Principes de base de la sélection génomique

Christèle Robert-Granié, Andres Legarra, Vincent Ducrocq

INRA Prod. Anim.,
2011, 24 (4), 331-340

Principes de base de la sélection génomique

C. ROBERT-GRANIÉ[1], A. LEGARRA[1], V. DUCROCQ[2,3]
[1] INRA, UR631 Station d'Amélioration Génétique des Animaux, F-31326 Castanet-Tolosan, France
[2] INRA, UMR1313 Génétique Animale et Biologie Intégrative, F-78352 Jouy-en-Josas, France
[3] AgroParisTech, Génétique Animale et Biologie Intégrative, 16 rue Claude Bernard, F-75231 Paris, France
Courriel : Christele.Robert-Granie@toulouse.inra.fr

Les progrès technologiques récents en matière de génotypage (disponibilité de puces SNP chez plusieurs de nos espèces d'intérêt) ouvrent de nouvelles perspectives dans le domaine de la sélection animale. L'information moléculaire dense de ces marqueurs permet une estimation plus précise de la valeur génétique d'un individu dès la naissance. Les programmes d'évaluation génétique des reproducteurs doivent être revisités pour prendre en compte ces nouvelles informations.

L'objectif des évaluations génomiques, décrites par extension sous le terme de «sélection génomique», est de prédire la valeur génétique d'un individu à partir de marqueurs denses couvrant l'ensemble du génome. Ces marqueurs sont des SNP dont le polymorphisme est caractérisé par la substitution d'une base d'ADN par une autre (cf. encadré 1 et figure 1). Les hypothèses sous-jacentes à cette approche sont les suivantes : la majeure partie de la variance génétique est expliquée par de nombreux QTL (*Quantitative Trait Loci*), la plupart à petits effets ; ces QTL sont en déséquilibre de liaison avec des marqueurs ; seuls les effets additifs sont considérés ; grâce à la loi des grands nombres, une bonne prédiction de la valeur génétique d'un individu (somme de l'effet de tous les marqueurs) est attendue, sans qu'il soit nécessaire de connaître précisément les effets individuels des marqueurs ou des QTL.

Dès 1990, Lande et Thomson (1990) imaginaient la sélection génomique en proposant le calcul d'un «score moléculaire» pour estimer la valeur génétique des animaux. Ce score était basé sur l'utilisation de marqueurs proches des segments chromosomiques influençant un caractère quantitatif (les QTL). Vers la fin des années 90, Haley et Visscher (1998) proposaient d'utiliser plusieurs milliers de marqueurs moléculaires couvrant le génome pour estimer la valeur génétique des individus. L'idée sous-jacente était qu'en considérant les marqueurs sur l'ensemble du génome, on doit être capable de suivre la transmission d'une génération à l'autre de tous les gènes intervenant sur un phénotype d'intérêt. L'estimation de la valeur génétique d'un individu consiste ainsi à calculer l'effet de chaque fragment de son génome. On commence alors à parler d'évaluation génomique.

En 2001, Meuwissen *et al* (2001) ont proposé les premiers modèles d'évaluation génomique qu'ils ont alors testés à partir de données simulées. Ils ont montré qu'on pouvait atteindre une précision d'évaluation génétique élevée dès la naissance de l'animal. A partir de 2007, la disponibilité des puces bovines 54K a permis de tester la sélection génomique en grandeur réelle. La densité de marqueurs utilisés est en effet telle que les QTL sont obligatoirement en déséquilibre de liaison avec les marqueurs les plus proches, au moins intra-race.

1 / La sélection génomique

La sélection génomique est basée sur les mêmes principes que la sélection classique : à partir de mesures phénotypiques, on prédit à l'aide d'une méthode statistique appropriée la valeur génétique des individus, y compris de ceux qui n'ont pas de phénotypes ou de descendance. Classiquement, cette prédiction est faite à partir de la connaissance de la parenté entre individus, qui permet de modéliser statistiquement leur covariance génétique et ainsi de faire des prédictions. Ces prédictions utilisent soit la théorie des index de sélection, soit plus couramment et plus généralement, le BLUP (meilleure prédiction linéaire non biaisée) qui conduit à une estimation simultanée des effets génétiques et de milieu (Ducrocq 1990, 1992).

L'originalité de la sélection génomique réside dans la manière de réaliser cette prédiction statistique : l'information abondante sur la transmission des marqueurs s'ajoute ou se substitue à l'information de parenté classique. La sélection génomique consiste dans un premier temps à génotyper et phénotyper un grand nombre d'individus (en général, et pour des raisons de coût, des mâles testés sur descendance) et à établir une relation statistique (décrite par une équation de prédiction) entre les génotypes aux marqueurs et les phénotypes. Les animaux génotypés et phénotypés sur lesquels se base l'équation de prédiction forment la «population de référence». De manière générale, on estime que cette population doit être au moins de l'ordre de 1000 individus pour que la relation entre génotype et phénotype soit suffisamment fiable et précise. La figure 2, extraite de l'article de Hayes *et al* (2009a), illustre le nombre de phénotypes nécessaire dans la population de référence pour atteindre une précision des valeurs génomique de 0,5 ou 0,7 en fonction de l'héritabilité du caractère étudié. Il est clairement mis en évidence que la population de référence doit être d'autant plus grande que l'héritabilité du caractère d'intérêt est faible. Notons également que la précision ne dépend pas ou peu de l'individu, elle est identique chez les mâles et les femelles. Une fois la relation entre les génotypes aux marqueurs et les phénotypes établie, il est alors possible de prédire un index de sélection génomique pour des

Encadré 1. Quelques définitions

Les **Single Nucleotide Polymorphisms** (SNP) : ce sont des mutations qui n'impliquent le changement que d'un nucléotide (A,G,T,C) en un locus donné. Ils sont de manière générale bialléliques (m, M). Chaque individu sera donc porteur au niveau d'un SNP d'un des trois génotypes possibles, les deux génotypes homozygotes (mm et MM) et le génotype hétérozygote (mM ou Mm indiscernables l'un de l'autre). Les SNP apparaissent en moyenne tous les 100 à 1000 bases. Ce sont donc des marqueurs qui représentent une source d'information riche et abondante : on peut trouver plusieurs millions de SNP dans le génome d'une même espèce. Leur connaissance permet de caractériser l'état du segment de chromosome où ils se trouvent.

La diminution à la fois du coût et du temps de séquençage a induit un intérêt croissant pour l'utilisation de cette information. Les SNP localisés dans les régions codantes d'un gène peuvent par ailleurs jouer un rôle direct en altérant la forme et ainsi la fonction de la protéine produite à partir du gène en question

Figure 1. *Un saut technologique : les puces à SNP.*

SNP = «*Single Nucleotide Polymorphism*» = polymorphisme de l'ADN en une base

- Les SNP sont très généralement bialléliques
- Ils existent par millions
- Ils sont «faciles» à détecter
- On dispose de génotypages à haut débit

Technique de génotypage haut-débit : pour déterminer le génotype des SNP d'un individu, son ADN est fragmenté et marqué par un fluorochrome dont la couleur dépend de l'allèle présent sur la séquence d'ADN. Cet ADN sonde est ensuite déposé sur une plaque ou «puce à ADN» qui contient l'ADN cible sur lequel il s'hybride. Chaque puits de cette puce contient l'ADN cible correspondant à un locus donné. La puce est ensuite «lue» à l'aide d'un scanner et en fonction du signal de fluorescence observé, on détermine le génotype de l'individu pour le locus considéré. Par exemple, on peut imaginer qu'une couleur «noire» correspondra au génotype homozygote AA, «blanche» au génotype homozygote aa, et un mélange des deux couleurs («gris») au génotype hétérozygote aA. Chaque locus est représenté plusieurs fois sur la puce de façon à améliorer la fiabilité des résultats.

Puce bovine 54K : La principale puce commerciale actuelle chez les bovins, la «*Illumina BovineSNP50™ beadchip*» contient de l'ordre de 54 000 SNP (d'où le nom usuel de «puce 54K») avec une bonne couverture du génome (un marqueur tous les 45 000 bases (45 Kb) en moyenne, pas d'intervalle inférieur à 22 Kb entre deux marqueurs et peu d'intervalles supérieurs à 70 Kb). Elle possède une informativité élevée : la fréquence moyenne de l'allèle rare de chaque SNP est supérieure à 20% chez la plupart des races *bos taurus*.

jeunes animaux (sans phénotypes), uniquement sur la base de leur génotype aux marqueurs SNP.

Ces nouvelles avancées permettent ainsi de prédire la valeur génétique d'un animal dès sa naissance, avant même de connaître ses performances ou celles de ses descendants avec des niveaux de précision relativement élevés (mais généralement plus faibles qu'un testage sur descendance) et en tout cas plus grande que celle permise par l'information sur ascendance. Ceci peut entraîner potentiellement un raccourcissement de l'intervalle de génération et un gain de précision des valeurs génétiques estimées pour les femelles. Notons toutefois que la validité de cette prédiction repose sur la fiabilité de la relation entre performances et génotypes aux marqueurs SNP.

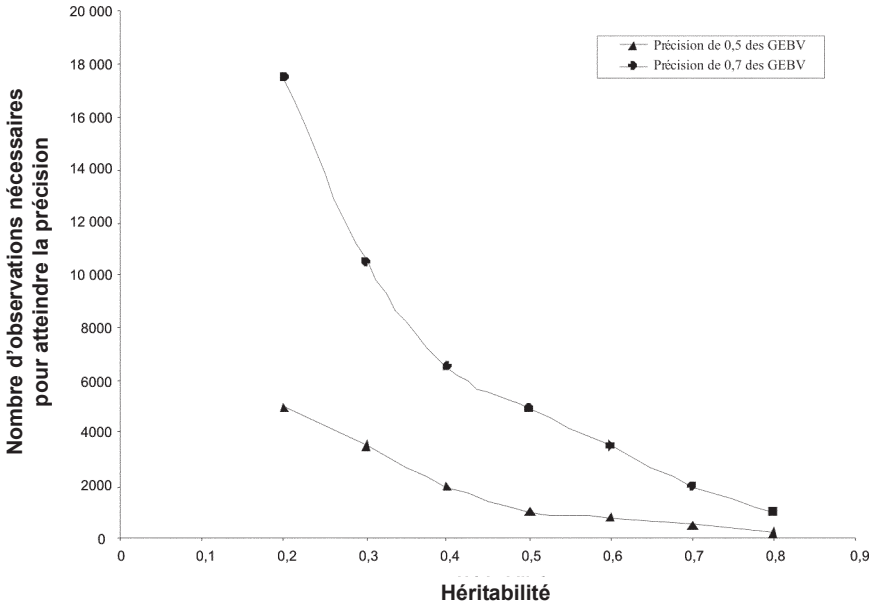
La sélection génomique permet également d'envisager l'analyse de nouveaux caractères peu héritables ou difficilement mesurables (caractères de santé et de bien-être animal, caractères mesurés à l'aide d'automates de traite, qualité de viande, composition fine du lait…).

1.1 / Phénotypes

Les phénotypes peuvent être soit des caractères directement mesurables sur l'individu lui-même (par exemple : un poids, une taille, une croissance), soit des «pseudo-performances» (non mesurables sur un individu donné mais pour lequel on souhaite lui attribuer une «mesure») obtenues à partir des évaluations génétiques classiques pour des animaux apparentés à l'individu considéré (par exemple : la performance moyenne des productions laitières des filles d'un taureau, corrigée par un ensemble d'effets du modèle, représentera la pseudo-performance du taureau). Une autre possibilité qui a été parfois retenue est l'utilisation des valeurs génétiques estimées (les «index») issues des évaluations génétiques officielles. Cela pose des problèmes de redondance, car ces index utilisent déjà une information génétique que l'on voudrait capturer à travers les SNP. Dans les populations de référence actuelles chez les bovins et ovins laitiers, on utilise couramment comme phénotype pour les mâles la performance moyenne des filles du mâle, appelée DYD (*Daughter Yield Deviation*, VanRaden et Wiggans 1991). Cette mesure est définie comme la moyenne des performances des filles du taureau, corrigées pour l'ensemble des facteurs non génétiques inclus dans le modèle d'évaluation génétique et pour la moitié de la valeur génétique de leur mère. Il est analogue à une performance propre du taureau, d'héritabilité égale à la précision de l'index sur descendance. Il a l'avantage de concentrer l'information phénotypique de nombreuses filles sur un seul individu génotypé. Lorsque les DYD sont utilisés dans les modèles d'évaluation, une précision (ou un poids) leur est associée qui est fonction de la contribution des filles (EDC pour *Equivalent Daughter Contribution*). Dans le cas où les DYD ne sont pas disponibles (taureaux étrangers par exemple), ils sont remplacés par les «index dérégressés» qui sont en principe équivalents aux DYD (Thomsen *et al* 2001) : par construction, si on les utilise comme données de base dans une évaluation BLUP, on obtient les mêmes évaluations génétiques des mâles qu'à partir du modèle complet.

Pour les femelles, on utilise les YD (*Yield Deviation*) qui sont définies comme la moyenne des performances de la femelle, corrigées pour les facteurs non génétiques du modèle. Comme pour les DYD, cela permet une évalua-

Figure 2. *Nombre de phénotypes nécessaire pour atteindre une précision des valeurs génomiques (GEBV) de 0,5 ou 0,7, en fonction de l'héritabilité du caractère.* La taille effective de la population (Ne) est de 1000 et une distribution normale des effets des QTL est supposée. Figure extraite de l'article de Hayes *et al* (2009a).



tion plus simple, car déjà corrigée pour les effets de milieu.

L'avantage de travailler à partir des DYD, lorsqu'ils sont disponibles, est la possibilité de diminuer la taille requise de la population de référence puisqu'ils sont précis (ce qui peut être utile pour des petites populations). En effet, cela permet de disposer de phénotypes pour des caractères variés avec des précisions comprises entre 0,5 et 0,9, précision plus élevée et plus homogène que les phénotypes bruts mesurés chez les femelles (0,01 à 0,5 environ).

1.2 / Modèles d'évaluation basés sur l'information moléculaire

Le modèle de base utilisé en évaluation est l'équation classique de la génétique quantitative *Phénotype = Génétique + Environnement*. Meuwissen *et al* (2001) ont proposé une extension simple de ce modèle aux marqueurs moléculaires avec, pour l'animal *j* :

$$y_j = \mu + \sum_i \mathbf{x}_{ij}\, \mathbf{g}_i + \mathbf{e}_j$$

ou sous forme matricielle, pour l'ensemble des animaux de la population de référence :

$$\mathbf{y} = \mu\, \mathbf{1} + \mathbf{X}\, \mathbf{g} + \mathbf{e}$$

où $\mathbf{y}$ est le vecteur des *n* phénotypes $\{y_j\}$considérés (les YD ou DYD) ; μ est la moyenne des phénotypes ; $\mathbf{1}$ est un vecteur unité de dimension *n* ; $\mathbf{x}_{ij}$ est un terme reliant le phénotype y_j de chaque individu *j* aux allèles *i* des marqueurs SNP qu'il possède ; $\mathbf{g}_i$ est l'effet de l'allèle i sur le phénotype. $\mathbf{X}$ est une matrice d'incidence dont les lignes sont les valeurs de $\mathbf{x}_{ij}$ pour chaque individu *j*, chaque colonne correspondant à un marqueur i. La valeur génétique d'un individu *j* est estimée par $\hat{u}_j = \sum_i \mathbf{x}_{ij}\, \mathbf{g}_i = \mathbf{x}_j\mathbf{g}$, où $\mathbf{x}_j$ est le vecteur ligne qui contient le génotype aux marqueurs de l'individu *j*, et $\mathbf{g}$ est l'ensemble des effets estimés des marqueurs.

Dans sa formulation initiale, le modèle décrit ci-dessus n'est pas limité aux marqueurs bialléliques. Dans un modèle n'estimant que des effets de SNP, le modèle se simplifie sans perdre en précision si on définit le génotype (les $\mathbf{x}_{ij}$) comme le nombre de copies (0, 1 ou 2) portées par un individu à chaque SNP pour l'un des deux allèles choisi comme référence.

Une fois le modèle formulé, il est nécessaire de choisir une méthode pour l'estimation statistique de ses inconnues. Une estimation naïve, par la méthode des moindres carrés par exemple, n'est pas souhaitable pour deux raisons, l'une pratique, l'autre théorique : tout d'abord, le nombre d'inconnues est beaucoup plus grand que la quantité de données disponible, rendant l'estimation très peu précise. De plus, il a été établi en sélection animale (Goffinet et Elsen 1984, Gianola et Fernando 1986) que pour un progrès génétique optimal, il est souhaitable de prendre en compte la distribution des effets génétiques. L'inclusion de cette distribution revient à ajouter une information a priori sur ces effets, dits aléatoires. Les méthodes statistiques optimales estiment les effets à travers l'espérance conditionnelle d'une distribution «a posteriori» qui combine information a priori et vraisemblance des données.

Parmi les nombreuses approches proposées pour estimer les effets des SNP, les plus courantes sont brièvement décrites ci-dessous.

1.3 / BLUP appliqués aux effets des SNP/Régression Ridge

Dans les modèles d'évaluations génétiques classiques, ce sont les effets génétiques qui sont conceptuellement tirés aléatoirement dans une «population» d'effets dont on connaît au moins partiellement la distribution. Dans le contexte d'une évaluation génomique, on peut supposer que les effets des SNP proviennent d'une loi normale centrée sur 0. On estime alors ces effets de SNP en utilisant le BLUP d'Henderson (Henderson 1963, 1973). Ce dernier est un estimateur adapté aux modèles mixtes, combinant effets fixes et aléatoires et possède certaines propriétés optimales (estimateur linéaire non biaisé, d'erreur minimale). Le BLUP appliqué à des données génomiques est souvent appelé GBLUP, pour «*Genomic BLUP*» et conduit au système d'équations suivant :

$$\begin{pmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'\mathbf{X} \\ \mathbf{X}'\mathbf{1} & \mathbf{X}'\mathbf{X} + \sigma_e^2/\sigma_g^2\mathbf{I} \end{pmatrix} \begin{pmatrix} \hat{\mu} \\ \hat{\mathbf{g}} \end{pmatrix} = \begin{pmatrix} \mathbf{1}'\mathbf{y} \\ \mathbf{X}'\mathbf{y} \end{pmatrix}$$

dans lequel σ^2_g représente la variance de la distribution des effets des marqueurs et σ^2_e est la variance résiduelle du modèle. Ces variances peuvent être calculées par des méthodes paramétriques (méthode du maximum de vraisemblance ou méthodes Bayésiennes) ou par des techniques non paramétriques basées sur la validation croisée, ou bien encore elles peuvent être déduites à partir des variances utilisées dans les évaluations génétiques classiques. Moyennant une organisation judicieuse des calculs (Legarra et Misztal 2008, Strandén et Garrick 2009), la résolution itérative d'un tel système est relativement aisée.

Ce même système d'équations peut également être obtenu en utilisant une technique statistique complètement différente : la régression Ridge, introduite par Hoerl et Kennard (1970). Elle consiste à modifier la méthode des moindres carrés (MC) de manière à stabiliser les effets estimés (ici, les SNP sont supposés être des effets fixes), en rajoutant un terme sur la diagonale du système d'équations. Dit autrement, on utilise la méthode des moindres carrés en y ajoutant une contrainte, une «pénalisation», sur la valeur de la somme des carrés des effets. De ce fait, cette

méthode introduit un biais sur les estimations des paramètres (alors que les paramètres obtenus par MC sont non biaisés). Ce léger inconvénient est compensé par la réduction de la variance des paramètres, et même par la réduction de leur erreur quadratique moyenne. On notera simplement que le rapport $\sigma_e^2/\ \sigma_g^2$ du GBLUP représente le terme rajouté sur la diagonale de X'X dans le cadre de la régression Ridge.

1.4 / GBLUP/Parenté génomique

Nous avons vu qu'à partir de l'expression du GBLUP décrite ci-dessus, la valeur génétique d'un individu *j* est égale à $\hat{u}_j = \sum_i \mathbf{x}_{ij}\, \mathbf{g}_{i.}$. Il est possible de transformer le modèle initial en un modèle totalement équivalent de façon à obtenir directement les mêmes valeurs génétiques individuelles $\hat{u}_j$. Le système d'équations du BLUP s'écrit alors de la manière suivante :

$$\begin{pmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}' \\ \mathbf{1} & \mathbf{I}+\sigma_e^2/\sigma_g^2\left(\mathbf{XX}'\right)^{-1} \end{pmatrix}\begin{pmatrix} \hat{\mu} \\ \hat{\mathbf{u}} \end{pmatrix}=\begin{pmatrix} \mathbf{1}'\mathbf{y} \\ \mathbf{y} \end{pmatrix}$$

On peut justifier l'utilisation de l'appellation «GBLUP» pour cette expression aussi : à titre de comparaison, avec une évaluation BLUP classique n'utilisant que les parentés classiques, le système d'équations du BLUP serait le suivant :

$$\begin{pmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}' \\ \mathbf{1} & \mathbf{I}+\alpha\mathbf{A}^{-1} \end{pmatrix}\begin{pmatrix} \hat{\mu} \\ \hat{\mathbf{u}} \end{pmatrix}=\begin{pmatrix} \mathbf{1}'\mathbf{y} \\ \mathbf{y} \end{pmatrix}$$

où **A** est la matrice de parenté «espérée» entre les individus et α est le rapport entre la variance résiduelle et la variance génétique additive. Différents auteurs (VanRaden 2008, Goddard 2009) ont montré que, pour un codage particulier (centré sur 0) des SNP, il est possible de construire une matrice dite «de parenté génomique» $\mathbf{G} = \mathbf{XX}' / 2\sum p_i q_i$ analogue à la matrice de parenté **A** et jouant le même rôle dans les équations BLUP. Le terme p_i représente la fréquence de l'allèle de référence du SNP *i* et $q_i = 1 - p_i$.

Les éléments de la matrice **G** mesurent la proportion moyenne d'allèles partagés par deux individus pour tous les SNP, pondérés par leur fréquence : le partage d'allèles plus rares est plus indicatif de la parenté. A condition de la calculer avec une quantité suffisante de marqueurs, cette «parenté génomique» est plus précise que celle basée sur le pedigree, car cette dernière est basée sur une généalogie tôt ou tard incomplète et elle ne prend pas en compte les écarts à la théorie dus à la liaison entre loci et à la taille réelle (finie) du génome.

Contrairement au cas de l'évaluation génétique basée sur la généalogie, l'inverse de la matrice **G** est dense, mais de nouveau des astuces peuvent éviter son calcul explicite par une résolution itérative du système (Strandén et Garrick 2009).

1.5 / Les méthodes Bayésiennes

Le GBLUP suppose implicitement que les effets des SNP suivent une loi normale, comme cela a été constaté lors de simulations, mais aussi à partir de données réelles. Si cette hypothèse n'est pas respectée, les prédictions ne sont pas optimales. Intuitivement, les effets des marqueurs devraient être faibles voire très faibles dans leur grande majorité (on ne s'attend pas à ce qu'un grand nombre de SNP soient des mutations causales ou en très fort déséquilibre de liaison avec les mutations causales), et symétriques (existence d'effets positifs ou négatifs). Le principal «problème» du GBLUP est qu'il interdit de fait des effets de SNP de grande taille. Or ce type de SNP, bien que peu fréquent, existe : c'est le cas par exemple, de la mutation sur le gène DGAT1 chez les bovins (Grisart *et al* 2002) qui influence fortement la composition en matière grasse du lait. Les méthodes Bayésiennes lèvent cette restriction, en permettant l'introduction de distributions plus adaptées à la variabilité des effets des SNP.

Bien qu'attractives parce que plus générales, les méthodes Bayésiennes ont la réputation d'être plus compliquées. Cependant les calculs associés deviennent aujourd'hui abordables à l'aide de techniques adaptées basées sur des simulations répétées, les MCMC (*Monte Carlo Markov Chain*) (Metropolis *et al* 1953, Robert 1996). Elles sont ainsi de plus en plus utilisées pour estimer les variances nécessaires pour les évaluations BLUP ou GBLUP.

Les différentes méthodes Bayésiennes utilisées en sélection génomique se distinguent par les hypothèses faites concernant la distribution des effets de SNP. Elles reposent sur des modèles hiérarchiques : on décrit par exemple la forme générale de la distribution d'un effet qui dépend d'un paramètre, par exemple une variance, qui provient elle-même d'une distribution générale des variances d'effets des SNP, etc.

BayesA : la méthode BayesA mise en avant par Meuwissen *et al* (2001) stipule ainsi que les effets des SNP proviennent d'une distribution normale avec une variance spécifique, différente d'un SNP à l'autre, de manière à ce que l'ordre de grandeur potentiel des effets des marqueurs soit variable. Gianola *et al* (2009) ont démontré qu'en fait, cela revient à postuler pour les effets des SNP une distribution *t* multivariée de faible degré de liberté. Cette distribution ressemble à une distribution normale «écrasée» avec des queues plus épaisses : de «gros» effets de SNP deviennent possibles contrairement au GBLUP.

BayesB : cette approche également proposée par Meuwissen *et al* (2001) suppose qu'une proportion donnée des SNP a un effet nul (avec une variance nulle) et ne contribue donc pas au caractère. Les autres effets de SNP (on ne sait pas lesquels à ce stade) sont tirés d'une distribution *t* comme pour le BayesA. Sur la base des résultats de Meuwissen *et al* (2001) et de nombreux travaux ultérieurs, la méthode BayesB est souvent considérée comme la référence en termes d'efficacité de prédiction génomique, mais elle est extrêmement coûteuse en temps de calcul. De plus, cette supériorité observée dans le cadre de simulations n'apparaît pas toujours lors de l'analyse de données réelles. Plusieurs méthodes apparentées existent, avec des efficacités plus ou moins similaires (BayesC, BayesCpi, BayesD, BayesR) mais des vitesses de convergence plus élevées. Pour en savoir plus sur ces méthodes, le lecteur peut se référer aux articles de Kizilkaya *et al* (2010) et Habier *et al* (2010). Elles ont été développées dans le but de réduire ou simplifier les calculs. Notons que la méthode BayesB et ses approximations gagnent en intérêt lorsque le nombre de marqueurs augmente. En effet, plus le nombre de marqueurs est élevé, plus la proportion de SNP avec un effet nul vrai diminue. A contrario, l'efficacité du GBLUP tend à plafonner lorsque le nombre de marqueurs augmente car la parenté n'est pas mieux estimée.

Lasso Bayésien : il fait l'hypothèse que les effets des marqueurs suivent une loi de Laplace (ou «double exponentielle»). Cette loi reflète le fait qu'un grand nombre de SNP sont supposés avoir un effet pratiquement nul et que potentiellement quelques marqueurs ont un grand effet et sont moins «contraints» que pour la loi normale. En pratique, les propriétés du Lasso Bayésien (un peu différent du Lasso classique décrit plus loin) sont similaires à l'approche BayesB, avec une réduction notable de la complexité des calculs (De los Campos *et al* 2009). Différentes formulations existent :

celle originale de Tibshirani (1996), et la modification de Park et Casella (2008) pour laquelle les variances des résidus et des effets des marqueurs ne sont pas indépendantes (Legarra *et al* 2011).

1.6 / Les méthodes de sélection de variables

Une autre manière de surmonter la difficulté d'estimer un très grand nombre d'effets de SNP (p variables) à partir d'une population de référence de taille limitée (n observations) – ce que les statisticiens appellent le problème de «$p >> n$» – est de faire appel aux techniques spécifiques développées dans d'autres domaines, et en particulier celles visant à réduire le nombre de paramètres à estimer, en sélectionnant les plus pertinents.

Lasso : le Lasso («*Least Absolute Shrinkage and Selection Operator*») est une méthode de régression ajoutant à la méthode des moindres carrés une contrainte ou une pénalisation sur la valeur absolue des effets de SNP. Cette contrainte a un effet de sélection de variable et de régularisation sur l'estimateur (Tibshirani 1996). D'une part, les coefficients sont régressés vers 0 : l'introduction d'un biais entraîne une réduction de la variance. D'autre part, certains coefficients sont annulés (mis à zéro), par conséquent, l'estimation et la sélection de variables sont effectuées simultanément. Un des inconvénients de cette méthode est que s'il y a un groupe de prédicteurs très corrélés entre eux, le Lasso tend à n'en sélectionner qu'un seul et ce prédicteur est choisi de façon quelconque dans le groupe. Cet inconvénient a peu d'impact lorsque l'on recherche le meilleur modèle prédictif (c'est le cas de l'évaluation génomique), ce qui n'est pas nécessairement le cas dans un cadre de statistique inférentielle (estimation et interprétation de chacune des variables du modèle). Le Lasso revient à postuler la même loi de Laplace pour les effets des marqueurs et à estimer non plus l'espérance conditionnelle de la distribution a posteriori des effets mais sa valeur maximale (le mode). Malgré sa faible utilisation, l'efficacité du Lasso a été démontrée (Usai *et al* 2009). Son interprétation est simple et il bénéficie d'un algorithme de calcul extrêmement efficace.

Elastic Net : L'Elastic Net (ou EN) (Zhou et Hastie 2005) est un estimateur qui combine deux méthodes, le Lasso et la régression Ridge, dans des proportions respectives τ et $1\text{-}\tau$ où τ est un paramètre à déterminer. Les deux méthodes font intervenir une pénalisation, la première sur la somme des valeurs absolues des effets, la deuxième sur la somme des carrés de ces effets. Les avantages de l'Elastic Net sont sa capacité à pallier certains comportements limitants du Lasso (comme des problèmes numériques en présence de forte colinéarité entre les effets) et de la régression Ridge (qui considère obligatoirement que tous les SNP ont un effet) ainsi que sa facilité de calcul. Comme pour le Lasso, la capacité de choisir un sous-ensemble de marqueurs pertinents est tout à fait intéressante. Néanmoins, cet ensemble peut varier au cours du temps avec l'ajout de nouvelles données. D'autre part, le calibrage de l'Elastic Net peut être délicat (Croiseau *et al* 2010). Il est utilisé, dans le contexte de l'évaluation BLUP-QTL française en bovins laitiers, pour sélectionner l'ensemble des SNP «à effet faible» qui complètent les QTL à effet plus fort.

Régression Partial Least Squares (PLS) et Sparse PLS : la régression PLS, introduite par Wold (1966), est la technique principale de modélisation prédictive dans les situations où les prédicteurs sont plus nombreux que les observations, fortement corrélés, et avec de nombreuses données manquantes. Son champ d'application initial est la chimiométrie, mais la régression PLS est très générale, et se développe rapidement dans tous les domaines (économie, médecine, psychologie, génétique…). Elle peut être perçue comme une généralisation de la Régression Linéaire Multiple mais également de la Régression sur Composantes Principales (RCP) et de l'Analyse Canonique. La régression PLS remplace l'espace initial des (nombreuses) variables explicatives par un espace de faible dimensionnalité, décrit par un petit nombre de variables appelées «facteurs» ou «variables latentes» qui sont construites l'une après l'autre de façon itérative. Ces facteurs seront les nouvelles variables explicatives d'un modèle de régression linéaire classique. Les variables latentes (ou facteurs), ainsi construites, sont orthogonales (non corrélées), et sont des combinaisons linéaires des variables explicatives initiales. A ce titre, elles ressemblent beaucoup aux Composantes Principales de la RCP. Mais alors que ces dernières ne sont calculées qu'à partir des variables initiales (et donc sans référence à la variable à expliquer y), les variables latentes de la régression PLS prennent en compte leur utilité individuelle pour prédire y en maximisant leurs corrélations successives avec y, tout en maintenant la contrainte d'orthogonalité avec les facteurs déjà construits (Tenenhaus 1998). Le nombre de variables latentes introduites dans le modèle de régression final peut être déterminé par des techniques de validation croisée.

La régression Sparse PLS, définie par Le Cao *et al* (2008), est une variante de la méthode PLS. Elle introduit une étape supplémentaire permettant de sélectionner simultanément les variables initiales entrant dans la composition des variables latentes construites, par introduction d'une pénalité de type valeur absolue, comme dans le Lasso. Cette approche, développée initialement dans le contexte de l'analyse des données transcriptomiques, a été appliquée à la sélection génomique des données bovines et ovines laitières françaises, avec des résultats très similaires aux approches citées précédemment (GBLUP, BayesCpi, Lasso, EN ; Colombani *et al* 2010, 2011a, Robert-Granié *et al* 2011).

Les SNP les plus importants mis en évidence par les méthodes de sélection de variables (BayesCpi, EN, sparse PLS) peuvent être intéressants pour cibler des régions chromosomiques ayant un effet sur les caractères étudiés (Colombani *et al* 2011b)

1.7 / Une approche «en une étape»

Une question controversée en sélection génomique est comment combiner les informations phénotypique et génomique chez les animaux génotypés hors population de référence. Pour cela, des approches adhoc ont été proposées et appliquées, mêlant les résultats des évaluations génomiques et ceux des évaluations classiques en utilisant la théorie des index de sélection (Van Raden *et al* 2009) ou le BLUP (Ducrocq et Liu 2009) en essayant d'éviter l'inclusion d'informations redondantes, mais elles ne sont pas entièrement satisfaisantes, car elles impliquent plusieurs étapes successives et indépendantes (évaluation classique, calcul des DYD, évaluation génomique puis combinaison des évaluations).

Comme déjà indiqué, l'évaluation génomique repose sur des «pseudo-phénotypes» d'individus génotypés. Mais si cette procédure est bien établie pour les mâles des filières laitières avec l'utilisation de DYD, c'est loin d'être le cas pour d'autres espèces, par exemple en bovins allaitants avec des effets maternels difficiles à mesurer ou en porcs avec des performances mesurées soit sur l'individu, soit sur ses apparentés. De plus, la génération des «pseudo-phénotypes» suppose que l'indexation de base est robuste et non biaisée. Or, ce ne sera plus le cas si la sélection se base fortement sur les index génomiques (Patry et Ducrocq 2011) : une des hypothèses de base des évaluations classiques est qu'en espérance, l'aléa de méiose, c'est-à-

dire l'écart moyen entre la valeur génétique d'un animal et la moyenne de ses parents est nul. Cette hypothèse n'est plus du tout respectée si seuls les jeunes animaux dont l'aléa de méiose est élevé sont retenus après évaluation génomique. Il en résulte des évaluations génétiques classiques (et donc des pseudo-performances) biaisées.

Enfin, il paraît naturellement souhaitable de faire bénéficier les animaux non génotypés de l'information génomique de leurs apparentés, ce qui n'est pas le cas avec les méthodes précédemment décrites.

Pour éviter l'ensemble de ces difficultés, la solution passe par une analyse simultanée de toutes les données quelle que soit leur origine, phénotypique ou génomique. C'est dans ce contexte qu'a été proposée la méthode dite «en une étape» (*single step*). Dans cette méthode (Aguilar *et al* 2010), on prédit de façon probabiliste les marqueurs d'individus non génotypés à partir du génotype d'animaux apparentés. Cette information s'intègre dans une matrice de parenté «généalogique-génomique» dont l'expression est la suivante :

$$\mathbf{H} = \begin{bmatrix} \mathbf{A}_{11} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\left(\mathbf{G} - \mathbf{A}_{22}\right)\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G} \\ \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G} \end{bmatrix}$$

où les indices 1 et 2 correspondent respectivement aux animaux non génotypés et génotypés, et **A** aux parentés «classiques» basées sur la généalogie uniquement. Ici, la matrice **G** peut être la matrice de parenté «génomique» définie précédemment, ou une autre matrice décrivant les covariances entre individus créée par exemple à partir du BayesB ou du Lasso Bayésien.

Or, cette matrice a un inverse extrêmement simple (Aguilar *et al* 2010), inverse dont on a besoin dans le système des équations du BLUP :

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$

De cette manière, on combine divers avantages : un calcul correct et sans biais et une information génomique qui est combinée naturellement aux données phénotypiques, que les animaux soient génotypés ou non. Néanmoins, le calcul explicite des inverses de **G** et $\mathbf{A_{22}}$ reste coûteux en temps de calcul, mais des travaux en cours nous laissent espérer des expressions équivalentes pouvant alléger les calculs (Ducrocq et Legarra 2011).

2 / Efficacité de la sélection génomique

2.1 / Validation des évaluations génomiques

La précision des évaluations génomiques est le plus souvent appréciée par validation croisée, à travers des études rétrospectives : supposons que nous sommes en 2010 avec un ensemble d'individus génotypés, c'est-à-dire une population de référence. On dispose d'un estimateur relativement précis de leur valeur génétique vraie, typiquement un index sur descendance ou les DYD. On se place quelques années en arrière, par exemple en 2007, et on réalise une évaluation génomique à partir de la population de référence de l'époque, qu'on appellera population «d'apprentissage» ou de «calibration». Puis on s'intéresse aux animaux sans phénotypes associés en 2007 (très généralement ce sont les plus jeunes en 2007) qui constituent la population de «validation». La qualité de l'évaluation génomique est mesurée en comparant dans la population de validation les évaluations génomiques de 2007 avec les index sur descendance ou les DYD obtenues en 2010 (étude de corrélation, ou pente de régression).

De telles études rétrospectives ont été entreprises sur données simulées et données réelles chez les bovins et ovins laitiers, la poule ou la souris. De nombreux d'articles ont été publiés, dressant un panorama désormais assez clair. On présentera ici quelques appréciations tirées des articles de Legarra *et al* (2008), Van Raden *et al* (2009), Hayes *et al* (2009a et b), Luan *et al* (2009), Habier *et al* (2010), Su *et al* (2010), Legarra *et al* (2011), Wolc *et al* (2011).

2.2 / Comparaison des méthodes d'évaluation

Généralement, on peut observer que les méthodes de prédiction génomique présentent toutes des corrélations entre valeurs génétiques prédites et observations nettement meilleures que celles obtenues par les index sur ascendance, seuls disponibles auparavant. Les différences entre les meilleures méthodes sont généralement faibles avec des écarts de corrélation entre méthodes le plus souvent inférieurs à 2 ou 3%. Néanmoins, ce n'est pas le cas pour les caractères avec des QTL à effet fort, comme par exemple les taux et quantité de matière grasse chez les bovins laitiers, pour lequel le QTL DGAT1 explique une partie conséquente de la variance génétique. Dans ce cas, un modèle prenant en compte les écarts à la normalité (hypothèse de base du GBLUP) peut entrainer un gain allant jusqu'à 20% en précision. Certaines méthodes sont toujours parmi les meilleures (c'est le cas des méthodes BayesA, BayesCpi et Lasso Bayésien par exemple) tandis que d'autres sont plus sensibles (BayesB, probablement parce qu'il est alors nécessaire de fixer a priori la proportion de SNP à effets nuls). Les résultats présentés dans le tableau 1 illustrent les performances de quelques méthodes de sélection génomique sur les populations bovines laitières Holstein et Montbéliarde françaises.

D'autre part, de nombreux travaux ont mis en évidence que même pour la puce 54K, les marqueurs ne permettent généralement pas de retrouver la totalité de la variance génétique additive d'un caractère. La raison principale est qu'à part dans le cas de SNP responsables de mutation causale, le déséquilibre de liaison entre marqueurs et QTL est très rarement total. Il en résulte une qualité d'ajustement de la régression des valeurs génétiques observées sur leur prédiction qui n'est pas optimale, en particulier avec une pente inférieure à 1. Cela pose des problèmes de fiabilité de la comparaison entre valeurs génétiques d'animaux génotypés et non génotypés, par exemple entre jeunes taureaux et taureaux évalués sur descendance. L'adéquation de cette pente est d'ailleurs le critère privilégié par Interbull, l'organisme d'évaluations internationales des taureaux de races laitières, pour valider les méthodologies d'évaluation génomique utilisées dans chaque pays (Nilforooshan *et al* 2010). Un moyen d'améliorer le modèle consiste à ajouter un effet polygénique «résiduel» utilisant les informations de parenté classique pour estimer ce qui n'a pas pu l'être par les SNP (Liu *et al* 2011). Une autre solution consisterait à augmenter le déséquilibre de liaison entre marqueurs et QTL, en augmentant la densité de marqueurs ou encore en utilisant des haplotypes.

2.3 / Influence de la population de référence

Dès le début des évaluations génomiques, il est apparu que la taille de la population de référence joue un rôle considérable sur la précision des évaluations : tout accroissement du nombre d'individus génotypés se traduisait par une augmentation de la corrélation entre valeurs génétiques prédites et observées. De plus, il était alors sous-entendu que la précision de la prédiction génomique était la même pour tous les candidats. Il s'avère que le panorama est un peu plus nuancé. Il a été ainsi montré (Legarra *et al* 2008, Habier *et al*

Tableau 1. *Performances de quelques méthodes de sélection génomique : corrélations entre valeurs génétiques prédites et observations (DYD) sur les populations bovines laitières Holstein et Montbéliarde françaises.*

Caractère	BLUP	GBLUP	Elastic Net	PLS	Sparse PLS	BayesCp
Lait	*0,38* 0,28	*0,56* 0,44	*0,57* 0,45	*0,53* 0,44	*0,48* 0,41	*0,57* 0,44
Matière Protéique	*0,44* 0,27	*0,55* 0,46	*0,57* 0,46	*0,55* 0,46	*0,51* 0,45	*0,56* 0,46
Taux Butyreux	*0,44* 0,40	*0,72* 0,51	*0,80* 0,59	*0,70* 0,54	*0,65* 0,61	*0,80* 0,62
Taux de conception	*0,29* 0,43	*0,35* 0,43	*0,33* 0,47	*0,33* 0,43	*0,29* 0,44	*0,34* 0,42

Première ligne (en italique) : en race Holstein ;
Seconde ligne : en race Montbéliarde
En rouge : méthode conduisant à la meilleure corrélation
Populations d'apprentissage/validation : 2976/964 en Holstein, 950/222 en Montbéliarde

2010) que plus les candidats sont apparentés aux individus génotypés et phénotypés, plus précise est la prédiction de son évaluation génomique. En effet, les relations de parenté proches induisent de très forts déséquilibres de liaison entre QTL et marqueurs. Ces déséquilibres sont de courte durée car ils peuvent être rompus à chaque recombinaison, c'est-à-dire à chaque génération. Ceci compromet la capacité prédictive d'une évaluation génomique au-delà d'un nombre faible de générations, même si certaines approches (BayesB) semblent légèrement plus robustes que d'autres telles que le GBLUP (Habier *et al* 2007, 2010). Ceci conduit donc à la nécessité d'entretenir une population de référence relativement jeune et de ré-estimer régulièrement les effets des SNP. Cependant, on peut imaginer que si la population de référence est très grande et le déséquilibre de liaison suffisamment fort, le lien avec l'apparentement devrait diminuer.

De ce fait, on peut donner certaines règles empiriques à suivre pour faire une prédiction génomique plus précise : d'abord, il parait souhaitable de génotyper et phénotyper les apparentés «informatifs» du candidat (idéalement, son père et sa mère, ou bien son père et son grand-père maternel). Pour que les caractéristiques de ces parents soient bien estimées, il est nécessaire de bien les comparer au reste de la population en génotypant et phénotypant également des animaux contemporains (ou un sous-ensemble représentatif de ceux-ci). Dans des espèces laitières, cela revient à génotyper des séries entières de testage sur descendance, de façon à ce que les «bons» comme les «mauvais» soient bien représentés.

3 / Compléments et perspectives

3.1 / Concernant les méthodes d'évaluation génomique

On remarquera que les méthodes d'évaluation génomique décrites ci-dessus estiment des effets de SNP isolément. Une alternative est d'étudier des effets d'haplotypes, c'est-à-dire de blocs de quelques SNP (souvent de 4 à 6), ce qui améliore l'informativité des marqueurs en accroissant le déséquilibre de liaison entre haplotypes de marqueurs et QTL (Guillaume *et al* 2009). Les évaluations génomiques doivent être adaptées à ce contexte où le nombre d'allèles différents est supérieur à 2. Ceci est possible par exemple avec l'Elastic Net.

Par ailleurs, il existe un autre moyen de mieux considérer les marqueurs (ou haplotypes de marqueurs) proches de QTL à effet fort : si on connaît leur localisation, par exemple à partir d'une analyse de cartographie fine, on peut les considérer individuellement – avec leurs caractéristiques propres et en particulier leur localisation et leur variance associée – dans une évaluation BLUP sur QTL, comme ce qui a été mise en place depuis le début des années 2000 en France pour la sélection assistée par marqueurs (Guillaume *et al* 2009). En pratique, il s'agit là d'une approche plus compliquée puisqu'elle nécessite une étape préliminaire de recherche de QTL contrairement à l'évaluation génomique pure. En France, elle a servi de base à la première méthode d'évaluation des bovins laitiers utilisant les génotypages de la puce 54K, la SAM2 (pour «Sélection Assistée par Marqueurs de seconde génération») mise en place en octobre 2008. Dans la SAM2, seuls ces gros QTL (de 20 à 40 suivant les caractères) étaient considérés, en plus d'une composante génétique résiduelle représentant au moins 50% de la variabilité génétique totale. Ces QTL étaient suivis à travers des haplotypes de 4 à 6 SNP.

La combinaison des deux approches (suivi explicite des gros QTL et évaluation génomique des autres marqueurs) donne de très bons résultats (les meilleurs). C'est cette approche – que l'on a baptisé SAMG (Sélection Assistée par Marqueurs et sélection Génomique) – qui a été mise en application en France depuis 2009 (Fritz *et al* 2010). De 300 à 700 QTL sont ainsi suivis à l'aide de 10 à 15 haplotypes de marqueurs chacun. Le choix des plus gros repose toujours sur les résultats de cartographie fine alors que les autres ont été retenus par l'Elastic Net.

3.2 / Concernant les populations de référence

Lorsque la population de référence est de taille trop réduite (cas des races à effectifs limités), il a été suggéré de constituer des populations de référence multiraciales, c'est-à-dire, de combiner l'information de plusieurs races. En pratique, les résultats observés sont en général décevants lorsque les génotypes sont issus de la puce 54K, avec des variations selon la combinaison «caractère-méthode» (Hayes *et al* 2009c). Des simulations (De Roos *et al* 2009) ont montré qu'en fait, pour capturer le déséquilibre de liaison lointain (antérieur à la création des races) entre marqueurs et QTL, il faut un génotypage plus dense, ce qui n'est possible que depuis peu, avec l'arrivée d'une puce bovine 777K. Mais cette puce haute densité n'est pas sans conséquence sur le choix des méthodes de sélection génomique. Ainsi, elle n'améliorera pratiquement pas la qualité des évaluations de type GBLUP, qui de par leur simplicité de mise en place, sont actuellement utilisées dans pratiquement tous les pays, à l'exception de la France et des Pays-Bas essentiellement. En effet, la parenté génomique entre deux individus n'est pas modifiée quand elle est calculée à partir de 54 000 ou 777 000 SNP. D'autres approches déjà très coûteuses en temps de calcul, en particulier la plupart des méthodes Bayésiennes, peuvent devenir carrément inapplicables à cause de la forte augmentation du nombre de paramètres à estimer. En fait, il est probable que les méthodes de sélection de variables – avec l'approche française

comme cas particulier – soient les mieux à même d'offrir un compromis entre accroissement de la précision des évaluations multiraciales et coûts de calcul acceptables.

D'importants projets de recherche en cours en France et dans d'autres pays visent à démontrer l'intérêt des génotypages haute densité et d'une approche multiraciale.

Conclusion

L'utilisation des informations moléculaires à haut débit est une révolution formidable dans le domaine de la génétique animale, révolution qui dépasse largement le cadre des évaluations génétiques (cf. Institut de l'Elevage et INRA 2011). Mais son impact en sélection est d'ores et déjà considérable car il permet l'obtention d'une estimation de la valeur génétique relativement précise à un âge très précoce et/ou sans attendre des mesures phénotypiques sur l'animal ou ses apparentés. Bien sûr, le gain permis par la sélection génomique par rapport à la sélection pratiquée jusqu'à maintenant sera très variable selon les espèces, les caractères étudiés, la nature et la quantité d'information moléculaire disponible. Il sera d'autant plus important que les caractères sont peu héritables, difficiles, tardifs, compliqués ou coûteux à mesurer. Plus généralement, il dépendra de la possibilité de constituer une population de référence de taille suffisante à un coût abordable.

Le foisonnement des méthodes d'évaluations génomiques ne doit pas surprendre pour une discipline aussi récente. A court terme, on peut s'attendre à d'une part là aussi un processus progressif de sélection de la (ou les) meilleure(s), d'autre part à des progrès encore importants permettant l'utilisation de données encore plus nombreuses (puces haute densité ou séquences). Les défis ne seront alors plus simplement méthodologiques (recherche du meilleur modèle génétique adapté, puissances de calcul et algorithmes adaptées aux quantités importantes d'informations à analyser…) mais concerneront l'utilisation optimale de ces informations dans de nouveaux schémas d'amélioration génétique plus ambitieux.

Remerciements

Nous remercions l'Agence Nationale pour la Recherche ainsi que APISGENE pour leur soutien financier au projet AMASGEN («Approches Méthodologiques et Application de la Sélection Génomique chez les bovins laitiers»).

Références

Aguilar I., Misztal I., Johnson D.L., Legarra A., Tsuruta S., Lawlor T.J., 2010. Hot Topic: A unified approach to utilize phenotypic, full pedigree and genomic information for genetic evaluation of Holstein final score. J. Dairy Sci., 93, 743-752.

Colombani C., Legarra A., Croiseau P., Guillaume F., Fritz S., Ducrocq V., Robert-Granié C., 2010. Application of PLS and Sparse PLS regression in genomic selection. 9th World Congr. Genet. Applied Livest. Prod., Leipzig, Germany, 4p.

Colombani C., Legarra A., Croiseau P., Fritz S., Guillaume F., Ducrocq V., Robert-Granié C., 2011a. BayesCp versus GBLUP, PLS regression, Sparse PLS and Elastic Net methods for genomic selection in French dairy cattle, 62th EAAP Ann. Meet., Stavanger, Norway, 1p.

Colombani C., Croiseau P., Hozé C., Fritz S., Guillaume F., Boichard D., Legarra A., Ducrocq V., Robert-Granié C., 2011b. Could genomic selection methods be efficient to detect QTL? A study in French dairy cattle, 15th QTLMAS Workshop, Rennes, France, 1p.

Croiseau P., Colombani C., Legarra A., Guillaume F., Fritz S., Baur A., Dassonneville R., Patry C., Robert-Granié C., Ducrocq V., 2010. Improving genomic evaluation strategies in dairy cattle through SNP selection. 9th World Congr. Genet. Applied Livest. Prod., Leipzig, Germany, 4p.

De los Campos G., Naya H., Gianola D., Crossa J., Legarra A., Manfredi E., Weigel K., Cotes J.M., 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics, 182, 375-385.

De Roos, A.P.W., Hayes, B.J., Goddard, M.E., 2009. Reliability of genomic predictions across multiple populations. Genetics, 183, 1545-1553.

Ducrocq V., 1990. Les techniques d'évaluation génétique des bovins laitiers. INRA Prod. Anim., 3, 3-16.

Ducrocq V., 1992. Du modèle génétique au modèle Statistique. In : Numéro Hors-série, Eléments de Génétique Quantitative et applications aux populations animales. INRA Prod. Anim., 75-82.

Ducrocq V., Liu Z., 2009. Combining genomic and classical information in national BLUP evaluations, Interbull Bull., 40, 172-177.

Ducrocq V., Legarra A., 2011. An iterative implementation of the single step approach for genomic evaluation which preserves existing genetic evaluation models and software. Interbull Meeting, Stavanger, Norway, 5p.

Fritz S., Guillaume F., Croiseau P., Baur A., Hoze C., Dassonneville R., Boscher M.Y., Journaux L., Boichard D., Ducrocq V., 2010. Mise en place de la sélection génomique dans les trois principales races françaises de bovins laitiers. Renc. Rech. Rum., 455-458.

Gianola D., Fernando R.L., 1986. Bayesian methods in animal breeding theory. J. Anim. Sci., 63, 217-244.

Gianola D., De los Campos G., Hill W.G., Manfredi E., Fernando R.L., 2009. Additive genetic variability and the Bayesian alphabet. Genetics, 183, 347-363.

Goddard M., 2009. Genomic selection: prediction of accuracy and maximization of long term response. Genetica, 136, 245-257.

Goffinet B., Elsen J.M., 1984. Critère optimal de sélection : quelques résultats généraux. Génét. Sél. Evol., 16, 307-318.

Grisart, B., Coppieters W., Farnir F., Karim L., Ford C., Berzi P., Cambisano N., Mni M., Reid S., Simon P., Spelman R., Georges M., Snell R., 2002. Positional candidate cloning of a QTL in dairy cattle: Identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Res., 12, 222-231.

Guillaume F., Fritz S., Legarra A., Croiseau P., Robert-Granié C., Colombani C., Patry C., Boichard D., Ducrocq V., 2009. Modèles d'évaluation génomique : application aux populations bovines laitières françaises. Renc. Rech. Rum., 399-406.

Habier D., Fernando R.L., Dekkers J.C.M., 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics, 177, 2389-2397.

Habier D., Fernando R.L., Kizilkaya K., Garrick D.J., 2010. Extension of the Bayesian alphabet for genomic selection. 9th World Congr. Genet. Applied Livest. Prod., Leipzig, Germany, 4p.

Haley C.S., Visscher P.M., 1998. Strategies to utilize marker-quantitative trait loci associations. Proc. Symp., Breeding objectives and strategies, 1997. J. Dairy Sci., 81 (Suppl. 2), 85-97.

Hayes B.J., Bowman P.J., Chamberlain A.J., Goddard M.E., 2009a. Invited review: Genomic selection in dairy cattle: Progress and challenges. J. Dairy Sci., 92, 433-443.

Hayes B.J., Visscher P.M., Goddard M.E., 2009b. Increased accuracy of artificial selection by using the realized relationship matrix. Genet. Res., 91, 47-60.

Hayes B.J., Bowman P.J., Chamberlain A.C., Verbyla K., Goddard M.E., 2009c. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet. Sel. Evol., 24, 41-51.

Henderson C.R., 1963. Selection index and expected genetic advance. In: Statistical genetics and plant breeding. Hanson W.D., Robinson H.F. (Eds.). National Academy of Science. National Research Council, Washington, DC, Publ. 982, 141-163.

Henderson C.R., 1973. Sire evaluation and genetic trends. Proc. Anim. Breed. Genet. Symp. in honor of Dr Jay L. Lush, 10-41.

Hoerl A.E., Kennard R.W., 1970. Ridge regression: biased estimation for non orthogonal problems. Technometrics, 12, 55-67.

Institut de l'Elevage, INRA, 2011. La révolution génomique animale, Ed. France

Agricole, Ouvrage collectif, Collection Agriproduction, 161p.

Kizilkaya K., Fernando R.L., Garrick D.J., 2010. Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. J. Anim. Sci., 88, 544-551.

Lande R., Thompson R., 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics, 124, 743-776.

Legarra A., Robert-Granié C., Manfredi E., Elsen J.M., 2008. Performance of genomic selection in mice. Genetics, 180, 611-618.

Legarra A., Misztal I., 2008. Genome-wide selection computing strategies. J. Dairy Sci., 91, 360-366.

Legarra A., Robert-Granié C., Croiseau P., Guillaume F., Fritz S., 2011. Improved Lasso for Genomic Selection. Genet. Res., 93, 77-87.

Le Cao K.A., Rossouw D., Robert-Granié C., Besse P., 2008. A Sparse PLS for variable selection when integrating omics data. Stat. Appl. Genet. Mol. Biol., 7, 32.

Liu Z., Seefried F.R., Reinhardt F., Rensing S., Thaller G., Reents R., 2011. Impacts of both reference population size and inclusion of a residual polygenic effect on the accuracy of genomic prediction. Genet. Sel. Evol., 43, 19.

Luan T., Woolliams J.A., Lien S., Kent M., Svendsen M., Meuwissen T.H.E., 2009. The accuracy of Genomic Selection in Norwegian red cattle assessed by cross-validation. Genetics, 183, 1119-1126.

Meuwissen T.H.E., Hayes B.J., Goddard M.E., 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics, 157, 1819-1829.

Metropolis N., Rosenbluth A.W., Rosenbluth M.N., Teller A.H., Teller E., 1953. Equation of state calculation by fast computing machines. J. Chem. Phys., 21, 1087-1092.

Nilforooshan M.A., Zumbach B., Jakobsen J., Loberg A., Jorjani H., Dürr J.W., 2010. Validation of national genomic evaluations. Interbull Bull., 42, 56-61.

Park T., Casella G., 2008. The Bayesian Lasso. J. Am. Statist. Ass., 103**,** 681-686.

Patry C., Ducrocq V., 2011. Evidence of biases in genetic evaluations due to genomic pre-selection in dairy cattle. J. Dairy Sci., 94, 1011-1020.

Robert C., 1996. Méthodes de Monte Carlo par Chaînes de Markov. Economica, Paris, France, 350p.

Robert-Granié C., Duchemin S., Larroque H., Baloche G., Barillet F., Moreno C., Legarra A., Manfredi E., 2011. A comparison of various methods for the computation of genomic breeding values in French Lacaune dairy sheep breed. 62th EAAP Ann. Meet., Stavanger, Norway, 1p.

Strandén I., Garrick D.J., 2009. Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. J. Dairy Sci., 92, 2971-2975.

Su G., Guldbrandtsen B., Gregersen V.R., Lund M.S., 2010. Preliminary investigation on reliability of genomic estimated breeding values in the Danish Holstein population. J. Dairy Sci., 93, 1175-1183.

Tenenhaus M., 1998. La régression PLS : théorie et pratique. Ed Technip, Paris, France, 254p.

Thomsen H., Reinsch N., Xu N., Looft C., Grupe S., Kuhn C., Brockmann G.A., Schwerin M., Leyhe-Horn B., Hiendleder S., Erhard G., Medjugorac I., Russ I., Forster M., Brening B., Reinhardt F., Reents R., Blumel J., Averdunk G., Kalm E., 2001. Comparison of estimated breeding values, daughter yield deviations and de-regressed proofs within a whole genome scan for QTL. J. Anim. Breed. Genet., 118, 357-370.

Tibshirani R., 1996. Regression shrinkage and selection *via* the Lasso. J. Royal Stat. Soc., Series B, 58, 267-288.

Usai M.G., Goddard M.E., Hayes B.J., 2009. Lasso with cross-validation for genomic selection. Genet. Res., 91, 427-436.

VanRaden P.M., 2008. Efficient methods to compute genomic predictions. J. Dairy Sci., 91, 4414-4423.

VanRaden P.M., Wiggans G.R., 1991. Derivation, calculation, and use of national animal model information. J. Dairy Sci., 74, 2737-2746.

VanRaden P.M., VanTassell C.P., Wiggans G.R., Sonstegard T.S, Schnabel R.D, Taylor J.F., 2009. Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci., 92, 16-24.

Zhou H., Hastie T., 2005. Regularization and variable selection *via* the Elastic Net. J. Roy. Stat. Soc., Serie B, 67, 301-320.

Wold H., 1966. Estimation of principal components and related models by iterative least squares. In: Multivariate Analysis. Krishnaiah P.R. (Ed.). Academic Press, New York, Wiley, UK, 391-420.

Wolc A., Arango J., Settar P., Fulton J.E., O'Sullivan N.P., Preisinger R., Habier D., Fernando R., Garrick D.J., Dekkers J.C.M., 2011. Persistence of accuracy of genomic estimated breeding values over generations in layer chickens. Genet. Sel. Evol., 43, 23.

Résumé

Avec l'arrivée de données de génotypage à haut débit, il est maintenant possible d'estimer la valeur génétique d'animaux candidats à la sélection dès leur naissance, sans attendre la collecte de phénotypes. La sélection génomique bouleverse complètement les perspectives en amélioration génétique. Elle nécessite la constitution d'une population de référence formée d'animaux génotypés (jusqu'à récemment, il s'agissait principalement de mâles) et ayant des performances précises, par exemple la performance moyenne de leurs filles. Les évaluations génomiques consistent à prédire les phénotypes dans cette population de référence comme la somme des effets des marqueurs moléculaires. Le problème méthodologique principal est que le nombre d'effets à estimer est typiquement beaucoup plus élevé que le nombre de phénotypes disponibles. Nous décrivons les idées générales de diverses familles de méthodes proposées : BLUP génomique basé sur une parenté entre individus calculée à partir des marqueurs, méthodes Bayésiennes plus flexibles mais aussi plus coûteuses, méthodes de sélection de variables, méthode en une seule étape qui combine évaluation génétique nationale et évaluation génomique. La précision des évaluations génomiques est faite par validation croisée chez les animaux les plus jeunes de la population de référence. La taille de la population de référence, la manière de prendre en compte les QTL à effet fort et le degré d'apparentement entre candidats à la sélection et animaux de la population de référence ont un impact non négligeable sur l'efficacité des méthodes de sélection génomique.

Abstract

Basic principles of genomic selection

With the advent of high throughput genotyping, it is now possible to estimate breeding values of selection candidates at birth without waiting for phenotypic data collection. Genomic selection is deeply changing future prospects in genetic improvement. It requires the creation of a reference population made of genotyped animals (generally males) with precise phenotypes, such as average daughter performances. Genomic evaluations consist in predicting phenotypes in this reference population as the sum of molecular marker effects. The main methodological challenge is the large number of effects to estimate, usually much larger than the number of available phenotypes. We briefly describe the various families of proposed methods: genomic BLUP based on relationships

computed from marker information, Bayesian methods which are more flexible but more tedious, variable selection methods and a single step method which combines national genetic evaluation and genomic evaluation. The precision of genomic selection is done via cross validation among the youngest animals of the reference population. The size of the reference population, the way QTL with large effects are modeled, the relationship between selection candidates and the reference population have a significant impact on the efficiency of genomic selection methods.

ROBERT-GRANIÉ C., LEGARRA A., DUCROCQ V., 2011. Principes de base de la sélection génomique. In : Numéro spécial, Amélioration génétique. Mulsant P., Bodin L., Coudurier B., Deretz S., Le Roy P., Quillet E., Perez J.M. (Eds). INRA Prod. Anim., 24, 331-340.