Andrologie (1999), 9, n° 3, 342-346

Le projet " Génome Humain " et la caractérisation des étiquettes (E.S.T.) de testicule humain

G. GUELLAËN

Unité INSERM 99, Hôpital Henri Mondor 94010 Créteil

RÉSUMÉ

L'une des approches développées dans le cadre du projet " Génome Humain ", consiste à caractériser les ARNm d'un tissu donné par la séquence partielle des ADNc correspondants. Les séquences obtenues, ou " Expressed Sequence Tag " (E.S.T.), constituent une source d'information majeure, pour l'identification de nouveaux gènes et leur localisation chromosomique. A ce jour 1 200 000 E.S.T. ont été déposés dans la base de données "dbest " par la communauté scientifique, dont 50 000 ont été obtenus à partir de testicule humain. A partir de ces E.S.T., nous avons caractérisé de nouveaux gènes exprimés spécifiquement dans le testicule humain et notamment une nouvelle mono ADP ribosyl transférase ainsi que deux protéines liant les queues polyA des ARNm.

Mots Clés : *Testicule, projet Génome Humain, séquence d'ADN, base de données, gènes*

INTRODUCTION

Le programme " Génome Humain " a comme objectif de séquencer l'intégralité du génome afin d'en déterminer tous les gènes ainsi que leur localisation chromosomique. L'ampleur de la tâche est considérable. Au rythme actuel de la production des séquences, il faudrait encore 25 ans pour séquencer l'intégralité des 3 milliards de paires de base. Néanmoins des efforts importants vont être mis en oeuvre, tant sur le plan des techniques que sur le potentiel humain, afin de ramener cette échéance à l'an 2003....

Dès la mise en oeuvre de ce projet, des stratégies complémentaires à la séquence du génome ont vu le jour afin d'accélérer l'obtention des résultats. Le développement des banques génomiques en YAC et le clonage d'un grand nombre de microsatellites humains (séquence polyAC) ont permis l'établissement de cartes génétique et physique de grande précision aboutissant à la localisation de plus de la moitié des gènes humains . Simultanément l'analyse de l'expression des gènes a été abordée de manière originale par Sydney Brenner. La stratégie proposée comprend trois phases :

i) le séquençage des extrémités 3' ou 5' d'ADNc correspondant à des ARNm d'un tissu donné ;

ii) la cartographie physique de ces séquences ou Expressed Sequence Tags (E.S.T.) ;

iii) à plus long terme l'identification des protéines et des gènes associés à ces E.S.T.

La première production d'E.S.T. a été obtenue à partir du cerveau humain par le groupe de Craig Venter , suivi par les productions d'autres équipes travaillant sur différents tissus humains ou d'autres espèces. Très rapidement, les retombées financières potentielles d'une telle production ont favorisé le développement de grands centres de séquençage comme The Institute of Genomic Research

(TIGR), Incyte et l'Université de Washington en collaboration avec la société Merck. Si une partie des E.S.T. produits par ces centres (TIGR, Incyte) est d'accès limité pour la communauté scientifique, un grand nombre sont déposés dans la base de données de séquence " dbest " accessible à l'ensemble de la communauté scientifique .

A la fin de l'année 1998, cette base de données contient plus de 2 millions d'E.S.T. dont 1.,2 million produit chez l'homme à partir de plus de 300 banques d'ADN complémentaire différentes. Sur ce nombre, un peu plus de 50 000 E.S.T. ont été produits à partir d'ARNm de testicule humain. Ces séquences présentent une source d'information importante pour les projets axés sur l'étude de l'expression des gènes dans le testicule.

Ces E.S.T. peuvent faire l'objet de recherches spécifiques permettant de cibler une population de gènes. Ainsi l'interrogation de la base de données " dbest " soit directement (http://www.ncbi.nlm.nih.gov/dbEST/), par l'intermédiaire du système Entrez du NCBI (http://www.ncbi.nlm.nih.gov/Entrez/) ou le système SRS (http://www.infobiogen.fr/srs/) permet de sélectionner plus précisément des d'E.S.T. d'intérêt en se basant sur les annotations qui leur sont associées (tissu d'origine, localisation chromosomique, homologie de séquence, etc.). Ces E.S.T. peuvent également être comparés avec des séquences de gènes connus dans d'autres espèces dont on recherche des homologues humains. Les clones correspondant à ces E.S.T. peuvent être obtenus auprès du consortium IMAGE (http://www-bio.llnl.gov/bbrp/image/iresources.html) afin de faire l'étude de nouveaux gènes.

Dans notre laboratoire nous avons mis en œuvre une telle stratégie de production d'E.S.T. pour l'identification de nouveaux gènes exprimés dans le testicule humain. Nous avons préparé une banque d'ADNc de testicule humain de 200 000 clones indépendants à partir d'ARNm de testicules d'un homme de 27 ans. Après élimination des clones correspondant aux ARNm les plus abondants nous avons purifié 8000 clones et préparé 3300 plasmides ayant un insert de taille supérieure à 700 paires de bases. Nous avons sélectionné 2200 plasmides sur lesquels nous avons effectué 2653 séquences. Deux mille d'entre elles ont été comparées aux bases de données et réparties en quatre groupes (figure 1). Une partie de ces séquences a été déposée dans la base "dbest" (n° accès T18854-T19462) et a donné lieu à une publication .

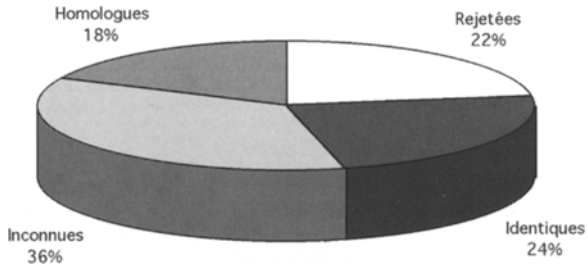

Figure 1 : Répartition de 2000 E.S.T. de testicule humain en fonction de leur comparaison avec les gènes présents dans les bases de données. Les séquences sont identiques à des gènes connus si elles présentent plus de 98 % d'identité avec des gènes connus, elles sont homologues à des gènes connus si elles présentent entre 40 et 98 % d'identité et inconnues en dessous de 40 %. Les séquences rejetées sont soit redondantes ou contiennent des séquences répétées.

Nous avons ciblé la suite de notre travail sur la caractérisation de clones homologues à des gènes connus. Ce groupe de clones représente la population la plus intéressante à court terme, surtout ceux qui ont des homologies avec des gènes exprimés dans des espèces éloignées (tableau 1). En effet, dans la plupart des cas, les faibles pourcentages d'homologie détectés par comparaison de séquence indiquent qu'il n'aurait pas été possible de cloner ces gènes chez l'homme par simple hybridation croisée avec une sonde correspondant au gène homologue. Pour la suite de notre travail, nous avons préférentiellement ciblé des gènes susceptibles d'intérêt pour la physiologie du testicule.

La démarche que nous avons adoptée comprend :

i) Une analyse de la littérature. Les informa-

Tableau 1 : Homologies de séquence entre des E.S.T. de testicule et des gènes d'espèces inférieures présents dans les bases de données (% id : pourcentage d'identité ; % sim : pourcentage de similitude) (voir pour plus de détails)

Gènes homologues	**%id**	**%sim**	**Espèces**
Serine protease	48	57	A. Salmonicida
P-glycoprotein	54	68	A. Thaliana
Ubiquitin activating enzyme E1	45	75	A. Thaliana
Ubiquitin conjugating enzyme E2	36	50	A. Thaliana
Serine acetyltransferase	40	64	B. Subtilis
Hyptothetical 64.2 KD protein	37	58	C. Elegans
Early embryogenesis protein Zyg 11	42	62	C. Elegans
Diff6 protein homolog	89	93	Drosophila
DEAD box protein	52	63	Drosophila
Goliath protein	32	50	Drosophila
Furin-like protease 2 precursor	52	56	Drosophila
Polynucleotide phosphorylase	37	53	E. Coli
Cysteine aminopeptidase	57	69	L. Lactis
ATPase	28	48	P. Falciparum
Tropomyosin	33	73	S. Mansoni
Mer operon ORF2 hypothetical protein	36	59	S. Marescens
Double-strand-break repair protein	65	81	S. Pombe
RNA polymerase II (sub 5)	59	78	Glycine max (Soja)
ATPase	44	52	T. Brucei
Chromosome segregation protein	39	64	S. Cerevisiae
Cell division control protein	43	66	S. Cerevisiae
Cell division cycle protein	54	72	S. Cerevisiae
Pré mRNA splicing factor PRP8	75	87	S. Cerevisiae
Translocation protein SEC62	30	61	S. Cerevisiae
UDP glucose 4 epimerase	62	83	S. Cerevisiae
NAM7 protein	62	76	S. Cerevisiae
Ubiquitin conjugating enzyme E2	68	78	S. Cerevisiae

tions recueillies sur les séquences présentes dans les bases de données (nom, numéro d'accès) ne sont pas suffisamment informatives. Ceci d'autant plus que certains noms de séquence, pour attractifs qu'ils soient, ne sont pas reliés directement à une fonction (i.e. Enigma, Goliath, etc....). Pour chacun des clones homologues nous avons donc effectué une recherche bibliographique en nous servant de bases de données comme Medline, Entrez, SRS, etc....

ii) Une étude du niveau et de la spécificité d'expression. L'expression des ARNm correspondant à certains des clones sélectionnés a été analysée par Northern blots sur de l'ARNm de

testicule humain, de souris ainsi que sur d'autres organes. Cela nous a permis de déterminer : la taille, l'abondance et la spécificité tissulaire des ARNm correspondants. La détection d'une hybridation croisée avec des messagers de souris, laisse entrevoir la possibilité de développer et d'utiliser des modèles animaux.

iii) La séquence complète du clone. Les clones sélectionnés sont séquencés complètement afin d'identifier la phase ouverte de lecture de la protéine correspondante. Ceci est évalué par comparaison des séquences obtenues avec la séquence de l'ARNm identifié initialement dans les autres espèces.

iv) La spécificité cellulaire dans le testicule. Le testicule comprend quatre grands types cellulaires (Leydig, Sertoli, péritubulaires et germinales). Le lieu d'expression des transcrits est identifié par hybridation *in situ* sur des coupes de testicules humains obtenues après cryocongélation ou inclusion du tissu en paraffine.

v) L'expression des protéines et préparation d'anticorps. Une fois que la phase ouverte de lecture est déterminée, la protéine est produite *in vitro* afin d'obtenir des anticorps et de tester éventuellement son activité potentielle. L'obtention de ces anticorps est très importante. Ils démontrent l'existence de la protéine par Western blot et confirment sa taille. Deuxièmement, ils permettent de localiser exactement l'antigène dans le testicule humain. Au cours de la spermatogenèse, certains transcrits sont traduits à un stade cellulaire différent de celui où ils ont été synthétisés.

Actuellement ces différentes étapes sont en cours sur plusieurs protéines. Ils nous ont permis d'identifier une nouvelle mono ADP ribosyl transférase absolument spécifique du testicule humain (figure 2). Plus récemment nous avons pu mettre en évidence deux nouvelles formes de protéines liant la queue poly A des ARNm. Ces protéines pourraient contribuer de manière déterminante à la stabilité et à la traductibilité des ARNm testiculaires. Leurs caractérisations sont en cours.

En conclusion, l'analyse des E.S.T. offre un champ d'investigation très utile pour l'identification et la caractérisation de nouveaux gènes

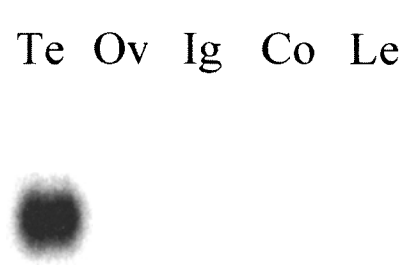

Figure 2 : Northern blot de la mono ADP ribosyl transférase htMART sur de l'ARNm humain de Rate (Ra), de Thymus (Th), de Prostate (Pr), de Testicule (Te), d'Ovaire (Ov), d'Intestin grêle (Ig), de Colon (Co) et de Leucocyte (Le).

dans le testicule humain. Cette caractérisation est une nécessité car quelle que soit la puissance de la séquence génomique et l'identification des gènes qui en découle, seule la caractérisation des ARNm correspondants démontre leur transcription. Bien sûr il ne faut pas considérer les E.S.T. comme une fin en soit, il est indispensable ensuite de " redescendre " à la protéine pour démontrer le rôle physiologique des nouveaux gènes ainsi identifiés.

RÉFÉRENCES

1. ADAMS M. D., DUBNICK M., KERLAVAGE A. R., et al. Sequence identification of 2,375 human brain genes [see comments]. Nature, 1992, 355 : 632-4.
2. BOGUSKI M. S., TOLSTOSHEV C. M., and BASSETT D. E., Jr. Gene discovery in dbEST [letter]. Science, 1994, 265 : 1993-4.
3. COHEN D., CHUMAKOV I., WEISSENBACH J. A first-generation physical map of the human genome. Nature, 1993, 366 : 698-701.
4. COLLINS F. S., PATRINOS A., JORDAN E., et al. New goals for the U.S. Human Genome Project: 1998-2003. Science, 1998, 282 : 682-9.
5. DELOUKAS P., SCHULER G. D., GYAPAY G., et al. A physical map of 30,000 human genes. Science, 1998, 282 : 744-6.
6. LEVY I., WU Y. Q., ROECKEL N., et al. Human testis specifically expresses a homologue of the rodent T lymphocytes RT6 mRNA. FEBS Lett, 1996, 382 : 276-80.
7. PAWLAK A., TOUSSAINT C., LEVY I., et al. Characterization of a large population of mRNAs from human testis. Genomics, 1995, 26 : 151-8.

8. WEISSENBACH J., GYAPAY G., DIB C., et al. A second-generation linkage map of the human genome [see comments]. Nature, 1992, 359 : 794-801.

ABSTRACT

Sequencing of the human genome and characterization of expressed sequence Tags (E.S.T.) in human testis

G.GUELLAEN

The " Human Genome " is devoted to the identification of all the human genes and their chromosomal localization. Besides the sequencing of the genome, complementary strategies have been developed, including the characterization of mRNA by single pass sequencing of the corresponding cDNA. The partial sequences generated during this process represent expressed sequence tags (E.S.T.) of the corresponding genes. We used this strategy, associated with Northern blot and *in situ* hybridization, in order to characterize genes expressed in human testis. A cDNA library was prepared using mRNA purified from testes of a 27 years old man. We isolated, and stored in glycerol 7750 recombinant clones. 2200 clones with an insert over 700bp were single-pass sequenced, mainly at the 5' end, generating 2563 sequences products which were compared, with the sequences present in public databases. Out of 2000 sequences, 481 were found identical to known human sequences; 358 were homologous to sequences from human or other species; 719 represented new sequences; 442 were rejected. Some of the clones corresponding to mRNAs transcribed from new genes were further analyzed on Northern blot of human, rat and mouse testes mRNA along with RNA of kidney, liver and brain. This allowed us to identify several new genes expressed in human testis of potential interest for the physiology of this tissue. Among others, they include a new mono-ADP ribosyl transferase and two testis specific poly A binding proteins.

Key words : *Testis, human genome project, DNA sequencing, databases, genes*