

# INTERVENANTS

## Mardi 20 Mai

- Génomique structurale : état de la cartographie des 4 génomes (bovin, porc, poule, truite) *André Eggen* Page 2
- Les dispositifs d'appui : Centre de ressources et plates-formes *P. Chardon* Page 3
- Présentation de SIGENAE (Système d'information d'AGENAE) *Ch .Klopp* Page 4
- Roger, au-delà d'un prénom, une volonté affirmée de FEDERER *P. Martin* Page 5
- « Stratégie pour l'identification de gènes impliqués dans le contrôle de la croissance et de la composition corporelle chez le poulet », programme en collaboration entre l'Université du Delaware (USA) et l'INRA. *Michel Duclos* Page 6
- Création d'une macro-array spécifique du système immunitaire du porc : application à l'étude de la réponse colibacillaire chez le porcelet *Isabelle Oswald* Page 7

## Mercredi 21 Mai

- La physiologie à l'ère du post-génome *Philippe Monget* Page 8
- Biologie intégrative et bioinformatique *Jean-François Gibrat* Page 9
- A SAGE saga : the human nephron gene expression database *J-M Elalouf* Page 10
- Exploitation informatique et interprétation des données SAGE *J. Marti* Page 11

# Génomique structurale : état de la cartographie des 4 génomes (bovin, porc, poule, truite)

André Eggen, Laboratoire de Génétique biochimique et de Cytogénétique,  
Département de Génétique Animale, INRA, Jouy-en-Josas

Au cours de la dernière décennie, la génomique animale a connu un essor considérable tant au niveau structural, que fonctionnel. Ces avancées sont en grande partie à attribuer aux progrès technologiques et stratégiques du Projet Génome Humain. Ainsi, outre l'intérêt fondamental de contribuer à une meilleure connaissance de la structure des génomes et de ses gènes exprimés, l'objectif majeur d'identifier des gènes d'intérêt agronomique et de caractériser, au niveau moléculaire, la variabilité génétique présente dans les différentes populations, est devenu réalité.

Pour la génomique structurale, les avancées se sont appuyées sur de nouveaux outils venant enrichir la panoplie du cartographe :

- Les études de polymorphisme de l'ADN ont pris une nouvelle dimension avec la découverte et l'identification des marqueurs génétiques de type microsatellite, et depuis peu des SNP (Single Nucleotide Polymorphism).
- Les progrès technologiques pour la détection de molécules d'ADN facilite une montée en puissance dans les études sur des populations spécifiques (primo-localisation de gènes d'intérêt) ou des populations de référence (construction de cartes génétiques) : il est enfin possible de génotyper de nombreux individus pour de nombreux marqueurs.
- Le développement de panels d'hybrides irradiés pour plusieurs des espèces animales permet la construction de cartes de marqueurs et de gènes de meilleure résolution que les cartes génétiques classiques avec des facilités technologiques évidentes.
- La construction de collections de grands fragments d'ADN, particulièrement des chromosomes artificiels de bactérie (BAC), favorise grandement la caractérisation d'une région particulière et l'identification de nouveaux marqueurs spécifiques d'une région d'intérêt.
- Les collections de BAC autorisent aujourd'hui la réalisation, dans un délai raisonnable, d'une carte physique globale du génome d'une espèce d'intérêt. Cette carte physique, ensemble de fragments chevauchants pour tout le génome, donne ainsi un accès immédiat à un quelconque segment du génome. Elle constitue de plus le point de départ pour l'étude de l'organisation à grande échelle d'un génome et constitue également l'étape de base pour le séquençage complet de ce génome.
- La génomique comparative enfin joue un rôle essentiel : elle permet de tirer profit des génomes déjà bien caractérisés, en particulier les génomes humain et murin pour lesquels les séquences complètes sont disponibles ainsi que de nombreuses données fonctionnelles et physiologiques. Des passerelles ont ainsi été construites entre les génomes humain et murin et les génomes de nos animaux domestiques, permettant une extrapolation de données.

Tous ces outils de génomique structurale contribuent à accélérer la caractérisation moléculaire de la variabilité génétique. Il convient toutefois de les intégrer aux outils récents de la génomique fonctionnelle et de permettre ensuite une meilleure gestion de la diversité génétique de nos espèces et de nos populations animales.

## Les dispositifs d'appui : Centre de ressources et plates-formes

*P Chardon, F. Piumi, C. Rogel Gaillard.*

Durant les deux dernières années, plusieurs plateaux techniques ont été mis en place comme soutien logistique des programmes développés dans le cadre d'AGENAE. L'objectif est de faciliter l'accès à un ensemble d'outils moléculaires et informatiques pour approfondir les connaissances sur les génomes, sur le polymorphisme des populations et sur le niveau d'expression des gènes.

Un centre de ressources biologiques a ainsi été ouvert à Jouy en Josas. C'est un lieu de conservation et de gestion centralisée des collections de clones bactériens contenant des fragments d'ADN des espèces d'intérêt. Actuellement près de deux millions de clones ont été rassemblés et sont à disposition des scientifiques pour l'étude des principales espèces de mammifères d'élevage, du poulet et de la truite. Les collections produites pour l'essentiel par des laboratoires de l'INRA ont été enrichies avec des clones construits par des équipes étrangères. Elles sont d'une aide précieuse pour l'étude de la structure des génomes, la construction des cartes génétiques ou le clonage de position et sont à la base des recherches sur l'expression des gènes. La deuxième mission du centre de ressources est la préparation de l'ADN et la fabrication des réseaux nécessaires aux études fonctionnelles. Quatre espèces majeures sont concernées : le porc, le bovin, le poulet et la truite.

Des plateaux techniques ont également été développés dans plusieurs centres INRA, à proximité des équipes utilisatrices. Jouy en Josas, Limoges, Theix, Toulouse, Tours et Rennes bénéficient ainsi des équipements nécessaires à l'utilisation des réseaux d'ADN et les moyens informatiques nécessaires à l'analyse. Plus que de simples supports techniques, ces plates-formes regroupent aussi savoir-faire, compétences et expertise pour l'étude du transcriptome.

Le fonctionnement du centre de ressources et des plates-formes est coordonné par un comité mis en place par AGENAE. Il est chargé du suivi de l'état des équipements et des nouveaux investissements indispensables à la conduite des travaux. La compatibilité des développements bio-informatiques et la mise en place de plans de formations font partis de ses attributions. Le comité procède aussi à l'évaluation de la faisabilité des projets soumis au programme AGENAE, compte tenu des moyens technologiques disponibles

Les plateaux techniques et le centre de ressources sont encore à un stade de montée en puissance, tant au niveau de l'équipement que des ressources humaines ou de l'acquisition des compétences. Ils ne peuvent fonctionner qu'en étroite interaction et ont tout bénéfice à s'appuyer sur des réseaux déjà existants, le réseau des Génopoles par exemple.

## **Présentation de SIGENAE (Système d'Information d'AGENAE)**

La mise en oeuvre de la génomique à grande échelle nécessite une infrastructure informatique capable de stocker et de traiter les données générées.

AGENAE, pour se faire, s'est doté d'une équipe de bio-informaticiens et de bio-analystes regroupés au sein de SIGENAE (Système d'Information d'AGENAE).

L'équipe se compose actuellement de six personnes dont quatre permanents. L'équipe est animée par Christophe Klopp ([christophe.klopp@toulouse.inra.fr](mailto:christophe.klopp@toulouse.inra.fr)).

Les membres de l'équipe sont répartis sur les sites hébergeant les pôles les plus importants de la génomique animale à l'INRA (Jouy-en-Josas, Rennes et Toulouse).

SIGENAE met à la disposition des biologistes, au travers d'un serveur centralisé <http://sigena.jouy.inra.fr> (d'accès restreint), des données et des modules de traitements nécessaires à leur recherches.

Les premiers travaux de l'équipe ont été ciblés sur le nettoyage, l'assemblage, l'annotation et la publication des séquences d'étiquettes.

Plus de 900 000 séquences figurent actuellement dans la base de données ; dont 113 314 appartenant aux équipes d'AGENAE. Les autres sont issues des banques de données publiques.

Les nouveaux outils mis à disposition en 2003 sont un environnement d'annotation manuelle des séquences et un logiciel de gestion des données des puces à ADN.

SIGENAE assure aussi la formation des utilisateurs sur les outils installés et réalise des travaux à façon suite aux demandes des équipes.

**Christophe KLOPP**

## ROGER, au -delà d'un prénom, une volonté affirmée de FEDERER

Patrice Martin\*, Sophie Blanchet, Isabelle Cassar-Malek, Yves Chilliard, Séverine Degrelle, Anne Listrat, Jean François Hocquette, Isabelle Hue, Aude Laisné, Johan Laubier, Fabrice Lepage, Fabienne Le Provost, Christine Leroux, Elisabeth Petit, Geneviève Piétu, Sophie Pollet, Jean Paul Renard, Gaël Rolland, Naomi Seely, Karine Sudre, Cristina Veltri, Emmanuelle Zalachas & Charles Auffray.

\*INRA, Département de Génétique animale, Unité Génomique & Physiologie de la Lactation (GPL)

CRJ, Bâtiment 22, Domaine de Vilvert, 78352 Jouy-en-Josas Cedex

L'objectif initial du programme ROGER, lancé début 1998, était de produire un premier répertoire ordonné de 1000 gènes représentatifs du génome bovin exprimé dans la glande mammaire, l'embryon et le muscle, de façon à permettre l'établissement de profils transcriptionnels de ces tissus dans différentes situations physiopathologiques, en utilisant la technologie des macro-réseaux d'ADNc (filtres à haute densité ou « macroarrays »). La finalité de ces recherches est d'aider à l'identification de gènes impliqués dans le développement et l'expression de certains phénotypes ou caractères d'intérêt pour la filière bovine.

Organisé en réseau de collaborations, impliquant initialement 3 laboratoires INRA (Génétique Biochimique et Cytogénétique, Biologie du Développement et Biotechnologie, INRA, Jouy-en-Josas, et Unité de Recherche sur les Herbivores, INRA, Theix), et une équipe du CNRS (Genexpress/IMAGE, FRE 2376, Villejuif), ROGER a, dans un second temps, été étendu à 2 autres unités INRA (Station d'Amélioration Génétique des Animaux, Castanet-Tolosan-Toulouse, et Pathologie Infectieuse et Immunologie, Nouzilly-Tours), engagées dans des programmes de recherches sur les Encéphalopathies Spongiformes Subaiguës Transmissibles.

Cette seconde phase du programme (ROGER-2), a permis l'obtention de premiers "profils d'expression" (en hétérologue notamment, ( Le Provost *et al.*, 2000, Blanchet *et al.*, 2000, Sudre *et al.*, 2003, ) mais aussi sur "oligoarrays" en fluorescence) dont quelques exemples seront présentés. Elle a également contribué à renforcer les interactions entre les équipes participantes, et abouti à la résolution de produire une membrane commune (MEMbrane pilote) permettant de collecter près de 2000 signatures expressionnelles dans le Muscle, l'Embryon et la Mamelle.

Dans la continuité de cette action un programme ROGER-3, plus particulièrement consacré au thème « différenciation tissulaire » a été proposé<sup>1</sup>, dans le cadre du lancement d'AGENAE en 2001. Ce programme, actuellement en cours, se décompose en 2 phases. La première vise à développer des répertoires génériques de 2000 et 5000 sondes d'ADNc, respectivement construites à partir de collections spécifiques INRA (Mamelle, Embryon et Muscle) et de collections USDA (multi-tissus). La seconde phase concerne la recherche de gènes (éventuellement communs) impliqués dans les mécanismes de différenciation des tissus musculaire, adipeux et mammaire et dans le développement de l'embryon pré-implantatoire. Même si des avancées significatives ont été enregistrées dans la connaissance des processus myogéniques, développementaux, et de différenciation terminale de la cellule épithéliale mammaire, bon nombre d'effecteurs et de mécanismes intervenant dans la différenciation et la relation avec l'environnement cellulaire, restent à découvrir.

---

<sup>1</sup> A la faveur de cette opération, deux nouveaux programmes, orientés respectivement sur la tremblante ovine (SCRAP'ARRAY), et sur la qualité de la viande bovine (MUGENE), se sont individualisés de ROGER. Un troisième projet portant sur la compréhension des phénomènes de mortalité embryonnaire (répertoire ordonné de transcrits extra-embryonnaires, fœtaux et endométriaux représentatifs des premiers mois de gestation chez le bovin) est en cours d'élaboration (Génanimal 2003). *SCRAP'ARRAY a pour objectif de produire un répertoire ordonné de gènes pour la production, à haut débit, de profils transcriptionnels et l'identification de marqueurs de l'étiologie de la tremblante naturelle ovine. Quant à MUGENE, dans sa nouvelle configuration (Génanimal 2003), il propose une approche intégrée combinant la génétique, la génomique et la biologie du muscle pour gérer la qualité de la viande bovine.*

**"Stratégie pour l'identification de gènes impliqués dans le contrôle de la croissance et de la composition corporelle chez le poulet",  
programme en collaboration entre l'Université du Delaware (USA) et l'INRA.**

**Michel J. Duclos & Jean Simon**  
SRA - INRA - 37380 Nouzilly

Un programme intitulé « Consortium pour la cartographie fonctionnelle des gènes régulant la croissance des poulets de type chair » est en cours. Il est financé par l'USDA dans le cadre de l'Initiative for Future Agriculture and Food Systems (IFAFS) et coordonné par le Pr L.A. Cogburn (Université du Delaware, Newark, Delaware, USA).

Trois universités américaines (Delaware, Georgia et Maryland) et trois unités INRA y participent : Station de Recherches Avicoles de Nouzilly, UMR Génétique Animale de Rennes, Laboratoire de Génétique Cellulaire de Toulouse (responsable français J. Simon, SRA, INRA).

Les étapes du projet sont les suivantes :

1. Réalisation de banques normalisées par tissu (foie, muscle, tissu adipeux, hypophyse et hypothalamus)
2. Séquençage de 5000 gènes par tissu et réalisation de filtres à haute densité pour l'étude du transcriptome de ces différents tissus
3. Comparaison du transcriptome dans les mêmes tissus à différents stades du développement entre 4 génotypes expérimentaux INRA de poulets sélectionnés de façon divergente pour la croissance (poulets lourds ou légers) ou l'engraissement (poulets maigres ou gras)
4. Recherche de QTL (croissance, composition corporelle, qualité de la viande et quelques paramètres physiologiques) sur des poulets issus des deux croisements : lignée lourde \* lignée légère d'une part, lignée grasse \* lignée maigre, d'autre part

Les deux premières étapes sont achevées, une collection de 40,000 clones d'ADNc est disponible avec environ 19,000 séquences non redondantes.

Une première génération de filtres à haute densité portant l'empreinte de 3000 ADNc hépatiques a été utilisée dans différentes études pilotes.

L'une d'elle a permis une première évaluation de la régulation nutritionnelle du transcriptome hépatique.

Les résultats suggèrent qu'un nombre assez important de gènes soit affecté différemment par l'état nutritionnel selon le génotype, lourd ou léger.

Très prochainement des puces à ADN portant l'empreinte d'une collection d'environ 10,000 ADNc communs aux tissus adipeux, musculaire et hépatique seront disponibles pour étudier le transcriptome de ces tissus.

Les gènes candidats fonctionnels révélés par cette stratégie seront ensuite évalués comme candidats positionnels à la lumière des résultats des études QTL.

## Création d'une macro-array spécifique du système immunitaire du porc : application à l'étude de la réponse colibacillaire chez le porcelet.

Neil Ledger<sup>1</sup>, Ionelia Taranu<sup>1</sup>, Anna Malik<sup>2</sup>, Béla Nagy<sup>2</sup> et Isabelle Oswald<sup>1</sup>

<sup>1</sup>INRA, Laboratoire de Pharmacologie-Toxicologie, Toulouse, FRANCE, <sup>2</sup>Veterinary Medical Research Institute, Budapest HONGRIE.

Les cytokines sont des médiateurs de nature glycoprotéique impliquées dans les communications intercellulaires, notamment dans le fonctionnement et la régulation du système immunitaire. Elles interviennent aussi dans les relations entre les cellules de l'immunité et le système nerveux ainsi que dans les mécanismes physiologiques de la reproduction.

Les nouvelles technologies (SAGE, macro- et micro-array...) permettent une approche systématique du transcriptome. Les collections d'expressions géniques mesurent simultanément le niveau d'expression de centaines voire de milliers de gènes. Le but de ce projet est de constituer une collection d'ADN complémentaires porcins spécifique du système immunitaire sur membrane de Nylon. En effet si de tels outils sont déjà commercialement disponibles pour analyser la réponse immunitaire de l'homme et des rongeurs de laboratoire (Atlas System de Clontech, GEM Microarray de Genome Systems...) il n'en est pas de même chez les animaux domestiques. Ces macro-array constituent un outil d'analyse très puissant de la réponse de l'hôte. Chez le porc, elles permettront une analyse globale de l'expression des ARNm des gènes de la réponse immune : chimiokines, cytokines inflammatoires ou caractéristiques des réponses Th1, Th2..., avec un gain de temps et d'échelle.

En utilisant les séquences porcines disponibles dans GenBank et les EST porcines publiées dans TIGR, nous avons cloné dans un vecteur commercial classique, les gènes de la plupart des cytokines porcines ainsi qu'une trentaine de gènes porcins impliqués dans la réponse immunitaire. La séquence de certaines cytokines porcines (IL-3, -9 ou -17), de la plupart des chimiokines et de leur récepteurs ne sont pas actuellement disponibles. Pour ces gènes, nous avons aligné les séquences obtenues dans d'autres espèces, puis dessiné des amorces consensus dans les zones ayant une forte homologie inter-espèce. Tous ces clones ont ensuite été séquencés afin de confirmer la nature de l'insert. A ce jour, notre collection comprend plus de 80 gènes :

35 gènes codant pour des cytokines/chimiokines et leur récepteur

24 gènes codant pour des marqueurs d'activation et autres gènes impliqués dans la réponse immunitaire

23 gènes de ménage et contrôles

Nous avons ensuite validé cette macro-array en comparant l'expression des gènes sur des cellules de sang périphérique porcins. Ces cellules composées principalement de lymphocytes et de macrophages ont été stimulées ou non par un mélange mitogénique (PMA + Ionomycine). Comme attendu nous avons montré que la stimulation mitogénique induit l'expression de nombreuses cytokines (IFN-g, TNF, IL-10...).

Nous utilisons actuellement cette puce pour analyser les réponses intestinales chez des porcelets infectés par différentes souches d'*Escherichia coli*. Nos résultats indiquent qu'*E. coli* stimule spécifiquement la production d'IL-8 au niveau de l'épithélium intestinal et que cette stimulation est liée au niveau de pathogénicité des souches. Ce résultat a été confirmé par RT-PCR.

## La physiologie à l'ère du post-génome

Les grands programmes de séquençage des génomes modèles (nématode, drosophile, homme, bientôt souris) et des étiquettes chez différentes espèces dont les animaux domestiques (bovins, porcins, poule, truite), changent progressivement notre manière de concevoir l'étude des fonctions physiologiques.

Comme dans tous les cas où des avancées technologiques majeures sont réalisées, cette nouvelle ère de la biologie, appelée post-génome, a enfanté de nouvelles disciplines et des termes nouveaux qui leur sont associés.

Après le génome, sont apparus dans le vocabulaire les termes nouveaux de transcriptome, protéome, métabolome, l'interactome etc... termes censés donner une dimension "massive" ou "à haut débit" aux disciplines correspondantes.

La "génomique fonctionnelle", qui sous-tend une étude rapide, globale, exhaustive et massive de la fonction des gènes, a pour but, entre autres, de réconcilier le décalage qui existe entre la vitesse et la relative facilité avec laquelle des milliers de séquences sont obtenues, et les efforts qu'il est nécessaire de déployer pour connaître le rôle exact d'une seule protéine.

De même, le terme de "Biologie intégrative" a récemment été inventé, sans qu'il n'y ait un réel consensus sur sa signification exacte.

L'intégration en question peut en effet être celle des différents niveaux d'analyse (moléculaire, cellulaire, organisme entier), ou celle des différentes stratégies (génétique, physiologie conventionnelle, bioinformatique) désormais à la disposition des biologistes pour étudier une fonction.

Il y a donc une certaine confusion dans la terminologie nouvelle qui émerge.

Plutôt que de s'étendre sur ces questions de sémantique, nous essaierons d'apporter ici quelques éléments de réflexion sur les questions qui nous semblent réellement pertinentes :

- Qu'est ce que les données du génome apportent à l'étude de la physiologie ? En d'autres termes, que pouvons nous faire aujourd'hui grâce à ces nouvelles données, que nous ne pouvions faire hier ?
- Comment extraire de ce nombre incalculable de données nouvelles, pour la plupart stockées dans les bases de données, les informations pertinentes ?

Parmi les stratégies nouvelles ou à développer issues des programmes de séquençage, au moins trois grands types semblent se dégager :

1. L'étude du génome expressionnel
2. L'exploitation généralisée et systématique de la variabilité phénotypique et l'étude de son déterminisme génétique
3. La biologie in silico

**Philippe MONGET**

## **Biologie intégrative et bioinformatique**

Les projets de biologie intégrative produisent des masses énormes de données hétérogènes. Ces données, en tant que telles, n'ont qu'un intérêt limité.

Le but ultime poursuivi par les biologistes qui entreprennent ces projets est de comprendre comment le génome d'un organisme particulier permet d'expliquer ses propriétés biologiques.

Il s'agit donc, d'une façon générale, de passer de données à des connaissances biologiques ou, exprimé d'une manière différente, d'aller du génome au phénotype en explorant tous les niveaux intermédiaires d'organisation du vivant : la cellule, le tissu, l'organe, etc.

La bioinformatique joue un rôle central dans ces projets de biologie intégrative.

Sa fonction première, qui est le rôle traditionnellement dévolu à l'informatique, est de gérer, organiser, stocker, bref, permettre de manipuler d'une manière efficace les données produites.

Cela concerne principalement le développement de bases de données et d'interfaces homme-machine.

La deuxième fonction de la bioinformatique consiste à développer des méthodes permettant d'analyser les données produites de façon à en extraire des connaissances (ce que les Anglo-Saxons nomment <<computational biology>>).

Le séminaire présentera ces 2 aspects en essayant de montrer l'apport des techniques de la bioinformatique pour répondre à certaines questions que peuvent se poser les biologistes.

**Jean-François GIBRAT**

## **A SAGE saga: the human nephron gene expression database**

Département de Biologie Joliot-Curie,  
Service de Biochimie et de Génétique Moléculaire,  
CEA Saclay, 91191 Gif-sur-Yvette cedex, France.

The human kidney consists of one million nephrons functioning in parallel to ensure efficient body fluid homeostasis.

This vital function rests on sequential blood filtration by the glomerulus, and specific transport processes accomplished by the successive nephron segments.

Because of axial nephron functional segmentation, studies carried out at the whole kidney level cannot define sites and mechanisms of physiological processes.

To obtain comprehensive transcriptomes for the human nephron, we isolated mRNA tags from the glomerulus and seven different nephron segments microdissected from fresh kidney pieces.

The 414,000 mRNA tags sequenced included 33,000 tags detected twice or more in the dataset. Expression of genes responsible for nephron transport and permeability properties was evidenced through transcripts for 116 solute carriers, 75 ion channels, 43 ion-transport ATPases, and 12 claudins.

Searching for differences between expression profiles, we found 998 transcripts greatly varying in abundance from one nephron portion to another. Approximately 60% of them correspond to characterized genes, including 75 morbid genes.

This systematic large-scale analysis of individual structures of a complex human tissue reveals sets of genes underlying specifically located functions. It also helps in sorting candidate genes for diseases uncharacterized at the molecular level.

**Jean-Marc ELALOUF**

## Exploitation informatique et interprétation des données SAGE

Institut de Génétique Humaine, UPR CNRS 1142 Montpellier.

La méthode SAGE est basée sur l'analyse séquentielle d'un grand nombre de signatures courtes (tags) extraites de l'extrémité 3' des ADNc. Les tags assemblés par paires (ditags) sont amplifiés par PCR et assemblés en concatémères pour l'analyse séquentielle.

L'exploitation des données comprend 3 étapes: extraction et dénombrement des séquences de tags des fichiers de séquence, identification des gènes, interprétation fonctionnelle. Nous avons développé des outils pour chaque étape.

La première consiste à caractériser concatémères et ditags, puis à dénombrer les tags, chaque ditag étant borné par les 4 bases du site de coupure initialement utilisé pour libérer le tag.

Le logiciel (Digitag) que nous avons développé fournit, outre le recensement des tags, des critères de qualité des banques SAGE. L'analyse de taille des concatémères permet, dès les 1000 premiers tags, d'évaluer le coût ultérieur du séquençage, incitant éventuellement à reconstruire la banque pour obtenir des séquences plus longues à même coût unitaire.

Du fait de la complexité des populations d'ARNm, chaque combinaison de tags est en principe unique. L'observation répétée de ditags identiques indiquera une faible complexité de la banque et des concatémères identiques signaleront une amplification excessive des clones bactériens, amenant à sous-estimer la complexité naturelle des ARNm. Enfin, la présence de ditags deux fois trop longs, dus à la perte du site séparateur par erreur de séquence, donne une mesure de ces erreurs.

Pour identifier les gènes correspondant, nous développons une base de données (Preditag) regroupant tous les tags virtuels susceptibles d'être observés d'après la connaissance du transcriptome de l'espèce considérée. La version initiale utilise les séquences de référence d'UniGene. Cependant, l'assemblage des séquences de GeneBank en clusters représentatifs des produits de transcription (ADNc, ESTs) est imparfait et des gènes connus ne sont pas représentés par la séquence attendue.

Face à l'instabilité chronique d'UniGene, nous avons construit une base (UniSage) regroupant les tags obtenus à partir de 6 millions de séquences en accès public, soit 437 625 tags différents pour *Homo sapiens*. Ce nombre n'est pas incompatible avec un nombre de gènes de l'ordre de 30 000.

Parmi les causes de cet écart, certaines sont méthodologiques, les tags observés une seule fois pouvant provenir d'erreurs de séquence, tandis que des tags en amont résultent d'une coupure incomplète sur le premier site en 3'.

D'autres causes sont naturelles : diversité génétique accroissant le nombre de signatures, sites multiples de polyadénylation engendrant plusieurs transcrits pour un même gène; pseudo-gènes exprimés.

Enfin, nous avons montré que l'existence naturelle de transcrits en orientation inverse (anti-sens) accroît la complexité de la population de tags. Ces facteurs contribuent à l'instabilité d'UniGene et le problème du nombre exact de transcrits de référence n'est pas résolu.

Parmi les gènes identifiés chez les mammifères, environ 12 000 sont annotés sur le plan fonctionnel, notamment grâce aux projets HUGO, G.O. et KEGG, qui nous ont permis de développer un système d'annotation automatique facilitant l'interprétation des données SAGE. En effet, celle-ci s'avère très laborieuse si on s'en tient aux informations individuelles associées à chaque gène.

La prochaine étape consistera à dresser des tables de gènes et de tags orthologues pour comparer les données SAGE obtenues à partir de différentes espèces animales.

Jacques MARTI