D. MILAN

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

e-mail : Denis.Milan@toulouse.inra.fr

4 - Recherche de gènes associés à des fonctions

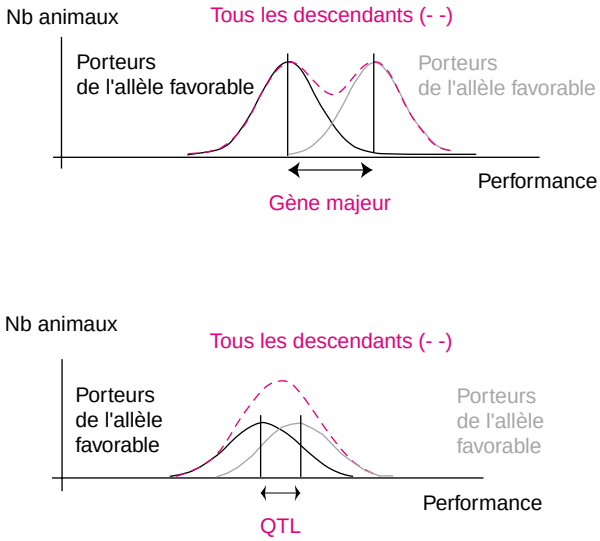
Notion de gène candidat

Résumé. *Lorsque l'analyse de performances permet de montrer qu'un gène gouverne une part importante de la variabilité d'un caractère, diverses approches sont possibles pour l'identifier : l'étude de candidats physiologiques afin d'identifier le gène recherché parmi les gènes connus intervenant dans ce caractère ; la cartographie fine de la région pour déterminer très précisément la position du gène recherché, jusqu'à ne plus trouver qu'un seul gène à cet endroit (démarche de clonage positionnel). Le plus souvent, le gène responsable est identifié par une combinaison de ces deux approches, en trouvant un candidat positionnel situé dans l'intervalle restreint où il est attendu. Dans certains cas cependant, l'étude de gènes connus ayant un effet similaire dans une autre espèce, permettra d'identifier rapidement le gène responsable.*

Les techniques de la génétique quantitative considèrent que les caractères sont gouvernés par un nombre infini de gènes, tous génétiquement indépendants, ayant chacun un effet infinitésimal sur le caractère d'intérêt. Pour chacun de ces gènes, il existerait deux allèles, l'un favorable et l'autre défavorable. La valeur génétique d'un animal serait alors fonction de la proportion de gènes pour lequel l'animal possède un allèle favorable. Basés sur ce modèle, les programmes d'indexation génétique ont pour but d'estimer la valeur génétique des candidats reproducteurs, afin d'identifier les animaux améliorateurs.

Cette approche statistique ignore donc les gènes réellement impliqués dans le caractère et suppose que chacun a un effet infime. Il est évident que ce modèle n'est pas exact, mais il a permis et permet de réaliser un fort progrès génétique dans les races domestiques, principalement pour les caractères à forte héritabilité pour lesquels les performances des animaux sont fortement liées à la valeur génétique additive ainsi estimée. Pour certains caractères cependant, il est évident que certains des 50 000 à 100 000 gènes de l'animal ont un effet prépondérant pour expliquer les performances des animaux. Ceci est clair si l'on regarde une population dans laquelle un gène de nanisme, par exemple, est en ségrégation. D'une manière générale, on parle dans ce cas de gènes à effet majeur ou, par raccourci, de gène majeur, lorsque l'effet du gène est observable sur les performances des animaux (figure 1). C'est par exemple le cas des gènes de nanisme chez la poule, d'hypertrophie musculaire chez le bovin, de cornage chez la chèvre, d'hypertrophie musculaire/sensibilité à l'halothane chez le porc, de prolificité Booroola chez le mouton ... Dans de tels cas, la démonstration de la ségrégation d'un gène à effet majeur peut être faite en mesurant les performances des descendants d'un animal hétérozygote.

Figure 1. *Descendance d'un animal hétérozygote pour un gène majeur ou pour un QTL.*



Dans d'autres cas, certains gènes ont des effets intermédiaires sur le caractère, mais ne peuvent être mis en évidence par la seule analyse des performances des descendants d'un animal hétérozygote. Ils sont appelés QTL (pour Quantitative Trait Locus), indiquant que ce locus intervient dans le déterminisme d'un caractère semblant purement quantitatif. On emploie ici le terme de locus pour indiquer que cette zone du génome a un effet sur le caractère étudié, sans que l'on sache si un seul ou plusieurs gènes formant un haplotype sont responsables de l'effet observé. Pour mettre en évidence ces locus à effet moyen, outre la mesure des performances, il faut

disposer des génotypes des marqueurs génétiques sur le(s) parent(s) et sur les descendants au voisinage immédiat du QTL. L'analyse consiste alors à déterminer si les animaux ayant reçu un allèle parental ont des performances significativement différentes des descendants ayant reçu l'autre allèle (figure 1).

Lorsqu'un gène (gène majeur ou QTL) a un effet significatif sur un caractère, il est évident que la prise en compte du génotype des animaux pour ce gène est importante pour sélectionner les meilleurs reproducteurs. L'un des objectifs de la génétique moléculaire est d'identifier ces gènes ayant un effet significatif sur les caractères d'intérêt zootechnique. Il faut pour cela identifier un ou plusieurs gènes candidats, puis tester leur éventuel effet sur le caractère étudié. On distingue deux types principaux de gènes candidats en fonction des informations conduisant à étudier ce gène : les candidats physiologiques choisis en fonction de la connaissance du caractère et du rôle du gène candidat, et les candidats positionnels choisis en fonction de leur position sur le génome à l'endroit où le gène majeur non identifié a été cartographié.

1 / Notion de candidat physiologique

Face à un gène majeur donné, l'approche physiologiste revient à rechercher parmi les gènes connus si l'un d'entre eux est susceptible d'expliquer les variations de performances observées. Cette approche repose donc sur une étude physiologique fine de l'effet du gène majeur. Un exemple classique d'une telle approche est, chez l'Homme, l'identification du gène responsable de la phénylcétonurie induisant un retard mental chez les enfants atteints. Chez ces enfants, il a été montré que cette maladie était due à la non fonctionnalité de l'enzyme phénylalanine-hydroxylase. Chez les animaux domestiques, une telle étude de candidats physiologiques a permis de montrer que la mutation du récepteur de l'hormone de croissance était responsable du nanisme chez la poule. Néanmoins, dans la plupart des cas, une telle approche ne permet pas d'identifier le gène responsable des effets majeurs observés. Dans le cas de l'étude du gène RN (Milan *et al* 2000, cet ouvrage), l'étude biochimique de l'activité de quatre enzymes clés du métabolisme du glycogène n'a pas permis d'identifier le gène responsable (Estrade 1994).

Ce criblage de candidats physiologiques, s'il peut permettre dans certains cas d'identifier très rapidement le gène responsable, peut également être très laborieux, voire infructueux si le gène responsable n'est pas connu.

Dans les cas décrits jusqu'à présent, cette étude de candidats est réalisée sur une population dans laquelle la ségrégation d'un gène majeur a été mise en évidence. Par extension, l'étude de candidats physiologiques peut être entreprise sur des populations où aucun gène majeur n'a été suspecté, mais où l'on espère la présence de QTL. Dans ces études, les auteurs recherchent une mutation dans ou au voisinage immédiat d'un gène candidat, puis testent une éventuelle différence de performances entre les animaux présentant l'un ou l'autre allèle du candidat. Max Rothschild a popularisé cette approche en montrant sur certaines lignées de porc un effet du polymorphisme du gène ESR (estradiol receptor) sur la prolificité des truies, avec un effet de 0,42 à 1,15 porcelet par portée dans certaines lignées (Rothschild *et al* 1996). Dans ce cas, la mutation responsable de l'effet n'est toutefois pas clairement identifiée.

2 / Notions de clonage positionnel et de candidat positionnel

Au cours des dix dernières années, le développement des cartes génétiques des principales espèces (humaine et murine, mais aussi bovine, ovine, caprine, porcine, et poule), permet de disposer de marqueurs génétiques très polymorphes couvrant le génome (Pitel et Riquet 2000 et Vaiman 2000, cet ouvrage). En étudiant les familles informatives pour la ségrégation des allèles du gène majeur ou du QTL étudié, il est alors possible de cartographier le gène d'intérêt.

Pour des gènes d'intérêt ayant un effet majeur, il est possible d'entreprendre une cartographie fine du gène afin de déterminer le plus précisément possible la position du gène recherché (voir la 4e partie de cet ouvrage). Chez l'Homme, le cas du gène DTD responsable de la dysplasie diatrophique est fameux : des études de déséquilibre de liaison avait prédit sa présence à 64 kb du marqueur génétique disponible le plus proche ; il a finalement été identifié à 70 kb de ce marqueur (Hastbacka *et al* 1994). Ce gène n'était pas connu auparavant, il a été identifié par une démarche dite de clonage positionnel.

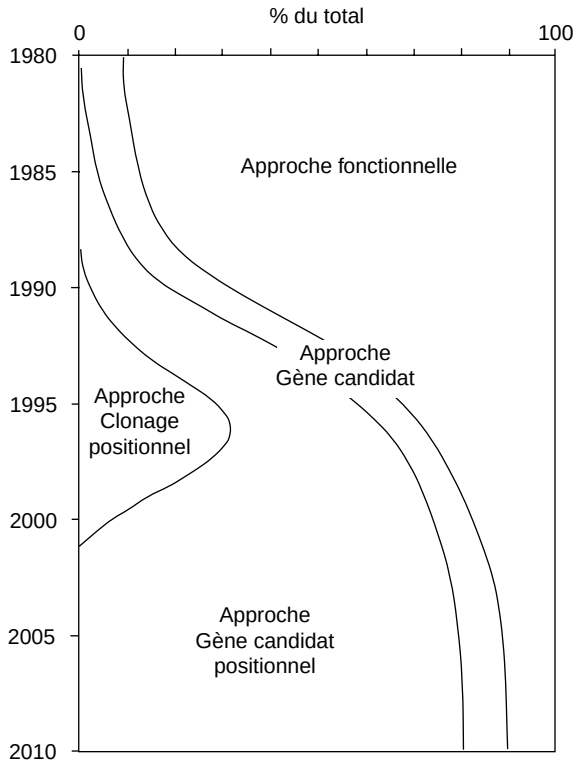
Lorsque l'on a déterminé la position du gène recherché, et que l'on teste un gène déjà connu situé à cet endroit, on parle d'une démarche de candidat positionnel. La mise en évidence d'un gène candidat positionnel nécessite que le gène en question soit connu et cartographié à l'endroit attendu, ce qui justifie les démarches de cartographie systématique entreprises sur de nombreuses espèces (voir la 3e partie de cet ouvrage). Alternativement, les avancées des travaux de cartographie comparée donnent maintenant une bonne idée de l'organisation comparée des génomes de l'Homme et des mammifères d'intérêt zootechnique. Dans la plupart des cas, si la position d'un gène est connue sur le génome humain, il est possible de prédire la position de ce gène dans l'espèce d'intérêt et donc de déterminer si le gène étudié se positionne ou non dans la région candidate.

La simple mise en évidence de l'existence d'un QTL dans un intervalle très grossier de l'ordre de 60 cM représentant seulement 2 % du génome, permet d'éliminer 98 % des gènes qui ne peuvent ainsi être responsable de l'effet observé. Par rapport à une étude physiologique qui s'intéresserait à l'ensemble des gènes connus impliqués dans une fonction, la connaissance de la localisation, même imprécise, du gène recherché permet donc de restreindre fortement le nombre de gènes susceptibles d'être responsables des variations observées sur le caractère étudié.

En 1995, Francis Collins, donnait son sentiment sur le succès des identifications de gènes par approche physiologique (fonctionnelle ou candidate selon qu'un seul ou plusieurs gènes étaient testés),

positionnelle ou candidat positionnel (figure 2). Quatre ans après, ses prévisions semblent assez réalistes, avec une importance croissante des découvertes par étude de candidat positionnel dans un intervalle plus ou moins fin, au fur et à mesure que le nombre de gènes connus et cartographiés augmente.

Figure 2. *Succès comparé des différentes approches pour cloner un gène impliqué dans une maladie génétique chez l'Homme (Collins 1995).*



Au sein du Département de Génétique Animale de l'INRA, les démarches de cartographie fine ont permis de cartographier le gène PIS dans un intervalle de l'ordre de 100 kb chez la chèvre, le gène RN dans un intervalle de l'ordre de 100 kb chez le porc, le gène Polled dans un intervalle de l'ordre de 600 kb chez le bovin, le gène Booroola dans un intervalle de l'ordre de 1000 kb chez la brebis. Si l'on attend en moyenne un gène tous les 30 kb, ces intervalles ne comprennent environ que 3 à 30 gènes. Dans certains de ces cas, le séquençage de fragments de ces régions (ou toute autre technique) permettra d'identifier un gène déjà connu non encore cartographié, ou un nouveau membre d'une famille de gènes connus, qui deviendra un candidat positionnel ; dans d'autres cas, le gène en question sera encore totalement inconnu, et une démarche de clonage positionnel complet sera alors nécessaire pour identifier le gène recherché.

3 / Des gènes ayant un effet démontré dans une autre espèce peuvent fournir des candidats précieux

Dans certains cas, la connaissance des gènes pour lesquels un effet à été montré dans une autre espèce est également un élément important pour l'identification de candidats. Un gène pour lequel un effet majeur a été montré dans une espèce est un candidat pour le même effet ou un effet voisin dans une autre espèce.

Un exemple intéressant est ainsi donné par les gènes impliqués dans l'hypertrophie musculaire. Le gène RYR1 responsable, chez le porc Piétrain, d'une hypertrophie musculaire a été identifié grâce aux travaux réalisés parallèlement chez l'Homme et le porc. Dans ces deux espèces, ce gène induit une hyperthermie maligne après exposition au gaz anesthésique halothane, et il a été montré qu'une mutation similaire du gène RYR1 est impliquée dans les deux espèces. Chez les bovins, des travaux avaient montré que le gène MH responsable de l'hypertrophie musculaire était cartographié sur le chromosome BTA2, dans une région non orthologue à la région contenant RYR1, il fallait donc rechercher un autre gène. Le clonage positionnel de ce gène MH était donc en cours, lorsque des travaux chez la souris ont montré l'existence d'un nouveau gène GDF8 (Growth Differentiation Factor 8), membre de la superfamille du TGFb (Transforming Growth Factor b). L'inactivation de ce gène par recombinaison homologue chez la souris avait mis en évidence une augmentation de la masse musculaire de 35 % des souris homozygotes, et un doublement du poids de certains muscles. A ce moment, il existait, dans les banques de données, une étiquette humaine homologue à la séquence du gène GDF8, cartographiée chez l'Homme dans la région orthologue au segment bovin où MH était attendu. Grobet *et al* (1997) ont alors montré que les animaux Blanc Bleu Belge culards présentaient une délétion de 11 nucléotides de la séquence codante de GDF8, induisant une inactivation de cette protéine par rupture du cadre de lecture. A nouveau, c'est la connaissance de l'effet du gène dans une autre espèce qui avait permis de proposer un gène candidat et d'identifier ainsi le gène responsable de l'hypertrophie musculaire chez les bovins.

Même si les comparaisons doivent être faites avec prudence, les banques de données humaines sont une source très importante de telles informations. Au 20 juillet 1999, la base de données OMIM (Online Mendelian Inheritance in Man, http://www.ncbi.nlm.nih.gov/Omim/) recense ainsi des données génétiques, physiologiques et cliniques sur 8 459 gènes ou locus. En interrogeant la banque avec le mot «disease» ou «disorder» ou «syndrome», on obtient 4 914 réponses.

4 / Comment démontrer l'implication d'un candidat positionnel ?

Dans la phase terminale d'un programme de clonage positionnel, il est nécessaire d'étudier un ou plusieurs candidats positionnels. Lorsque aucun gène candidat connu n'a pu être mis en évidence, il faut étudier un ou plusieurs gènes encore inconnus présents dans la région. Pour guider ce travail, une connaissance du caractère étudié la plus fine possible est nécessaire. Dans le cas de l'étude du gène RN, ce gène avait originellement été mis en évidence par son effet défavorable sur le rendement technologique lors de la cuisson du jambon. Les études physiologiques ultérieures ont montré que l'allèle défavorable induisait une augmentation de la quantité de glycogène musculaire, sans modifier la quantité de glycogène hépatique (Estrade 1994). Un gène exprimé spécifiquement dans le muscle et ayant un

effet sur le métabolisme du glycogène pourrait donc être un «bon candidat». En second lieu, la prise en compte du déterminisme génétique du caractère est importante : une mutation récessive a ainsi plus de chances de correspondre à une inactivation d'un gène par rupture du cadre de lecture, ou par substitution d'un acide aminé affectant la structure ou l'activité de la protéine.

Plusieurs approches peuvent être poursuivies pour tester le statut de candidat du gène étudié : étude du polymorphisme du gène afin de mettre en évidence une mutation causale candidate, étude de l'expression du gène pour déterminer les tissus dans lesquels il s'exprime, étude des homologies des séquences avec des séquences déjà connues pour tenter de prédire la fonction des gènes encore inconnus, étude de la fonction du gène candidat par inactivation de gène (knock-out) chez la souris.

Discussion

Lorsqu'un gène majeur a été détecté, on cherche en général à identifier le gène responsable des différences de performances observées. Deux approches différentes, physiologique et positionnelle, peuvent être utilisées, mais, dans la réalité, la plupart des projets combinent les deux approches, en cherchant à cartographier au mieux le gène recherché, puis en cherchant d'éventuels candidats au sein du segment ainsi défini.

A force de travail constante, l'analyse d'un candidat se faisant au détriment des travaux de cartographie fine, il faut bien réfléchir avant d'entreprendre un travail ciblé sur un candidat, quand la poursuite des travaux de cartographie fine pourrait permettre de préciser la position du gène recherché et, éventuellement, de positionner le candidat à l'extérieur du segment, rendant par là même son étude inutile. Même si la comparaison ne vaut que ce qu'elle vaut, on peut dresser un parallèle entre choisir d'entreprendre l'étude de gènes candidats ou d'entreprendre une démarche de clonage positionnel, et choisir de miser sa fortune du jour au tiercé pensant bénéficier d'un bon «tuyau» plutôt que de mettre de côté tous les jours une somme équivalente, en se sachant un peu plus riche à la fin de l'année.

Les projets de séquençage du génome humain mettent à disposition une masse colossale d'informations. Les programmes de séquençage d'étiquettes et de valorisation de ces séquences permettent aujourd'hui de disposer dans la base Unigene (http://www.ncbi.nlm.nih.gov/UniGene/) de plus de 19 000 séquences codantes de gènes connus ou d'ARNm complets ainsi que de la séquence partielle de plus de 74 000 gènes putatifs complémentaires. Outre le séquençage complet du génome, un des prochains défis est l'étude de l'expression de ces gènes, afin de déterminer dans quels tissus et à quels moments ces gènes s'expriment. Ces études seront facilitées par le développement des nouvelles technologies de filtres à haute densité et de puces à ADN (voir Hatey *et al* 1998 et Hatey 2000, cet ouvrage). On devrait donc disposer de répertoires de gènes par position sur le génome et par tissus où ils s'expriment, tout d'abord chez l'Homme puis progressivement dans les espèces animales. Ces outils faciliteront bien évidemment l'identification de candidats positionnels pertinents.

Les projets actuellement en développement chez l'Homme nous font une fois de plus changer d'échelle. Pour ne plus dépendre de collections de cas familiaux toujours difficile à collecter, les généticiens humains envisagent d'utiliser des études de déséquilibre de liaison à l'échelle de la population mondiale, permettant de mettre en évidence des haplotypes ancestraux rencontrés chez les individus souffrant d'une même affection génétique. Dans ce but, il est envisagé de produire 500 000 marqueurs polymorphes ponctuels (SNP, Single Nucleotide Polymorphism) afin de disposer d'un marqueur génétique situé au plus à 3kb de la mutation recherchée (Kruglyak 1999). Des études de séquençage systématique de grandes régions du génome humain ont en effet conduit à une estimation d'une mutation rencontrée toutes les 2000 bases séquencées environ dans les séquences codantes, et toutes les 800 bases séquencées dans les séquences non codantes (Halushka *et al* 1999). Ces outils devraient permettre de cartographier très précisément les locus impliqués et d'identifier des candidats susceptibles d'être responsable des maladies étudiées.

Aux banques de séquences et aux données fonctionnelles sur les gènes correspondants s'ajouteront donc des collections de points polymorphes dans les séquences codantes ou non codantes. Pourrons-nous transposer aux espèces animales d'intérêt zootechnique au moins une partie de ces marqueurs, devons-nous développer des approches similaires, quelle densité de carte est-elle nécessaire pour des études de déséquilibre de liaison dans les espèces animales ? Une fois de plus, les avancées de la génétique humaine nous font réfléchir !

Références

Collins F., 1995. Positional cloning moves from perditional to traditional. Nature Genetics, 9, 347-350.

Estrade M., 1994. Etude de l'expression métabolique du gène RN-. Thèse de doctorat de l'université Blaise Pascal (Université d'Auvergne) N° 605.

Halushka M.K., Fan J.B., Bentley K., Hsie L., Shen N., Weder A., Cooper R., Lipshutz R., Chakravarti A., 1999. Patterns of single nucleotide polymorphisms in candidate genes for blood pressure homeostatis. Nature Genetics, 22, 239-247.

Hastbacka J., de la Chapelle A., Mahtani M.M., Clines G., Reeve-Daly M.P., Daly M., Hamilton B.A., Kusumi K., Trivedi B., Weaver A., *et al* , 1994 The diastrophic dysplasia gene encodes a novel sulfate transporter: positional cloning by fine-structure linkage disequilibrium mapping. Cell., 78, 1073-1087.

Hatey F., 2000. Recherche de gènes associés à des fonctions : l'approche fonctionnelle. INRA Productions Animales, numéro hors série « Génétique moléculaire : principes et application aux populations animales », 153-160.

Hatey F., Tosser-Klopp G., Clouscard-Martinato C., Mulsant P., Gasser F., 1998. Expressed sequence tag for genes : a review. Genetics Selection Evolution, 30, 521-541.

Kruglyak L., 1999. Prospects for whole genome linkage disequilibrium mapping of common disease genes. Nature Genetics, 22, 139-144.

Milan D., Robic A., Chardon P., Iannuccelli N., Caritez J.C., Yerle M., Gellin J., Looft C., Andersson L., Elsen J.M., Le Roy P., 2000. Exemple de cartographie fine : le cas du gène RN chez le porc. INRA Productions Animales, numéro hors série « Génétique moléculaire : principes et application aux populations animales »,137-139.

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.

Rothschild M., Jacobson C., Vaske D., Tuggle C., Wang L., Short T., Eckardt G., Sasaki S., Vincent A., McLaren D., Southwood O., van der Steen H., Mileham A., Plastow G., 1996. The estrogen receptor locus is associated with a major gene influencing litter size in pigs. Proceedings of the National Academy of Sciences of the USA, 93, 201-205.

Vaiman D., 2000. Etablissement des cartes génétiques. INRA Productions Animales, numéro hors série « Génétique moléculaire : principes et application aux populations animales », 73-78.