Pascale LE ROY

INRA Station de Génétique Quantitative et Appliquée 78352 Jouy-en-Josas Cedex

Les bases de la génétique quantitative

Les méthodes de mise en évidence des gènes majeurs

Résumé. *S'il existe un gène à effet majeur sur un caractère, la distribution des phénotypes est un mélange de distributions élémentaires dans des proportions obéissant aux lois de Mendel. Différents tests statistiques basés sur ce principe sont présentés. Les données exploitées peuvent provenir d'expériences de croisements spécialement conçues pour détecter un gène majeur, mais aussi de fichiers de contrôle des performances de structure quelconque. Les méthodes utilisées sont des tests d'hypothèses permettant de choisir entre l'absence et la présence d'un locus majeur à partir des informations disponibles. Il s'agit soit d'indicateurs numériquement très simples à obtenir, soit de méthodes plus complexes faisant appel aux techniques du maximum de vraisemblance. Dans l'un et l'autre cas, les tests construits prennent plus ou moins en compte la structure généalogique de la population. La méthode la plus élaborée est l'analyse de ségrégation qui permet de synthétiser toute l'information disponible pour comparer différentes hypothèses de transmission du caractère et qui fournit des estimations des paramètres du modèle génétique retenu. Cependant, elle est numériquement très coûteuse et nécessite des simplifications pour être applicable aux populations d'animaux domestiques. Des critères moins puissants mais plus simples ont donc un intérêt non négligeable en permettant en première approche une recherche systématique des gènes à effet majeur.*

La détection d'un gène à effet majeur sur un caractère est un sujet qui, bien que classique en génétique quantitative, a pris une importance grandissante au cours des 15 dernières années. Ce regain d'intérêt est tout d'abord à rapprocher de la découverte, plus ou moins fortuite, de plusieurs gènes influant sur des caractères quantitatifs économiquement importants chez différentes espèces domestiques : le gène de nanisme chez la poule, le gène de la sensibilité à l'halothane chez le porc, le gène culard chez les bovins ou le gène de prolificité Booroola chez le mouton en sont quelques exemples célèbres. Par ailleurs, la mise en évidence, déjà ancienne, de très nombreux locus majeurs chez la souris a conduit à la recherche de tels gènes chez les animaux domestiques, certains gènes connus agissant sur des caractères dont l'amélioration est une préoccupation en productions animales : résistance aux maladies, croissance, viabilité embryonnaire, taux de muscle... Mais l'élément, sans conteste le plus déterminant de cette évolution, est le développement des biotechnologies qui permet d'espérer une efficacité grandissante dans l'utilisation des gènes majeurs pour l'amélioration génétique des animaux domestiques.

Si dans les exemples qui viennent d'être cités le qualificatif de "majeur" est facilement justifiable, il n'en va pas toujours ainsi. Il est en effet difficile, face à la continuité des phénomènes biologiques en cause, de donner une définition exacte de la notion d'effet majeur d'un gène, la limite arbitraire souvent posée étant l'existence d'une différence de 1 écart-type phénotypique entre les valeurs moyennes associées aux génotypes extrêmes. Cependant, de meilleures définitions peuvent être données en terme de pourcentage de la variance génétique du caractère dû au locus considéré ou, de façon plus synthétique, en terme de gain consécutif à la prise en considération d'un déterminisme simple dans une stratégie générale d'amélioration génétique.

L'avantage essentiel de l'existence d'un gène majeur est qu'une valeur génétique élevée peut être obtenue de façon immédiate et stable par la fixation d'un génotype favorable au locus majeur. Par rapport à une sélection classique, au gain de temps potentiel s'ajoute éventuellement un affranchissement vis-à-vis d'antagonismes génétiques entravant la réalisation d'un objectif global. L'introduction dans une population d'accueil d'un allèle particulièrement intéressant, tout en conservant le reste du génome receveur en l'état, constitue l'étape suivante aujourd'hui réalisée par croisements en retour successifs. Toutefois, si les premières applications envisagées du transfert de gènes ont essentiellement une finalité pharmaceutique, l'utilisation de cette technique pour l'élevage est d'ores et déjà d'actualité et, dans la mesure où leur(s) allèle(s) à effet favorable ne sont pas totalement récessifs, les gènes majeurs identifiés sont bien sûr les premiers candidats au transfert.

Toutes ces considérations ont donc stimulé la mise au point de nombreuses méthodes pour la mise en évidence de l'effet, que nous appellerons majeur, d'un

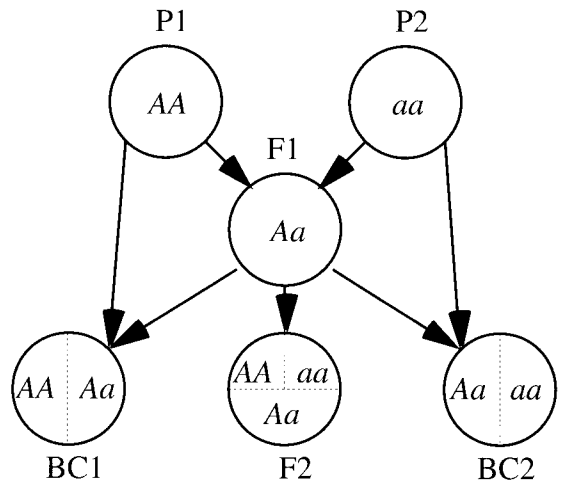
gène sur un caractère quantitatif. Il s'agit de tests d'hypothèses qui consistent, à partir d'un ensemble de mesures d'un caractère, à effectuer un choix par des techniques statistiques entre 2 hypothèses génétiques, l'une postulant l'absence de gène majeur affectant le caractère étudié, l'autre supposant sa présence. L'idée de base est que, s'il existe une ségrégation au locus majeur, la distribution des observations est un mélange de distributions élémentaires dans des proportions prévisibles d'après les lois de Mendel (tableau 1). Au contraire, s'il n'y a pas de gène majeur, bien que ces proportions soient valables pour n'importe quel gène, il n'y a pas de mélange visible de la distribution des données du fait des effets très faibles, et donc non individualisables, des polygènes. De plus, le gène majeur étant entouré sur le chromosome par d'autres gènes, l'étude de la liaison entre la distribution du caractère et la ségrégation d'allèles à des locus marqueurs répartis sur le génome peut venir compléter les données phénotypiques afin d'identifier plus aisément la ségrégation au locus majeur.

Trois grands groupes de méthodes de mise en évidence d'un gène majeur, respectant cette démarche générale, peuvent être arbitrairement définis selon le type de données prises en compte. Un premier groupe est constitué des méthodes utilisant des données issues d'expérimentations spécialement conçues pour la mise en évidence d'un gène majeur ; ces méthodes ont été essentiellement développées chez les végétaux et les animaux de laboratoire. Un second groupe rassemble les méthodes tirant profit d'informations recueillies sur des populations non expérimentales ; ces méthodes ont été surtout mises en oeuvre sur des populations humaines. Un troisième groupe enfin est constitué par les méthodes utilisant des marqueurs génétiques dont la liaison avec d'éventuels gènes ayant un effet identifiable sur un caractère quantitatif (de tels gènes majeurs sont appelés Quantitative Trait Loci) est recherchée. Cet article portera sur des méthodes appartenant aux 2 premiers groupes, la recherche de QTL à l'aide de marqueurs génétiques étant abordée dans l'article de Chevalet et Boichard (1992).

Tableau 1. *Ségrégation mendélienne de 2 allèles A et a à un locus autosomal : génotype des descendants en fonction des génotypes parentaux.*

Génotypes des parents	AA	Aa	aa	Totaux
AA	AA	1/2AA 1/2Aa	Aa	1/2AA 1/2Aa
Aa	1/2AA 1/2Aa	1/4AA 1/2Aa 1/4aa	1/2Aa 1/2aa	1/4AA 1/2Aa 1/4aa
aa	Aa	1/2Aa 1/2aa	aa	1/2Aa 1/2aa

Figure 1. *Schéma expérimental pour la mise en évidence d'un gène majeur.*



1 / Les données

1.1 / Populations expérimentales

L'analyse de résultats provenant d'expériences de croisements entre lignées homozygotes différant en 1 locus est la méthode de référence pour mettre en évidence l'effet d'un gène sur un caractère (Elston et Stewart 1973). En effet, en théorie, ce dispositif expérimental est optimal pour démontrer une ségrégation mendélienne. Le protocole classique prévoit de disposer de mesures sur les 2 lignées parentales (P_1 et P_2), sur les croisements F_1 (P_1xP_2) et F_2 (F_1xF_1) et sur les croisements en retour BC_1 (F_1xP_1) et BC_2 (F_1xP_2). Les 2 lignées P_1 et P_2 étant supposées fixées aux 2 états homozygotes AA et aa, la F_1 est alors homogène hétérozygote Aa et la ségrégation du gène peut être mise en évidence par le caractère composite des distributions des croisements F_2, BC_1 et BC_2, les distributions élémentaires étant celles des populations parentales, P_1 et P_2, et de la F_1 (figure 1).

Chez les animaux domestiques, ce schéma expérimental est réalisé, non pas avec des lignées consanguines, mais avec des lots d'animaux de performances extrêmes, animaux sensibles et résistants à une maladie par exemple. Si le raisonnement sur la ségrégation au locus majeur demeure tout à fait valable dans ce cas, certains écarts aux hypothèses du protocole théorique peuvent mener à une conclusion erronée. Ainsi, le plus souvent, les 2 lots parentaux correspondent en fait à deux races différentes et les paramètres du croisement entre ces races, notamment les effets directs des races et l'hétérosis, interviennent sur les moyennes des types génétiques croisés. La référence aux distributions des performances dans les races parentales et dans la F_1 n'est alors plus correcte pour rechercher les proportions d'un mélange en seconde génération de croisement.

Par exemple, en ce qui concerne l'effet d'hétérosis, sous l'hypothèse de présence d'un gène majeur, les moyennes des distributions intra génotype en F_2 sont décalées par rapport aux moyennes de référence (des AA en P_1, aa en P_2 et Aa en F_1), d'où des erreurs éventuelles de tri des animaux de la F_2 et par suite une perte de puissance de la méthode de détection. A l'inverse, un effet direct de la race peut avoir sur la moyenne des distributions intra type génétique des conséquences équivalentes à celles de la ségrégation de 2 allèles à un locus majeur. En considérant, par exemple, que les 2 races parentales P_1 et P_2 ont des performances moyennes respectivement égales à 0 et à 2, en l'absence d'effet d'hétérosis, l'espérance de la

distribution des performances en F_1 est de 1 et elle est de 1/2, 1 et 3/2 en BC_1, F_2 et BC_2. L'hypothèse d'absence de gène majeur est mal représentée si ces effets race ne sont pas pris en compte dans l'analyse, et la confusion entre une différence entre races et une ségrégation à un locus majeur peut alors se produire. En effet, un rapide calcul permet de voir que les mêmes espérances des distributions intra-type génétique sont attendues dans le cas de la ségrégation de 2 allèles codominants à un locus majeur tel que l'effet du génotype AA est 0 et l'effet du génotype aa est 2. Cependant, ces déviations par rapport aux hypothèses du modèle théorique ont l'avantage d'être tout à fait prévisibles et modélisables : les méthodes, pour conserver leur puissance et leur robustesse, doivent simplement tenir compte des paramètres du croisement entre les races utilisées.

Moins prévisible, mais sans doute très fréquente, la non fixation des populations parentales P_1 et P_2 aux 2 états homozygotes est un autre type de déviation par rapport aux hypothèses posées dans le protocole théorique. La conséquence directe est l'existence d'un biais entre les distributions des performances en P_1, P_2 et F_1 et les distributions réelles intra génotype AA, Aa et aa. Les proportions du mélange des types P_1, P_2 et F_1 attendues en seconde génération dans ce cas ne sont plus connues a priori. La perte d'efficacité est ici inévitable mais peut varier d'une méthode à l'autre, les différents tests étant plus ou moins robustes à cet écart aux hypothèses.

Cependant, malgré ces quelques complications par rapport au schéma idéal, ce dispositif reste le meilleur pour tester l'existence d'un gène majeur et pour la prouver si la mise en évidence a eu lieu sur des données "tout venant". Dans ce cas, des animaux ayant une forte probabilité d'être homozygotes AA et aa sont repérés dans la population où le gène a été mis en évidence pour constituer les 2 lots parentaux P_1 et P_2.

1.2 / Populations non expérimentales

Lorsqu'il n'est pas possible de disposer de données issues d'un protocole expérimental, la mise en évidence d'un gène majeur peut tout de même être envisagée dans une population quelconque, notamment dans une population panmictique. La démarche adoptée est similaire à celle qui vient d'être décrite, la ségrégation de 2 allèles A et a au locus majeur étant cette fois observée intra famille. Chaque famille de l'échantillon, par exemple les 2 parents et leurs descendants (famille nucléaire), est alors l'équivalent d'un protocole expérimental où les types parentaux pourraient être des 3 génotypes AA, Aa ou aa. Connaissant les génotypes des parents, il est en effet possible de prévoir les proportions attendues des 3 génotypes chez les descendants d'après les lois de Mendel (tableau 1).

La taille obligatoirement réduite de chacun des mini protocoles que représentent les familles nucléaires est un premier inconvénient par rapport à l'utilisation de données expérimentales. Cependant, celui-ci est souvent atténué chez les animaux domestiques par l'adoption d'un modèle d'analyse hiérarchique, père, mère intra-père, qui permet de considérer des familles de demi-frères de père qui sont d'effectif beaucoup plus important. Un second inconvénient est de devoir raisonner avec des probabilités de génotypes parentaux et non plus avec une structure génotypique initiale parfaitement connue. Toutefois, des effectifs très importants, comparativement à ce qui est réalisable en expérimentation, permettent de compenser en partie ces incertitudes lors de l'étude de données tout venant. Ainsi, la notion de variabilité entre familles peut être exploitée : l'existence d'une ségrégation au locus majeur s'accompagne en effet d'une hétérogénéité des types de distributions intra famille liée à l'existence de combinaisons différentes des génotypes parentaux (tableau 2).

Tableau 2. *Caractéristiques des distributions intra-famille de pleins frères d'un caractère quantitatif soumis à l'effet d'un gène majeur avec 2 allèles A et a. (Hypothèses : équilibre de Hardy-Weinberg ; variances intra-génotype égales).* μ_1, μ_2, μ_3 *: valeurs moyennes des animaux AA, Aa, aa.* σ^2 *: variance intra-génotype (AA, Aa ou aa) ; p, q : fréquences des allèles A et a.*

Génotypes des parents	Fréquence de la famille	Moyenne	Variance	Nombre de modes
AA x AA	p^4	μ_1	σ^2	1
AA x Aa	$2p^3q$	$\frac{\mu_1 + \mu_2}{2}$	$\sigma^2 + \left(\frac{\mu_1 - \mu_2}{2}\right)^2$	2
AA x aa	p^2q^2	μ_2	σ^2	1
Aa x Aa	$4p^2q^2$	$\frac{\mu_1 + 2\mu_2 + \mu_3}{4}$	$\sigma^2 + (3(\mu_1 - \mu_3)^2 + 4(\mu_2 - \mu_1)(\mu_2 - \mu_3)) / 16$	3
Aa x aa	$2pq^3$	$\frac{\mu_2 + \mu_3}{2}$	$\sigma^2 + \left(\frac{\mu_3 - \mu_2}{2}\right)^2$	2
aa x aa	q^4	μ_3	σ^2	1

Il faut par ailleurs signaler que ces populations non expérimentales sont très souvent, dans le cas des animaux d'élevage, des populations soumises à sélection. Si un gène majeur influence le caractère sélectionné, la fréquence du ou des allèles favorables augmente alors au cours des générations. De plus, s'il existe des gènes marqueurs en déséquilibre de liaison avec ce locus, les allèles marqueurs liés préférentiellement aux allèles favorables du gène majeur sont également entraînés par la sélection. Un tel effet est dit d'auto-stop (hitch-hicking effect), l'observation au cours de la sélection d'une évolution de la fréquence d'un allèle marqueur, non explicable par la seule dérive génétique, révélant l'éventuelle présence d'un locus majeur.

L'augmentation de la fréquence des allèles favorables au locus majeur peut ainsi être très marquée dans le cas de la sélection de lignées dites "hyper". Le principe de ces lignées est d'utiliser une base de sélection extrêmement large, la population nationale par exemple, afin de pouvoir appliquer une intensité de sélection importante, souvent inférieure à 1 p.mille, en rassemblant dans un petit noyau les meilleurs reproducteurs mâles ou/et femelles. Avec une telle procédure, il est possible de retenir des animaux rares porteurs, à un locus majeur, d'allèles favorables qui ne seraient pas détectés par l'observa-

tion de l'évolution de la moyenne du caractère dans la population. Il s'agit donc là d'un schéma très puissant pour la mise en évidence de l'effet majeur d'un gène, qui a déjà permis de révéler l'existence de 2 gènes d'hyperovulation chez les ovins : le gène Booroola (rassemblement en 1953 d'animaux extrêmes dans une lignée de sélection du CSIRO en Australie) et récemment le gène Inverdale lié au sexe (criblage d'animaux hyperprolifiques par le MAF en Nouvelle Zélande). Bien entendu, la sélection d'une lignée "hyper" ne prouve en rien l'existence d'un gène à effet majeur. Seuls des accouplements raisonnés entre des animaux extrêmes et des animaux de niveau moyen, éventuellement suivis par des accouplements en retour vers les extrêmes ou la moyenne, permettent de tester cette hypothèse : l'équivalent du protocole classique décrit plus haut est alors réalisé.

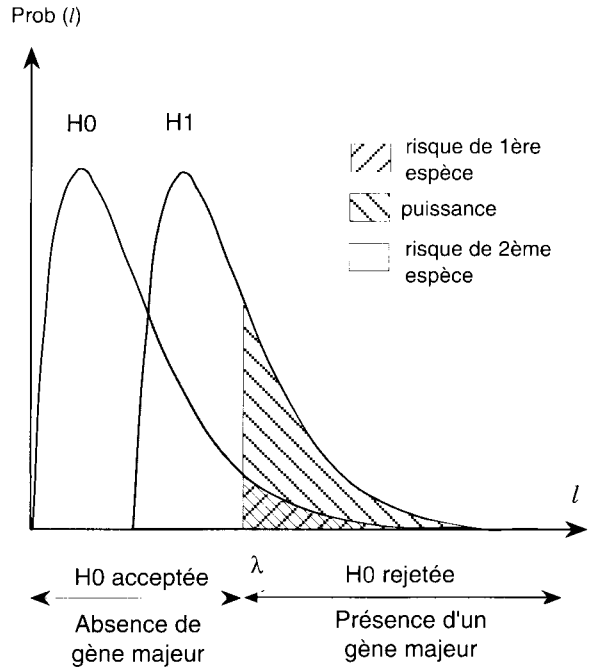
2 / Les méthodes

2.1 / Principe

Les données consistent en un ensemble d'observations $(y_1,y_2, \ldots y_n)=\mathbf{y}$ qui sont considérées comme les réalisations d'une variable aléatoire **Y**. Le problème posé est de tester une hypothèse portant sur la distribution de **Y** sachant **y**. Pour cela, la démarche générale est celle appliquée dans tout test statistique. La première étape est de choisir une fonction, $l(\mathbf{Y})$ dite statistique de test, qui permet de résumer l'information disponible par une valeur numérique. Il faut ensuite connaître, ou éventuellement établir, la loi de probabilité suivie par l sous l'hypothèse particulière à tester, c'est à dire savoir pour toute valeur de la statistique l quelle est sa probabilité de réalisation si cette hypothèse est vraie. L'utilisateur du test, s'étant fixé le risque de première espèce α qu'il accepte de prendre (risque de conclure que l'hypothèse testée est fausse alors qu'elle est vraie), peut déduire de cette loi le domaine de rejet de l'hypothèse. Dans le cas des tests que nous allons envisager ici, cela revient à connaître la valeur de $l(\mathbf{Y})$ à partir de laquelle il rejettera l'hypothèse, c'est à dire le seuil λ tel que $\text{Prob}(l(\mathbf{Y})>\lambda)=\alpha$. La dernière étape consiste à calculer la valeur prise par l à partir des données et à confronter le résultat numérique obtenu à cet ensemble de rejet : si $l(\mathbf{y})>\lambda$ l'hypothèse est rejetée, si $l(\mathbf{y})<\lambda$ l'hypothèse est acceptée. Dans la plupart des cas, les méthodes employées pour la mise en évidence d'un gène majeur relèvent de la théorie des tests d'hypothèses emboitées, c'est à dire que l'hypothèse particulière à tester, dite hypothèse nulle et notée H_0, est testée contre une hypothèse générale (H_1) qui la contient. En pratique, l'hypothèse nulle H_0 que nous voulons tester est presque toujours celle d'un déterminisme polygénique du caractère, classiquement caractérisée par la valeur du coefficient d'héritabilité. Cette hypothèse particulière est emboitée dans l'hypothèse générale H_1 d'un déterminisme mixte, où l'effet d'un gène majeur s'ajoute à l'hérédité polygénique "classique" du caractère (figure 2).

Dans ces conditions, le risque de conclure à l'existence d'un gène majeur alors que cela est faux est le risque de première espèce qui est fixé par l'utilisateur. Par contre, le risque de conclure à l'absence de gène majeur alors qu'il en existe un, risque de seconde espèce et complément à 1 de la puissance du test (figure 2), n'est pas contrôlable par l'utilisateur mais

Figure 2. *Distributions de la statistique de test l sous H_0 et sous H_1 : définitions des risques de 1ère et 2ème espèces, du seuil de rejet λ de H_0 et de la puissance du test. Règle de décision.*



varie avec les paramètres caractérisant le déterminisme génétique du caractère (valeurs moyennes des génotypes au locus majeur, variances intra génotype, fréquences alléliques, héritabilité ...), le nombre, la taille et la structure des familles constituant l'échantillon de données, l'adéquation entre le modèle génétique posé pour écrire le test et la réalité biologique.

La qualité statistique d'une méthode donnée peut être appréhendée au travers de 3 caractéristiques : son niveau, sa robustesse et sa puissance. Le plus souvent, seule la loi asymptotique de la statistique de test est connue; en dehors de ces conditions, par exemple pour des nombres ou des tailles de familles limités, utiliser le seuil de rejet donné dans les tables numériques peut impliquer que le risque α' réellement accepté est supérieur au risque α choisi. La première qualité d'une méthode est donc d'avoir un niveau que l'utilisateur du test peut contrôler à distance finie. Par ailleurs, les tests construits reposent presque toujours sur l'hypothèse de normalité des distributions des variables aléatoires étudiées : la seconde qualité d'une méthode est d'être robuste lorsqu'il existe un écart à la normalité de ces distributions, c'est à dire notamment d'avoir des seuils de rejet de H_0 qui restent corrects dans ce cas. Enfin, à niveau donné, les différentes méthodes peuvent être évaluées en terme de puissance, c'est à dire sur leurs capacités respectives à détecter tel ou tel type de gène majeur lorsqu'il est présent.

2.2 / Diversité des méthodes

Les nombreuses méthodes qui ont été proposées pour détecter un gène majeur peuvent être classées en fonction du critère statistique qu'elles utilisent.

a / Indicateurs simples

Un premier groupe rassemble des méthodes intuitives reposant sur la valeur prise par diverses statistiques élémentaires lorsqu'il existe une ségrégation à un locus majeur. Ces indicateurs simples sont soit des critères classiques pour étudier la forme de la distribution des performances, soit des critères spécifiquement construits pour la mise en évidence d'une ségrégation mendélienne. Ils prennent plus ou moins en considération la structure génétique des données.

Ainsi, la détection d'un gène majeur affectant un caractère à variation continue est traditionnellement associée à la recherche d'écarts à la normalité de la distribution des phénotypes dans la population, notamment d'une asymétrie ou d'un aplatissement, révélateurs de l'existence d'un mélange de lois sous-jacent à la distribution totale observée. Des tests classiques de normalité, test D de Kolmogorov-Smirnov, tests sur les coefficients de skewness (asymétrie) ou de kurtosis (aplatissement) notamment, sont alors utilisés. Cependant, ils ne peuvent être considérés que comme une première approche car leur manque de robustesse est évident, la distribution des phénotypes pouvant s'écarter de la normalité pour bien d'autres raisons que la présence d'un gène à effet fort.

La prise en compte d'une information généalogique peut être envisagée si des données concernant un ensemble d'individus répartis en familles, de pleins-frères ou de demi-frères, sont disponibles. Le but des différents tests proposés est de tester l'hétérogénéité, évoquée précédemment (tableau 2), des types de distribution intra-famille afin de révéler l'existence de 2 groupes de familles : les familles où des allèles sont en ségrégation au gène majeur (plusieurs génotypes existent dans la même famille) et les familles où le gène est fixé à un état homozygote (tous les membres de la famille ont le même génotype). Divers critères existent pour tester soit l'hétérogénéité des variances intra-famille (Bartlett 1937), soit l'hétérogénéité des formes de distribution intra-famille (asymétrie ou aplatissement) (Mérat 1968), soit la curvilinéarité de la relation entre moyenne et variance de la famille (Fain 1978).

La possession d'informations sur les performances propres des parents est à l'origine d'un dernier ensemble de méthodes simples. L'idée générale est qu'un descendant ressemble moins à la moyenne de ses parents qu'à l'un d'entre eux lorsqu'un gène majeur est présent. Ce principe est utilisé par des tests de régression entre les performances des parents et des descendants (Hanset et Michaux 1985) et par l'analyse exploratoire structurée des données (SEDA=Structured Exploratory Data Analysis) (Karlin *et al* 1979). Cette dernière utilise 3 critères destinés à tester l'hypothèse spécifique d'un déterminisme mendélien : l'index MGI (Major Gene Index), la fonction OBP (Offspring Between Parents) et la corrélation MPCC (Midparental Pairwise Correlation Coefficient).

Parmi toutes ces propositions, les tests d'hétérogénéité des variances intra-familles, sur lesquels nous reviendrons, ou de curvilinéarité de la régression moyenne variance intra-famille sont souvent les plus puissants. En ce qui concerne le type de gène majeur présent, la puissance des tests est d'autant meilleure que les fréquences alléliques sont proches et l'écart entre effets moyens des génotypes important, ces 2 tests étant particulièrement adaptés pour la mise en évidence de gènes caractérisés par une dominance complète ou par une distribution du caractère plus variable chez les homozygotes de moyenne élevée. Ces 2 critères présentent par ailleurs l'avantage important d'avoir des distributions asymptotiques, sous H_0, connues et des seuils de rejet tabulés. Cependant, leur robustesse demeure très faible car ils sont très dépendants de l'hypothèse de normalité de la distribution des phénotypes intra-génotype. Ceci limite beaucoup leur portée, et ce d'autant plus qu'ils sont plutôt utilisés pour l'étude de la transmission de caractères dont la distribution dévie de la normalité (ce qui est observé lorsqu'il existe effectivement un gène majeur) (Le Roy et Elsen 1992).

b / Tests du maximum de vraisemblance

Un second groupe est constitué par des méthodes, numériquement plus complexes, qui sont des tests dits du maximum de vraisemblance. Etant donnée une hypothèse, l'information disponible est intégrée dans une fonction de vraisemblance, notée M, qui est la probabilité d'observer les données sous cette hypothèse : en reprenant les notations précédentes, $M_0(\mathbf{y})$ est la probabilité d'observer $\mathbf{y}$ sous H_0 et $M_1(\mathbf{y})$ la probabilité d'observer $\mathbf{y}$ sous H_1. Le principe du test est de retenir l'hypothèse sous laquelle la vraisemblance des données est la plus grande, c'est-à-dire le modèle génétique expliquant le mieux les distributions observées. Les fonctions de vraisemblance dépendent d'un certain nombre de paramètres inconnus utiles à la définition de l'hypothèse (moyennes, variances, fréquences alléliques, héritabilité ...) : $\theta_0=(\theta_{01},\theta_{02} \ldots \theta_{0n})$ sous H_0 et $\theta_1=(\theta_{11},\theta_{12} \ldots \theta_{1n})$ sous H_1. La démarche adoptée consiste à rechercher les valeurs $\hat{\theta}_0$ et $\hat{\theta}_1$ appelées estimations du maximum de vraisemblance, qui rendent maximales les fonctions $M_0(\mathbf{y})$ et $M_1(\mathbf{y})$ et à déclarer que H_0 est vraie si $\hat{M}_0(\mathbf{y})$ est significativement plus grande que $\hat{M}_1(\mathbf{y})$. La statistique de test employée pour prendre cette décision est le rapport des maximums de vraisemblance $l(\mathbf{y})=-2\mathrm{Log}(\hat{M}_0(\mathbf{y})/\hat{M}_1(\mathbf{y}))$ dont la distribution asymptotique est une loi de χ^2 avec un nombre de degrés de liberté égal au nombre de paramètres de H_1 qu'il faut fixer pour revenir sous H_0.

De nombreux tests de mise en évidence d'un gène majeur appartiennent à cette catégorie de statistiques, depuis les tests de multimodalité de la distribution des phénotypes dans la population jusqu'aux méthodes très complexes utilisées en génétique épidémiologique et connues sous le nom d'analyse de ségrégation. Par rapport aux indicateurs simples présentés plus haut, ces méthodes sont en principe plus puissantes et il existe diverses techniques pour leur assurer une bonne robustesse. De plus, elles présentent le gros avantage de fournir des estimations des différents paramètres qui permettent, lorsqu'un gène majeur est détecté, de le caractériser, première étape indispensable à son utilisation. Enfin, elles sont applicables quelles que soient les données, expérimentales ou non, et peuvent prendre en compte de façon très fidèle la structure génétique de la population.

Les méthodes les plus simples se rapprochent dans leur conception des indicateurs présentés précédemment car elles reposent sur la mise en évidence d'un écart à la normale d'un critère caractérisant la distribution des phénotypes. Ainsi, l'hétérogénéité des variances intra-famille est le plus souvent testée par

un test du maximum de vraisemblance, le test de Bartlett (1937). La recherche d'une bimodalité de la distribution des données est généralement entreprise en comparant les vraisemblances de l'échantillon sous les modèles d'unimodalité (H_0) et de bimodalité (H_1) (Titterington *et al* 1985) :

$$M_0(\mathbf{y})=\prod_{i=1}^{n} f_0(y_i)$$

$$M_1(\mathbf{y})=\prod_{i=1}^{n} (pf_1(y_i)+(1-p)f_2(y_i))$$

où $f_k(y_i)$ est une loi normale de moyenne μ_k et de variance σ^2.

L'analyse de ségrégation (Elston 1980) repose quant à elle sur l'élaboration de modèles plus complets, incluant des facteurs génétiques et environnementaux, qui permettent de tester différentes hypothèses de transmission du caractère : modèle sporadique (pas de gène agissant sur le caractère), monogénique (1 seul gène à effet fort), oligogénique (quelques gènes à effets individualisables), polygénique (une infinité de gènes à effets faibles), mixte (1 gène majeur et des polygènes)... La fonction de vraisemblance permet de décrire, en théorie, n'importe quelle structure de population bien que des simplifications numériques soient souvent indispensables pour rendre la méthode opérationnelle. Trois types de paramètres interviennent dans l'écriture de la vraisemblance, décrivant 3 distributions : (1) la distribution des génotypes dans la population, (2) la distribution des génotypes des descendants conditionnellement aux génotypes des parents, (3) la distribution des phénotypes conditionnellement aux génotypes. La probabilité du phénotype de chaque individu du pedigree est décomposée en utilisant le théorème des probabilités conditionnelles selon :
Prob(phénotype) = $\Sigma_{\text{génotype}}$ Prob(génotype)Prob(phénotype/génotype)

La probabilité du génotype est accessible par l'estimation des paramètres des distributions (1) et (2), selon que l'individu est un fondateur (individu de parents inconnus) ou un non fondateur du pedigree. La probabilité d'observer le phénotype sachant le génotype, encore appelée pénétrance du génotype, est donnée par la distribution (3).

Par exemple, dans le cas simple du test de l'hypothèse d'un déterminisme sporadique contre celle d'un déterminisme monogénique, avec des données issues d'un protocole de croisement entre 2 lignées P_1 et P_2, les vraisemblances s'écrivent :

sous H_0 (déterminisme sporadique) :

$$M_0(\mathbf{y})=\prod_{i=1}^{n} f_0(y_i)$$

sous H_1 (déterminisme monogénique) :

$$M_1(\mathbf{y})=\prod_{i=1}^{n_{P1}} f_{AA}(y_i)\prod_{i=1}^{n_{P2}} f_{aa}(y_i)\prod_{i=1}^{n_{F1}} f_{Aa}(y_i)$$

$$\prod_{i=1}^{n_{BC1}} (f_{AA}(y_i)+f_{Aa}(y_i))/2\prod_{i=1}^{n_{BC2}} (f_{Aa}(y_i)+f_{aa}(y_i))/2$$

$$\prod_{i=1}^{n_{F2}} (f_{AA}(y_i)+2f_{Aa}(y_i)+f_{aa}(y_i))/4$$

où

n_j est le nombre d'animaux de type génétique j, n est le nombre total d'animaux,

$f_k(y_i)$ est la distribution des performances des animaux de génotype k, par exemple une loi normale de moyenne μ_k et de variance, commune à tous les génotypes, σ^2.

L'approche, présentée ici dans une situation très simple, se généralise sans difficulté théorique au cas d'un pedigree quelconque. Cependant, l'existence de relations de parenté complexes dans les pedigree animaux, notamment de boucles de consanguinité et de mariage, rend le calcul exact de la vraisemblance des observations numériquement très coûteux. Sous l'hypothèse d'une hérédité mixte du caractère, ce calcul devient rapidement impossible du fait de l'accumulation des sommations sur les génotypes au locus majeur et des intégrations sur les valeurs polygéniques des animaux. Plusieurs simplifications numériques ont donc été proposées pour adapter ces méthodes à l'étude de populations d'animaux domestiques (Le Roy *et al* 1989).

Conclusion

L'analyse de ségrégation est la méthode de référence pour la mise en évidence d'un gène à effet majeur sur un caractère. Elle permet de synthétiser toute l'information disponible, y compris les données relatives à des gènes marqueurs, pour comparer différents modèles de transmission du caractère. En plus d'une méthode de test d'une hypothèse particulière d'hérédité, elle est également une méthode d'estimation des paramètres caractérisant un déterminisme génétique. Cependant, elle nécessite de gros moyens de calculs d'où l'intérêt non négligeable de certains critères simples dans une logique de recherche systématique, par exemple à partir des fichiers nationaux du contrôle des performances, d'éventuels gènes à effet majeur.

La détection d'un gène à effet majeur par des techniques statistiques telles que l'analyse de ségrégation ne constitue qu'une première étape. En effet, tous ces tests acceptent, par définition, un risque non nul de commettre une erreur en acceptant l'hypothèse de l'existence d'un gène majeur. Ils nécessitent donc toujours une confirmation expérimentale car aucune méthode statistique, même très évoluée, ne peut remplacer un dispositif "bien fait" pour l'obtention des données. De plus, avant d'utiliser un gène majeur pour l'amélioration d'un caractère, il faut vérifier ses effets sur les autres caractères économiquement importants. Enfin, l'utilisation d'un tel gène est facilitée par l'existence de gènes marqueurs proches qui permettent d'identifier individuellement les génotypes des reproducteurs. La mise en place d'un protocole expérimental peut donc être également l'occasion de rechercher ces marqueurs.

Références bibliographiques

Bartlett M.S., 1937. Some examples of statistical methods of research in agriculture and applied biology. J. R. Soc. (suppl), 4, 137-170.

Chevalet C., Boichard D., 1991. Utilisation de marqueurs pour localiser les gènes responsables de la variabilité des caractères quantitatifs ("QTL"). Séminaire, "Eléments de génétique quantitative et application aux populations animales", Port d'Albret, 14-18 Octobre 1991, II : Les bases de la génétique quantitative, - , Département de Génétique Animale, INRA.

Elston R.C., 1980. Segregation analysis. In : Mielke J.H., Crawford M.H. (ed.), Current developments in anthropological genetics, 1, 327-354, Plenum Publishing Corporation, New York.

Elston R.C., Stewart J., 1973. The analysis of quantitative traits for simple genetic models from parental, F1 and backcross data. Genetics, 73, 695-711.

Fain P.R., 1978. Characteristics of simple sibship variance tests for the detection of major loci and application to height, weight and spatial performance. Ann. Hum. Genet., 42, 109-120.

Hanset R., Michaux C., 1985. On the genetic determinism of muscular hypertrophy in the Belgian White and Blue cattle breed. II. Population data. Génét. Sél. Evol., 17, 369-386.

Karlin S., Carmelli D., Williams R., 1979. Index measures for assessing the mode of inheritance of continuously distributed traits. I. Theory and justifications. Theor. Pop. Biol., 16, 81-106.

Le Roy P., 1989. Méthodes de détection de gènes majeurs. Application aux animaux domestiques. Thèse de Doctorat, Université Paris Sud-Orsay, 229pp.

Le Roy P., Elsen J.M., 1992. Simple test statistics for major gene detection : a numerical comparison. Theor. Appl. Genet., 83, 635-644.

Le Roy P., Elsen J.M., Knott S., 1989. Comparison of four statistical methods for detection of a major gene in a progeny test design. Genet. Sel. Evol., 21, 341-357.

Mérat P., 1968. Distributions de fréquences, interprétation du déterminisme génétique des caractères quantitatifs et recherche de "gènes majeurs". Biometrics, 24, 277-293.

Titterington D.M., Smith A.F.M., Makov U.E., 1985. Statistical analysis of finite mixture distributions, John Wiley and Sons, New York.