

F. HATEY

INRA, Laboratoire de Génétique Cellulaire,  
BP 27, 31326 Castanet-Tolosan cedex

e-mail : Francois.Hatey@toulouse.inra.fr

## 4 - Recherche de gènes associés à des fonctions

# Recherche de gènes associés à des fonctions : l'approche fonctionnelle

**Résumé.** L'objectif est ici de partir de la fonction pour identifier les gènes d'intérêt. Cette identification passe par le séquençage des extrémités des ADN complémentaires pour donner des « étiquettes ». Cette stratégie est très efficace, même si, compte tenu des mécanismes de transcription et des contraintes expérimentales, des étiquettes différentes peuvent correspondre à un même gène. L'analyse de l'expression différentielle entre des animaux extrêmes pour un caractère permet d'accéder directement aux gènes d'intérêt pour ce caractère. Les analyses peuvent être ciblées sur les gènes régulés ou être systématiques ; c'est le cas des filtres haute densité et des micro-réseaux qui permettent d'étudier un répertoire très large de transcrits. Le projet Asteroger lancé par l'INRA a pour objectif d'établir un tel répertoire dans quatre espèces animales. Ce répertoire va constituer un outil précieux aussi bien en physiopathologie qu'en génétique où il facilitera la mise en œuvre de la stratégie d'identification des gènes majeurs et des QTL par l'approche 'candidat positionnel'.

La démarche proposée ici est basée sur la prise en compte des caractéristiques fonctionnelles, c'est-à-dire des données de physiologie ou de pathologie au sens large : l'expression d'un caractère donné résulte de l'action d'un certain nombre de gènes, et il est facile de comprendre que la variabilité de ces gènes va entraîner celle du caractère considéré. L'amélioration de ce caractère passe donc par la prise en compte des gènes impliqués.

Le premier objectif est donc d'identifier les gènes dont l'expression est liée à cette fonction ou à ce caractère : gènes exprimés dans la glande mammaire pour la production de lait, gènes exprimés dans l'ovaire pour le taux d'ovulation, gènes exprimés dans le muscle pour la qualité de la viande.

Cette information d'expression n'est toutefois pas suffisante car de très nombreux gènes interviennent dans la mise en place d'une fonction, d'un caractère donné : il n'est pas facile d'identifier, parmi les gènes de ce répertoire, ceux dont la variabilité va entraîner une modification du caractère étudié. Pour faciliter cette identification, nous allons comparer l'expression de ces gènes dans des situations où la fonction - le caractère - présente des valeurs très différentes : si un gène est exprimé différemment chez deux animaux extrêmes pour le caractère considéré, toutes choses étant égales par ailleurs, ce gène est très certainement impliqué dans le déterminisme de cette différence entre les animaux. Le deuxième objectif est donc de mesurer le niveau d'expression des gènes afin d'identifier ceux dont l'expression varie en fonction de la valeur du caractère.

Certaines notions présentées ici peuvent être retrouvées dans une mise au point récente, plus large, qui intègre également des aspects génétiques qui ne seront pas évoqués ici (Hatey *et al* 1998).

## 1 / Identification des gènes exprimés dans une fonction donnée

Une fonction donnée résulte de l'expression de certains gènes, c'est-à-dire de la présence de certaines protéines qu'il « suffit » d'identifier, soit directement, soit par le biais de leurs précurseurs, les ARN messagers. La grande variété des techniques d'analyse des acides nucléiques disponibles aujourd'hui (Riquet et Pitel 2000, cet ouvrage) rend cette deuxième approche plus performante, mais il ne faut pas perdre de vue que le niveau d'analyse le plus proche de la fonction est celui des protéines.

### 1.1 / Une cellule, des gènes

Chez les mammifères, on estime de 30 000 à 120 000 le nombre de gènes, dont 10 000 à 30 000 sont transcrits - exprimés - par cellule. Toutefois, tous les gènes ne sont pas exprimés avec la même intensité : le nombre de copies varie donc selon les ARN messagers et l'on distingue classiquement trois classes d'abondance (tableau 1).

**Tableau 1.** Répartition des ARN messagers (ARNm) dans les différentes classes d'abondance.

Classe	très abondants	abondants	rares
% des ARNm totaux	10 (10-20)	45 (40-45)	45 (40-45)
Nombre d'espèces d'ARNm	10	1 000	15 000
Nombre de copies par espèce	5 000	225	15
Nombre total d'ARNm	50 000	225 000	225 000

La présence de plusieurs copies d'un même transcrit constitue la redondance ; celle-ci, associée à l'existence d'espèces moléculaires différentes, est à l'origine du caractère complexe des populations d'ARN messagers, populations constituées d'environ 500 000 molécules d'ARN messenger par cellule.

### 1.2 / Un gène, des ARN messagers

Au cours de la transcription des gènes en ARN, différents processus pré- ou post-transcriptionnels peuvent intervenir, de telle sorte qu'à partir d'un même gène, plusieurs ARN messagers (ARNm) ou transcrits différents peuvent être produits ; ces possibilités sont liées, d'une part, à l'existence de sites alternatifs pour le démarrage (promoteur) ou l'arrêt (site de polyadénylation) de la transcription et, d'autre part, à l'existence de mécanismes d'épissage alternatif qui introduisent des variations dans la conservation et l'exclusion des différents exons et introns (figure 1 A et B).

### 1.3 / Un ARN messenger, des ADN complémentaires

Les ADN complémentaires (ADNc) sont obtenus par transcription inverse des ARN messagers, en utilisant comme amorce soit un oligo(dT) complémentaire de la séquence poly(A) de l'ARNm, soit des amorces de séquence aléatoire qui vont se fixer en différents points de l'ARNm. Cependant, même si la synthèse est amorcée en 3' sur le poly(A), les ADNc sont fréquemment incomplets du fait de la présence, dans l'ARN messenger, de structures secondaires stables qui bloquent la transcriptase. A partir d'un même transcrit, différents ADNc de longueurs différentes peuvent donc être obtenus ; cependant, dans le cas d'un amorçage en 3', tous les ADNc issus d'un même transcrit auront la même séquence en 3', ce qui permet de les regrouper (figure 1 C ; voir plus loin).

### 1.4 / Un ADN complémentaire, des étiquettes

Les ADNc sont identifiés par leur séquence, par comparaison avec celles répertoriées dans les bases de données (Vignal 2000 et Corpet et Chevalet 2000, cet ouvrage). Plutôt que de séquencer la longueur totale de chaque ADNc, ce qui nécessite plusieurs réactions de séquence avec des amorces intermédiaires différentes et donc plusieurs 'lectures', seules les extrémités de chaque ADNc sont séquencées : les données sont obtenues après une seule lecture et avec des amorces universelles (figure 1 C). De telles séquences constituent des étiquettes (ou EST pour Expressed Sequenced Tags) ; bien qu'elles soient incomplètes, elles sont souvent suffisantes pour permettre l'identification du gène correspondant. Comme les extrémités 3' et 5' d'un ARN messenger

n'ont pas les mêmes caractéristiques, les étiquettes correspondantes ne présentent pas le même intérêt. En 5', la séquence est le plus souvent dans la région codante, ce qui facilite l'identification des gènes. En 3', au contraire, la séquence se trouve généralement dans une région non traduite ; elle est peu conservée entre les espèces et présente rarement des introns. Ces caractéristiques les rendent très utiles pour la localisation des gènes sur les chromosomes en utilisant les panels d'hybrides somatiques (Yerle 2000, cet ouvrage).

### 1.5 / Un gène, des étiquettes

Si deux gènes différents produisent des étiquettes différentes, les différents mécanismes présentés indiquent qu'au contraire, deux étiquettes différentes ne correspondent pas systématiquement à deux gènes différents. Il est néanmoins possible d'essayer de regrouper ces séquences entre elles, soit parce qu'elles sont chevauchantes, soit parce qu'elles sont issues d'un même transcrit, comme le démontre la présence en 3' d'une séquence identique pour plusieurs clones. De tels regroupements d'étiquettes ou « clusters » permettent d'approcher la séquence complète des ADNc (figure 1 C).

## 2 / Analyse de l'expression différentielle des gènes

L'objectif est de comparer des animaux extrêmes ou des situations expérimentales différentes pour identifier les gènes qui sont exprimés différemment, à travers des approches

- ciblées, visant à identifier les ARN messagers régulés ;
- systématiques, visant à analyser « tous » les gènes.

### 2.1 / Approches ciblées

L'objectif est ici d'identifier les ARN messagers régulés par comparaison de profils d'expression ou par élimination (soustraction) des gènes dont l'expression est invariable.

#### a / Approche comparative : diviser pour régner

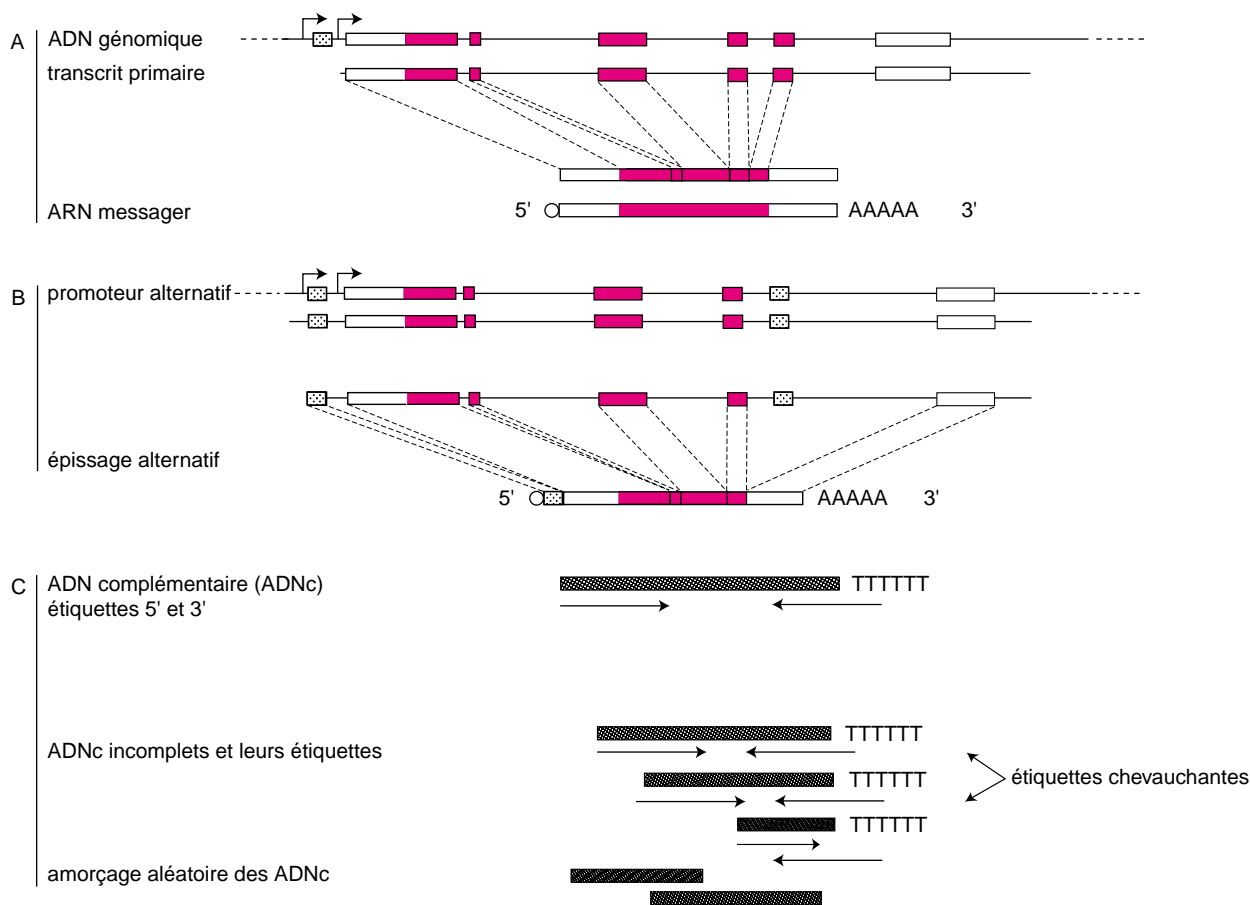
Dans une première approche permettant de mettre en évidence les transcrits spécifiques d'un état ou d'une situation donnée, la population des transcrits, sous forme d'ADN complémentaires (ADNc), est séparée en différents sous-groupes successifs de telle sorte que l'on puisse les individualiser par électrophorèse. Les profils obtenus dans les différentes situations analysées sont comparés, permettant ainsi d'identifier directement les transcrits qui sont régulés (figure 2).

**Figure 1.** Du gène aux étiquettes (d'après Hatey et al 1998).

A. Dans l'ADN génomique, un gène est composé d'introns (lignes) et d'exons (rectangles) que l'on retrouve dans le transcrit primaire. La transcription démarre au niveau du promoteur (flèche). Au cours de la maturation des ARN messagers, les introns sont éliminés par excision et les exons liés entre eux par épissage ; certains exons (en entier ou en partie) ne sont pas traduits (fonds blancs). L'ARN messenger subit encore deux autres modifications : l'addition en 5' d'une structure particulière, la coiffe (cercle), et en 3' l'addition d'une séquence de poly (A) (plusieurs dizaines de résidus).

B. L'existence d'un promoteur alternatif entraîne la présence d'un exon supplémentaire ; par contre, l'épissage alternatif entraîne ici l'absence d'un exon dans le messenger final.

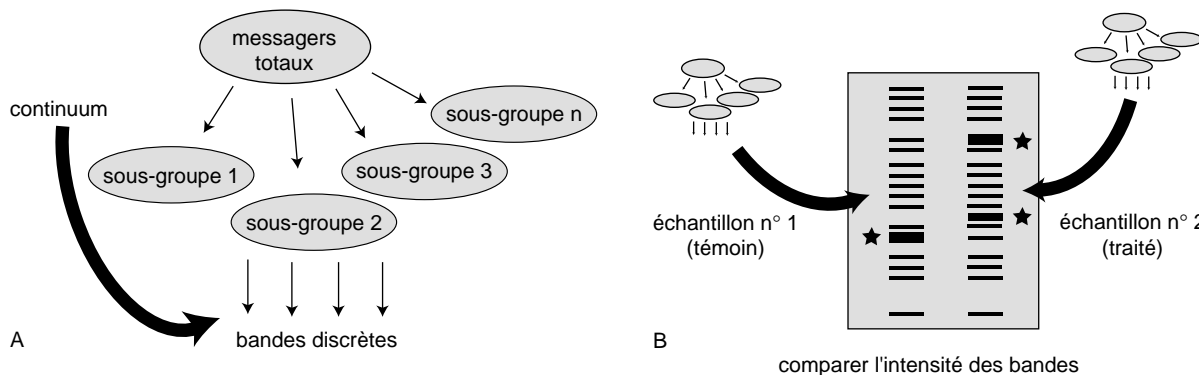
C. Les ADN complémentaires (ADNc) sont obtenus en utilisant comme amorce un oligo(dT) hybridé à la séquence poly(A). Les étiquettes (flèches) correspondent à la séquence de l'extrémité des ADNc. Pour un même messenger, à cause de la structure secondaire des ARN messagers qui fait obstacle à la reverse transcriptase, des ADNc incomplets, de longueurs différentes sont produits. Des étiquettes 5' différentes sont ainsi obtenues, qui peuvent se chevaucher, permettant de constituer une séquence consensus. Il en va de même si la synthèse des ADNc est amorcée avec des oligonucléotides aléatoires.



**Figure 2.** Diviser pour régner.

A. La population des ARN messagers présente une continuité de taille et il n'est pas possible d'identifier directement une espèce moléculaire particulière. La stratégie présentée ici consiste à fractionner cette population en différents sous-groupes qui seront à leur tour fractionnés afin d'obtenir des produits de taille différente. Ce fractionnement passe par une amplification PCR avec différents jeux d'amorces.

B. Les différents sous-groupes sont analysés par électrophorèse en déposant côte à côte, sur un même gel d'électrophorèse, les réactions correspondant aux différents échantillons. La comparaison de l'intensité des bandes révèle les variations d'abondance des ADNc, indiquant une régulation de l'expression des gènes correspondants.



Les sous-groupes sont constitués par amplification des ADN complémentaires à l'aide de différentes amorces.

Le mRNA Differential Display (DD ; Liang et Pardee 1992) est un concept simple et d'une mise en œuvre relativement facile : après transcription inverse en utilisant des amorces oligo(dT) ancrées sur une ou deux bases, les ADNc obtenus sont amplifiés par PCR en utilisant le même oligo(dT) ancré et un décimère de séquence aléatoire. Les produits d'amplification obtenus sont analysés par électrophorèse en gel d'acrylamide et l'on observe une certaine de bandes pour chaque combinaison d'amorces. Les profils obtenus à partir des différentes populations étudiées sont analysés sur un même gel : si une même bande présente des différences d'intensité selon les échantillons, c'est que les concentrations des ARN messagers correspondants sont différentes dans les populations d'origine. Les fragments d'ADNc présents dans ces différentes bandes sont alors élués, clonés et séquencés.

L'application aux ADNc de la technique d'AFLP (Amplified restriction Fragment Length Polymorphism ; Bachem *et al* 1996) permet, comme le Differential Display, de séparer les ADNc en différents sous-groupes. Dans une première étape, les ADNc sont digérés par deux enzymes de restriction différentes puis des adaptateurs correspondant aux sites de restriction sont fixés aux extrémités des fragments. Ceux-ci sont ensuite amplifiés par PCR avec des amorces complémentaires des adaptateurs et ancrées par addition de deux bases. Les différentes combinaisons d'amorces ancrées permettent d'obtenir différents sous-groupes. Les produits d'amplification sont séparés par électrophorèse et les profils obtenus sont comparés comme précédemment.

Ces deux techniques présentent comme avantages majeurs la possibilité d'effectuer des comparaisons multiples (plus de deux) et de pouvoir mettre en évidence simultanément des régulations positives et négatives. Par ailleurs, il semble bien que les messagers rares soient accessibles. Ces techniques ne nécessitent pas de clonage préalable ; elles n'exigent pas de compétences très pointues en biologie moléculaire et sont d'un coût modéré.

Des résultats partiels peuvent être obtenus avec des quantités d'ARN relativement faibles (moins d'un microgramme d'ARN total), mais, pour couvrir la totalité ou la quasi totalité du spectre des transcrits, de nombreuses réactions avec des amorces différentes sont nécessaires (pour le DD, 80 amorces aléatoires différentes couplées aux 3 oligo(dT) ancrés, soit 240 réactions différentes seraient nécessaires ; pour l'AFLP, 16 amorces différentes par site de restriction, soit 256 combinaisons pour les deux sites). Les différents contrôles nécessaires - répétitions, vérification du caractère différentiel - augmentent encore les quantités d'ARN utilisées pour exploiter la totalité de l'analyse.

Les inconvénients majeurs du Differential Display sont liés d'une part au nombre élevé de faux positifs et, d'autre part, à la petite taille (quelques centaines de bases au plus) des fragments isolés. Ce dernier point implique un travail complémentaire pour allonger ces fragments ou pour rechercher dans une génothèque un clone plus long ; de plus, la localisation en 3' de ces fragments, dans une région souvent

non codante du gène, ne facilite pas leur identification. Les progrès les plus significatifs à attendre sont donc certainement ceux qui permettront d'obtenir des fragments plus longs (> 1 kb). Malgré ces limites, le Differential Display est une technique très largement utilisée, comme en témoigne le nombre important de publications qui lui ont été consacrées.

La méthode cDNA-AFLP lève partiellement les limites du DD, en particulier parce que les amplifications par PCR sont conduites dans des conditions strictes, ce qui assure une meilleure reproductibilité et une diminution du nombre de faux positifs. De plus, les fragments isolés ne sont pas limités à l'extrémité 3' des transcrits et peuvent donc se trouver dans les régions codantes du gène.

### ***b / Approche soustractive : éliminer les ADNc superflus***

Une deuxième approche permettant de mettre en évidence des différences entre deux populations d'ARN utilise l'hybridation soustractive pour éliminer tous les ADNc correspondant aux transcrits présents en quantité équivalente dans les deux populations ; seuls sont ainsi conservés les transcrits différentiels.

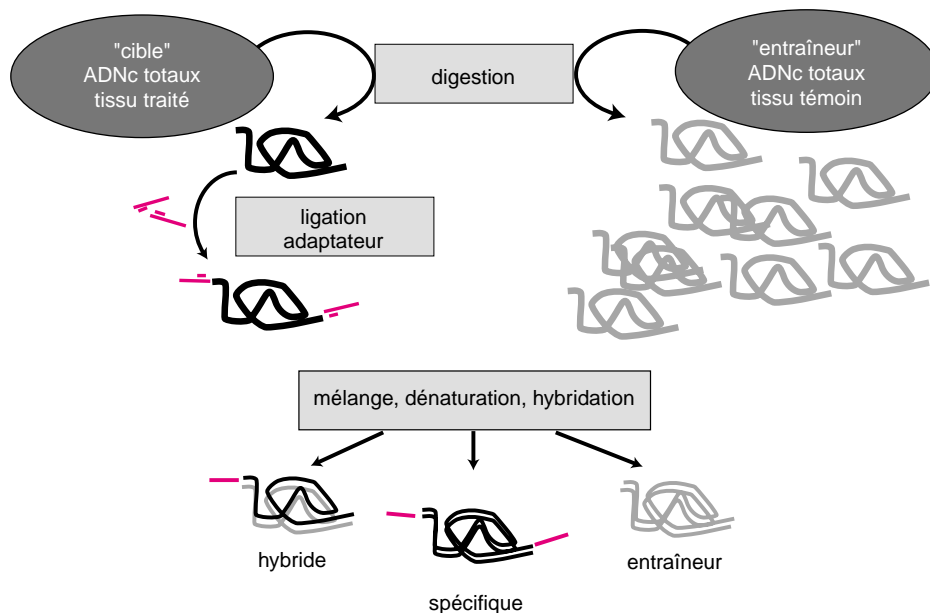
Cette approche est utilisée dans la « Representational Difference Analysis » appliquée aux ADNc (cDNA-RDA) et la « Suppression Subtractive Hybridization » (SSH ; Hubank et Schatz 1994, Diatchenko *et al* 1996). Toutes deux utilisent la PCR pour d'une part éliminer les ADNc communs aux deux populations d'ARN à comparer et, d'autre part, augmenter la sensibilité de la technique, rendant possible la détection des transcrits différentiels faiblement exprimés.

La RDA et la SSH reposent sur le même principe (figure 3). Dans un premier temps les ADNc double brin obtenus à partir des deux populations à comparer sont digérés par une enzyme de restriction afin d'obtenir des fragments d'environ 500 paires de bases. Des adaptateurs sont ensuite fixés aux extrémités des fragments de restriction obtenus à partir des ADNc de la population d'intérêt. Ces ADNc sont ensuite mélangés avec un large excès d'ADNc issus de la population témoin ; le mélange est dénaturé puis placé en conditions d'hybridation. Après hybridation, les ADNc sont soumis à une amplification par PCR en utilisant des amorces complémentaires de la séquence des adaptateurs. Parce qu'ils portent un adaptateur à chaque extrémité, les homoduplex correspondant aux transcrits plus abondants dans la population d'intérêt sont amplifiés de manière exponentielle. Les autres duplex ne sont pas amplifiables parce qu'ils n'ont pas d'adaptateurs, ou bien sont amplifiés linéairement parce qu'ils n'ont qu'un seul adaptateur. Après cette PCR, le nombre de copies des ADNc d'intérêt est considérablement augmenté.

Dans le cas de la SSH, la population d'ADNc d'intérêt est divisée en deux fractions qui reçoivent chacune un adaptateur différent. Après l'hybridation soustractive, les deux fractions sont mélangées et hybridées ; une PCR permet ensuite d'amplifier préférentiellement les hétéroduplex ayant un adaptateur différent sur chaque brin. En effet, la structure des adaptateurs ne permet pas l'amplification des homoduplex par un mécanisme de 'PCR suppressive' :

**Figure 3. Eliminer les ADNc superflus.**

Pour identifier les ADNc spécifiques de la population « cible », celle-ci reçoit, après digestion, un adaptateur à chaque extrémité puis est mélangée avec un large excès d'une population d'ADNc « entraîneur » ne contenant pas les séquences recherchées ; ces ADNc ont été digérés comme ceux de la cible, mais ne reçoivent pas d'adaptateur. Le mélange cible/entraîneur est dénaturé puis renaturé et amplifié par PCR en utilisant des amorces complémentaires de la séquence des adaptateurs. Parmi les trois types d'ADNc double brin obtenus, entraîneur/entraîneur, entraîneur/cible, cible/cible, seul le dernier possède un adaptateur à chaque extrémité et sera amplifié exponentiellement.



L'hybridation intramoléculaire de ces adaptateurs entraîne la formation d'une boucle qui empêche la synthèse d'un brin complémentaire. L'avantage de cette étape est de normaliser le produit soustrait puisque les séquences spécifiques des ADNc d'intérêt mais très abondantes seront sélectionnées par la soustraction, mais éliminées par la suppression.

Ces techniques présentent l'inconvénient de nécessiter deux soustractions pour explorer l'ensemble des régulations - positives et négatives - pour chaque paire de situations expérimentales ; de plus, la soustraction n'est jamais complètement efficace, et il faut ensuite éliminer les faux positifs. En revanche, les fragments obtenus ne sont pas nécessairement dans la région 3', ce qui facilite leur identification.

## 2.2 / Approches systématiques

L'objectif est ici d'identifier « tous » les transcrits d'un type cellulaire donné ou d'un tissu afin de définir le répertoire des gènes exprimés ou transcriptome. Ces approches, systématiques, doivent permettre la mise en évidence, par exemple, des cascades de signalisation ou des interactions entre gènes et non plus simplement la régulation d'un gène ou d'un petit nombre de gènes. Deux approches différentes sont présentées ici, une approche numérique dans laquelle on va compter le nombre de transcrits d'un gène donné, et une approche analogique dans laquelle on va mesurer la « quantité » de transcrit d'un gène donné.

### a / Approche numérique : des étiquettes en série

La Serial Analysis of Gene Expression (SAGE) permet l'évaluation simultanée, tant quantitative que qualitative, de l'expression de l'ensemble des gènes d'une lignée cellulaire ou d'un tissu donné.

Cette technique met à profit le fait qu'une étiquette aussi courte que 13 ou 14 paires de bases (pb) contient suffisamment d'information pour identifier un transcrit unique, sachant sa position définie dans un transcrit ; la concaténation de ces étiquettes permet une analyse en série, rapide et efficace, des transcrits (Velculescu *et al* 1995).

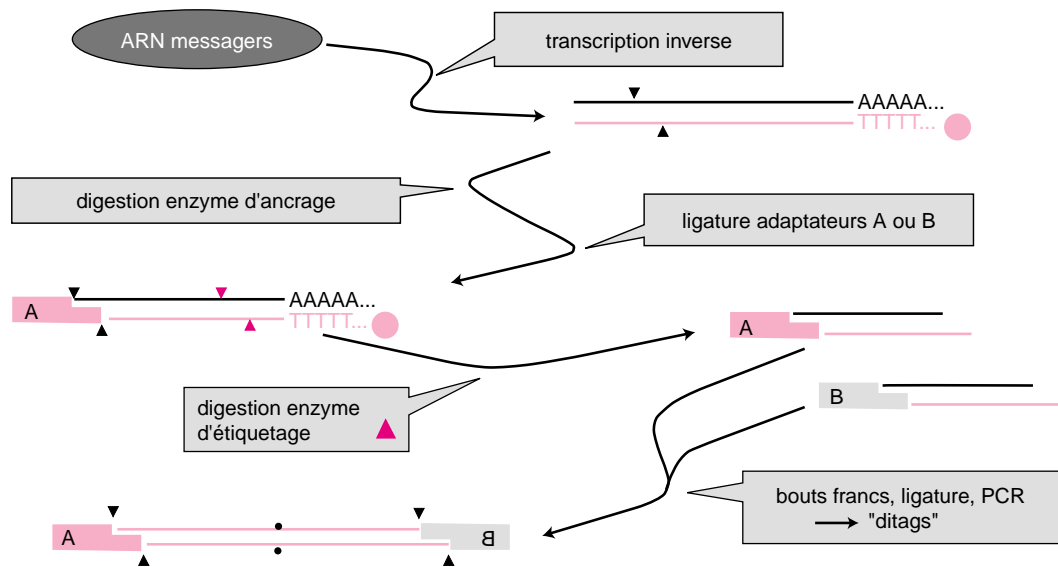
La première étape met en œuvre une digestion des ADNc par une enzyme de restriction ; le site de coupure le plus proche du poly(A), appelé site d'ancrage, définit la position de l'étiquette dans le transcrit. La deuxième étape consiste à découper une étiquette de longueur constante dans le fragment d'ADNc compris entre le site d'ancrage et le poly(A) ; cette coupure est réalisée par une enzyme particulière, dite de classe II S, qui coupe à distance de son site de reconnaissance. Ce site est apporté par un adaptateur que l'on fixe à l'extrémité du fragment d'ADNc. Après digestion, les fragments obtenus, débarrassés des adaptateurs, sont liés les uns aux autres, clonés et séquencés (figure 4). Une seule réaction de séquence sur chacun des clones obtenus permet alors d'identifier plus d'une vingtaine d'étiquettes et donc de gènes différents.

Le nombre d'occurrences d'une étiquette donnée reflète le niveau d'expression du gène correspondant ; celui-ci peut être identifié par comparaison avec les bases de données de séquence ; l'oligonucléotide correspondant à l'étiquette peut être utilisé pour cribler une génothèque d'ADNc, et le clone ainsi isolé va permettre de contrôler le niveau d'expression par Northern blot.

Performante, cette technique nécessite cependant une bonne pratique de la biologie moléculaire et l'accès à des moyens de séquençage conséquents car il faut analyser de 25 000 à 50 000 étiquettes pour obtenir des comparaisons significatives. De plus, les séquences obtenues correspondent à l'extrémité 3'

**Figure 4.** Technique SAGE.

Les ADN complémentaires sont synthétisés en utilisant comme amorce un oligo(dT) fixé sur des billes magnétiques ; après digestion par une enzyme de restriction, il est ainsi possible d'isoler les fragments 3'. La fraction correspondante est divisée en deux portions identiques et chacune reçoit un adaptateur différent, A ou B (A est seul présenté sur les deux étapes suivantes). Cet adaptateur porte la séquence de reconnaissance pour une enzyme de classe IIS, enzyme d'étiquetage qui va permettre de libérer un fragment d'une douzaine de bases. Les ditags finalement obtenus sont



des ARN messagers, ce qui ne facilite pas l'identification des gènes, surtout si l'on étudie des espèces peu représentées dans les banques, ce qui est le cas pour nos espèces d'élevage. En contre partie, pourvu que l'on y mette les moyens (séquençage), tous les gènes sont accessibles, quel que soit leur niveau d'expression. Les possibilités de cette technique sont illustrées par l'analyse approfondie du transcriptome humain qui vient d'être publiée (Velculescu *et al* 1999).

### **b / Approche analogique : gènothèque bien ordonnée...**

Dans l'approche analogique, la quantité d'ARN messagers présents est estimée par la mesure de l'intensité du signal d'hybridation qui lui est proportionnelle. Une goutte de la solution d'ADNc est déposée sur un support solide puis hybridée avec des sondes complexes c'est-à-dire obtenues à partir des ARN messagers totaux extraits du tissu ou du type cellulaire étudié (cf paragraphe 1.1). L'hybridation en parallèle de centaines ou de milliers d'ADNc regroupés sur un même support permet d'approcher ainsi l'analyse de 'tous' les gènes (Jordan 1998, Gerhold *et al* 1999).

#### *Filtres haute densité*

Le principe de base est d'organiser les gènothèques d'ADNc en plaçant les différents clones dans les puits de plaques de microtitration (96 ou 384 puits) puis de transférer ces mêmes clones ou leur produit d'amplification PCR sur une membrane (nylon) selon la même matrice (96, 384). En diminuant le diamètre des dépôts, il est possible de transférer plusieurs plaques sur une seule membrane aux dimensions de la plaque, en décalant légèrement les différentes séries de dépôts (figure 5). Réalisées avec l'aide de robots, de telles membranes peuvent être obtenues avec une densité de 9, 16 ou 25 dépôts par cm<sup>2</sup> soit par exemple 864, 1536 ou 2400 dépôts sur une surface de 8 x 12 cm. Chaque

hybridation avec une sonde complexe, radioactive, permet d'obtenir des informations sur l'expression de ces 2400 ADNc ; les données des différentes hybridations sont cumulatives pour les différents ADNc déposés sur ce même support.

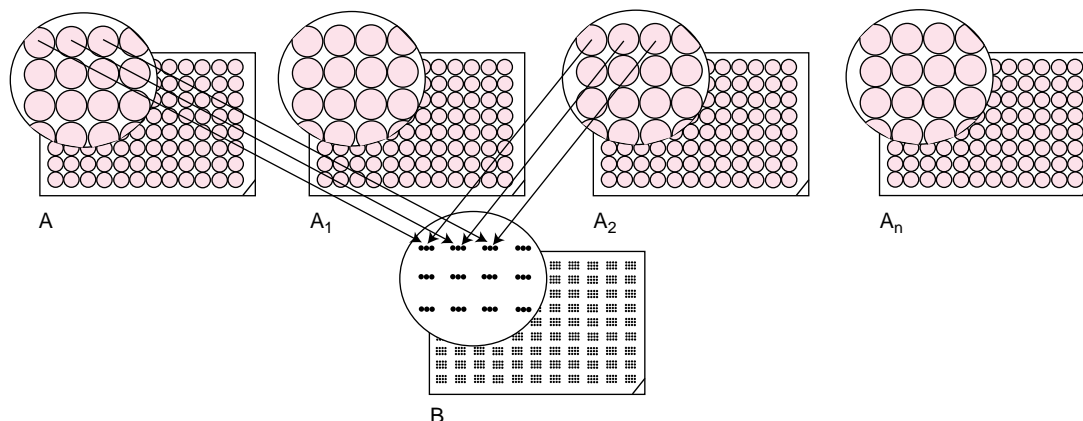
#### *Puces à ADN*

Selon le même principe, les puces à ADN permettent d'analyser en parallèle plusieurs milliers de gènes, toujours par hybridation de 'cibles', oligonucléotides ou ADN complémentaires, avec différentes sondes. Les supports sont constitués d'une membrane de nylon, d'une lame de verre ou de silicium, et la densité atteint actuellement quelques milliers de dépôts par cm<sup>2</sup>. Deux types de puces peuvent être distingués, les microarray ou microréseaux et les oligo chips.

Les microréseaux sont dérivés directement des filtres haute densité dont ils sont l'équivalent miniaturisé ; la membrane de nylon est remplacée par une lame de verre et les sondes sont marquées par un fluorophore. Cette approche est illustrée par les travaux réalisés sur la levure dans le laboratoire de Patrick Brown à Stanford : en utilisant un microréseau de 1,8 cm de côté, l'activité des 6 000 gènes de la levure a été analysée dans différentes conditions expérimentales, en particulier au cours de la croissance, lorsque le glucose du milieu de culture s'épuise, ou bien encore chez des mutants d'un facteur de transcription. L'hybridation est réalisée avec un mélange de deux sondes, marquées chacune par une molécule fluorescente différente, ce qui permet une comparaison directe des signaux obtenus dans deux conditions expérimentales différentes. Le groupe de Patrick Brown a ainsi pu mettre en évidence l'ensemble des gènes dont la transcription est activée ou inhibée au cours de l'épuisement en glucose du milieu de culture, ou bien les différents gènes placés sous la dépendance du facteur de transcription ayant fait l'objet de la mutation analysée (DeRisi *et al* 1997).

**Figure 5.** Filtres haute densité.

Dans une génothèque organisée, les clones d'ADNc ou leurs produits d'amplification sont placés individuellement dans les puits de plaques de microtitration (A...An). Une fraction de la suspension bactérienne ou de la solution est prélevée par un robot et déposée sur une membrane de nylon (B) sous la forme d'une tache de faible diamètre de telle sorte qu'un grand nombre de dépôts puisse être réalisé à partir de plusieurs plaques.



Ces informations d'expression dans différentes situations physiologiques, pathologiques ou expérimentales permettent de mieux cerner la fonction des gènes étudiés. Les résultats d'une telle approche, systématique, constituent ce qu'il est convenu aujourd'hui d'appeler un transcriptome.

Dans le deuxième type, puces à ADN ou oligo chips, ce sont des oligonucléotides de 20 résidus qui sont placés sur un support de silicium. En plaçant côte à côte des séries d'oligonucléotides ne différant que par une seule base et par hybridation dans des conditions très spécifiques, il est possible de mettre en évidence une différence de séquence entre deux échantillons. Cette technique de quasi séquençage permet donc de mettre en évidence très facilement des mutations dans un gène déjà connu (Pitel et Riquet 2000, cet ouvrage). La société Affymetrix, leader dans ce domaine, commercialise déjà des puces permettant l'identification des mutations de la protéine p53 impliquée dans de nombreux cancers, des cytochromes P450 2D6 et 2C19 ou de la protéase et de la reverse transcriptase du VIH. Toujours en utilisant des oligonucléotides, la société Affymetrix a développé des puces permettant d'analyser l'expression de quelques milliers de gènes, à la manière des microréseaux ; ces puces sont disponibles pour l'Homme (6 800 gènes différents), la souris (6 500 gènes) ou la levure (6 100 gènes). Affymetrix n'est pas la seule société sur le marché, même si elle le protège par un certain nombre de brevets ; en France, par exemple, le CEA développe des puces avec ses propres méthodologies.

L'utilisation de tels dispositifs nécessite des installations spécifiques permettant de produire les puces dans des conditions reproductibles, de les hybrider, et de lire des signaux fluorescents. Il faut ensuite gérer la quantité importante d'information que ces expériences vont produire.

## Conclusion et perspectives

La dernière décennie est marquée par un certain nombre d'avancées qui ont contribué au développement du concept de génomique :

- le séquençage complet du génome de plusieurs espèces bactériennes, de la levure, du nématode

*Caenorhabditis elegans*, en attendant celui de la drosophile et surtout celui de l'Homme dont une première version sera disponible dans le courant de l'année 2000 ;

- l'identification systématique des gènes exprimés dans de nombreux tissus ou types cellulaires différents chez l'Homme par le biais des étiquettes, donnant accès au transcriptome ;

- la mise au point de nouvelles méthodes d'analyse de l'expression des gènes (Differential Display, Subtractive Suppressive Hybridization, Serial Analysis of Gene Expression, puces à ADN).

Ces avancées rendent accessible l'étude simultanée chez l'Homme d'un très grand nombre de gènes ; par voie de conséquence, elles vont faciliter l'identification des gènes impliqués dans des maladies humaines et accélérer le criblage des molécules d'intérêt pharmacologique.

Chez les espèces d'élevage, l'application de ces avancées est possible : elle est envisagée à l'INRA dans le cadre du programme transversal Asteroger présenté plus loin (Hatey *et al* 2000, cet ouvrage) et dont l'objectif est d'obtenir un répertoire aussi large que possible des transcrits chez les bovins (ruminants), porcins, poulets et poissons. La construction de cet outil passe par la création de banques normalisées multi-tissus pour chacune des espèces concernées, le séquençage de ces banques et leur mise à disposition sous forme de filtres haute densité ou de microréseaux.

Ce répertoire n'est pas tout : il ressemble au tas de pierres dont Henri Poincaré (La science et l'hypothèse) rappelle qu'il n'est pas une maison. En d'autres termes, à côté des données de séquence, il faut des données de localisation et d'expression pour cerner la fonction des différents gènes et, au-delà, pour déterminer leur rôle. Comme chez l'Homme, une retombée immédiate de ces répertoires va être le développement de la carte des transcrits : de très nombreux gènes vont être localisés sur les chromosomes. Leur localisation dans des zones du génome où des gènes majeurs ou des QTL ont été identifiés par les méthodes génétiques classiques va simplifier considérablement l'identification de ces gènes majeurs et QTL en passant d'un laborieux clonage positionnel au test de candidats positionnels.

## Références

- Bachem C.W., van der Hoeven R.S., de Bruijn S.M., Vreugdenhil D., Zabeau M., Visser R.G., 1996. Visualization of differential gene expression using a novel method of RNA fingerprinting based on AFLP: analysis of gene expression during potato tuber development. *Plant Journal*, 9, 745-753.
- Corpet F., Chevalet C., 2000. Analyse informatique des données moléculaires. INRA Productions Animales, numéro hors série « Génétique moléculaire : principes et application aux populations animales », 191-195.
- DeRisi J.L., Iyer V.R., Brown P.O., 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278, 680-686.
- Diatchenko L., Lau Y.F., Campbell A.P., Chenchik A., Moqadam F., Huang B., Lukyanov S., Lukyanov K., Gurskaya N., Sverdlov E.D., *et al*, 1996. Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proceedings of the National Academy of Sciences of the USA*, 93, 6025-6030.
- Gerhold D., Rushmore T., Caskey C.T., 1999. DNA chips: promising toys have become powerful tools. *Trends in Biochemical Sciences*, 24, 168-173.
- Hatey F., Tosser-Klopp G., Cloucard-Martinato C., Mulsant P., Gasser F., 1998. Expressed sequence tags for genes: a review. *Genetics Selection Evolution*, 30, 521-541.
- Hatey F., Martin P., Douaire M., Le Gac F., Dambrine G., Herpin P., Monget P., 2000. Le programme Asteroger : vers un outil multifonctionnel pour les productions animales. INRA Productions Animales, numéro hors série « Génétique moléculaire : principes et application aux populations animales », 175-180.
- Hubank M., Schatz D.G., 1994. Identifying differences in mRNA expression by representational difference analysis of cDNA. *Nucleic Acids Research*, 22, 5640-5648.
- Jordan B., 1998. Voyage au pays des puces. *Médecine/sciences*, 14, 1097-1102.
- Liang P., Pardee A.B., 1992. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science*, 257, 967-971.
- Pitel F., Riquet J., 2000. Les marqueurs anonymes et la détection de leur polymorphisme. INRA Productions Animales, numéro hors série « Génétique moléculaire : principes et application aux populations animales », 45-53.
- Riquet J., Pitel F., 2000. Les techniques de base de la génétique moléculaire. INRA Productions Animales, numéro hors série « Génétique moléculaire : principes et application aux populations animales », 29-35.
- Velculescu V.E., Zhang L., Vogelstein B., Kinzler K.W., 1995. Serial analysis of gene expression. *Science*, 270, 484-487.
- Velculescu V.E., Madden S.L., Zhang L., Lash A.E., Yu J., Rago C., Lal A., Wang C.J., Beaudry G.A., Ciriello K.M., *et al*, 1999. Analysis of human transcriptomes. *Nature Genetics*, 23, 387-388.
- Vignal A., 2000. Bases de données en biologie. INRA Productions Animales, numéro hors série « Génétique moléculaire : principes et application aux populations animales », 187-189.
- Yerle M., 2000. Etablissement des cartes cytogénétiques et physiques. INRA Productions Animales, numéro hors série « Génétique moléculaire : principes et application aux populations animales », 87-93.