Modèles d'évaluation génomique : application aux populations bovines laitières françaises

GUILLAUME F. (1), FRITZ S. (2), CROISEAU P. (3), LEGARRA A. (4), ROBERT-GRANIÉ C. (4), COLOMBANI C. (4), PATRY C. (2,3), BOICHARD D. (3), DUCROCQ V. (3)

(1) Institut de l'élevage, 149 rue de Bercy, Paris, France

(2) Union nationale des coopératives agricoles d'élevage et d'insémination animale, service génétique, 149 rue de Bercy, 75595 Paris Cedex 12, France

(3) INRA, UMR 1313 génétique animale et biologie intégrative, 78352 Jouy-en-Josas, cedex, France

(4) INRA, UR631 station d'amélioration génétique des animaux, 78352 Jouy-en-Josas, cedex, France

RESUME - La disponibilité de puces denses de SNP a suscité une certaine révolution au niveau international dans le domaine de la sélection des bovins laitiers. En effet, l'information fournie par ces puces permet d'estimer la valeur génétique de reproducteurs dès leur naissance avec une précision nettement supérieure à celle obtenue par un modèle polygénique classique. Plusieurs méthodes sont envisageables et font l'objet de nombreux développements. La méthode actuellement utilisée en France (dite SAM2 pour sélection assistée par marqueurs de seconde génération) offre un bon compromis entre une évaluation rapide reposant sur une information limitée à quelques régions QTL bien confirmées et une bonne précision des valeurs génétiques estimées. Les méthodes d'évaluation génomique au sens strict considèrent quant à elles l'effet de l'ensemble du génome, au risque d'utiliser des régions qui n'ont pas d'effet réel. Leur avantage est de prendre en compte, par construction, la totalité de la variance génétique, là où la SAM2 se limite à une fraction expliquée par les QTL, comprise aujourd'hui entre 40 et 60 % selon les caractères.

La méthode actuellement la plus efficace (BayesB) pose des problèmes de temps de calcul et de définition des paramètres. C'est pourquoi des méthodes alternatives sont proposées. Dans un premier temps, nous passons en revue les différentes familles de modèle d'évaluation génomique, en mettant en avant leur efficacité, leurs avantages et inconvénients respectifs.

Dans un second temps, à l'aide d'un jeu de données issu du programme de sélection assistée par marqueurs SAM2, nous comparons la qualité prédictive d'un ensemble de modèles sur deux échantillons de reproducteurs, de race Holstein et Montbéliard, mis à l'épreuve en 2004 et avec des index sur descendance en 2008. Ces résultats sont comparés à la méthode actuellement mise en œuvre en France (SAM2). La comparaison est effectuée sur des caractères de production.

Les facteurs clés concourant le plus à la réussite d'un programme de sélection génomique sont également dégagés. Enfin, on présente des perspectives d'évolution pour augmenter la précision mais aussi pour disposer de méthodes adaptées aux différentes populations d'élevage.

Genomic selection models: Application to French dairy cattle

GUILLAUME F. (1), FRITZ S. (2), CROISEAU P. (3), LEGARRA A. (4), ROBERT-GRANIÉ C. (4), COLOMBANI C. (4), PATRY C. (2,3), BOICHARD D. (3), DUCROCQ V. (3)

(1) Institut de l'Elevage, 149 rue de Bercy, Paris, France

SUMMARY

The availability of dense SNP chips has led to quite a revolution on dairy cattle breeding at the international scale. Indeed, information provided by such a chip allows for the breeding value estimation of bulls and cows at birth with a much higher accuracy than with a polygenic model. The methodology currently applied in France (Second generation marker-assisted selection or MAS2) offers a good trade-off between a fast breeding value estimation relying on a reduced set of confirmed QTL regions and a good accuracy of estimated breeding values. In contrast, methodologies for genomic selection in a strict sense consider the effect of the whole genome, with a risk of involving regions with no real effect. Their main advantage is to account for the whole genetic variance, at least conceptually, whereas the fraction explained by MAS2 is presently limited to 40-60 % according to traits.

In terms of accuracy, the current theoretical best genomic approach (BayesB) remains too time consuming and its key parameters are difficult to choose. Therefore, several alternative models are proposed. We first present the different types of genomic evaluation models, focussing on their efficiency, and their respective pros and cons.

Using a data set from the MAS2 program, we compared the predictive ability of several approaches on two batches of sires from the Holstein and Montbéliarde breeds. Those sires entered progeny testing in 2004 and received first official proofs in 2008. The predictive ability of the different approaches is compared with the results from the MAS2 model currently used in France. The comparison is done for production traits.

For each model, advantages and drawbacks were analysed in order to select the model that would best fulfil the dairy breeding industry needs. Key factors for a successful genomic selection program are underlined. Finally, for each model, ways to improve accuracy or to fulfil other population needs are described.

INTRODUCTION

Mettre à profit l'information moléculaire pour améliorer la sélection est une idée assez ancienne (Lande et Thompson, 1990). Elle se base sur l'utilisation de marqueurs pour suivre la transmission de régions chromosomiques influençant les caractères, appelées QTL *(Quantitative Trait Locus)*. Ces marqueurs sont des fragments d'ADN, pour lesquels on peut identifier différentes formes (allèles) au sein de la population pouvant être associées à la présence de différents allèles d'un ou plusieurs QTL proches. Il existe de nombreux types de marqueurs (liés à des mutations, délétions, insertions, inversions, variations du nombre de copies, etc.) Les plus utilisés en pratique restent les marqueurs microsatellites (très en vogue dans les années 90) et plus récemment les marqueurs SNP *(Single Nucleotide Polymorphism)* qui sont des marqueurs abondants sur le génome pour lesquels les progrès technologiques ont grandement facilité le génotypage à haut débit.

À partir des années 90, de nombreuses études (Georges *et al*., 1995, Boichard *et al*., 2003) ont eu pour but de localiser des QTL à l'aide de marqueurs microsatellites couvrant l'ensemble du génome. Ainsi de grandes régions dont les variations étaient associées à des variations de phénotype au sein de familles de même père ont été mises en évidence. Mais la couverture peu dense du génome - quelques centaines de marqueurs pour suivre l'ensemble du génome - ne permettait en aucun cas d'identifier le (ou les) gène(s) responsables des variations de phénotype. Dans la plupart des cas, l'identification des gènes sous-jacents aux QTL était difficilement envisageable. De lourds travaux de densification en marqueurs des régions QTL étaient requis pour simplement identifier des marqueurs plus proches de ces gènes et présentant une association systématique (ou déséquilibre de liaison (DL)) entre allèle au marqueur et allèle au QTL.

Néanmoins, la connaissance des QTL était suffisante pour établir des associations utilisables intra famille. C'est pourquoi dès 2001, l'INRA, l'UNCEIA et Labogena ont mis en place un premier programme de sélection assistée par marqueurs (SAM1) permettant d'identifier les animaux les plus prometteurs intra famille (Fritz *et al*., 2003). L'efficacité du programme SAM1 s'est progressivement améliorée par l'augmentation du nombre d'animaux typés (environ 70 000 animaux de 2001 à 2008), la caractérisation plus fine des QTL initiaux et la découverte de nouveaux (Fritz *et al*., 2007).

Parallèlement, des travaux théoriques (Visscher *et al*.,1998) ont envisagé l'utilisation pour l'évaluation génétique de marqueurs moléculaires denses couvrant l'intégralité du génome (c'est-à-dire plusieurs milliers de marqueurs). L'idée développée par cette approche est qu'en suivant la transmission de l'ensemble du génome, on suit notamment la transmission de *tous* les gènes intervenant sur un phénotype d'intérêt. L'estimation de la valeur génétique d'un individu revient alors à calculer l'effet de chacun des fragments de son génome : on parle alors d'évaluation génomique. La difficulté de cette approche est de distinguer les régions ayant un effet réel sur le phénotype de la majorité des autres qui n'ont aucun effet.

En 2001, Meuwissen *et al*. (2001) ont proposé deux premiers modèles d'évaluation adaptés à modélisation. Le résultat principal de ces travaux est de démontrer qu'on peut ainsi atteindre une précision d'évaluation génétique élevée dès la naissance d'un animal. Ces travaux sont restés théoriques, jusqu'au séquençage du génome bovin et la mise en évidence d'un grand nombre de SNP en 2006. En 2007, la disponibilité de puces bovines de génotypage à haut débit (54 000 SNP) a permis de tester puis mettre en œuvre la sélection génomique. Avec ces puces, la densité de marqueurs utilisés est telle que les QTL sont obligatoirement en déséquilibre de liaison avec leurs marqueurs les plus proches.

Deux stratégies sont alors possibles : la première repose sur la cartographie fine de QTL puis une prédiction de la valeur génétique des animaux par le suivi des QTL, avec une approche d'évaluation assistée par marqueurs de seconde génération (SAM2) ; la deuxième consiste en une évaluation simultanée des effets de tous les marqueurs couvrant le génome sans étape préalable de localisation des QTL. C'est cette deuxième approche que Meuwissen *et al*. (2001) ont appelé un peu abusivement « sélection génomique » mais dont le terme est maintenant consacré.

La France s'est engagée en 2008 sur une évolution de la SAM1 vers la SAM2 qui a abouti à partir d'octobre 2008 à une évaluation mensuelle des candidats à la sélection à partir de leur génotype issu de la puce 54k d'Illumina (Fritz *et al*., 2008). Les évaluations de la SAM2 sont devenues officielles pour les jeunes taureaux en juin 2009.

Bien que nous ayons trouvé une précision des évaluations de la SAM2 du même ordre de grandeur que celle annoncée pour la sélection génomique dans d'autres pays (Van Raden *et al*., 2009 ; Guillaume *et al*., 2009), des études poussées sur l'intérêt d'une sélection génomique au sens strict ont démarré et les tout premiers résultats sont présentés dans cet article.

1. MODELES DE SELECTION GENOMIQUE

Nous allons tout d'abord exposer les principes des grandes familles de méthodes de sélection génomique.

1.1. MODELE PRINCEPS

Le modèle de Meuwissen *et al*. (2001) revient à aborder ensemble deux problèmes distincts pour l'évaluation génomique : d'une part l'identification des régions chromosomiques ayant un effet sur le caractère étudié et d'autre part le calcul pour ces régions des effets des différents variants (allèles) des marqueurs présents dans la population.

En pratique, on s'appuie sur une population de référence pour laquelle on dispose à la fois de génotypes (les allèles de chaque marqueur réellement présents chez un individu) et de phénotypes. Cette population de référence joue un rôle essentiel puisqu'elle permet l'estimation des relations génotype phénotype, relations que l'on utilise pour la prédiction de la valeur des animaux sans phénotype. Dans les populations de référence bovines laitières, on utilise un phénotype particulier des taureaux, la performance moyenne de ses filles (ou *Daughter Yield deviation* : DYD ; Van Raden et Wiggans, 1991). Ce phénotype est analogue à une performance propre du taureau, d'héritabilité égale à la précision de l'index sur descendance. Cette précision élevée améliore sensiblement l'efficacité du dispositif. L'utilisation des DYD est ici fondamentale : elle permet de concentrer l'information phénotypique de nombreuses filles sur un seul individu génotypé. C'est ce qui explique que la sélection génomique est presque aussi efficace quelle que soit l'héritabilité du caractère considéré. Cette population de référence permet d'estimer l'effet a_{jk} de chaque allèle k du marqueur j. La valeur de tout individu i est ensuite

prédite par $\hat{g} = \sum_j n_{ijk} \hat{a}_{jk}$, avec n_{ijk} le nombre d'allèles k du marqueur j chez l'individu i. Enfin, L'évaluation génomique peut éventuellement être complétée en combinant cette prédiction génomique avec la valeur polygénique des parents (index sur ascendance).

Pour estimer les effets des marqueurs, Meuwissen *et al*. (2001) ont privilégié une approche bayésienne comme souvent en génétique quantitative. On en rappelle ici les grands principes. Dans le monde bayésien, on cherche à décrire notre degré de connaissance de la valeur d'un effet à travers une distribution statistique. L'effet étudié est caractérisé par sa distribution a posteriori, obtenue en combinant deux sources d'information, une information a priori (connue avant l'obtention des données) et l'information apportée par les données. C'est par exemple le cas en évaluation génétique classique (le BLUP) où avant même l'obtention des performances des animaux, on connaît leur généalogie ainsi que les caractéristiques de leur valeur génétique additive, supposée suivre une distribution normale, avec des paramètres génétiques connus. La connaissance a posteriori des paramètres qui nous intéressent agrège de manière optimale ces deux sources d'information : performances et généalogie.

Malgré tout, la distribution conjointe a posteriori de tous les paramètres (par exemple l'effet de tous les SNP) est trop compliquée pour être accessible directement. Des méthodes calculatoires basées sur des simulations ont été développées pour estimer les paramètres d'intérêt. Pour cela, on divise l'estimation conjointe en une multitude de problèmes plus simples : si on connaît tous les paramètres sauf un, ce dernier est habituellement assez simple à simuler sachant la valeur des autres et celles des données. En simulant ainsi chaque paramètre séquentiellement sachant tous les autres, et en répétant l'opération un très grand nombre de fois, on aboutit à un échantillon de paramètres appartenant à la vraie distribution conjointe a posteriori. On peut en tirer toute l'information nécessaire (moyenne, variance, intervalle de confiance). C'est le principe de l'échantillonnage de Gibbs et plus généralement des méthodes MCMC *(Monte Carlo Markov Chain)*, qui ont rendu possible informatiquement de nombreux développements statistiques.

Meuwissen *et al*. (2001) calculent ainsi les caractéristiques (moyenne et variance) des distributions des effets des marqueurs et de la variance de chacun d'entre eux. Ils les estiment en supposant qu'ils ont tous un effet non nul sur le caractère considéré (méthode dite « BayesA »), ou bien ils supposent que seulement un pourcentage d'entre eux fixé à l'avance ont un effet, sans spécifier lesquels (« BayesB »). Ils ont comparé ces deux approches à une estimation des effets de tous les marqueurs par les méthodes classiques des moindres carrés et du BLUP.

En simulant une population de référence génotypée sur marqueurs microsatellites (chacun ayant en moyenne sept allèles) et avec performances pour deux cents parents (génération n) et deux mille descendants (n+1), Meuwissen *et al*. (2001) obtiennent des corrélations de l'ordre de 0,80 entre ce que l'on veut prédire (la valeur génétique vraie des candidats à la sélection de la génération n+2) et leur valeur estimée seulement sur la base de leur génotype. Cette corrélation correspond à une précision (coefficient de détermination ou CD) de 0,6, les résultats étant moins bons avec le BLUP et mauvais avec la méthode des moindres carrés.

L'avantage des approches bayésiennes ayant recours aux MCMC pour les calculs est leur flexibilité, permettant de mieux s'adapter aux données. Ceci a néanmoins un coût : un temps de calcul prohibitif pour une évaluation génomique fréquente et un volume de données amené à croître rapidement. De plus les méthodes BayesA et BayesB font appel à des connaissances préalables (information a priori) sur la distribution et la taille des QTL, par exemple à partir de la bibliographie. Or, cette information est fortement biaisée par le fait qu'on ne détecte que les QTL à effet fort. Ceci pose des problèmes sérieux pour la calibration des méthodes. Malgré tout, les très nombreuses comparaisons faites au niveau international montrent une robustesse relativement bonne des deux méthodes (Verbyla *et al.,* 2009).

Cette première étude n'avait pas anticipé l'arrivée des marqueurs denses bi-allèliques que sont les SNP. Ce type de marqueurs implique une moindre informativité de chaque marqueur. Ainsi, Solberg *et al*. (2006) ont répété l'étude précédente avec des SNP et ont obtenu des résultats plus faibles (CD de l'ordre de 0,4 à 0,5).

Cependant, depuis l'arrivée des puces à haut débit (plus de 50 000 SNP), cette moindre informativité est compensée par la densité importante de SNP, ce qui permet de regagner en précision, comme le prouvent Solberg *et al*. (2008). Mais une fois de plus, cela est obtenu au prix d'un temps de calcul supérieur.

1.2. PROBLEMATIQUE DE L'EVALUATION GENOMIQUE

Comme l'illustre l'exemple simulé par Meuwissen *et al.,* (2001), la principale difficulté de l'évaluation génétique est d'ordre statistique : comment estimer avec suffisamment de précision l'effet de p SNP à partir de n observations ou performances, lorsque p est beaucoup plus grand que n ? Notons tout d'abord qu'on sait très bien faire le contraire : estimer un petit nombre p d'effets (comme les effets « année » ou « troupeau ») à partir d'un grand nombre d'observations (par année ou par troupeau), par la méthode des moindre carrés par exemple. Mais cette même méthode donne une infinité de solutions si p>>n. Le problème du « large p, small n » est en fait assez classique en statistique et de nombreuses méthodes ont été proposées pour le résoudre. On peut en distinguer plusieurs familles :

1.2.1. Les méthodes résumant l'information des marqueurs en un nombre plus faible de prédicteurs

Parmi celles-ci, on retiendra les méthodes PLS *(Partial Least Square)* ou PCA *(Principal Component Analysis)* (Solberg *et al.*, 2009) très employées dans d'autres domaines. Elles consistent à concentrer l'information génomique d'une part et les performances d'autre part en un nombre réduit de combinaisons linéaires dont les coefficients sont calculés de façon à rendre maximale la corrélation globale entre les deux. Par rapport aux méthodes BayesA ou BayesB, les coûts de calcul sont très fortement réduits. Malheureusement, la qualité des prédictions est aussi moins bonne, même si certaines variantes semblent prometteuses.

1.2.2. Les méthodes dérivées du BLUP

Divers travaux (Habier *et al*., 2007, Van Raden *et al*., 2009) ont montré que sous certaines conditions, il était équivalent de décrire la composante génétique des performances des animaux soit comme une somme d'effets de marqueurs, soit comme une valeur génétique additive issue d'une distribution multinormale particulière. La structure de variance de cette distribution serait proportionnelle à une

matrice dont chaque élément i,j est une fonction des allèles des marqueurs présents à la fois chez i et j. Cette matrice, vite dénommée « matrice de parenté génomique » est ainsi analogue à la matrice de parenté classique utilisée dans les évaluations BLUP. Elle décrit la parenté « réelle » observée à travers les marqueurs plutôt que la parenté espérée. Cette dernière est égale par exemple à 0,5 pour deux plein frères ou 0,25 pour deux demi-frères, alors que la parenté réelle varie autour de ces valeurs, à cause de la ségrégation mendélienne : ainsi, deux frères peuvent avoir conceptuellement une parenté génomique de 1 s'ils ont hérité les mêmes chromosomes de leurs parents. Dans cette approche, on retient toujours la notion d'effet individuel de chaque SNP, même si on ne les calcule pas explicitement. Cette équivalence a conduit plusieurs pays (Etats-Unis, Canada, Allemagne) à mettre en place une telle évaluation, dite BLUP génomique (GBLUP).

1.2.3. Les méthodes nécessitant ou entraînant une présélection des marqueurs

On peut envisager par exemple, d'estimer de manière simple l'effet individuel de chaque SNP pris un à un, puis de ne retenir que les p SNP les plus significatifs ($p<n$) et enfin d'estimer conjointement leur effet par moindre carrés. Cette approche donne malheureusement des résultats très décevants car on retient trop de SNP avec des effets surestimés par hasard (biais de sélection).

Pour éviter ce problème, une catégorie de méthodes basées sur la réduction du nombre de variables nommée « régression pénalisée » apparaît plus adaptée à la génétique. Ces méthodes permettent de sélectionner les marqueurs les plus pertinents tout en estimant l'effet de ces marqueurs simultanément. Ainsi on évite au moins partiellement les problèmes inhérents à la présence de déséquilibre de liaison.

Parmi la grande quantité de marqueurs disponibles, il est très probable que seule une faible proportion d'entre eux soit réellement impliquée dans le caractère d'intérêt. Pour autant, la plupart des méthodes de sélection génomique (BayesA ou GBLUP) estiment l'effet de chacun de ces marqueurs.

Les méthodes de régression pénalisée régressent vers 0 (elles « pénalisent ») l'effet des marqueurs de telle sorte que seuls ceux qui ont un effet suffisamment important influencent l'estimation de la valeur génétique. Ainsi, les autres marqueurs – ceux qui ont un effet apparent faible, peut-être dû au hasard (bruit de fond) – perturberont moins l'estimation des effets des marqueurs les plus prometteurs.

Dans la littérature, différentes approches de régression pénalisées ont été décrites, se distinguant par la nature de la pénalisation des coefficients affectés à chaque polymorphisme. Ainsi, dans la *Ridge Regression* (Frank et Friedman, 1993), la pénalité appliquée à l'effet des marqueurs est proportionnelle à leur carré tandis que pour la méthode Lasso (*Least Absolute Shrinkage and Selection Operator* ; Tibshirani, 1996), cette pénalisation est fonction de la valeur absolue des coefficients. En fait, cette dernière revient à fixer à 0 les effets estimés les plus faibles et correspond donc à une sélection automatique des marqueurs. En présence d'un ensemble de marqueurs en fort déséquilibre de liaison, la RR a tendance à estimer un effet pour chacun des marqueurs tandis que le Lasso ne retient qu'un seul des marqueurs dans le modèle et élimine tous les autres. Pour éviter de choisir entre ces deux solutions extrêmes qui ne semblent pas adéquates au contexte génétique, une approche – *l'Elastic Net* (EN ; Zou et Hastie, 2005) – permet d'obtenir une pénalité qui est une combinaison linéaire des pénalités de la *Ridge Regression* et du Lasso.

1.3. MODELE FRANÇAIS : LA SAM 2

Le modèle d'évaluation assistée par marqueurs français de deuxième génération est quant à lui dans la continuité de la SAM1. L'évolution principale est l'utilisation de groupes (par exemple 5-6) de marqueurs successifs sur le génome, que l'on appelle « haplotypes » encadrant finement les QTL et présentant un fort déséquilibre de liaison avec eux. On peut identifier ainsi un haplotype associé de manière quasi unique à un allèle du QTL au niveau de la population. Le nombre d'haplotypes distincts au sein de la population reste limité à quelques dizaines et le nombre de performances d'animaux porteurs de ces haplotypes est en général élevé. Contrairement aux approches de sélection génomique stricte, on est donc dans un contexte où le nombre d'observations est grand par rapport au nombre de paramètres à estimer. Il en découle une estimation très précise de l'effet de chaque haplotype.

La nécessaire recherche initiale des QTL et des haplotypes associés conduit malgré tout à exclure la majeure partie des informations moléculaires. Seule une fraction de la variabilité génétique totale est ainsi expliquée à l'aide de marqueurs. La part de variance génétique non expliquée par les QTL est décrite dans le modèle par un terme polygénique classique, construit en se basant sur les relations de parenté.

Cette approche permet une évaluation très rapide une fois faite l'identification des marqueurs en déséquilibre de liaison avec les QTL. Elle nécessite cependant de construire les haplotypes de marqueurs. Par ailleurs, des détections de QTL doivent être effectuées à intervalles réguliers afin de tirer profit de l'accumulation de nouvelles données collectées.

1.4. COMPARAISON DES METHODES

1.4.1. Validation

À un problème apparemment bien défini , « comment évaluer la valeur génétique d'animaux à partir de leur information moléculaire ? », de nombreuses solutions techniques ont donc été envisagées. Il est nécessaire de pouvoir les comparer en termes d'efficacité à bien prédire la valeur génétique vraie. La principale difficulté dans cette comparaison est que la valeur génétique vraie nous sera toujours inconnue dans notre monde réel.

Une première approche est l'utilisation de simulations qui permettent de connaître exactement la valeur génétique vraie, comme dans le cas de Meuwissen *et al.* (2001). Néanmoins, les simulations dépendent des hypothèses faites et tendent souvent à simplifier la réalité. Elles nécessitent de connaître le déterminisme des caractères (nombre de QTL, parts de variance expliquées, distributions des effets, etc…) qui sont impossibles à estimer dans la pratique. Il s'ensuit des conclusions souvent trop optimistes par rapport à la réalité.

Une seconde approche est d'utiliser des données réelles pour une validation croisée. Dans ce cas, on évalue le modèle à travers sa capacité à prédire un phénotype. Dans la pratique, deux types de prédictions sont envisagées : soit la performance future d'un candidat à la sélection (par exemple les DYD qui lui sont attachées), soit sa valeur génétique estimée après testage sur descendance (son index classique). Notons tout de suite que la plupart des premiers travaux ont retenu la deuxième option, qui donne par ailleurs les résultats les plus optimistes. Or c'est bien la

première qui semble beaucoup plus juste : l'index classique combine à la fois l'information sur ascendance (connue, même pour les candidats à la sélection) et sur descendance - la seule qu'on cherche à prédire avec les marqueurs.
Quelle que soit l'option retenue, valeur génétique estimée ou DYD sont elles-mêmes entachées d'imprécision par rapport à la valeur génétique vraie, d'où des résultats de validation qui peuvent parfois paraître décevants.
Deux stratégies de validation croisée sont envisageables. Le ré-échantillonnage *(re-sampling)* consiste à scinder la population de référence en deux par échantillonnage, d'établir des équations de prédiction à partir du premier échantillon et d'examiner l'adéquation entre prédiction et valeur observée (de l'index ou des DYD) dans le second échantillon. Cette stratégie est dangereuse et a conduit à des résultats beaucoup trop optimistes à l'origine (*cf.* par exemple les premiers travaux néo-zélandais). En effet, à cause de l'existence du progrès génétique dans la population, les animaux de générations successives ont des valeurs génétiques pouvant être très différentes. Il est alors plus facile de trouver ce qui les distingue au niveau génomique, par rapport à des animaux aux caractéristiques génétiques beaucoup plus proches tels que des contemporains, voire des pleins frères.
La stratégie couramment admise est une validation a posteriori, qui consiste à se calquer sur l'utilisation future de l'évaluation génomique. Ainsi, la population de validation est constituée des tout derniers taureaux de la population de référence pour lesquels on dispose d'une évaluation génétique classique après testage sur descendance. On se place dans des conditions similaires à celles de leur naissance (pas de performances propres, information parentale moins précise) et on utilise les estimations des effets SNP obtenus à partir de la population d'apprentissage. Si dans de telles conditions, les prédictions sont bien corrélées avec les DYD (ou les index) effectivement obtenus quatre ans plus tard, on peut avoir une bonne confiance dans la qualité de la méthode d'évaluation génomique utilisée.

1.4.2. Précision des index

La validation de modèle apprécie la qualité d'une méthode à prédire la valeur génétique d'un échantillon précis d'animaux. Parce qu'il y a parfois ambiguïté dans les présentations, il est important de distinguer les résultats de cette validation du calcul de la précision des index (CD). En effet, le CD reflète une précision théorique, mesurée à partir de la quantité d'information disponible et des hypothèses sur le déterminisme génétique, en particulier le déséquilibre de liaison entre marqueurs et QTL. Dans les modèles de sélection génomique, les CD théoriques sont généralement très surestimés et n'ont pas grande signification. Ainsi, l'approche de VanRaden *et al.* (2009) basée sur l'inversion de la matrice des coefficients du « BLUP génomique » suppose de manière incorrecte que le déséquilibre de liaison entre tous les QTL et au moins un marqueur est total. Il est reconnu par ses auteurs qu'elle donne des résultats particulièrement optimistes, qu'il faut les multiplier par un facteur plus ou moins arbitrairement fixé (0,5 actuellement !) pour que les CD restent crédibles… La SAM2 reposant sur des concepts plus proches des évaluations classiques, les CD de la SAM2 ne présentent pas ces inconvénients (Ducrocq *et al*., 2009).

2. LE PROJET AMASGEN

Le projet AMASGEN (Approches Méthodologiques et Application de la Sélection GENomique) est un projet financé par l'ANR Génomique 2008 et ApisGène qui a débuté en janvier 2009. AMASGEN a pour objectif le développement de méthodes pour la sélection génomique chez les bovins laitiers, leur comparaison, en vue de faire évoluer l'outil actuel vers une méthodologie encore plus performante.
Les différentes méthodes actuellement utilisées dans le monde sont difficiles à comparer, car les critères utilisés ne sont pas homogènes et les structures des populations de validation mal connues et souvent différentes. Aussi est-il prévu de comparer les différentes approches à l'aide de jeux de données identiques. De tout premiers résultats sont présentés ci-dessous.

2.1. MATERIEL ET METHODES

La population d'apprentissage est constituée de 1217 taureaux Prim'Holstein et 452 taureaux Montbéliards nés avant 2000. La population de validation inclut 549 taureaux Prim'Holstein et 227 taureaux Montbéliards mis en testage en 2004 et recevant une première évaluation sur descendance en 2008. Pour cette validation, l'information considérée est celle disponible en 2004, à laquelle on ajoute les données de génotypage sur la puce de 54 000 SNP.
On compare trois approches différentes : le modèle polygénique qui correspond à la valeur sur ascendance classique, le modèle de la SAM2 (incluant l'effet de 20 à 35 QTL et un terme polygénique), et enfin, une évaluation par *Elastic Net* (et n'incluant à ce stade que l'information marqueurs).
On compare les valeurs génomiques estimées à partir des données de 2004 à l'aide des trois modèles, aux DYD obtenus après testage sur descendance en 2008. Il convient de noter que les DYD ne sont qu'une estimation très imparfaite de la valeurs génétique vraie. Les corrélations obtenues sont réduites : à l'imprécision des valeurs génétiques estimées de 2004 se cumule l'imprécision de l'estimation des DYD de 2008. On peut supposer toutefois que cela n'affecte pas le classement entre méthodes.

Tableau 1 : corrélations entre index de 2004 obtenus par le BLUP classique (modèle polygénique), l'évaluation SAM2, ou *l'Elastic Net* et les performances moyennes des filles (DYD) après testage sur descendance en 2008, pour les 227 taureaux Montbéliards et les 549 taureaux Prim'Holstein.

	Montbéliarde			Prim'Holstein		
	Polygénique	SAM2	Elastic Net	Polygénique	SAM2	Elastic Net
Quantité de Lait	0,273	0,420	**0,493**	0,423	**0,520**	0,443
Matière Protéique	0,276	0,383	**0,549**	0,330	**0,459**	0,373
Matière Grasse	0,355	0,438	**0,469**	0,317	**0,532**	0,503
Taux Protéique	0,214	**0,543**	0,392	0,449	**0,673**	0,635
Taux Butyreux	0,372	0,579	**0,614**	0,390	**0,755**	0,734

2.2. RESULTATS PRELIMINAIRES

Les résultats pour les caractères de production sont présentés dans le tableau 1 pour chacune des deux races. L'intégration d'informations moléculaires permet un accroissement moyen des corrélations de 0,19. Aussi bien la SAM2 que *l'Elastic Net* avec leurs paramétrages actuels permettent des gains substantiels en Montbéliarde, ce qui n'était pas forcément un fait acquis compte tenu de la faible taille de la population de référence.

Ainsi, pour cette race, on observe un gain moyen de corrélation de 0,17 point avec la SAM2 et de 0,21 point avec *l'Elastic Net*. Chez la Prim'Holstein, le constat global est le même avec un net gain de corrélation avec la SAM2 ou *l'Elastic Net.* Par contre, la SAM2 donne les meilleurs résultats avec un gain de corrélation moyen de 0,21 point contre 0,16 point avec *l'Elastic Net*.

La SAM2 et *l'Elastic Net* n'intègrent dans leur équations de prédiction qu'une faible proportion des 54 000 SNP de la puce. Ainsi, environ 150 SNP pour chaque caractère sont utilisés dans la SAM2 sous forme d'haplotypes de cinq à six marqueurs contre environ 500 SNP considérés individuellement dans *l'Elastic Net*. Logiquement, la plupart des SNP retenus pour la SAM2 le sont également par l'Elastic Net.

2.3. DISCUSSION

À ce stade, on ne note a priori pas de réelle supériorité systématique d'une méthode intégrant de l'information moléculaire par rapport à l'autre en race Montbéliarde. Par contre, pour les caractères étudiés, la SAM2 donne toujours de meilleurs résultats que *l'Elastic Net* en Prim'Holstein. Malgré tout, l'écart est faible pour la matière grasse et les taux. La performance de *l'Elastic Net* meilleure en Monbéliard qu'en Prim'Holstein malgré une population d'apprentissage nettement plus petite reste à éclaircir.

Il faut noter que les résultats de *l'Elastic Net* présentés ici n'utilisent aucune information issue du pedigree. L'ajout de cette information (index sur ascendance classique) pourrait permettre des gains supplémentaires en terme de corrélations, rapprochant ces deux méthodes en Prim'Holstein.

En outre, il semble intéressant de combiner les deux approches en utilisant *l'Elastic Net* pour « couvrir » la part de variance génétique représentée par le terme polygénique du modèle SAM2. La SAM2 permet de bien expliquer l'influence de gros QTL bien identifiés, *l'Elastic Net* estimant l'effet de tous les autres QTL. Cette combinaison constitue l'un des volets du projet AMASGEN.

D'autres approches d'évaluation génomique sont également à l'étude. Il est trop tôt pour tirer des conclusions claires, car pour chaque méthode, les résultats semblent beaucoup dépendre de la qualité du paramétrage. Sur ces mêmes données et avec les paramètres actuellement utilisés qui peuvent être sous-optimaux, des approches de type GBLUP et Lasso semblent aboutir à des corrélations oscillant entre celles obtenues avec un simple modèle polygénique et les meilleurs résultats obtenus avec *l'Elastic Net* ou la SAM2.

Ces premières observations semblent confirmées au niveau international. Ainsi, Verbyla *et al.* (2009) ont comparé un grand nombre de méthodes qu'ils ont appliquées à des données australiennes. La hiérarchie des meilleures approches parmi celles étudiées (qui n'incluaient ni *l'Elastic Net* ni la SAM2) variait d'un caractère à l'autre, le BLUP génomique donnant globalement de bons résultats, sauf dans la situation où la distribution des effets des QTL est atypique. C'est notamment le cas du taux butyreux où il existe un gène majeur, DGAT1 (Grisart *et al.,* 2002), avec un effet très marqué. Une approche de type BayesB est alors clairement supérieure. Par contre, aucun autre pays n'a actuellement considéré la possibilité de combiner SAM et sélection génomique.

Mécaniquement la collecte continue de nouvelles données de phénotypes et de génotypes devrait permettre d'atteindre rapidement une taille suffisante de population de référence pour obtenir des CD élevés avec des méthodes de type sélection génomique stricte. Les futurs enjeux seront donc la combinaison de la SAM et de la sélection génomique, et le maintien de temps de calcul raisonnables pour les évaluations.

3. PERSPECTIVES D'APPLICATIONS

Aujourd'hui, il semble donc que ce ne soit pas tant la méthode d'évaluation qui importe mais son adéquation aux données disponibles ainsi qu'au déterminisme génétique du caractère à évaluer. Ceci implique qu'avec l'accumulation de données et l'évolution des populations, l'intérêt de chacune des méthodes peut également évoluer. Le choix de la méthodologie d'évaluation ne doit donc pas être dissociée des caractères et de la structure de la population sur laquelle elle est appliquée.

3.1. PERSPECTIVES D'UTILISATION DE LA SELECTION GENOMIQUE

3.1.1. Disponibilité de la SAM2 dans les trois principales races françaises

Grâce à un programme de recherche mené en collaboration entre l'INRA, l'UNCEIA et LABOGENA, financé conjointement par l'ANR et ApisGène, la France dispose d'évaluations génomiques de type SAM2 dans les races Prim'Holstein, Normande et Montbéliarde depuis octobre 2008 (Fritz *et al.*, 2008). Tous les caractères avec évaluation officielle seront évalués en SAM2 en février 2010. Malgré une population de référence de l'ordre de 800 taureaux seulement (tableau 2), les races Normande et Montbéliarde disposent dès aujourd'hui d'évaluations génomiques efficaces, comme la race Prim'Holstein. Dans l'avenir, la collecte continue de génotypes et de phénotypes ouvre la voie à l'identification de nouveaux QTL permettant d'expliquer des parts de variance génétique de plus en plus importantes ou d'adopter un modèle d'évaluation de type génomique avec de meilleures garanties de succès. Les perspectives d'accroissement des tailles de population de référence des trois races sont d'ailleurs très bonnes (tableau 2) grâce aux investissements de la profession via ApisGene. Elles sont même excellentes en race Prim'Holstein qui va bénéficier d'un accord entre professionnels européens pour un échange de génotypes.

Tableau 2 : nombre de taureaux testés sur descendance et génotypés sur 54001 SNP en juin 2009

Race	Taille de la population d'apprentissage française en juin 2009	Taille de la population de référence espérée en décembre 2009
Holstein	2063	4000 (15000*)
Montbéliarde	804	1200
Normande	821	1200

* grâce à l'échange de génotypes avec l'Allemagne, les Pays-Bas et le Danemark

3.1.2. Vers une évaluation globale en une seule étape

L'évaluation génomique nécessite actuellement deux étapes : une évaluation nationale classique est tout d'abord réalisée à partir de laquelle sont tirées des informations résumées des performances des animaux (les DYD). Puis ces DYD sont utilisées pour l'évaluation génomique. Cette approche en deux étapes n'est pas optimale. De plus, l'évaluation génomique n'a actuellement pas de retombées sur les individus non génotypés proches d'animaux génotypés, par exemple, la mère d'un jeune taureau.

Pour pallier ces inconvénients, des méthodes sont en cours de développement (Misztal *et al*., 2009, Legarra *et al*., 2009) pour réaliser une évaluation globale incluant l'information génomique dans les évaluations nationales.

Dans le prolongement des méthodes GBLUP, la parenté génomique est étendue à tout individu (même non génotypé) à l'aide des règles de calcul de la parenté classique (Legarra *et al*., 2009 ; Christensen *et al*., 2009). Cette matrice complète permet de s'affranchir du calcul de DYD et de bénéficier de l'information moléculaire pour tous les individus. L'applicabilité d'une telle démarche reste encore à démontrer mais elle ouvre des perspectives très intéressantes pour l'inclusion systématique des génotypages des femelles dans les évaluations.

3.1.3. Perspectives pour les races à plus faible effectif

Pour les races laitières à plus faible effectif, la taille des populations de référence françaises ne permettra pas d'envisager à court terme des résultats significatifs. Pour certaines d'entre elles, des collaborations internationales sont ou peuvent être envisagées pour atteindre une taille de population critique pour la mise en place d'une évaluation génomique (cas de la Brune). La disponibilité d'une puce de génotypage à très haute densité (plus de 600 000 SNP) programmée pour 2010 devrait permettre des travaux de détection de QTL dont pourront profiter les races à faible effectif. En effet avec cette densité en marqueurs SNP, il devrait être possible d'identifier des associations entre allèles aux marqueurs et allèles aux QTL qui soient conservées entre différentes races et non plus simplement intra race : les informations des plus grandes races pourront alors bénéficier aux races d'effectif plus réduit.

3.2. CONSEQUENCES ET UTILISATION DES EVALUATIONS GENOMIQUES (OU SAM2) POUR LA SELECTION

3.2.1. Sélection équilibrée sur les différents caractères

Grâce aux évaluations génomiques, chez les bovins laitiers la précision des index est moins liée à l'héritabilité du caractère (Guillaume *et al*., 2008) que dans une approche classique. Comme on l'a vu (partie 1.1), cela est dû à la qualité du testage sur descendance qui aboutit à des CD élevés sur tous les caractères. Au moment de la sélection des reproducteurs, tous les caractères inclus dans les objectifs de sélection sont donc disponibles avec une précision comparable, même ceux faiblement héritables tels que la fertilité. Il devient donc possible aujourd'hui de sélectionner efficacement sur ce type de caractères fonctionnels et non plus de simplement limiter leur dégradation. Une autre nouveauté importante est la disponibilité d'index génomiques aussi précis pour les femelles que pour les mâles.

3.2.2. Diffusion des jeunes taureaux génotypés

En races Prim'Holstein, Normande ou Montbéliarde, une utilisation réfléchie des évaluations génomiques pour la sélection permet de pratiquement doubler le progrès génétique global grâce à une réduction importante de l'intervalle de génération, surtout sur la voie mâle. Dans ce contexte, on peut dès à présent prédire la disparition programmée du long et coûteux testage sur descendance. Un point important qui mérite une attention particulière est la gestion correcte de la parenté et de la consanguinité de la population. Colleau *et al.* (2009) démontrent que le doublement du progrès génétique peut être obtenu sans conséquences néfastes pour la consanguinité grâce à une utilisation plus équilibrée d'un plus grand nombre de reproducteurs mâles. Par ailleurs, pour la plupart des caractères évalués et dans les conditions actuelles, l'index d'un jeune taureau reste moins précis qu'après un testage sur descendance. Les index génomiques sont donc susceptibles de varier davantage qu'après testage. Il est donc important de ne surtout pas diffuser aussi intensément ces jeunes taureaux, qu'on le ferait avec taureaux testés actuels. On recommande de diffuser la semence des jeunes taureaux génotypés par groupe (ou « pack ») de façon à limiter les risques de chute d'index. En effet, si l'estimation du niveau génétique d'un jeune taureau est moins précise que celle d'un taureau testé sur descendance, l'estimation du niveau génétique moyen de quatre ou cinq taureaux est quant à elle bien plus précise. Cela permet de diffuser ces taureaux de façon homogène, de réaliser un supplément de progrès génétique et d'éviter les déceptions.

3.2.3. Maintenir le contrôle de performances

Bien qu'une évaluation génomique permette d'estimer assez précisément le niveau génétique des jeunes animaux sans performance, le contrôle de performances est essentiel pour garantir l'efficacité des évaluations génomiques à long terme. La collecte des performances de plusieurs centaines de filles par taureau permet en effet de vérifier la qualité des évaluations génomiques passées et de comparer les schémas de sélection. Surtout, elle est nécessaire pour reconstituer de nouvelles populations de référence, pour de nouvelles évaluations génomiques. Sans cela, il a été démontré que la précision des évaluations génomiques baisserait rapidement au cours des générations (Muir, 2007). Il est donc clair que la fin du testage sur descendance ne doit pas s'accompagner de l'abandon d'un contrôle de performances aussi exhaustif que possible.

La pression de sélection disponible, tant chez les mâles que les femelles, permet de prendre en compte des objectifs de sélection plus complexes. On peut donc imaginer que des caractères nouveaux pourront être sélectionnés à l'avenir, dès lors que leur mesure est possible, sur une fraction au moins de la population.

3.2.4. Corriger les biais des évaluations polygéniques dus à une présélection des animaux

L'efficacité à long terme des évaluations génomiques repose non seulement sur le maintien du contrôle de performances mais aussi sur des évaluations polygéniques classiques non biaisées. En effet, l'inclusion d'une étape de présélection génomique des taureaux diffusés ne permet pas de respecter certaines hypothèses essentielles du modèle polygénique. En particulier, le BLUP suppose que les animaux issus d'un même couple de parents ont une valeur génétique moyenne égale à la moyenne des valeurs génétiques des parents. Ce n'est clairement plus le cas quand seuls les meilleurs descendants sont retenus sur la base de leur évaluation génomique. Le BLUP requiert également l'inclusion de toutes les performances qui ont servi à la prise de décision de sélection ou d'élimination.

Sans ces conditions, la présélection des taureaux sur information génomique introduit un biais non négligeable :

les valeurs génétiques des animaux présélectionnés et de leurs filles sont sous-estimées et leurs précisions (CD) sont surestimées (Patry et Ducrocq, 2009). Les évaluations classiques – nationales et internationales - doivent donc être améliorées elles aussi. Il s'agit là d'un autre volet inclus dans le projet AMASGEN.

Par ailleurs, une autre source de biais souvent évoquée y compris dans les évaluations classiques est l'existence de traitements préférentiels sur certaines mères à taureau. Il est certain que la disponibilité d'évaluations génomiques de ces mères à taureau va permettre de détecter celles dont l'index classique a été biaisé. On peut se demander alors s'il est avantageux ou non d'inclure les performances des mères à taureau dans les évaluations génomiques. Cette interrogation est à l'ordre du jour en Amérique du Nord, avec une réponse différente pour les USA (qui incluent ces performances) et le Canada (qui les excluent). Le projet AMASGEN abordera également cette question.

3.3. PERSPECTIVES POUR LES BOVINS ALLAITANTS, LES OVINS ET LES CAPRINS

Pour les autres productions de ruminants, la mise en place d'une évaluation génomique ne bénéficie pas des avantages des bovins laitiers, à l'exception peut-être des ovins laitiers : l'insémination artificielle est moins développée et le nombre de mâles mis annuellement en testage est faible. Dans la plupart des cas, pour être suffisamment informatives, les populations de référence devront intégrer des animaux avec performances propres. Pour compenser la moindre information par rapport aux DYD, ces populations devront sans doute être de plus grande taille, ce qui augmente le coût de l'investissement initial. L'importance de la monte naturelle suppose une adaptation des outils pour assurer un progrès génétique plus élevé et sa bonne diffusion. Enfin, le coût des puces SNP (non disponibles pour l'instant en caprins) par rapport au coût des reproducteurs est nettement plus élevé en petits ruminants.

Malgré ces obstacles, il est certain qu'assez vite, cette situation va changer et qu'une sélection génomique pourra être mise en place. Parmi les facteurs favorisant cette évolution, on peut citer la baisse des coûts de génotypage, l'augmentation de la densité des puces autorisant des analyses multiraciales, les progrès méthodologiques dans la détection des QTL et les évaluations génomiques, et l'analyse conjointe des performances individuelles et de l'information génomique permettant d'accroître considérablement la taille des populations de référence.

CONCLUSION

L'arrivée des nouveaux outils de génotypage de type puces denses de SNP a permis de lever un ensemble de facteurs limitants pour la sélection assistée par marqueurs et la sélection génomique. La France a tiré profit de cet outil pour mettre rapidement en place une évaluation adaptée aux trois principales races bovines laitières, en suivant un modèle éprouvé et validé. Cette solution originale permet à l'heure actuelle d'obtenir des évaluations aussi précises que les approches d'évaluation génomique au sens strict.

Alors que les informations génotypiques s'accumulent à un rythme régulier, des efforts sont faits par le biais du projet AMASGEN pour améliorer les outils existants. Abandon du testage, utilisation radicalement différente des taureaux, progrès génétique potentiellement doublé et nettement plus équilibré : oui, la sélection génomique est à l'origine d'une vraie révolution dans le monde de la génétique animale !

Boichard *et al.*, 2003, *Genet, Sel, Evol,, 35, 77-101*
Colleau *et al.*, 2009, 3R 16, sous presse
Ducrocq *et al.*, 2009, *Interbull Bull, 39, 17-22*
Christensen *et al.*, 2009 *Proc 60th EAAP, 15, 299*
De Roos *et al.*, 2007 *J Dairy Sci, 90, 4821–4829*
Frank et Friedman, 1993, *Technometrics, 35, 109-135*
Fritz et *al.*, 2003 3 R 10, 53-56
Fritz et *al.*, 2007, 3 R 14, 129-132
Fritz et *al.*, 2008, 3 R 15, 423-426
Georges *et al.*, 1995 *Genetics, 139, 907–920,*
Grisart *et al.*, 2002, *Genome Res, 12, 222–231*
Guillaume *et al.*, 2009, *Genet Sel Evol, soumis*
Habier *et al.*, 2007, *Genetics, 177, 2389–2397*
Lande et Thompson, 1990, *Genetics, 124, 743–756*,
Legarra et al., 2009, *J, Dairy Sci, 92 :4656–4663*
Meuwissen *et al.*, 2001, *Genetics 157:1819–1829,*
Misztal *et al.*,, 2009, *Proc 60th EAAP, 15, 298.*
Muir, 2007, J Anim Breed Genet, 124 ,345-355
Patry et Ducrocq , 2009, *Interbull Bull, 40*, sous presse
Solberg *et al.*, 2006, *Proc 8th WCALP, 22-13*
Solberg *et al.*, 2008, *J, Anim, Sci, 86:2447–2454*
Solberg *et al.*, 2009, *Genet Sel Evol, 41, 29*
Tibshirani, 1996, *J Royal Stat Soc, 58, 267-288*
Van Raden et Wiggans, 1991, *J Dairy Sci, 74, 2737–2746*
Van Raden *et al.*, 2009**,** *J Dairy Sci, 92, 16-24,*
Verbyla *et al.*, 2009, *Proc 60th EAAP, 15, 294.*
Visscher et Haley, 1998*, Proc 6th WCALP, 23–503–510*
Zou et Hastie, 2005, *J Royal Stat Soc, 67, 301-320*