans le cas des retards mentaux non syndromiques. En conséquence, seuls onze gènes responsables de retards mentaux non syndromiques liés à l'X ont été identifiés [4-6], mais aucun gène associé à un retard mental isolé récessif autosomique (→). La gravité et la fréquence de ces affections font du démembrement des gènes responsables de ces maladies un des enjeux majeurs de la génétique médicale.

Afin d'identifier certains des gènes autosomiques, nous avons choisi d'utiliser une

méthode de localisation utilisant l'homozygotie par filiation. Cette stratégie, décrite en 1987 par E.S. Lander et D. Botstein, consiste à étudier des familles consanguines « multiplex » [7]. Dans de telles familles, tous les sujets atteints sont homozygotes par descendance non seulement pour la mutation morbide mais aussi pour les marqueurs avoisinants. D'autres régions sont également retrouvées homozygotes, mais sont différentes d'un individu atteint à l'autre. L'étude, par cette stratégie, d'une famille consanguine comprenant quatre enfants atteints de retard mental sévère et quatre enfants sains nous a permis de localiser le premier gène de retard mental autosomique récessif sur le bras long du chromosome 4. Parmi les nombreux gènes portés par cette partie du chromosome, nous nous sommes plus particulièrement intéressés au gène *PRSS12*.

Ce gène, très fortement exprimé dans le cerveau, code pour la neurotrypsine, un membre de la famille des protéases à sérine [8]. Des études menées chez la souris suggéraient que cette protéine puisse être impliquée dans

(→) m/s  
2000, n° 3,  
p. 363