m/s DOSSIER TECHNIQUE

médecine/sciences 1995 ; 11 : 1707-16

De l'ADN aux gènes : un éventail de techniques

Anouk Courseaux, Michel Fontès, Pierre Szepetowski

Différentes méthodes permettent aujourd'hui d'isoler spécifiquement des gènes à partir de grandes régions d'ADN génomique. Leur principe repose, soit sur l'identification d'ADN complémentaire correspondant aux séquences transcrites à partir de l'ADN génomique cible, soit sur la reconnaissance d'éléments fonctionnels impliqués dans l'expression des gènes. Les progrès réalisés dans ce domaine ont déjà permis d'isoler des gènes associés à des maladies tant monogéniques que multifactorielles (cancers familiaux du sein et du côlon, obésité, maladie de Huntington, syndrome de Menkès, ataxie-télangiectasie, etc.).

Depuis quelques années, de nombreux gènes impliqués dans des maladies héréditaires tant monogéniques (mucoviscidose, chorée de Huntington) que multifactorielles (obésité, cancer du sein) ont pu être identifiés grâce aux progrès réalisés dans le domaine du « clonage positionnel ». Le principe de cette méthode repose, dans un premier temps, sur la mise en évidence d'une liaison génétique entre le phénotype d'une maladie donnée et des marqueurs génétiques dont la localisation chromosomique est connue, permettant ainsi de circonscrire la région dans laquelle doit être recherché le(s) gène(s) candidat(s). Les efforts de recherche investis dans la construction de cartes génétiques détaillées du génome humain ont permis, aujourd'hui, d'atteindre une résolution d'environ un centimorgan (1 cM correspond environ à 1 000 kilobases (kb)) [1, 2]. La localisation d'un *locus* morbide peut également être obtenue par l'identification de remaniements chromosomiques (perte d'hétérozygotie, translocation, amplification...). L'étape suivante consiste en l'obtention de clones d'ADN génomique représentatifs de la région d'intérêt, travail grandement facilité par l'existence de banques ordonnées de cosmides ou mieux de chromosomes artificiels de levure (YAC) couvrant tout ou partie du génome. La recherche de gènes est alors effectuée à partir de l'ensemble de clones chevauchants *(contig)* s'étendant généralement sur plusieurs centaines de kilobases. Les domaines exprimés ou exons ne représentant que 1 % à 5 % de l'ADN génomique total, tout le problème est d'isoler spécifiquement ces séquences géniques.

Il y a encore quelques années, on ne disposait que de moyens peu sensibles. La présence de séquences transcrites était généralement recherchée directement par hybridation de petits fragments d'ADN génomique sur *Northern blot* ou par la mise en évidence d'îlots CpG, séquences situées au voisinage de nombreux gènes. Une autre stratégie consistait à rechercher les exons sur la base de la conservation de leur séquence entre différentes espèces. Ainsi le gène de la mucoviscidose *(CFTR) (m/s n° 8, vol. 5, p. 589)* [3] mais aussi tout récemment le gène de l'hypoplasie surrénale congénitale liée à l'X *(m/s n° 4, vol. 11, p. 634) (DAX1)* [4] ont-ils pu être identifiés. Cependant, ces méthodes, lourdes et fastidieuses,

TIRÉS À PART

P. Szepetowski.

nécessitent notamment le sous-clonage systématique de grandes régions d'ADN en petits fragments. De telles difficultés ont conduit au récent développement de nouvelles techniques de recherche de gènes adaptées à l'étude de grandes portions du génome. Ces nouvelles méthodes peuvent être classées en deux grands groupes: celles permettant l'identification d'ADN complémentaires (ADNc) correspondant aux ARN messagers (ARNm) issus de la région d'intérêt, et celles fondées sur la reconnaissance d'éléments fonctionnels impliqués dans l'expression (exons, sites de polyadénylation, régions régulatrices...) *(figure 1)*.

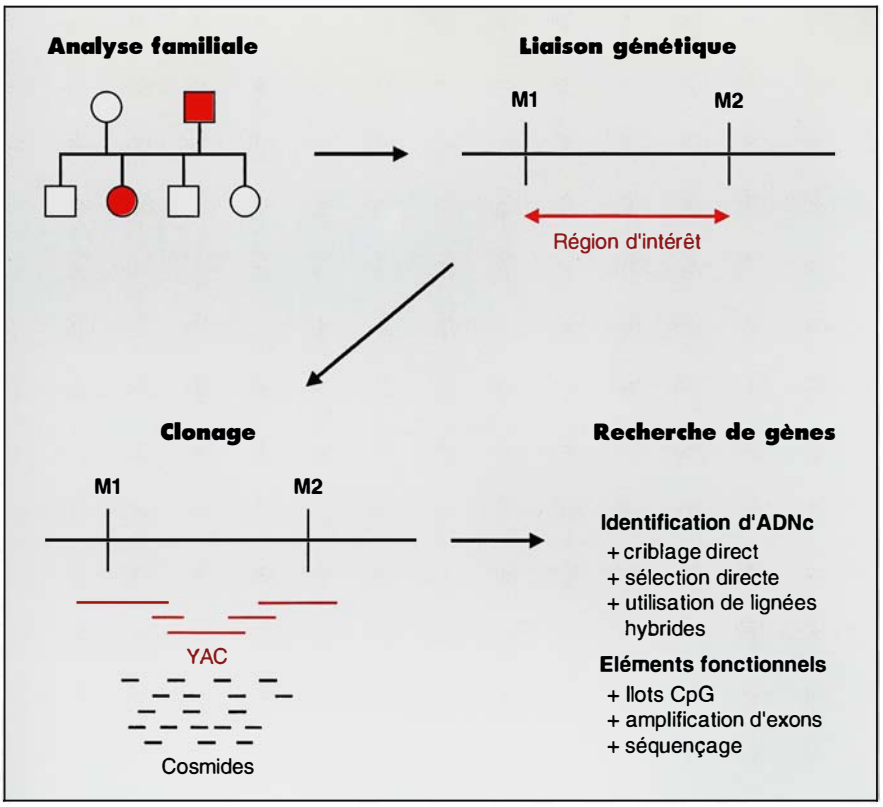

Figure 1. ***Principe du clonage positionnel.***

Le criblage direct

Les collections de phages ou de cosmides représentatives de régions chromosomiques définies, obtenues par différentes techniques (chromosome isolé par cytofluorométrie de flux, microdissection, hybrides somatiques, *contigs* de YAC...), sont de plus en plus fréquemment ordonnées sur filtres à haute densité, facilitant ainsi leur manipulation. La présence de séquences transcrites sur chaque clone peut être décelée directement par hybridation d'ADNc marqué [5], les clones positifs servant alors de support à la recherche d'exons *(figure 2)*. Des criblages successifs avec plusieurs types d'ADNc (origines tissulaires distinctes, différents stades de développement d'un organe) permettent l'obtention de cartes transcriptionnelles et différentielles de la région génomique considérée.

A l'inverse, il est possible d'utiliser de grands fragments d'ADN génomique marqués (YAC par exemple) pour cribler des banques d'ADNc. Une telle approche présente deux inconvénients majeurs: les fragments génomiques correspondant à des séquences exprimées ne représentent qu'une faible proportion de l'ADN génomique total utilisé comme sonde; par ailleurs, la grande densité en séquences répétées, qu'il n'est pas toujours aisé d'éliminer, contribue à la détection de nombreux faux-positifs. Cette stratégie s'est néanmoins avérée payante pour isoler le gène de la neurofibromatose de type I *(NF1)*, en utilisant l'ADN d'un YAC situé dans la région d'intérêt et couvrant un point de translocation chez un sujet atteint [6], ainsi que pour cloner *APC*, un des gènes de susceptibilité aux cancers colorectaux, à partir de YAC représentatifs de la région 5q21 [7, 8].

Néanmoins, et c'est là leur principal inconvénient, ces techniques d'hybridation demeurent peu sensibles, puisque seuls les gènes fortement exprimés seront détectés.

La sélection directe d'ADNc

C'est pourquoi fut proposée, il y a quelques années, une méthode permettant un enrichissement en ADNc correspondant aux gènes présents sur les YAC d'intérêt. De l'ADN génomique (YAC purifiés sur gel, par exemple) est hybridé avec le produit d'amplification par PCR d'une banque d'ADNc, après épuisement des séquences répétées et ribosomiques. L'ADN cible est immobilisé sur filtre avant hybridation [9, 10], ou sur billes magnétiques après hybridation [11] *(figure 3)*. Les ADNc appariés spécifiquement aux exons correspondants sur le(s) YAC sont alors élués et les produits d'élution amplifiés par une nouvelle PCR (oligonucléotides plus internes). Un second tour de sélection, réalisé dans les mêmes conditions, aboutit finalement à l'obtention d'une minibanque d'ADNc enrichie en séquences géniques présentes sur le(s) YAC de départ. L'intérêt de cette méthode est parfaitement illustré par le tout récent clonage du gène *BRCA1* responsable de la prédisposition à certaines formes familiales de

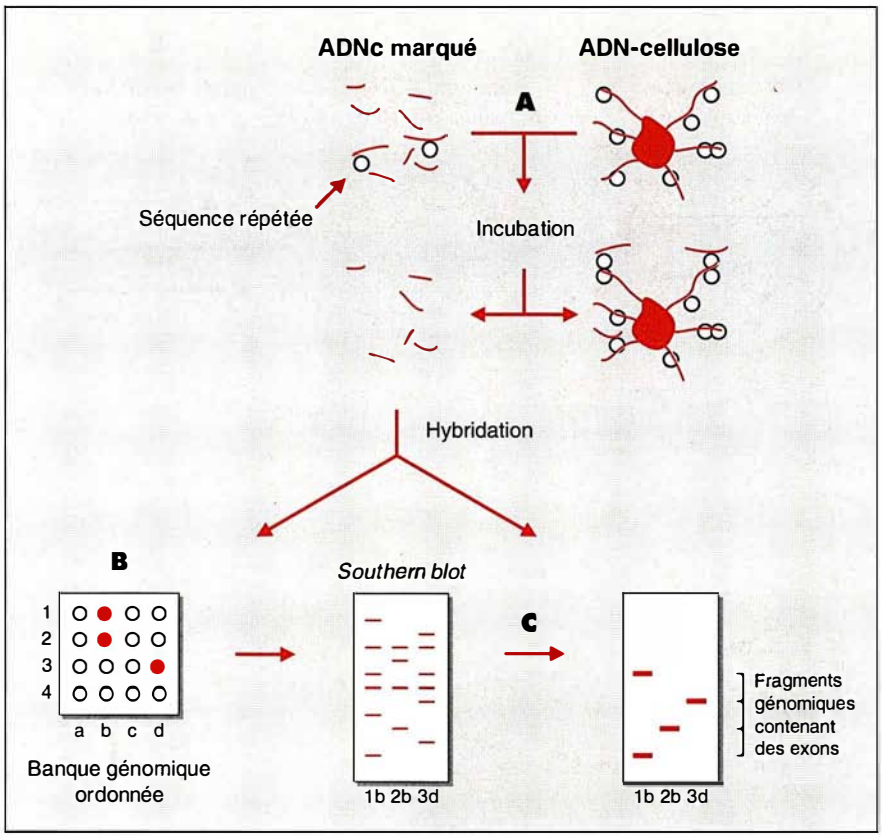

Figure 2. ***Criblage direct d'une banque génomique ordonnée. A.*** *Une sonde d'ADNc simple brin, synthétisée par transcription inverse d'ARN poly(A⁺), est incubée en présence d'ADN génomique de haut poids moléculaire couplé à de la cellulose diazotée (ADN-cellulose) pour éliminer les séquences fortement répétées.* ***B.*** *L'hybridation des filtres de la banque génomique ordonnée permet l'identification de clones positifs.* ***C.*** *Après digestion enzymatique des clones sélectionnés, les fragments de restriction contenant des exons sont détectés par hybridation de* Southern blots *avec la sonde d'ADNc précédemment utilisée.*

cancers du sein et de l'ovaire [12]. Un autre exemple est fourni par des travaux portant sur un ensemble de YAC chevauchants, construit dans le but d'isoler des gènes de la bande q13.3 du chromosome X humain [13, 14] ; soixante-dix pour cent des clones sélectionnés proviennent effectivement du *contig* utilisé, soit une densité d'une séquence transcrite tous les 180 kb. Parmi les cinq gènes finalement obtenus, l'un est impliqué dans le syndrome de Menkès, alors qu'un autre, *XNP*, vient d'être associé au syndrome de retard mental avec α-thalassémie [15].

L'enrichissement en ADNc est généralement de l'ordre de plusieurs milliers de fois (jusqu'à 100 000 fois), rendant ainsi possible la détection de transcrits présents au départ à moins d'une molécule pour un million. Qui plus est, la technique de sélection directe possède intrinsèquement un pouvoir de normalisation : les transcrits faiblement représentés dans la banque de départ sont enrichis à un taux supérieur à celui des transcrits initialement abondants [13]. Bien entendu, un gène ne pourra être isolé que s'il est représenté (ne serait-ce que très faiblement) dans la banque d'ADNc utilisée. Ce type de problème peut être contourné grâce à l'utilisation d'un mélange de banques d'ADNc d'origines tissulaires diverses et/ou correspondant à différents stades du développement embryonnaire [16]. Cette technique est applicable à des régions chromosomiques étendues, bien que, comme on pouvait s'y attendre, l'enrichissement décroisse avec l'augmentation de la taille de l'ADN génomique utilisé : une expérience de sélection directe réalisée à partir de la microdissection d'un amplicon localisé en 12q13 a permis d'obtenir des taux d'enrichissement variant selon les ADNc de 120 à 860 [17].

Le taux de clones correspondant effectivement à des gènes présents dans la région génomique d'intérêt dépend, bien évidemment, de la qualité des banques d'ADNc utilisées. En effet, la présence de clones « contaminants » (vecteur non recombinant, ARN ribosomique, ADN génomique copié par la transcriptase inverse) peut diminuer grandement le rendement de la sélection directe. Dans un autre ordre d'idées, il n'est pas rare d'isoler des ADNc qui, quoique n'étant ni ribosomiques, ni pourvus de séquences répétées, ne correspondent pas à un gène présent sur l'ADN génomique mais à un pseudogène (le gène pouvant être localisé

dans une autre région chromosomique) ou à un membre d'une même famille génique. Généralement, ce type de faux-positifs n'est éliminé qu'avec difficulté après analyse complète des ADNc sélectionnés.

Utilisation de lignées hybrides

Des techniques, reposant à la fois sur la reconnaissance d'éléments fonctionnels et sur l'identification d'ADNc, permettent d'isoler sélectivement les gènes d'une région génomique humaine donnée contenue dans un hybride cellulaire interspécifique. Après extraction des ARN totaux, la synthèse d'ADNc est amorcée, grâce à des oligonucléotides spécifiques des séquences consensus d'épissage, à partir des ARN messagers immatures non épissés (hnARN) [18] ; l'enrichissement en clones d'origine humaine est obtenu par criblage de la banque d'ADNc à l'aide de séquences répétées humaines généralement présentes dans les introns. La synthèse des ADNc humains peut également être amorcée en utilisant des oligonucléotides spécifiques des séquences répétées humaines *Alu* [19]. Toutefois, un gène ne pourra être isolé que si son expression est compatible avec le type cellulaire de l'hybride somatique, et ce à un niveau suffisant pour être représenté dans la banque d'ADNc, ce qui limite bien évidemment le champ d'application de ce type de méthode.

Les îlots CpG

La recherche de gènes par la localisation d'îlots CpG (encore appelés îlots HTF) a déjà fait l'objet d'une revue dans ce même journal [20]. Rappelons que ces séquences hypométhylées sont localisées au voisinage immédiat d'environ 60 % des gènes. De nouvelles techniques, permettant le clonage direct des séquences adjacentes aux îlots CpG, ont permis d'obtenir des résultats prometteurs. Cross *et al.* [21] décrivent une méthode *(figure 4)* consistant à fixer sur une matrice mixte d'agarose et de nickel la protéine de rat MeCP2. Cette protéine a la propriété de se lier aux fragments d'ADN fortement méthylés. Dans un premier temps, les sé-

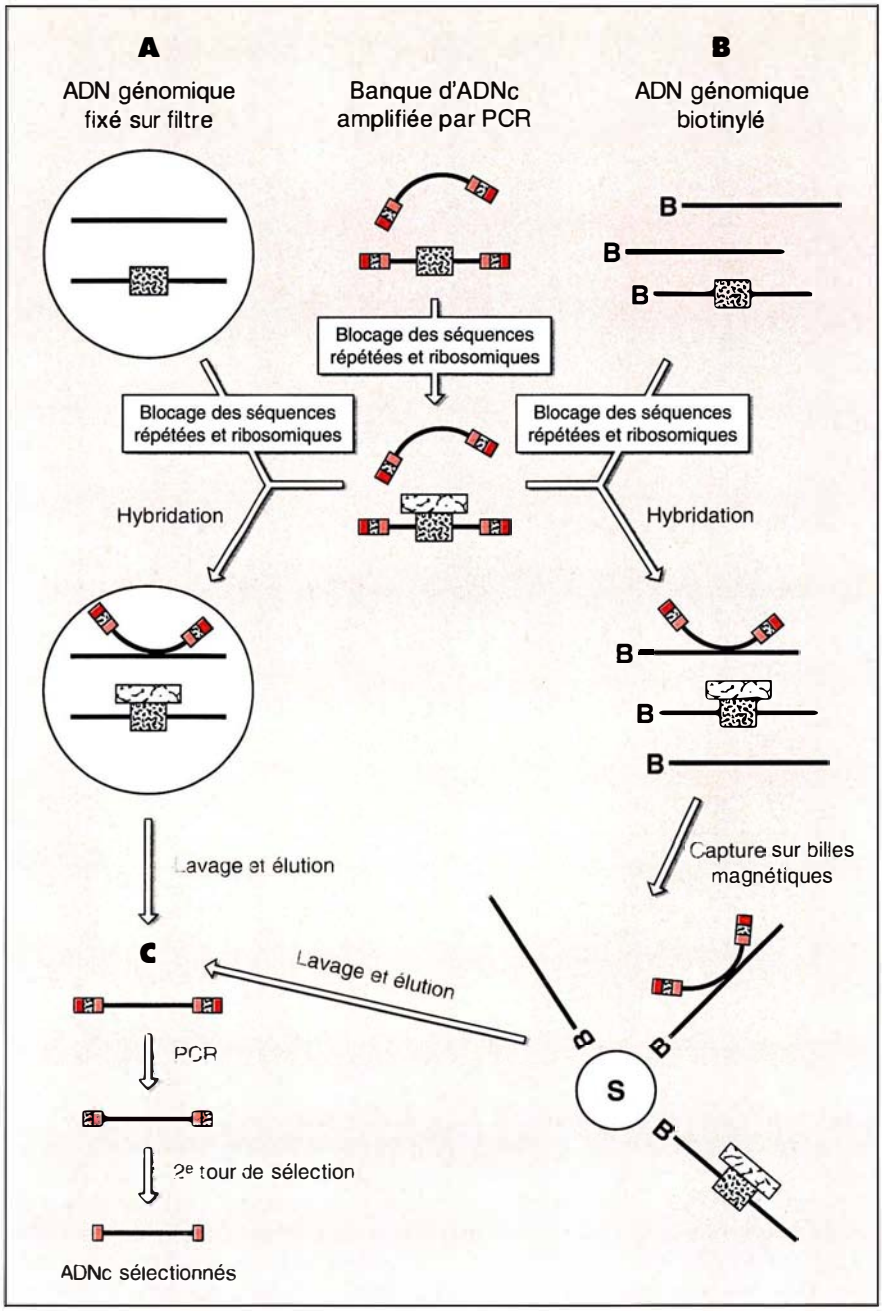

Figure 3. ***Illustration des techniques de sélection directe d'ADNc. A.*** *L'ADN génomique purifié et dénaturé est immobilisé sur un filtre de nylon, puis hybridé avec les produits d'amplification PCR de la banque d'ADNc.* ***B.*** *L'ADN génomique biotinylé (B) est hybridé en milieu liquide avec la banque d'ADNc amplifiée par PCR, puis capturé par des billes magnétiques recouvertes de streptavidine (S).* ***C.*** *Les ADNc sélectionnés après élution sont amplifiés par PCR, et subissent un second tour de sélection. Les séquences répétées à bloquer sont représentées par un rectangle x et les séquences bloquantes par un rectangle y.*

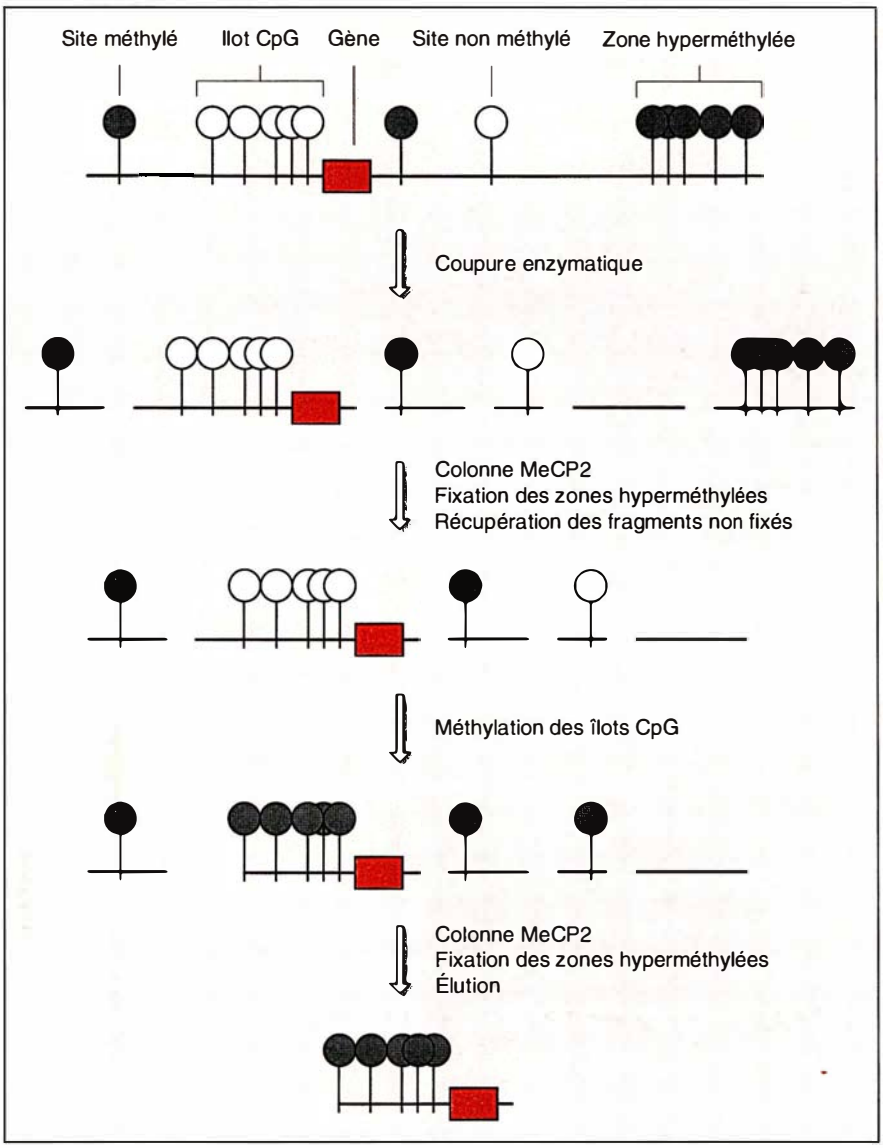

Figure 4. ***Purification de fragments contenant des îlots CpG par colonne d'affinité.***

quences génomiques naturellement très méthylées sont éliminées par adsorption spécifique sur la matrice. L'étape suivante consiste en la méthylation spécifique *in vitro*, à l'aide d'une méthyl-transférase bactérienne, des îlots CpG. Les fragments contenant ces séquences artificiellement hyperméthylées, contrairement au reste de l'ADN génomique, peuvent alors se fixer sur une nouvelle matrice. Suit alors une phase d'élution à forte concentration saline permettant de récupérer une collection de fragments enrichie en îlots CpG et donc potentiellement en séquences exoniques. A partir d'ADN humain total, les auteurs montrent que parmi vingt clones analysés, cinq correspondent effectivement à des transcrits présents dans le tissu lymphoblastoïde.

Une autre méthode, décrite par Valdes *et al.* [22], repose sur le principe du système «vectorette» traditionnellement utilisé pour isoler les extrémités des YAC. Il s'agit ici de cliver spécifiquement l'ADN d'un YAC à l'aide d'une enzyme dépendante de la méthylation, reconnaissant les îlots CpG. La ligation de cassettes aux extrémités des fragments de restriction ainsi produits permet, ensuite, l'amplification par PCR de séquences d'ADN situées entre, d'une part, un oligonucléotide spécifique de la cassette et, d'autre part, un oligonucléotide spécifique des séquences répétées humaines *Alu*. Bien entendu, le succès de l'opération dépend de la distance séparant l'îlot CpG d'une séquence *Alu*; néanmoins, cette restriction doit être relativisée du fait de l'avènement de la « PCR longue distance ».

Ces résultats, certes encourageants, demandent cependant à être confirmés par des études ultérieures. Quoi qu'il en soit, les méthodes de clonage des îlots CpG présentent un avantage commun, à savoir que la détection des gènes est indépendante de leur spectre d'expression tissulaire. En revanche, seuls 60 % des gènes sont potentiellement accessibles. De plus, les régions 5' terminales de nombreux gènes sont sous-représentées*, voire non représentées, dans les banques d'ADNc. En outre, de par leur richesse en CG, les sondes dérivées de tels clones donnent souvent des signaux parasites (notamment sur les contaminants ARNr), entraînant la détection de nombreux faux-positifs.

L'amplification d'exons

Chez les mammifères, les gènes sont généralement constitués d'exons et d'introns; lors de la maturation des

** Le fragment génomique contenant des îlots CpG est censé contenir également une partie (le plus souvent la partie 5') de la séquence transcrite du gène potentiellement adjacent. Ce fragment est utilisé pour cribler une banque d'ADNc. Les banques d'ADNc sont souvent sous-représentées pour les parties 5' des gènes, d'où le risque de cribler cette banque sans détecter la séquence d'ADNc correspondante.*

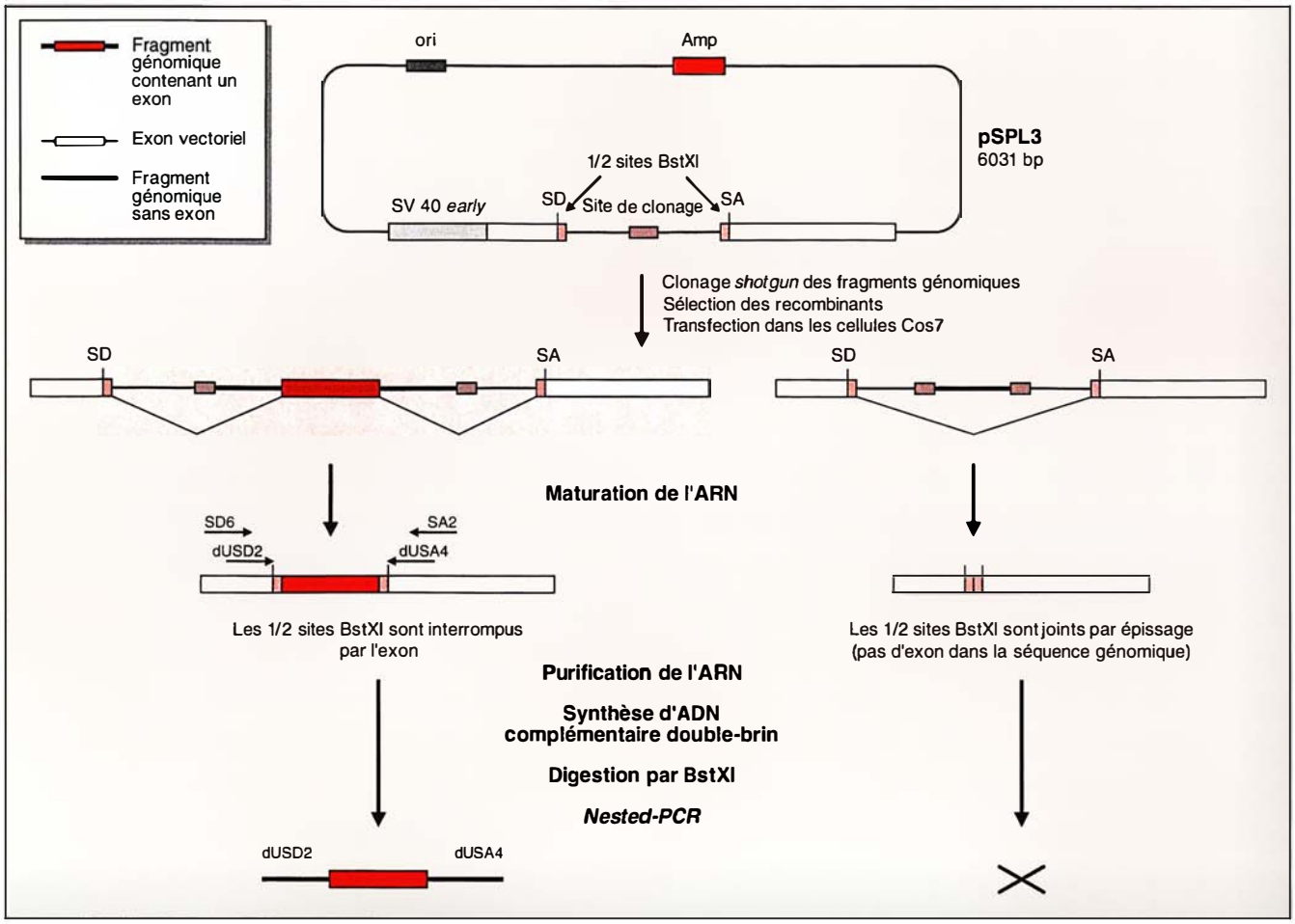

Figure 5. ***Représentation schématique de la méthode d'amplification des exons internes.*** *L'ADN génomique est purifié après électrophorèse en champs pulsés, puis digéré et sous-cloné de façon au hasard* (shotgun) *dans le site de clonage multiple localisé dans l'intron du vecteur de « capture » (pSPL3). Les recombinants sélectionnés sont groupés et transfectés dans des cellules Cos7. Les ARN totaux, produits grâce à la présence du promoteur des gènes à expression précoce de SV40, sont récupérés 24 heures plus tard. Des demi-sites BstXI sont présents de chaque côté de l'intron flanquant les sites donneur (SD) et accepteur (SA) de pSPL3. Une digestion BstXI permet d'éliminer les faux-positifs, c'est-à-dire les molécules ne contenant pas d'exon interne. Un premier brin d'ADN complémentaire (ADNc) est synthétisé par transcription inverse, puis amplifié par 2 tours de PCR successifs en utilisant des oligonucléotides spécifiques du vecteur (SD6, SA2 et dUSD2, dUSA4). On parle de nested PCR car les deux couples successifs d'amorces nucléotidiques sont emboîtés.*

ARN messagers, les mécanismes d'épissage cellulaires induisent la jonction des exons et l'élimination des introns. La méthode d'amplification d'exons, appelée également *exon-trapping*, est fondée sur la reconnaissance de sites d'épissage fonctionnels et permet d'isoler des exons directement à partir d'ADN génomique. Les améliorations successives apportées à la technique originelle [23] ont abouti à la génération de deux méthodologies qui diffèrent par la séquence (cible) génomique reconnue.

La première technique permet de « capturer » sélectivement des exons internes complets [24], ceux-ci étant encadrés par des sites accepteur et donneur d'épissage *(figure 5)*. L'étape initiale consiste à sous-cloner des fragments d'ADN génomique dans un plasmide de « capture » ; les plasmides recombinants sont alors sélectionnés dans *E. coli*. Le vecteur le plus communément utilisé, pSPL3, possède une cassette contenant un promoteur transcriptionnel, ainsi que 2 exons artificiels (un exon 5' avec un site donneur et un exon 3' avec un site accepteur d'épissage) encadrant le site de clonage. La présence d'un exon interne complet au sein d'un fragment génomique sous-cloné permet la formation d'un gène chimérique constitué de cet exon et des deux exons vectoriels, ceux-ci jouant les rôles d'exon initial et terminal. L'unité de transcription ainsi

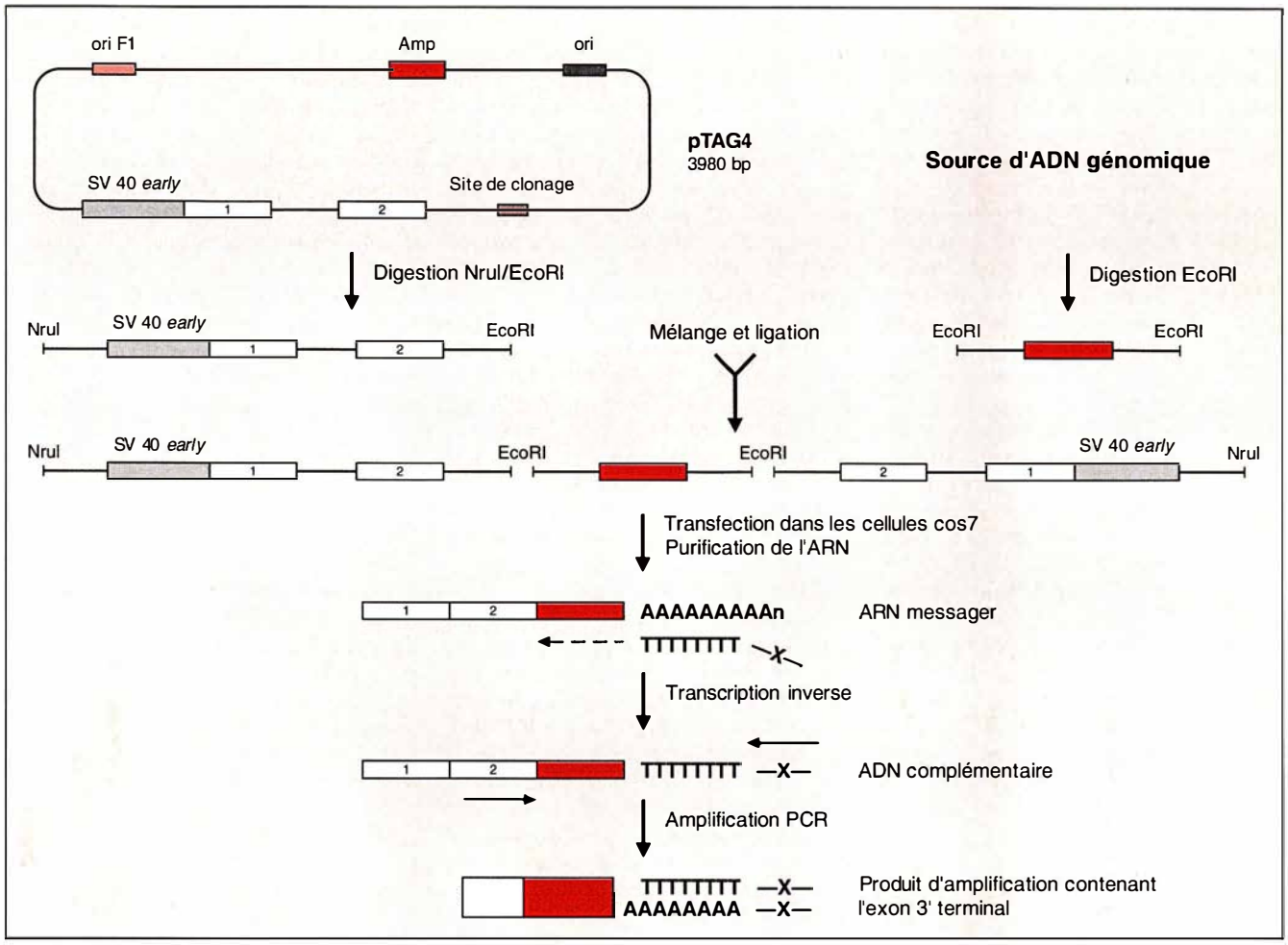

Figure 6. ***Représentation schématique de la méthode d'amplification des exons terminaux.*** *Des fragments de restriction d'ADN génomique sont sous-clonés dans le vecteur pTAG4 digéré par NruI et la même enzyme que celle utilisée pour la digestion de l'ADN cible. La double digestion de pTAG4 permet de laisser une extrémité franche en amont du promoteur des gènes précoces de SV40 et une extrémité cohésive en aval du second exon. Le site NruI est très difficile à relier, tandis que l'extrémité cohésive se lie facilement à la même extrémité cohésive, ce qui permet de réaliser des concatémères d'ADN cible flanqué de chaque côté par pTAG4. Le produit de ligation est directement transfecté dans les cellules Cos7 pour permettre la transcription du concatémère linéaire. Un premier brin d'ADNc est synthétisé par transcription réverse grâce un oligonucléotide poly(dT), puis amplifié par PCR.*

formée conduit à la production d'un ARNm hybride lorsque transfectée dans des cellules mammifères compatibles avec le système de transcription du vecteur. L'exon interne ainsi capturé est amplifié sélectivement par RT/PCR à partir de l'ARN total des cellules transfectées, les amorces utilisées étant spécifiques des deux exons du vecteur.

Cette méthode présente néanmoins quelques imperfections, l'une d'elles étant que la petite taille des exons internes capturés (20-200 bases) rend délicate leur utilisation directe pour des expériences d'hybridation ultérieures; une autre étant le «risque» d'isoler indépendamment plusieurs exons d'un même gène. Pour remédier à ces problèmes, Krizman et Berget [25] ont développé un autre type de vecteur, pTAG4, permettant non plus la capture des exons internes mais celle de l'exon terminal d'un gène. La cassette de «capture» est, ici, une unité de transcription incomplète dont le dernier exon est absent *(figure 6)*. Les fragments d'ADN génomique sont censés être donateurs d'un exon terminal (flanqué d'un site accepteur d'épissage et d'un signal poly(A)) pour permettre au vecteur d'engendrer un ARNm hybride stable après transfection dans des cellules cos 7. La majorité des gènes des vertébrés ne contient qu'un seul exon terminal, ce qui permet de n'obtenir qu'un seul type de clone par gène. Cette méthode présente,

en outre, les avantages d'éviter l'étape de sélection des plasmides de «capture» recombinants, et de permettre l'isolement des séquences transcrites directement à partir d'un YAC [26]. Plus étendues que les exons internes, ces séquences peuvent aller de 200 à 2500 paires de bases. Il faut cependant souligner que les exons terminaux sont constitués en partie de la région 3' non traduite des gènes, certes spécifique de ceux-ci, mais riche en séquences répétées humaines telles que des éléments Alu et LINE.

En résumé, l'approche par amplification d'exon interne ou terminal présente l'avantage d'être indépendante de l'expression naturelle du gène à partir duquel un exon est dérivé. Le rendement de cette technique dépend principalement de la complexité de l'ADN de départ; en moyenne on peut espérer isoler un exon tous les 25 à 85 kb. *A priori*, seuls les gènes sans intron ne sont pas détectables; à l'inverse, des faux-positifs peuvent être obtenus du fait de la présence de sites cryptiques d'épissage.

Cette méthode a déjà été utilisée avec succès pour isoler de nombreux gènes, parmi lesquels ceux qui sont responsables du syndrome de Menkès *(m/s n° 3, vol. 9, p. 316)*, de la maladie de Huntington *(m/s n° 4, vol. 9, p. 488)*, de la neurofibromatose de type 2 *(m/s n° 4, vol. 9, p. 484)*, d'un certain type d'obésité murine *(m/s n° 12, vol. 10, p. 1337)* [27-30].

Méthodes bio-informatiques

Le séquençage à grande échelle peut finalement représenter une alternative tentante, du fait de l'émergence de nouveaux programmes informatiques de traitement des données obtenues. Deux types d'approches sont utilisés pour prédire la présence de séquences codantes: une approche globale reposant sur la détection de phases ouvertes de lecture et une approche locale fondée sur l'identification de motifs fonctionnels tels que les promoteurs, les sites d'épissage, les sites initiaux et terminaux de traduction, les signaux poly(A). Plusieurs algorithmes sont d'ores et déjà disponibles et accessibles par courrier électronique (*GRAIL*, *BLASTX*, *GeneID*, *SORFIND* entre autres). Les utilisateurs peuvent soumettre leurs séquences et recevoir en retour une analyse de la position d'exons potentiels avec une attribution à un brin, une phase de lecture préférentielle, et une évaluation qualitative de la prédiction. Dans la mesure où la séquence génomique est déterminée dans le seul but de localiser des exons candidats dans une région associée à un phénotype donné, Claverie [31] propose de ne plus ordonner les séquences obtenues les unes par rapport aux autres, évitant ainsi le sous-clonage successif en phages λ puis en M13 et l'étape de traitement des données correspondant. Il est en effet possible de sous-cloner directement de façon *shot-gun* en phage monobrin M13 des régions de 100 à 500 kb et d'utiliser les séquences sans les ordonner pour une recherche informatique d'exons potentiels au sein de chaque séquence. Une couverture de séquence de deux fois serait suffisante pour atteindre une probabilité supérieure à 0,9 de détecter les exons d'une région génomique donnée; cela est parfaitement illustré par le cas du gène du syndrome de Kallman, précèdemment identifié grâce aux protocoles de séquences classiques *(m/s n° 9, vol. 7, p. 980)* [32]. Bien que paraissant être une méthode lourde, cette stratégie présente l'intérêt d'utiliser des protocoles d'expériences bien établis pouvant être entièrement automatisés, la seule limitation étant en fait la capacité de séquençage d'un laboratoire.

Que choisir?

Si nous nous replaçons dans le cadre du clonage positionnel, quelle méthode faut-il adopter lorsque la région d'intérêt a été définie et les clones génomiques obtenus? Et faut-il n'en utiliser qu'une? A l'heure actuelle, il est encore difficile de définir une attitude logique en étudiant les choix des équipes qui ont récemment identifié des gènes associés à des maladies héréditaires. Le gène *BRCA1* a été isolé par sélection directe [12], le gène *obese* par amplification d'exons [30], et le gène *DAX1* par les anciennes méthodes de localisation d'îlots CpG et de recherche de conservations de séquence [4]... L'essentiel est-il donc de choisir une méthode, et de s'y tenir? En fait, quand c'est possible, il est sans doute avantageux d'utiliser simultanément plusieurs méthodes différentes. Les études actuellement en cours montreront-elles la complémentarité des techniques de sélection directe et d'amplification d'exons? En fait, certaines des méthodes que nous avons décrites ont déjà été employées ensemble: par exemple, après criblage direct avec des ADNc marqués, les cosmides positifs sont étudiés par amplification d'exons afin d'y détecter les séquences géniques [5]. De plus, les concepts sous-tendant la mise au point de ces techniques peuvent aussi s'interpénétrer: Sedlacek *et al.* [33], en appliquant à des cosmides la sélection directe, non pas d'ADNc, mais de fragments d'ADN génomique d'une espèce différente, ne font finalement rien d'autre que rechercher des conservations de séquence entre deux espèces... En attendant que le séquençage, qui est le but ultime des «projets Génome», et son traitement informatique, soient plus performants, on ne peut que tenter, très schématiquement, de définir des situations extrêmes, entre lesquelles les nuances les plus subtiles seront envisageables. Si l'on n'a aucune idée précise du type de tissu dans lequel le gène recherché peut être exprimé, les méthodes reposant sur les ADNc ne semblent pas très indiquées. Si la région d'intérêt est très étendue (*contig* de YAC, région microdisséquée), les techniques d'amplification d'exons semblent actuellement moins bien adaptées que celles de sélection directe. Ces deux exemples, volontairement très simplifiés, ne doivent pas faire oublier que le problème ne se pose pas seulement pour le choix de la méthode de recherche de gènes, mais aussi en amont, lors du clonage positionnel dans les maladies multigéniques, et en aval, pour déterminer, parmi les gènes isolés, lequel est impliqué dans la maladie étudiée ■

Note ajoutée aux épreuves

Depuis la rédaction de cette revue, plusieurs articles sont parus, qui montrent effectivement la complémentarité des techniques de sélection directe et d'amplification d'exons (Harshman *et al. Hum Mol Genet* 1995; 4: 1259-66 et Yaspo *et al. Hum Mol Genet* 1995; 4: 1291-304). Les deux techniques ont également permis l'isolement d'exons différents d'un même gène, celui de l'ataxie-télengiectasie (Savitsky *et al. Science* 1995; 208: 1749-53).

Remerciements

Nous remercions Patrick Gaudray pour ses critiques et ses conseils pendant la préparation de ce manuscrit.

A. Courseaux
Étudiante en thèse.

P. Szepetowski
Chargé de recherche au Cnrs.
LGMCH, Ura Cnrs 1462, UFR de médecine, avenue de Valombrose, 06107 Nice Cedex 2, France.

M. Fontès
Directeur de recherches au Cnrs.
Inserm U. 406, UFR de médecine, 27, boulevard Jean-Moulin, 13385 Marseille Cedex 5, France.

RÉFÉRENCES

1. Cooperative Human Linkage Center (CHLC): Murray JC, Buetow KH, Weber JL, Ludwigsen S, Scherpbier-Heddema T, Manion F, Quillen J, Sheffield VC, Sunden S, Duyk GM; Généthon: Weissenbach J, Gyapay G, Dib C, Morissette J, Lathrop GM, Vignal A; University of Utah: White R, Matsunami N, Gerken S, Melis R, Albertsen H, Plaetke R, Odelberg S; Yale University: Ward D; Centre d'Étude du Polymorphisme Humain (CEPH): Dausset J, Cohen D, Cann H. A comprehensive human linkage map with centimorgan density. *Science* 1994; 265: 2049-54.

2. Weissenbach J. Le génome humain entre médecine et science. *médecine/sciences* 1995; 11: 317-23.

3. Rommens JM, Iannuzzi MC, Kerem B, Drumm ML, Melmer G, Dean M, Rozmahel R, Cole JL, Kennedy D, Hidaka N, Zsiga M, Buchwald M, Riordan JR, Tsui L, Collins FS. Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science* 1989; 245: 1059-65.

4. Zanaria E, Muscatelli F, Bardoni B, Strom TM, Guioli S, Guo W, Lalli E, Moser C, Walker AP, McCabe ERB, Meitinger T, Monaco AP, Sassone-Corsi P, Camerino G. An unusual member of the nuclear hormone receptor superfamily responsible for X-linked adrenal hypoplasia congenita. *Nature* 1994; 372: 635-41.

5. Lawrence BJ, Schwabe W, Kioschis P, Coy JF, Poustka A, Brennan MB, Hochgeschwender U. Rapid identification of gene sequences for transcriptional map assembly by direct cDNA screening of genomic reference libraries. *Hum Mol Genet* 1994; 3: 2019-23.

6. Wallace MR, Marchuk DA, Andersen LB, Letcher R, Odeh HM, Saulino AM, Fountain JW, Brereton A, Nicholson J, Mitchell AL, Brownstein BH, Collins FS. Type 1 neurofibromatosis gene: identification of a large transcript disrupted in three NF1 patients. *Science* 1990; 249: 181-6.

7. Kinzler KW, Nilbert MC, Su LK, Vogelstein B, Bryan TM, Levy DB, Smith KJ, Preisinger AC, Hedge P, McKechnie D, Finniear R, Markham A, Groffen J, Boguski MS, Altschul SF, Horii A, Ando H, Miyoshi Y, Miki Y, Nishisho I, Nakamura Y. Identification of FAP locus genes from chromosome 5q21. *Science* 1991; 253: 661-5.

8. Thomas G. Dix ans de recherche sur les prédispositions génétiques au développement des tumeurs. *médecine/sciences* 1995; 11: 336-48.

9. Parimoo S, Patanjiali SR, Shukla H, Chaplin DD, Weissman SM. cDNA selection: efficient PCR approach for the selection of cDNAs encoded in large chromosomal DNA fragments. *Proc Natl Acad Sci USA* 1991; 88: 9623-7.

10. Lovett M, Kere J, Hinton LM. Direct selection: a method for the isolation of cDNAs encoded by large genomic regions. *Proc Natl Acad Sci USA* 1991; 88: 9628-32.

11. Tagle DA, Swaroop M, Lovett M, Collins FS. A magnetic bead capture of expressed sequences encoded whithin large genomic segments. *Nature* 1993; 361: 751-3.

12. Miki Y, Swensen J, Shattusk-Eidens D, Futreal PA, Harshman K, Tavtigian S, Liu Q, Cochran C, Bennett ML, Ding W, Bell R, Rosenthal J, Hussey C, Tran T, McClure M, Frye C, Hattier T, Phelbs R, Haugen-Strano A, Katcher H, Yakumo K, Gholami Z, Shaffer D, Stone S, Bayer S, Wray C, Bodgen R, Dayananth P, Ward J, Tonin P, Narod S, Bristow PK, Norris FH, Helvering L, Morrison P, Rosteck P, Lai M, Barreth JC, Lewis C, Neuhausen S, Cannon-Albright L, Goldgar D, Wiseman R, Kamb A, Skolnick MH. A strong candidate from the breast and ovarian cancer susceptibility genes *BRCA1*. *Science* 1994; 266: 66-71.

13. Gecz J, Pollard H, Consalez G, Villard L, Stayton C, Millasseau P, Khrestchatisky M, Fontès M. Cloning and expression of the murine homologue of a putative human X-linked nuclear protein gene closely linked to *PGK1* in Xq13.3. *Hum Mol Genet* 1994; 3: 39-44.

14. Villard L, Gecz J, Colleaux L, Lossi A, Chelly J, Ishikawa-Brush Y, Monaco AP, Fontès M. Construction of a YAC contig spanning the Xq13.3 sub-band. *Genomics* 1995; 26: 115-22.

15. Gibbons RJ, Picketts DJ, Villard L, Higgs DR. Mutations in a putative global transcriptional regulator cause X-linked mental retardation with a thalassaemia (ATR-X syndrome). *Cell* 1995; 80: 837-45.

16. Abe K. Rapid isolation of desired sequences from lone linker PCR amplified cDNA mixtures: application to identification and recovery of expressed sequences in cloned genomic DNA. *Mammalian Genome* 1992; 2: 252-9.

17. Su YA, Trent JM, Guan XY, Meltzer PS. Direct isolation of genes encoded within a homogeneously staining region by chromosome microdissection. *Proc Natl Acad Sci USA* 1994; 91: 9121-5.

18. Liu P, Legerski R, Siciliano MJ. Isolation of human transcribed sequences from human-rodent somatic cell hybrids. *Science* 1989; 246: 813-5.

19. Corbo L, Maley JA, Nelson DL, Caskey CT. Direct cloning of human transcripts with hnRNA from hybrid cell lines. *Science* 1990; 249: 652-5.

20. Jordan B. Ilots HTF: le gène annoncé. *médecine/sciences* 1991; 7: 153-60.

21. Cross SH, Charlton JA, Nan X, Bird AP. Purification of CpG islands using a methylated DNA binding column. *Nature Genet* 1994 ; 6: 236-44.

22. Valdes M, Tagle DA, Collins FS. Island rescue PCR: a rapid and efficient method for isolating transcribed sequences from yeast artificial chromosomes and cosmids. *Proc Natl Acad Sci USA* 1994; 91: 5377-81.

23. Duyk GM, Kim S, Myers RM, Cox DR. Exon trapping: a genetic screen to identify candidate transcribed sequences in cloned mammalian genomic DNA. *Proc Natl Acad Sci USA* 1990; 87: 8995-9.

24. Church DM, Stotler CJ, Rutter JL, Murrell JR, Trofatter JA, Buckler AJ. Isolation of genes from complex sources of mammalian genomic DNA using exon amplification. *Nature Genet* 1994; 6: 98-105.

25. Krizman DB, Berget SM. Efficient selection of 3'-terminal exons from vertebrate DNA. *Nucleic Acids Res* 1993; 21: 5198-202.

26. Krizman DB, Hofmann TA, DeSilva U, Green ED, Meltzer PS, Trent JM. Identification of 3' terminal exons from yeast artificial chromosomes. *Nucleic Acids Res* 1995 (sous presse).

27. Vulpe C, Levinson B, Whitney S, Packman S, Gitschier J. Isolation of a candidate gene for Menkes diseases and evidence that it encodes a copper-transporting ATPase. *Nature Genet* 1993; 3: 7-13.

28. The Huntington's Disease Collaborative Research Group. A novel gene containing a trinucleotide repeat that is expended and instable on Huntington's disease chromosomes. *Cell* 1993; 72: 971-83.

29. Trofatter JA, Mac Collin M, Rutter JL, Murrell JR, Duyao MP, Parry DM, Eldridge R, Kley N, Menon AG, Pulaski K, Haase VH, Ambrose CM, Munroe D, Bove C, Haines JL, Martuza RL, Mac Donald ME, Seizinger BR, Short MP, Buckler AJ, Gusella JF. A novel moesin-, ezrin-, radaxin-like gene is a candidate for the neurofibromatosis 2 tumor suppressor. *Cell* 1993; 72: 791-800.

30. Zhang Y, Proenca R, Maffei M, Barone M, Leopold L, Friedman JM. Positional cloning of the mouse *obese* gene and its human homologue. *Nature* 1994; 372: 425-32.

31. Claverie JM. A streamlined random sequencing strategy for finding coding exons. *Genomics* 1994; 23: 575-81.

32. Legouis R, Hardelin JP, Levilliers J, Claverie JM, Compain S, Wunderle V, Milasseau P, Le Paslier D, Cohen D, Caterina D, Bougueleret L, Delemarre-Van de Waal H, Lutfalla G, Weissenbach J, Petit C. The candidate gene for the X-linked Kallman syndrome encodes a protein related to adhesion molecules. *Cell* 1991; 67: 423-35.

33. Sedlacek Z, Konecki DS, Siebenhaar R, Kioschis P, Poustka A. Direct selection of DNA sequences conserved between species. *Nucleic Acids Res* 1993; 21: 3419-25.

Summary

From DNA to cDNA: how to isolate genes from large genomic regions

While considerable progress has been made in mapping of the human genome, approaches are now needed to identify the genes contained within large genomic regions. Traditional techniques, including identification of CpG islands, detection of conserved sequences and direct screening of cDNAs libraries, are effective but very laborious. More powerful methods, such as exon trapping, cDNA selection, and computational analysis of genomic sequences, are now available. A combination of several different methods should be required to isolate most transcribed sequences from a given region.