Intégration de l'information moléculaire dans l'évaluation génétique

François Guillaume

N° /__/__/__/__/__/__/__/__/__/__/

THÈSE

pour obtenir le grade de

Docteur

de

l'Institut des Sciences et Industries du Vivant et de l'Environnement (Agro Paris Tech)

Spécialité : Génétique animale

présentée et soutenue publiquement par

François GUILLAUME

le 06/10/2009

INTÉGRATION DE L'INFORMATION MOLÉCULAIRE DANS L'ÉVALUATION GÉNÉTIQUE

Directeurs de thèse : Didier BOICHARD et Tom DRUET

Travail réalisé : INRA, UMR de génétique animale et biologie intégrative, F-78352 Jouy-en-josas

Devant le jury :

Etienne VERRIER	Professeur**, AgroParisTech, Paris**	**Président**
Pascale LE ROY	Directeur de recherche, **Inra, Rennes**	**Rapporteur**
Frédéric FARNIR	Professeur**, Université de Liège, Liège**	**Rapporteur**
Pierre-Louis GASTINEL	Docteur, **Institut de l'Élevage, Paris**	**Examinateur**
Didier BOICHARD	Directeur de recherche, **Inra, Jouy-en-josas**	**Examinateur**
Tom DRUET	Docteur**, Université de Liège, Liège**	**Examinateur**

Résumé

L'évaluation génétique de reproducteurs dans les espèces d'intérêt agronomique repose depuis des années sur un modèle polygénique. Ce modèle utilise exclusivement l'information des généalogies et des phénotypes. La disponibilité de marqueurs moléculaires permettant de suivre les gènes influençant les performances, a ouvert la voie à une amélioration des évaluations. Différents modèles d'intégration de l'information moléculaire sont décrits dans la littérature, s'adaptant à l'état des connaissances et aux informations moléculaires disponibles. Ainsi, les gènes peuvent être parfaitement identifiés et l'effet de leurs allèles connu ; plus fréquemment, les gènes ne sont pas identifiés mais localisés dans une région chromosomique dite QTL, grossièrement ou finement. Une évaluation des reproducteurs intégrant des informations moléculaires a été mise en place en 2001 pour l'ensemble des trois principales races bovines laitières françaises et cette thèse présente d'abord un premier bilan de la qualité de ces évaluations, à partir de travaux de simulation et de données réelles. Ce travail présente dans un second temps les résultats obtenus dans le cadre d'un projet de cartographie fine de QTL à l'aide de marqueurs SNP à haut débit. La meilleure localisation des QTL et leur meilleur suivi permettant une évolution des modèles d'évaluation, les gains de fiabilité ainsi permis sont présentés dans une troisième partie avant de conclure sur les répercussions de ces nouvelles technologies en sélection.

Abstract

Integration of molecular information in breeding value estimation. For many years, breeding value estimation in domestic animal species has been based on the polygenic model. This model relies exclusively on pedigree and phenotypic informations. Molecular markers make it possible to follow the transmission of genes affecting phenotypes and, therefore, has opened new perspectives for evaluation improvements. Several models integrating molecular information have been described in literature, suited to the molecular information available. In some situations, the genes underlying the genetic variability of phenotypes are known, as well as their alleles and effects. More frequently, these genes are still unknown and are roughly or finely mapped in a QTL region. A breeding value evaluation with molecular information has been set up in 2001 in the three main French dairy cattle breeds. This thesis presents an overview of this original scheme and of its efficiency, based on both simulated and real data. In the next chapter, we present the results of a large scale QTL fine mapping experiment using high throughput SNP genotyping. These results have been used in a new and more efficient evaluation model described in the next chapter. Finally, the tremendous impact of these new technologies in animal breeding are discussed in the conclusion.

Mots clés : Sélection assistée par marqueurs - Bovins laitiers - QTL
Keywords : Marker assisted selection - Dairy cattle - QTL
Thèse réalisée à : INRA Centre de Recherche de Jouy-en-Josas
UMR de Génétique Animale et Biologie Intégrative (GABI, UMR1313)
78352 Jouy-en-Josas, France

Ce travail a été effectué dans le cadre d'une thèse CIFRE (n° 20050682) réalisée dans le cadre d'un projet soutenu par le CAS DAR (AAP 04 90) associant :

– L'INRA - GABI
– L'institut de l'élevage

Les données utilisées dans le cadre de cette thèse proviennent du programme SAM français associant :

– L'INRA
– Labogena
– L'UNCEIA

À mes parents,
à ma famille

Remerciements

> La plupart des hommes au moment de s'embarquer ne songent pas à la tempête
>
> De la tranquilité de l'âme, XI, 8.
> Sénèque

UNE thèse est comme un voyage en mer, plein de surprises, alternant périodes de calme plat et raz de marée ! Heureusement, dans chaque voyage il y a un équipage, que je tiens donc ici à remercier.

MES pensées se tournent tout d'abord aux capitaines du navire qui auront conduit ce travail de thèse à bon port. Tout d'abord, merci Tom, pour m'avoir fait confiance et proposé ce sujet de thèse. Malgré ton départ, ton encadrement à distance, ton aide rapide et efficace lors de la mise en place de la SAM2, m'ont permis de garder confiance malgré le challenge qu'ont été ces 10 derniers mois. Merci également Didier, pour avoir su m'accorder du temps, pour m'aiguiller dans les moments de doute, malgré la charge de travail qui est la tienne.

JE tiens également à remercier les armateurs du navire. Merci à Jean-Pierre Bidanel pour m'avoir accueilli au sein de la SGQA ....et de GABI depuis 2009. Merci à Jean-Claude Mocquot, Pierre-Louis Gastinel, Sophie Mattalia de m'avoir recruté au sein de l'Institut de l'Élevage. Merci également à Vincent Ducrocq, de m'avoir accueilli au sein de l'équipe Bovin Laitiers...puis G^2b.

MA gratitude se tourne également vers les membres du jury Étienne Verrier, pour accepter la présidence de ce jury, Pascale Le Roy et Frédéric Farnir les rapporteurs de ce travail ainsi que, Pierre-Louis Gastinel, Didier Boichard et Tom Druet pour avoir avoir accepté d'être examinateurs de ce travail.

JE remercie également tous les acteurs du programme SAM Français, l'UNCEIA, l'INRA et Labogena pour la confiance qu'ils ont su témoigner dans la mise en place de la SAM2.

MES pensées vont également vers l'ensemble de mes collègues. Merci à Sébastien Fritz, tout d'abord pour avoir su rendre mon travail tous les jours motivant. Merci à mes collègues successifs de bureau, Stéphanie Minéry, Mickael Brochard, Joaquim Tarrès et enfin Aurélia Baur, pour m'avoir supporté au jour le jour pendant presque 5 ans.

JE remercie également l'ensemble des membres du groupe Forza QTL et notamment André, Mathieu, Mekkhi et Slim pour avoir su si bien s'adapter et nous informer des multiples rebondissements qu'a pu connaitre le projet cartofine.

CETTE liste n'ayant pas la prétention d'être exhaustive, je tiens à remercier l'ensemble des personnes que j'ai pu cotoyer durant ces années de thèse. Notamment mes compagnons de RER, Florence, Gilles, Guillemette, Francis pour avoir parfois souffert sans mot dire de mon trop plein d'énergie du matin ... et parfois du soir, Alban et les autres membres de Doc'J pour avoir enduré les affres de la vie associative, les informaticiens de la station (Sylvie, Hervé et Fabien) pour mes demandes et discussions de geeks....les usagers de la salle café cochon, pour avoir parfois subi ces mêmes discussions de geeks. Une pensée particulière pour le cercle des caféinomanes de la pause de 5 heures, Thierry, Denis, Pierre, Marie-Pierre, Denis, Hélène, Aurélie pour nos discussions toujours amusantes et instructives. Merci également à Sophie et Clotilde pour la relecture de ce manuscrit et ainsi qu'à Serge pour la reliure de ce document.

ENfin, j'ai une pensée particulière à tous ceux à qui je n'ai pu accorder suffisamment de temps à cause de la fin de cette thèse, ma famille surtout, mon colloc' Manu, mes amis (les morellistes : Benjamin, Jocelyn et Jérémy), les amis de l'ESA Romain, Franck, François, Jean-Denis, Catherine, Aude, Claire... J'espère que vous êtes prêts à m'aider à rattraper le temps perdu.

EN guise de conclusion, on ne craint aucune tempête lorsque l'on est aussi bien accompagné que j'ai pu l'être durant ces trois années. C'est pourquoi je vous adresse à tous encore un grand ...

Merci

Table des matières

Chapitre 1

Introduction

Depuis que l'homme a domestiqué les animaux pour l'élevage, il les a sélectionnés. Constatant que les descendants bénéficient au moins partiellement des caractéristiques des parents, il a choisi pour la reproduction les individus présentant les phénotypes recherchés, ou leurs descendants. À partir du 18-19ème siècle, ces pratiques se sont organisées en différentiant des individus homogènes à l'origine des races, chaque race présentant des caractéristiques propres pour son apparence extérieure et pour certaines qualités d'élevage recherchées. Au 20ème siècle, ces approches deviennent scientifiques avec les apports de Fisher (1918) avec le modèle polygénique, puis de Hazel (1943), avec la théorie des index de sélection. Cette approche formalise l'apport de la connaissance des généalogies et des phénotypes pour prédire la valeur génétique des reproducteurs potentiels. Depuis cette période et pratiquement jusqu'à aujourd'hui, c'est cette approche qui est utilisée quasi-exclusivement, même si elle a été rénovée et perfectionnée avec les progrès de la statistique et de l'informatique. Parmi les évolutions marquantes, le BLUP (Henderson, 1975) formalise le modèle de description des données, permettant de distinguer les effets génétiques des effets de milieu qui contribuent simultanément au phénotype. Les évolutions plus récentes ont concerné la modélisation du phénotype pour prendre en compte des distributions non normales, des hétérogénéités de variance, des données répétées en séries chronologiques, des données manquantes ou censurées, mais aucune de ces évolutions n'a remis en cause le concept du modèle polygénique qui suppose que la valeur génétique est la somme des effets d'un nombre supposé infini (en tout cas élevé) de gènes.

Sous cette hypothèse, la valeur génétique d'un individu i (g_i) est la somme de la moitié de la valeur de chaque parent (g_p, g_m) et d'un aléa de méiose (ϕ_i).

$$g_i = \frac{1}{2} g_p + \frac{1}{2} g_m + \phi_i \tag{1.1}$$

Les propriétés de cet aléa de méiose ϕ_i sont les suivantes :
- Son espérance est nulle, de sorte que la valeur d'un individu est en espérance la valeur moyenne de ses parents.
- Sa variance est égale à la moitié de la variance génétique (pour des parents non consanguins). Cette variabilité intra famille explique que deux pleins frères ne soient pas identiques.
- Il est indépendant de la valeur des parents et ne peut donc pas être prédit à partir de leur information généalogique ou phénotypique. On peut l'illustrer intuitivement avec l'exemple suivant : chez un parent hétérozygote A/B à un locus, le descendant reçoit aléatoirement l'allèle A ou l'allèle B avec une probabilité $\frac{1}{2}$, et cette transmission ne peut pas être prédite à partir du seul parent.

Cette hypothèse du modèle polygénique est évidemment biologiquement fausse car le nombre de gènes est fini. Toutefois, ce modèle s'avère robuste et opérationnel dans la mesure où les prédictions réalisées sont précises et engendrent très généralement le progrès génétique espéré. Il permet de construire des programmes de sélection de façon rationnelle, optimiser la sélection, raisonner les coûts en fonction des gains espérés.

Retenons en particulier l'expression du progrès génétique espéré $E(\Delta_g)$ dans le cadre du modèle polygénique :

$$E(\Delta_g) = \frac{i \quad R \quad \sigma_g}{T} \tag{1.2}$$

où :
- i est l'intensité de sélection, égale à la différence entre la moyenne des animaux sélectionnés et la moyenne des candidats à la sélection ;
- R est la corrélation entre valeur génétique estimée et valeur génétique vraie ;
- σ_g est l'écart type génétique du caractère sélectionné ;
- T est l'intervalle de générations, c'est-à-dire l'âge moyen des parents à la naissance de leurs descendants.

Ce modèle polygénique présente pourtant des limitations importantes :
- Il ne prend pas en compte les avancées des connaissances sur le génome et, de ce fait, il est sans doute sous-optimal, au moins dans certaines conditions. On peut supposer que la connaissance au moins des gènes les plus importants dans le déterminisme d'un caractère doit améliorer le pouvoir prédictif du modèle.
- Il est peu efficace quand le caractère est peu héritable, du fait d'une faible valeur de R, égale à la racine de l'héritabilité en sélection massale. Pour augmenter R, il faut augmenter la quantité d'information, ce qui passe souvent par un contrôle sur descendance long, complexe et coûteux.
- Il nécessite une organisation lourde pour disposer des informations nécessaires, phénotypes et généalogies, ce qui est toujours coûteux à mettre en œuvre et parfois difficile, par exemple en milieu extensif ou quand la valeur des animaux est réduite. On contourne partiellement la difficulté par une structuration pyramidale entre sélectionneurs, multiplicateurs et utilisateurs, mais à nouveau cette organisation est complexe.

– Enfin, le modèle polygénique perd beaucoup en efficacité quand le phénotype n'est pas mesurable chez le candidat. En effet, dans le modèle polygénique, le phénotype de l'individu ou de ses descendants représente la seule possibilité de prédire l'aléa de méiose, c'est-à-dire l'écart à la moyenne familiale. Or cette situation est fréquente : caractère limité à un sexe, exprimé tardivement dans la vie ou nécessitant l'abattage, mesure incompatible avec le statut de reproducteur, etc.

On conçoit aisément que la connaissance du génome permet de prédire la valeur génétique d'un individu (notamment son aléa de méiose) sans recours aux phénotypes. Ainsi, dans le cas précédent du parent A/B, la connaissance du génome du descendant nous indique bien entendu s'il a reçu A ou B.

Une première approche pourrait donc consister à identifier tous les gènes et leurs mutations causales impliqués dans le déterminisme génétique de la variabilité d'un caractère, ou tout au moins les plus importants d'entre eux, et prédire la valeur d'un individu en fonction des allèles causaux qu'il porte, en supposant l'absence d'interactions entre ces gènes. En pratique, sauf exception, cette approche est difficile. Dans les cas les plus favorables où le déterminisme est dû à un gène majeur, ces gènes et leurs mutations causales peuvent être identifiés, ce qui permet ensuite une sélection directe sur le génotype. Lorsque ce gène n'explique qu'une partie de la variabilité génétique, une sélection combinant les informations du gène majeur et du phénotype est préférable pour tirer profit de toute la variabilité génétique (Lande et Thompson, 1990).

En pratique, les mutations causales sont difficiles à mettre en évidence. Les généticiens ont recours à une approche indirecte, reposant sur les propriétés des chromosomes. Au cours de la méiose pour la production des gamètes haploïdes, les gènes ne sont pas transmis aléatoirement les uns par rapport aux autres, mais de longs segments chromosomiques, contenant des centaines de gènes, sont transmis en bloc du parent au descendant. En moyenne, un chromosome ne subit guère plus d'un crossing-over par méiose. À l'échelle moléculaire, les recombinaisons sont donc rares : 1 centiMorgan (cM), soit 1% de recombinaison, correspond à 1 million de bases et à 10 gènes chez les mammifères. On considère donc que les ségrégations entre parents et descendants peuvent être efficacement suivies par quelques centaines de marqueurs qui sont capables de résumer les transmissions des larges segments chromosomiques qui les entourent. On définit un marqueur génétique comme n'importe quelle séquence polymorphe dont on sait discerner les allèles, par une méthode ou une autre.

Cette approche est mise en œuvre chez les animaux d'élevage depuis le milieu des années 90. La disponibilité de marqueurs microsatellites s'est traduite par la réalisation de programmes de détection et de cartographie de quantitative trait locus (QTL) dans de nombreuses espèces. Nous n'allons pas décrire les différentes possibilités mais retenons que ces approches ont reposé sur la détection de ségrégations intra famille, dans des dispositifs de croisement ou dans de grandes familles.

Quand les QTL sont localisés avec des marqueurs, deux grandes situations peuvent se présenter :

1. La localisation est suffisamment précise pour que l'on observe une association des marqueurs les plus proches du QTL avec le niveau de performance, dans l'ensemble de la population. On parle de déséquilibre de liaison (DL) populationnel entre marqueurs et QTL. Cette situation permet de sélectionner sur les QTL en sélectionnant sur certains allèles des marqueurs. En cas de déséquilibre très fort, c'est équivalent à une sélection directe sur le QTL. La précision de localisation requise dépend de l'effectif génétique de la population. Plus cet effectif est élevé, plus le déséquilibre de liaison diminue rapidement avec la distance, et plus la cartographie doit être fine. En pratique, dans les populations d'élevage bovines qui ont subi le goulot d'étranglement de la domestication puis la constitution récente des races, des localisations de 0,5 à 1 cM sont nécessaires intra race, et de l'ordre de 10 à 20 kbases à l'échelle de l'espèce.

2. La localisation est moins précise, de quelques cM à quelques dizaines de cM. Dans ce cas, il n'existe pas d'association au niveau populationnel, de sorte qu'une sélection sur certains allèles marqueurs n'induirait aucun progrès sur la fréquence des allèles des QTL. Par contre, du fait de la grande taille des segments chromosomiques transmis du parent aux descendants, il existe un déséquilibre de liaison familial utilisable en sélection, mais ce déséquilibre est rapidement détruit avec les générations par l'accumulation des recombinaisons. Dans cette situation, les marqueurs permettent de suivre les QTL d'une génération à l'autre et donc d'affiner la matrice de parenté au QTL. Cette information est utilisable en sélection, même si c'est plus complexe et moins efficace que le cas précédent.

Pendant de nombreuses années, il a été admis que l'utilisation de l'information moléculaire reposait sur l'utilisation des QTL les plus importants à l'aide de marqueurs liés, le reste de la variabilité génétique étant négligé ou valorisé par l'intermédiaire des performances. Il était en effet considéré que la détection des nombreux QTL expliquant une faible part de variance était hors de portée des dispositifs classiquement utilisés. Cette approche a un inconvénient important si les QTL pris en compte n'expliquent qu'une petite part de la variabilité génétique. Et en pratique, on constate que c'est fréquemment le cas. Hayes et Goddard (2001) ont montré que les dispositifs les plus fréquents, manquant de puissance, ne sont capables de mettre en évidence qu'une petite fraction des QTL et tendent à surestimer leurs parts de variance. Ainsi, la comparaison des résultats de la bibliographie montre généralement quelques résultats communs, correspondant à des QTL forts, et une faible redondance pour la majorité des autres, même dans des populations proches.

Certains auteurs ont alors proposé d'augmenter le nombre de QTL pris en compte (Stella *et al.*, 2002). Cette solution a l'avantage d'augmenter la part de variance vraie expliquée par les QTL, mais l'inconvénient d'introduire une proportion croissante de faux QTL réduisant l'efficacité de l'approche. Les résultats sont parfois divergents, mais il est généralement conclu qu'une certaine prise de risque est intéressante.

D'autres auteurs ont proposé d'utiliser des marqueurs couvrant tout le génome pour estimer le niveau d'apparentement vrai entre individus, plutôt que de supposer un apparentement moyen. Ainsi, alors que la parenté moyenne entre deux pleins frères est de $\frac{1}{2}$, la parenté vraie peut varier théoriquement de 0 à 1.

Ces deux idées ont été reprises et formalisées par Meuwissen *et al.* (2001) qui ont proposé le concept de sélection génomique. L'objectif est de prédire la valeur génétique d'un individu à partir de marqueurs denses couvrant tout le génome. La sélection génomique repose sur des principes proches de la sélection assistée par marqueurs (SAM), la différence essentielle résidant dans le nombre de QTL pris en compte dans la prédiction de la valeur génétique : quelques QTL à effet fort, bien localisés et caractérisés, dans le cas de la SAM ; un nombre inconnu et a priori élevé dans le cas de la sélection génomique, une fraction importante de ces QTL ayant un effet faible.

Quelques idées essentielles sont sous-jacentes à la sélection génomique :

1. la majeure partie de la variance génétique est expliquée par des QTL nombreux à petits effets ;
2. on peut obtenir une bonne prédiction de la valeur génétique d'un individu (définie comme la somme de l'effet de tous ses QTL), sans connaître précisément les effets individuels des QTL ;
3. on cherche les marqueurs les plus prédictifs, ce ne sont pas forcément les plus proches des QTL.

Le cadre de cette thèse est la sélection des bovins laitiers. Cette décennie a vu des évolutions très rapides du contexte et des applications dans cette espèce. Disposant de schémas de sélection classique solides reposant sur le testage sur descendance des mâles, ces populations ont été le socle de plusieurs programmes intégrant l'information moléculaire : détection puis cartographie fine de QTL ; sélection assistée par marqueurs de première génération utilisant le DL intra famille, puis de seconde génération utilisant le DL populationnel, et très récemment sélection génomique. Ces évolutions sont dues à une révolution de la technique de génotypage et le passage des microsatellites aux puces de Single Nucleotide Polymorphism (SNP) à haut débit.

Chez les bovins laitiers, le dispositif de choix est un dispositif petites-filles (Weller *et al.*, 1990) constitué de familles de taureaux demi-frères de père. Il est directement disponible sans surcoût et correspond à la population en sélection. Ces taureaux sont évalués sur descendance, ce qui leur confère un phénotype très particulier, la performance moyenne de leurs filles (Van-Raden et Wiggans, 1991). C'est une estimation précise de la valeur génétique du taureau car le nombre de filles est généralement élevé, de l'ordre de 100. Ce phénotype est équivalent à une performance propre pour un caractère d'héritabilité égale à la précision de l'index (R^2=0,5 à 0,9 selon les caractères). Les phénotypes sont ceux présents dans les bases de données du contrôle de performances et concernent généralement 20 à 30 caractères différents.

En France, un ambitieux programme de Sélection Assistée par Marqueurs de première génération (SAM1) a été mis en place en 2001 dans les trois principales races bovines laitières (Holstein, Normande et Montbéliarde). Il constituait un matériel de choix pour étudier l'apport de l'information moléculaire pour la sélection. Dans le même temps, des avancées importantes ont eu lieu dans la connaissance du génome et dans la disponibilité des outils d'analyses. Ainsi, en 2007-2008, nous avons bénéficié de génotypes SNP à haut débit, permettant de développer des approches basées sur le génotypage dense.

Objectifs de cette thèse

La première partie de cette thèse présente les outils et notions sur lesquels se base l'évaluation assistée par marqueurs. Nous détaillons différents modèles intégrant l'information moléculaire ainsi que les principales conclusions tirées des travaux les ayant étudiés.

Dans une seconde partie, après avoir décrit le programme SAM français de première génération, nous présentons une analyse de son efficacité.

La troisième partie présente des travaux de cartographie fine (sans, puis avec données SNP) réalisés pour la mise en place d'une sélection assistée par marqueurs de seconde génération.

Enfin le dernier article présente des travaux de validation du programme de sélection assistée par marqueurs de seconde génération.

La conclusion discutera des perspectives offertes par cette nouvelle situation qui est une véritable révolution pour le domaine, tant en France qu' au niveau international.

Chapitre 2

Méthodes de sélection assistée par marqueurs

Introduction

L'intégration d'information moléculaire dans la sélection serait grandement facilitée par l'identification des gènes responsables de la variabilité génétique des caractères. Jusqu'à présent, malgré quelques succès spectaculaires comme DGAT1 (Grisart *et al.*, 2002) sur le chromosome 14, GHR (Blott *et al.*, 2003) sur le chromosome 20, ABCG2 (Cohen-Zinder *et al.*, 2005) sur le chromosome 6, le nombre de mutations causales identifiées chez les bovins laitiers reste encore très réduit. De plus, des interrogations persistent quant à l'existence d'autres mutations dans ces mêmes gènes (Gautier *et al.*, 2007; Bennewitz *et al.*, 2004b).

En pratique, dans la très grande majorité des cas, les gènes responsables et a fortiori les mutations causales restent inconnus et l'on dispose seulement d'une localisation plus ou moins fine d'une région, dite QTL pour "quantitative trait locus ", contenant un ou plusieurs de ces gènes, ainsi qu'une estimation de leur effet. Bien que ces informations soient sous optimales, certains modèles d'évaluation génétique peuvent tirer parti de cette information, pour conduire à des estimations plus fiables.

Dans un premier temps, nous rappelons les conditions d'obtention de ces QTL et les propriétés associées. Dans un second temps, nous décrivons les principaux modèles d'évaluation possibles. Enfin, nous présentons en détail le programme français de sélection assistée par marqueurs .

2.1 Détection de QTL

Un QTL est caractérisé par sa position sur le génome (ou l'intervalle de confiance de sa position), la part de variance génétique du caractère qu'il explique et, dans les cas les plus simples, le nombre d'allèles et leurs effets.

Chez les bovins laitiers, Weller *et al.* (1990) a proposé d'utiliser la structure de population existante dans le cadre des programmes de sélection classique pour détecter les QTL. À cette époque, seules des cartes de marqueurs à faible densité sont envisageables et le dispositif petites-filles est constitué de familles de taureaux demi-frères de père.

Comme expliqué précédemment, les taureaux sont évalués sur descendance et ces index sont une excellente prédiction de leur valeur génétique. Ces index sur descendance sont utilisés comme phénotype en détection de QTL. Grâce à une forte réduction de la variance environnementale (du moins quand le nombre de petites-filles est suffisant), les dispositifs petites-filles présentent une bonne puissance en dépit de leurs défauts intrinsèques (le dispositif n'est que partiellement informatif car seule une partie des grands-pères est hétérozygote aux QTL ; l'analyse est intra-famille de père et les méioses maternelles ne sont pas utilisées ; la taille des familles, contrainte par les choix des sélectionneurs, est souvent en dessous de l'optimum).

Ces dispositifs sont en général analysés par régression intra père du phénotype sur la probabilité d'avoir reçu l'un ou l'autre segment chromosomique paternel à un locus donné, conditionnellement à l'information marqueurs. Pour maximiser l'information disponible, cette probabilité est estimée à partir des marqueurs informatifs flanquant la position testée, après avoir reconstitué les phases des pères, c'est-à-dire la répartition des allèles des marqueurs sur les deux chromosomes du père (Knott, 1994).

Après les travaux pionniers de Georges *et al.* (1995), de nombreux pays ont conduit des programmes de ce type, avec des dispositifs variant de quelques centaines à plusieurs milliers de taureaux. En France, le programme conduit de 1996 à 1999 comprend 1554 taureaux (Boichard *et al.*, 2003). On peut se référer à la synthèse de Khatkar *et al.* (2004) pour une revue des dispositifs en bovins laitiers et de leurs résultats.

Les résultats issus de ces programmes peuvent être résumés comme suit :
- Des QTL sont détectés en très grand nombre : plusieurs centaines de QTL sont décrits, ce qui contraste avec le petit nombre de mutations causales identifiées.
- Les parts de variance génétique estimées sont assez élevées mais généralement fortement surestimées. Les analyses plus récentes avec de très gros dispositifs donnent des estimations beaucoup plus réduites, confirmant ainsi que le nombre de QTL pour un caractère donné est plus élevé qu'initialement supposé.
- Mis à part quelques QTL majeurs retrouvés dans de nombreuses analyses, beaucoup de résultats semblent spécifiques à chaque étude et/ou à chaque race.
- La précision de localisation des QTL est très mauvaise, souvent de l'ordre de plusieurs

dizaines de cM. Conséquence directe, ces premiers résultats ont été suivis par des travaux importants de cartographie fine, en vue de réduire l'intervalle de confiance de la localisation. La stratégie utilisée repose avant tout sur un enrichissement en marqueurs, sur des méthodes d'analyse extrayant plus d'information (Bolard et Boichard, 2002) et parfois (mais pas assez souvent) sur une augmentation des effectifs. En France, le programme de sélection assistée par marqueurs a largement alimenté l'augmentation de taille du dispositif.

Une première évolution majeure dans la cartographie de QTL est la prise en compte du déséquilibre de liaison populationnel. Cette approche permet de tirer profit des recombinaisons historiques qui ne maintiennent une association entre marqueurs et QTL qu'à courte distance génétique. Dans l'approche linkage disequilibrium and linkage analysis (LD-LA), Meuwissen et Goddard (2001) proposent de combiner l'analyse de liaison (intra famille) et d'association (intra population) pour bénéficier des avantages de chacune de ces deux méthodes, la robustesse de l'analyse de liaison et la puissance et la résolution de l'analyse d'association. Cette approche permet de conserver le même dispositif petites-filles et de combiner l'analyse de liaison intra famille de père et l'analyse d'association entre parents, tirant ainsi parti (entre autres) de l'information maternelle. Toutefois, cette approche ne donne pas tout de suite tous les résultats potentiels, limitée par la densité en marqueurs (en général microsatellites) souvent trop faible pour bénéficier pleinement de l'information de déséquilibre de liaison.

2.2 Sélection assistée par marqueurs en déséquilibre de liaison familial

2.2.1 Modèles de description

Un premier type de modèle a été proposé par Fernando et Grossman (1989). L'idée sous-jacente est de décomposer la valeur génétique additive d'un animal en la somme d'un effet polygénique et la somme des effets aux QTL supposés additifs. On distingue pour chaque QTL l'effet de l'allèle au QTL d'origine paternelle et l'effet de l'allèle d'origine maternelle.

Le modèle de description est donc équivalent à celui de l'équation 2.1.

$$y = u_i + \sum_{k=1}^{N_{QTL}} (v_{ik}^p + v_{ik}^m) + e_i \tag{2.1}$$

où : – u_i est l'effet polygénique de l'individu i,
– v_{ik}^p etv_{ik}^m sont les effets de l'allèle au QTL k, d'origines paternelle et maternelle respectivement de l'individu i
– e_i est un terme résiduel.

Pour un animal i, l'effet de l'allèle au QTL reçu du parent j peut s'exprimer en fonction des effets des deux allèles au QTL de ce parent par la formule suivante 2.2.

$$v_i^j = p_1^j v_j^p + (1 - p_1^j) v_j^m + \varepsilon_i^j \tag{2.2}$$

où :
- v_i^j est l'effet de l'allèle au QTL reçu par l'individu i du parent j
- v_j^p et v_j^m sont respectivement les effets à l'allèle paternel et maternel du QTL pour l'individu j
- p_1^j est la probabilité d'identité entre l'allèle v_j^p d'origine paternelle de j et l'allèle v_j^i reçu par i de j
- ε_i^j est une erreur de modèle mais n'a pas de sens biologique. Elle reflète notre méconnaissance sur la transmission du QTL. En l'absence d'information marqueur, l'effet du descendant est l'effet moyen entre les deux effets parentaux et ε_i^j est égal à l'écart à cette moyenne. Au contraire, si p_1^j est égal à 0 ou 1, ε_i^j est nul.

Un modèle de ce type crée une structure de covariance entre effets de QTL, ce qui suppose donc que ces effets sont aléatoires. ε_i^j a une espérance nulle, les résidus sont indépendants les uns des autres. Leur variance dépend des probabilités d'identité des QTL ($p(q_i^p \equiv q_s^p)$), de la consanguinité des parents (F_s) et de la variance des effets du QTL (σ_q^2) :

$$Var(\varepsilon_i^p) = 2\sigma_q^2 p(q_i^p \equiv q_s^p)(1 - p(q_i^p \equiv q_s^p))(1 - F_s) \tag{2.3}$$

Les probabilités de transmission des allèles au QTL sont calculées à partir des informations moléculaires proches du QTL. Dans l'approche de Fernando et Grossman (1989), le modèle est simplifié à plusieurs niveaux : les parents sont supposés non consanguins ($F_s = 0$), l'origine parentale du QTL est connue et le modèle ne compte qu'un seul marqueur supposé informatif. Dans ce cas, p_1^j vaut $(1 - r)$ ou r, selon que i a reçu de j, l'allèle marqueur d'origine grand-paternelle ou grand-maternelle , r étant le taux de recombinaison entre le marqueur et le QTL. La partie suivante abordera les situations plus réelles.

Il est important de noter que dans ce modèle, aucun déséquilibre de liaison populationnel n'est supposé. Le modèle tire profit du déséquilibre intra famille, les marqueurs servant à suivre les QTL au sein d'un pedigree. Les marqueurs permettent de construire une matrice de parenté locale au QTL.

Fernando et Grossman montrent que :
- la méthodologie BLUP est applicable à ce modèle, ce qui permet d'utiliser tout l'arsenal du modèle polygénique ;

– le résidu ε_i est indépendant des valeurs parentales v_j^p et v_j^m, ce qui permet de construire la matrice $\mathbf{A_q}$ de probabilité d'identité (et surtout son inverse $\mathbf{A_q^{-1}}$) entre allèles QTL de façon simple ;
– $\mathbf{A_q^{-1}}$, l'inverse de cette matrice de probabilité d'identité, est creuse, ce qui simplifie considérablement les calculs.

Les équations du modèle mixte correspondantes sont les suivantes (réduites, pour des raisons de simplicité, aux parties polygénique et QTL) :

$$\begin{bmatrix} \mathbf{Z'Z} + \lambda_u \mathbf{A^{-1}} & \mathbf{Z'Z_q} \\ \mathbf{Z'_q Z} & \mathbf{Z'_q Z_q} + \lambda_q \mathbf{A_q^{-1}} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{u}} \\ \hat{\mathbf{v}} \end{bmatrix} = \begin{bmatrix} \mathbf{Z'y} \\ \mathbf{Z'_q y} \end{bmatrix} \tag{2.4}$$

Avec : – $\mathbf{Z}$ une matrice (p × n) d'incidence reliant chaque performance à un individu
– $\mathbf{Z_q}$ une matrice (p × 2n) d'incidence reliant chaque performance aux allèles QTL d'un individu
– $\mathbf{A^{-1}}$ est l'inverse de la matrice de parenté polygénique
– $\mathbf{A_q^{-1}}$ est l'inverse de la matrice gamétique.
– λ_u est le rapport de la variance résiduelle du modèle sur la variance polygénique (hors QTL)
– λ_q est le rapport de la variance résiduelle du modèle sur la variance QTL

Cette approche peut paraître curieuse dans la mesure où on cherche à estimer autant d'effets de QTL qu'il y a de copies, sous des formes alléliques variables, dans la population. Le nombre d'effets indépendants est bien inférieur, puisqu'en l'absence de mutations, il est au plus égal à 2 fois le nombre de fondateurs, les autres exemplaires n'étant que des copies transmises au reste de la population. L'approche BLUP permet d'estimer un grand nombre d'effets de QTL, un certain nombre d'entre eux étant très corrélés .

Matrice gamétique

Cette matrice est l'élément distinctif de l'évaluation assistée par marqueurs par rapport à une évaluation polygénique classique. $\mathbf{A_q}$ est la matrice des probabilités conditionnelles que les 2 allèles au QTL soient identiques par descendance, sachant le génotype au niveau des marqueurs. Cette matrice a une ligne et une colonne pour chacun des 2 allèles au QTL de chaque individu i.

Reprenons le cas simple supposé par Fernando et Grossman (cf figure 2.1). Notons q_i^p et q_i^m les copies du QTL porté par un individu et reçues de son père s et de sa mère d, et v_i^p et v_i^m leurs effets sur le caractère. De même, nous notons v_s^p et v_s^m les effets du QTL portés par son père s, et v_d^p et v_d^m les effets du QTL portés par sa mère d. L'information marqueur permet d'estimer les 4 probabilités $p(q_i^p \equiv q_s^p)$, $p(q_i^p \equiv q_s^m)$, $p(q_i^m \equiv q_d^p)$ et $p(q_i^p \equiv q_d^p)$. Le QTL est flanqué d'un marqueur M dont les taux de recombinaison avec le QTL est r. Le produit a reçu du père l'allèle

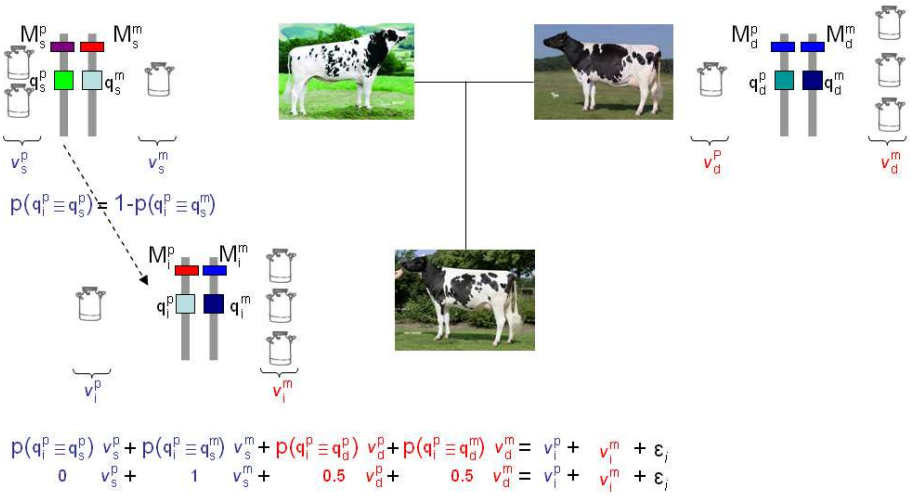


FIGURE 2.1 – Illustration du calcul des effets gamétiques dans un modèle Fernando et Grossman

M_s^p d'origine grand-paternelle, il a donc reçu q_s^p s'il n'y a pas eu de recombinaison (avec une probabilité $1-r$), et q_s^m sinon (avec une probabilité r).

Connaissant ces probabilités d'identité, on peut modéliser les effets de QTL (par exemple v_i^p) de la façon suivante (Fernando et Grossman, 1989) :

$$v_i^p = p(q_i^p \equiv q_s^p)v_s^p + p(q_i^p \equiv q_s^m)v_s^m + \varepsilon_i^p \tag{2.5}$$

Les résidus ont une espérance nulle et ils sont indépendants entre eux. Leur variance dépend des probabilités d'identité, de la consanguinité des parents (F_s) et de la variance des effets du QTL σ_q^2 (équation 2.3). Elle est bien sûr maximale en l'absence de données de marquage et nulle lorsqu'une probabilité d'identité est de 1. La formule 2.5 est donc très comparable à la formule relative aux valeurs polygéniques 1.1, de sorte que l'inverse de la matrice de parenté $\mathbf{A}_q$ entre effets de QTL peut être construite simplement à partir d'un fichier de pedigree et des probabilités d'identité des segments chromosomiques. Fernando et Grossman (1989) ont indiqué les règles de construction de $\mathbf{A}_q^{-1}$ lorsque l'origine parentale des allèles marqueurs est connue. Dans ce cas, pour chaque triplet individu (i), père (s) et mère (d), la contribution à $\mathbf{A}_q^{-1}$ se limite à deux matrices 3 x 3. La matrice 3 x 3 correspondant au chromosome de i (q_i^p) d'origine paternelle et aux deux chromosomes du père (q_s^p et q_s^m) reçoit les contributions suivantes :

Termes	Valeurs
(q_i^p, q_i^p)	d_i^p
(q_i^p, q_s^p) et (q_s^p, q_i^p)	$-pd_i^p$
(q_i^p, q_s^m) et (q_s^m, q_i^p)	$-(1-p)d_i^p$
(q_s^p, q_s^m) et (q_s^m, q_s^p)	$p(1-p)d_i^p$
(q_s^p, q_s^p)	$p^2 d_i^p$
(q_s^m, q_s^m)	$(1-p)^2 d_i^p$

où : – p = $p(q_i^p \equiv q_s^p)$
– d_i^p est l'inverse de la variance résiduelle, c'est-à-dire $d_i^p = \frac{1}{2p(1-p)(1-F_s)}$

La matrice 3 x 3 correspondant au chromosome de i (q_i^m) d'origine maternelle et aux deux chromosomes de la mère (q_d^p et q_d^m) est construite de façon identique. Lorsque aucune information de marquage n'est disponible, p=0.5. Lorsque le parent (s par exemple) est inconnu, la seule contribution restante est le terme (q_i^p, q_i^p) égal à 1.

Goddard (1992) étend ces formules au cas de marqueurs flanquants. Si le produit a reçu les allèles M_s^p et N_s^p grand-paternels du père (c'est-à-dire un haplotype parental), il a donc reçu l'allèle au QTL de cet haplotype (q_s^p) sauf s'il y a eu double recombinaison.

$$p(q_i^p \equiv q_s^p) = \frac{(1-r_{MQ})(1-r_{NQ})}{1-r_{MN}} \text{ et } p(q_i^p \equiv q_s^m) = \frac{r_{MQ}r_{NQ}}{1-r_{MN}} \tag{2.6}$$

Si le produit a reçu un haplotype recombinant (par exemple M_s^p et N_s^m) du père, l'expression est la suivante :

$$p(q_i^p \equiv q_s^p) = \frac{(1-r_{MQ})r_{NQ}}{r_{MN}} \text{ et } p(q_i^p \equiv q_s^m) = \frac{r_{MQ}(1-r_{NQ})}{r_{MN}} \tag{2.7}$$

La matrice gamétique se construit de la même façon que précédemment, en remplaçant les probabilités par leurs nouvelles valeurs.

Wang *et al.* (1995) étendent l'approche de Fernando et Grossman (1989) dans la situation où l'origine parentale des allèles marqueurs est incertaine, ce qui est un cas très général. Ils formalisent également le calcul de la consanguinité au QTL, alors que Fernando et Grossman l'approchaient par la consanguinité prédite par le pedigree. Enfin, Wang *et al.* proposent des solutions pour construire la matrice gamétique lorsque certaines informations au marqueur sont manquantes chez des individus. Pour construire la matrice gamétique, l'approche est tabulaire, à l'instar de la construction de la matrice de parenté. Comme pour l'inverse de la matrice de parenté

avec les règles d'Henderson, l'inverse de la matrice gamétique peut être construite directement par accumulation de contributions des triplets (individu, père, mère) successivement, à condition de connaître préalablement la consanguinité au QTL (ou de la négliger). Ces contributions sont des matrices 6 x 6.

L'approche de Wang *et al.* est limitée à un seul marqueur. Il n'existe pas, à notre connaissance, de méthode déterministe exacte pour calculer une matrice gamétique dans le cas général. Des méthodes MCMC ont été proposées pour prendre en compte des situations complexes de façon flexible, par exemple dans le logiciel LOKI, (Heath, 1997), mais elles peuvent être très longues en temps de calcul et la validation des résultats est problématique.

Pong-Wong *et al.* (2001) ont proposé une méthode approchée dérivée de celle de Wang *et al.* mais utilisant l'information de marqueurs flanquants. Ils montrent que leur approche est une très bonne approximation des valeurs vraies obtenues par MCMC. C'est cette méthode qui a été utilisée dans le programme SAM français de première génération.

La matrice $\mathbf{A_q^{-1}}$ permet d'estimer les effets de QTL avec un modèle BLUP. Par ailleurs, elle ouvre la voie à une estimation de la variance QTL par REML. Ainsi, tout logiciel REML autorisant l'inclusion d'une structure de variance-covariance comme ASREML (Gilmour, 1999) permet cette estimation.

En pratique, quand les probabilités d'identité par descendance , en anglais identity by descent (IBD) sont proches de 1 entre QTL, la matrice des coefficients n'est pas inversible et le système est surparamétré. Il convient alors de réduire le nombre d'équations, en regroupant les effets de QTL identiques en un seul effet. Cette approche a été proposée par Goddard (1992) et étendue au cas où le segment chromosomique compris entre deux marqueurs flanquants et proches semble être transmis sans recombinaison, ce qui revient à négliger les doubles recombinaisons. Cette approche devient indispensable avec l'utilisation de marqueurs denses, la majorité des probabilités IBD non nulles étant proches de 1.

L'approche de Fernando et Grossman (1989), avec ou sans ses évolutions, conduit à des systèmes d'équations de grande taille, avec deux effets à estimer par QTL et par individu. Le regroupement des effets QTL fortement corrélés contribue à diminuer cette taille. Une seconde solution repose sur l'utilisation d'un modèle animal réduit, par l'intégration de l'information phénotypique des animaux non génotypés dans leurs parents (Meuwissen et Goddard, 1999) et plus généralement par la définition de phénotypes intégrant l'information des apparentés non génotypés (Boichard *et al.*, 2002). Une autre possibilité, moins rigoureuse mais fréquente, consiste à limiter l'analyse à une sous population d'animaux génotypés (Boichard *et al.*, 2002).

Plus récemment, l'émergence du marquage à haut débit a permis la prise en compte du déséquilibre de liaison populationnel. À l'intérieur d'un pedigree, c'est-à-dire entre parents et descendants, l'information est déjà prise en compte par les méthodes décrites. Le déséquilibre de liaison se traduit donc par l'estimation d'un apparentement entre QTL des fondateurs. Une approche a été décrite par Meuwissen et Goddard (2001) pour la cartographie de QTL par LD-LA

mais elle est applicable pour l'évaluation. Ces auteurs proposent une méthode de calcul de la probabilité d'identité au QTL entre fondateurs à partir de l'identité par état d'un haplotype de marqueurs comprenant le QTL, en faisant des hypothèses sur l'histoire de la population ayant engendré les fondateurs connus. Au final, cette méthode se traduit par la construction d'une matrice de probabilité d'identité entre QTL de fondateurs. À nouveau, si ces probabilités sont élevées, un regroupement des QTL fondateurs est possible et souhaitable. C'est cette méthode qui est utilisée dans le programme SAM français de seconde génération.

2.2.2 Résultats

Les travaux publiés sur l'efficacité des évaluations utilisant le déséquilibre de liaison familial se sont généralement basés sur des travaux de simulation. Les paramètres de ces simulations étant rarement identiques entre publications, il est difficile de quantifier le gain réel permis par l'utilisation de marqueurs pour l'évaluation génétique – les chiffres rapportés variant d'un gain génétique augmenté de 64 % par rapport au gain dans un cadre polygénique (Meuwissen et Goddard, 1996) à des gains quasi nuls. Ces grandes variations sont dues à 2 points essentiels :

1. les paramètres du modèle utilisé : nombre de QTL, héritabilité du caractère, part de variance expliquée ;
2. les paramètres techniques de la sélection : quantité d'informations considérée, conditions d'utilisation des index assistés par marqueurs, paramètres de la sélection.

Paramètres du modèle

Il est communément admis que les informations moléculaires apportent peu de gain de progrès génétique dans les cas où les évaluations polygéniques sont déjà fiables : phénotypes déjà disponibles (Meuwissen et Arendonk, 1992), héritabilité forte (Lande et Thompson, 1990).

Plus la part de variance génétique expliquée par un QTL est importante plus les gains de progrès génétiques sont importants (Kashi *et al.*, 1990; Meuwissen et Goddard, 1996; Ruane et Colleau, 1995), de même plus la variance expliquée par l'ensemble des QTL est élevée, plus les gains de progrès génétique sont importants (Meuwissen et Goddard, 1996; Spelman *et al.*, 1999). Par exemple, Spelman *et al.* (1999) observent un progrès génétique amélioré d'un pourcent pour chaque pourcent de variance génétique expliquée par les QTL supplémentaires.

La mauvaise estimation de la variance associée à un QTL entraîne des réductions mineures de progrès génétique (Ruane et Colleau, 1996; Spelman et van Arendonk, 1997). Par exemple, pour un caractère d'héritabilité égale à 0.5, selon Ruane et Colleau (1996) le progrès génétique permis par la SAM est diminué de 0.02 % en première génération pour une surestimation ou sous

estimation de moitié de la variance du QTL. Après 3 générations ces chiffres deviennent 0.17 % et 0.15 % pour, respectivement, la surestimation et la sous estimation de moitié de la variance du QTL. L'utilisation de QTL fantômes est, par contre, plus problématique. Les résultats sont parfois contradictoires. Il ressort cependant que, quand le nombre de QTL est faible, ces QTL doivent être bien réels. Mais il est souvent préférable d'augmenter le nombre total de QTL pris en compte pour augmenter la part de variance QTL expliquée, même si une certaine proportion (qu'il faut minimiser) correspond à des QTL fantômes.

Dans le cadre d'une évaluation assistée par marqueurs , la capacité à suivre l'allèle au QTL affecte les gains de progrès génétique (Ruane et Colleau, 1995; Meuwissen et Goddard, 1996). Ainsi, Spelman et Bovenhuis (1998) indiquent des réductions de gains de progrès génétique de 30 % dues à l'utilisation de marqueurs flanquants distants de 15 cM au lieu de seulement 2 cM. Après 5 générations, cette réduction augmente du fait des recombinaisons, la réduction est alors de l'ordre de 55 % si le QTL explique 5% de la variance phénotypique. En pratique, le bon suivi implique une localisation fine du QTL, Spelman et van Arendonk (1997) estiment ainsi qu'une erreur de localisation de 15 cM pouvait réduire de 25 % le progrès génétique permis par l'utilisation de la SAM.

Le progrès génétique dû à la SAM est également dépendant des fréquences des allèles favorables aux QTL, ainsi on obtient un gain de progrès génétique plus important si l'allèle favorable d'un QTL est initialement présent avec une fréquence faible (Spelman et Garrick, 1997; Larzul *et al.*, 1997). Par exemple, pour un QTL d'effet égal à 2 écart-types génétique, un gain de 2.28 % de progrès génétique est observé, grâce à l'ajout d'information moléculaire, à l'horizon de 10 ans, si la fréquence allélique au QTL est de 10 %, tandis que ce gain n'est que de 0.13 %, avec une fréquence allélique de 0.75 (Spelman et Garrick, 1997). Ce gain de progrès génétique est également tributaire de l'échelle de temps sur lequel on l'observe, par exemple, pour un QTL dont l'allèle favorable de fréquence égale à 10%, d'effet égal à 2 écart-types génétique, Spelman et Garrick (1997) notent un gain de progrès génétique de 2.28 % à l'horizon de 10 ans tandis que cette supériorité n'est plus que de 0.76 % à l'horizon de 30 ans .

Cet effet du temps s'explique par la diminution progressive de la variabilité génétique disponible au QTL, ainsi qu'à la perte de pression de sélection réalisée sur la composante polygénique. Ruane et Colleau (1995) observent pour un caractère d'héritabilité 0.5 un gain de progrès génétique de 2.1 % en première génération tandis que ce gain n'est plus que de 0.6 % après six générations. Tel que souligné par Meuwissen et Goddard (1996), la SAM exploite plus rapidement la variance génétique due au QTL qu'un modèle polygénique, ceci explique la convergence de progrès génétique observé sur le moyen et long terme. La part polygénique du progrès génétique obtenu est quant à elle diminuée par rapport à une sélection sans marqueur (Ruane et Colleau, 1995), des résultats inférieurs à une sélection polygénique classique peuvent ainsi être obtenus sur le long terme (Gibson, 1994; Verrier, 2001). Ce dernier point peut s'expliquer par l'effet Bulmer (Bulmer, 1971), qui crée un déséquilibre entre l'allèle favorable au QTL sélectionné et l'effet polygénique.

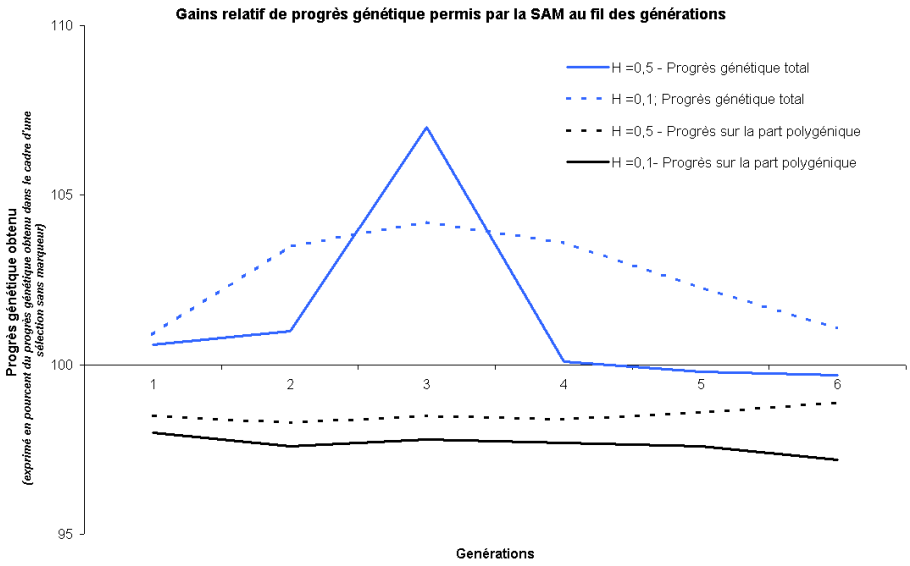


FIGURE 2.2 – Illustration de l'évolution sur 6 générations du gain de progrès génétique permis par la SAM, d'après Ruane et Colleau (1995). L'axe des ordonnées correspond au progrès génétique obtenu sans intégration de l'information moléculaire.

Mise en œuvre

La manière dont est mis en œuvre le programme d'évaluation assistée par marqueurs, ainsi que l'utilisation faite des évaluations enrichies de l'information moléculaire, ont un effet sur le progrès génétique.

Dans le cadre d'une évaluation assistée par marqueurs en équilibre de liaison, l'évaluation doit être utilisée dans les phases précoces, lorsque l'information des phénotypes n'est pas encore disponible. Ainsi, chez les bovins laitiers, l'utilisation d'une sélection assistée par marqueurs n'a pas d'intérêt après testage sur descendance car elle ne permet pas de gain de précision important (Meuwissen et Goddard, 1996). De plus, la SAM est plus efficace si elle est utilisée pour réaliser une présélection intra famille de pères (Kashi *et al.*, 1990) plutôt qu'au niveau de la population (Gibson, 1994).

La sélection des mères à taureaux permet également des gains de progrès génétique. Schrooten *et al.* (2005) montrent par exemple, qu'en réduisant de moitié le nombre de femelles, un gain de progrès génétique de 21 % est observé par rapport à un schéma de sélection sans information moléculaire . Dans le même ordre d'idée, le doublement du nombre d'embryons produits permet un gain de progrès génétique de 4.8 % par rapport à un schéma sans information moléculaire. Bien que cette étude soit simplificatrice (sélection uni caractère, un seul QTL), elle met en évidence le fait que l'adaptation des schémas de sélection, tout autant que le modèle d'évaluation, contribue à augmenter le progrès génétique.

2.3 Évaluation assistée par gènes

2.3.1 Modèle de description

Lorsque la mutation causale responsable du QTL est identifiée, elle est utilisable directement à la place des marqueurs. Cette situation est bien sûr beaucoup plus simple et les avantages sont nombreux : typages plus réduits en nombre de tests et nombres d'animaux ; nombre d'effets à estimer plus réduit et donc plus grande précision ; transposition aisée à d'autres populations ; possibilité de modéliser plus précisément l'effet, avec prise en compte de la dominance et éventuellement d'épistasie. Dans le modèle statistique, l'effet de l'allèle au QTL est remplacé par l'effet de l'allèle à la mutation causale du gène γ_p.

$$y = u_i + \sum_{N_{\text{Gènes}}} \sum_{phase=1,2} \gamma_p + e_i \tag{2.8}$$

où : – u_i est l'effet polygénique de l'individu i,
– γ_p est l'effet de l'allèle porté sur la phase p au gène de l'individu i
– e_i est un terme résiduel.

Sauf exception, comme les gènes soumis à empreinte, l'origine parentale des allèles n'impacte en rien le modèle. Ce modèle peut être considéré comme la situation extrême du cas précédent, avec des probabilités IBD égale à 1 et la constitution de clusters recouvrant exactement la définition des allèles. Les effets peuvent être considérés comme fixes ou aléatoires. Le modèle peut considérer l'effet des allèles supposés additifs ou l'effet des génotypes, incluant la dominance. Le modèle reste complexe quand la population n'est pas entièrement typée. Il convient alors d'imputer (en probabilité) le génotype de la fraction de la population non typée, en fonction de la fraction typée et des relations de parenté.

2.3.2 Résultats

Les résultats obtenus dans les études précédemment citées s'appliquent également au cas de la sélection assistée par gènes, aux nuances près que : les problèmes de distance entre marqueurs et QTL n'existent plus et que le risque de fixation des allèles favorables est plus important.

Bien que de nombreuses publications aient identifié des mutations supposées causales, rares sont les gènes pour lesquels ces mutations ont été confirmées (Ron et Weller, 2007). De plus, la mise en évidence d'une mutation causale dans un gène (Grisart *et al.*, 2002) n'exclut pas l'existence d'autres mutations pour un même gène ou cluster de gènes (Bennewitz *et al.*, 2004b; Gautier *et al.*, 2007).

Sur données réelles, quelques études ont évalué le gain de prédiction permis par une approche sélection assistée par gène. Ainsi, dans le cadre d'un modèle utilisant seulement la mutation causale de DGAT1 (K232A) pour évaluer la valeur génétique du taux butyreux de Roos *et al.* (2007) ont mis en évidence un gain de 0.25 point de corrélation par rapport à un modèle animal classique. Ce résultat est néanmoins à relativiser, d'une part parce qu'il concerne un cas très particulier dans lequel la mutation causale est connue et d'autre part parce que le gène explique une proportion importante de la variance génétique du caractère.

2.4 Modèles de sélection génomique

La disponibilité de puce pangénomique de SNP, a ouvert de nouvelles perspectives pour la sélection, parmi lesquelles la plus étudiée est sûrement la sélection génomique. Ce type d'évaluation regroupe en fait un ensemble de méthodes et modèles assez différents mais ayant en commun la prise en compte d'informations moléculaires couvrant l'ensemble du génome de façon dense. Cette idée a été émise par Visscher et Haley (1998), mais la première étude de faisabilité d'une évaluation génomique n'est apparue que plus tard (Meuwissen *et al.*, 2001). Par analogie aux méthodes d'évaluation précédemment présentées, le modèle utilisé considère que tout fragment du génome est un QTL (dont l'effet peut cependant être nul). De plus, du fait de la taille réduite des fragments chromosomiques, on suppose qu'il existe un déséquilibre de liaison entre fragment chromosomique et gène sous jacent au QTL. L'ensemble des effets de gène étant évalué par le biais des fragments chromosomiques, la composante polygénique devient a priori inutile (Habier *et al.*, 2007).

Deux idées sont alors sous-jacentes au concept de sélection génomique : 1) la majeure partie de la variance génétique est expliquée par des QTL à petits effets ; 2) on peut obtenir une bonne prédiction de la valeur génétique d'un individu (définie comme la somme de l'effet de tous ses QTL) sans connaître précisément les effets individuels des QTL.

2.4.1 Modèle de description

Ainsi, les idées sous jacentes à la sélection génomique peuvent être résumées par les modèles suivants :

$$\mathbf{y} = \mu \mathbf{1_n} + \sum_i \mathbf{X_i g_i} + \mathbf{e} \tag{2.9}$$

où : – $\mathbf{y}$ est un vecteur de phénotypes,
– μ une moyenne,

- $\mathbf{1_n}$ un vecteur unité de même longueur que $\mathbf{y}$
- X_i une matrice d'incidence reliant les phénotypes de chaque individu aux allèles au fragment chromosomique i ,
- g_i l' effet de l'allèle .

Dans l'article princeps de Meuwissen *et al.* (2001), plusieurs méthodes ont été utilisées pour estimer les inconnues du modèle 2.9. Notamment, une méthode simple de moindres carrés et un modèle de type BLUP dans lequel on considère que chaque région chromosomique a un effet aléatoire de variance σ_i^2, telle que $\sum_i^N \sigma_i^2 = V(G)$.

Deux méthodes bayesiennes (BayesA et BayesB) ont également été proposées et, de par leurs performances, se sont imposées comme les méthodes de référence. Le principe de ces méthodes est d'estimer conjointement l'effet des régions chromosomiques et leur variance.

Méthode BayesA Dans la méthode BayesA, la variance de l'effet de QTL est échantillonnée à partir d'une loi de chi 2 inversé. Elle permet donc de prendre en compte des contributions très différentes d'un QTL à l'autre à la variance génétique totale.

$$Post(\sigma_{gi}^2|g_i) = \chi^{-2}(\upsilon + n_i, S + \mathbf{g_i'g_i})$$

La variance d'erreur est également tirée d'une loi de chi 2 inversé .

$$Post(\sigma_e^2|e_i) = \chi^{-2}(n-2, \mathbf{e_i'e_i})$$

Le tirage des valeurs de variance est effectué des milliers de fois et on calcule les paramètres des modèles par intégration des paramètres successivement obtenus.

Méthode BayesB L'approche BayesB suppose que seule une (petite) proportion des segments chromosomiques portent effectivement un QTL, de variance non nulle. Dans la majorité des cas, l'effet estimé est donc nul, ce qui évite un bruit de fond important. La distribution des valeurs de σ_{gi}^2 suit alors un mélange de lois tel que :

$$\begin{cases} \sigma_{gi}^2 & = 0 & \text{avec une probabilité } \pi \\ \sigma_{gi}^2 & \sim \chi^{-2}(\upsilon + n_i, S) & \text{avec une probabilité de } 1-\pi \end{cases}$$

La supériorité de cette méthode par rapport à BayesA est de pouvoir considérer une proportion restreinte (1-π) d'effets à estimer. La méthode BayesB est actuellement la référence en terme d'efficacité de prédiction génomique. Elle présente cependant une importante limitation, son temps de calcul. Des méthodes approchées sont néanmoins proposées (Meuwissen *et al.*, 2009).

Méthode	corrélation	coefficient de régression
LS	0.318	0.285
BLUP	0.732	0.896
BayesA	0.798	0.827
BayesB	0.848	0.946

TABLE 2.1 – Comparaison des résultats de différentes méthodes de sélection génomique d'après Meuwissen *et al.* (2001). Pour la génération n+2,(animaux génotypés mais sans performances), la corrélation est calculée entre valeur génétique vraie et celle prédite par le modèle, de même le coefficient de régression est calculé en régressant la valeur génétique vraie sur la valeur génétique prédite.

2.4.2 Résultats

Les résultats obtenus par simulations dans l'article original sont récapitulés dans le tableau 2.1, ils reflètent la capacité des modèles à prédire pour une population d'animaux candidats (sans performances propres) leur valeur génétique vraie. Les niveaux de corrélations affichés sont de l'ordre de grandeur des corrélations attendues entre prédictions de la valeur génétique après un testage sur descendance et valeurs génétiques vraies. La question soulevée par ces résultats est donc l'intérêt de réaliser un testage sur descendance, si des estimateurs de qualité équivalente peuvent être obtenus dès la naissance d'un candidat.

On peut relativiser ces résultats. Ils sont tout d'abord issus de simulations qui postulent une homogénéité de l'information tant moléculaire que généalogique. De plus le caractère simulé présente une héritabilité assez élevée (h^2=0.3) alors que l'enjeu majeur de la sélection assistée par marqueurs reste les caractères à faible héritabilité (h^2<0.1). Enfin, la connaissance des paramètres utilisés pour la simulation permet de s'affranchir de la difficile étape de calcul des paramètres du modèle.

Avec la disponibilité de puces pangénomiques contenant des dizaines de milliers de SNP, la sélection génomique s'avère être une approche réaliste, la moindre informativité des SNP par rapport aux marqueurs microsatellites n'engendrant que de faibles baisses de précision largement compensées par une plus grande densité de marqueurs (Solberg *et al.*, 2008) ou l'utilisation d'haplotypes de SNP (Villumsen *et al.*, 2009).

2.5 Quelle différence entre sélection assistée par marqueurs et sélection génomique ?

Dans son principe, la sélection génomique repose sur des concepts proches de la sélection assistée par marqueurs. Mais, alors que la SAM ne s'intéresse qu'à un nombre limité de QTL à effet important pour la prédiction de la valeur génétique, chaque QTL étant préalablement cartographié, la sélection génomique considère un nombre inconnu et a priori élevé de QTL. Leur nombre n'étant pas spécifié, leur localisation et leur effet sont a fortiori inconnus. L'idée est donc de laisser parler la statistique pour détecter les marqueurs les plus prédictifs. En pratique, les marqueurs à plus fort effet ne sont pas forcément les plus proches du QTL (Villumsen *et al.*, 2009). En effet, les marqueurs sélectionnés sont ceux en plus fort déséquilibre de liaison avec le QTL, leur distance important peu. Une autre différence forte entre SAM et sélection génomique est la nature de l'information moléculaire prise en compte. En SAM, de plus en plus, les QTL sont suivis par des haplotypes denses, l'idée étant de constituer un marqueur composite présentant un nombre suffisant d'allèles pour obtenir le déséquilibre de liaison voulu, mais un nombre pas trop élevé pour éviter la perte de précision de l'estimation de son effet. Au contraire, la sélection génomique considère classiquement un QTL par intervalle, donc un QTL pour un ou deux marqueurs SNP.

On peut imaginer une convergence progressive entre sélection génomique et SAM, dès lors que la SAM prend en compte plus de QTL (détectés à partir de dispositifs de plus en plus puissants), la sélection génomique se base sur des haplotypes de marqueurs (Villumsen *et al.*, 2009) ; et, pourquoi pas, la sélection génomique pilote le choix des SNP retenus en forçant ceux qui sont placés sur les QTL connus.

2.6 Programme français de sélection assistée par marqueurs

La France est l'un des rares pays qui a mis en œuvre un programme de sélection assistée par marqueurs de grande envergure. Ce programme a débuté en 2001 (Boichard *et al.*, 2002), il visait à utiliser en sélection les QTL détectés dans les trois principales races bovines laitières françaises au cours des années 90.

2.6.1 Programme de détection de QTL

Le programme de détection de QTL était un protocole petites-filles classique constitué de 1548 taureaux d'insémination artificielle appartenant à 14 familles de pères de races Montbéliarde, Normande et Holstein (Boichard *et al.*, 2003). Ces taureaux disposaient d'évaluations génétiques après testage sur descendance pour 24 caractères d'intérêt et ont été typés à l'aide de 169 marqueurs, pour la plupart microsatellites, couvrant l'ensemble des autosomes. La détection de

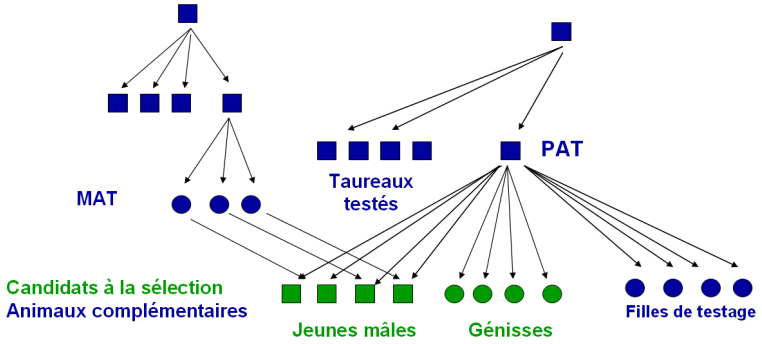


FIGURE 2.3 – Représentation des animaux typés dans le programme SAM

QTL a été réalisée à partir d'un modèle de régression intra-père (Haley *et al.*, 1994), les seuils de significativité des tests ont été obtenus par permutations (Churchill et Doerge, 1994).

Au total, 120 QTL avec un seuil de signification inférieur à 3% au niveau du chromosome ont été mis en évidence. Parmi ces QTL, 32 dépassaient le seuil de 1% au niveau du génome. Une majorité des QTL présentant les effets les plus importants confirmèrent d'autres résultats préalablement publiés (Georges *et al.*, 1995; Heyen *et al.*, 1999). Les estimations de parts de variance génétique expliquées par les QTL sont généralement comprises entre 5 et 15%, ce qui est suffisamment élevé pour envisager une utilisation en sélection.

2.6.2 Description du programme SAM

Le programme SAM est de première génération. La localisation des QTL est relativement imprécise, avec des intervalles de confiance d'au moins 5 cM et pouvant atteindre 20 cM. Avec une telle distance, aucun déséquilibre populationnel n'est envisageable et les marqueurs sont utilisés pour construire une matrice de parenté locale aux QTL, permettant de suivre les QTL entre parents et descendants.

Pour être efficace, le programme distingue 2 types d'animaux à génotyper, d'une part les candidats à la sélection – animaux jeunes et sans performance – et d'autre part les animaux complémentaires. Les animaux génotypés se répartissent en nombres comparables entre candidats et complémentaires. Les animaux complémentaires sont caractérisés par leurs performances propres ou celles de descendants. C'est cette information que les marqueurs transmettent jusqu'aux candidats. Compte tenu des probabilités de recombinaison à l'intérieur de l'intervalle de confiance, il est important que les animaux avec information phénotypique ne soient pas trop distants des candidats. La stratégie combine les deux approches "Bottom-up" et " Top Down" proposées par Mackinnon et Georges (1998). Autrement dit, les candidats sont évalués d'une part à partir de leurs demi-sœurs avec performances, d'autre part à partir de leurs oncles (côté paternel) et grands-oncles

(côté maternel) évalués sur descendance. Durant les 7 années du programme, 10 000 animaux ont été génotypés annuellement, sur un ensemble de 43 marqueurs microsatellites assurant le suivi de 14 régions, constituant une base de données de génotypages d'une taille inédite. Huit caractères ont été évalués mensuellement. Chaque caractère était prédit à partir de 3 à 5 QTL et chaque QTL était suivi par 3 à 4 marqueurs.

Deux types de phénotypes ont été utilisés, selon le sexe de l'individu : d'une part, pour les mâles, les « DYD » (pour daughter yield deviation) définis comme la moyenne des performances de ses filles non typées, corrigées pour l'ensemble des facteurs inclus dans le modèle d'indexation officielle et la moitié de la valeur génétique de leur mère (VanRaden et Wiggans, 1991) ; d'autre part, pour les femelles les « YD » (Yield Deviation), définis comme la moyenne des performances de la femelle, corrigées pour les facteurs non génétiques du modèle. DYD et YD sont des sous-produits de l'évaluation génétique officielle. Dans les cas où les DYD ne sont pas disponibles (taureaux étrangers par exemple), ils sont remplacés par des index « dérégressés », en principe équivalents même si ceux-ci sont moins indiqués (Thomsen *et al.*, 2001).

La stratégie adoptée était donc ambitieuse : apporter beaucoup d'information phénotypique au prix du génotypage de nombreux animaux non-candidats ; suivi de nombreux QTL pour prendre en compte une forte proportion de la variance génétique mais aussi compenser les pertes d'efficacité prévisibles du programme et éviter les situations de non-informativité complète, source d'incompréhension des utilisateurs.

Une organisation lourde a été mise en place, assurant un délai de 4 à 6 semaines entre la réception d'un échantillon à Labogena et la restitution des prédictions de valeur génétique pour les candidats. Le système permettait à la fois des résultats privatifs pour des entreprises de sélection en concurrence mais une mutualisation des informations pour maximiser la précision de l'évaluation.

L'évaluation génétique repose sur un modèle BLUP mono-caractère, multiQTL, estimant simultanément l'effet des QTL et la composante polygénique résiduelle. L'inverse de la matrice IBD est construite à partir de l'information multi-marqueurs, ce qui suppose la construction préalable des phases. Cette construction est facilitée par la forte information familiale mais elle peut être source de perte d'information dans certaines situations. En cas de phase non connue, le marqueur informatif le plus proche est pris en compte.

2.6.3 Évolutions du programme français

Le programme de sélection assistée par marqueurs a toujours été pensé comme un système évolutif. Dans son principe même, il était prévu qu'il gagne progressivement en efficacité au fur et à mesure de l'empilement des générations et de la concentration d'information. D'autre part, il était conçu en synergie du programme de cartographie fine : il alimentait les travaux de cartographie et réciproquement, il pouvait intégrer les résultats acquis. Le jeu de marqueurs utilisé a été modifié en 2004 pour permettre un meilleur suivi de certains QTL. Le typage des filles de

testage demi-sœurs des candidats a été encouragé du fait de son impact positif sur l'efficacité du dispositif, malgré la lourdeur organisationnelle induite. Les paramètres génétiques du modèle ont été réestimés à partir de l'ensemble de la base de données (Druet *et al.*, 2006), ce qui a conduit à une diminution des parts de variance attribuées à la plupart des QTL. Enfin, des caractères nouveaux ont été intégrés à l'évaluation mensuelle. La stratégie du programme était de progresser le plus possible sur la connaissance du génome afin de proposer un modèle reposant sur des informations vérifiées, et d'affiner la localisation de QTL (Gautier *et al.*, 2006; Guillaume *et al.*, 2007b) afin de tirer parti du déséquilibre de liaison, puis éventuellement des mutations causales (Boichard *et al.*, 2006) .

En 2007, la disponibilité de puces SNP bovines commerciales de 54 000 SNP couvrant l'ensemble du génome a considérablement accéléré l'évolution du programme, permettant à la fois de raffiner la localisation des QTL et d'intégrer le déséquilibre de liaison dans l'évaluation génétique. C'est ainsi qu'à partir de 2008, le programme SAM1 a cédé la place à un programme de SAM de deuxième génération, utilisant des haplotypes de SNP en déséquilibre de liaison avec les principaux QTL localisés finement lors d'un second projet de détection de QTL appelé CartoFine.

2.7 Conclusion de partie

La littérature sur la sélection assistée par marqueur est assez abondante, néanmoins les enseignements tirés de ces études manquent globalement de généralité, il reste ainsi difficile d'appréhender les effets que cet outil peut avoir sur des schémas réels de sélection. L'intégration du déséquilibre de liaison aux évaluations permet de lever plusieurs freins à l'efficacité d'un schéma de sélection assistée par marqueurs, il est néanmoins important de retenir que si le déséquilibre améliore la précision de l'évaluation génétique, les questions relatives à l'impact d'un tel outil sur le progrès génétique demeurent.

Chapitre 3

Validations du programme SAM de première génération

3.1 Introduction

Les premiers travaux de détection de QTL en bovins laitiers,(Georges *et al.*, 1995; Ashwell *et al.*, 2001; Boichard *et al.*, 2003; Bennewitz *et al.*, 2004a) ont permis de mettre en évidence un nombre important de QTL (Khatkar *et al.*, 2004) pour la majorité des caractères d'intérêt. Ces connaissances ont permis d'envisager la mise en œuvre d'une évaluation assistée par marqueurs, plusieurs programmes ont ainsi été mis en place, mais seulement deux ont été décrits dans la littérature : l'un en France (Boichard *et al.*, 2002) et l'autre en Allemagne (Bennewitz *et al.*, 2004c).

Le programme français constitue un exemple quasi unique de sélection assistée par marqueurs en conditions réelles et à grande échelle. Les premiers candidats du programme ayant été génotypés au début de l'année 2001, il a été possible au cours de l'année 2006 de mesurer l'efficacité du programme en comparant les estimations de valeurs génétiques intégrant ou non une information moléculaire avant testage sur descendance (en 2001) avec les évaluations génétiques après testage (2006).

Objectifs Les nombreuses études de simulations publiées sur l'efficacité de la sélection assistée par marqueurs avaient des visées plutôt générales et s'attelaient à quantifier les effets de certains facteurs. L'étude de l'efficacité du programme SAM français nécessitait par contre une approche particulière, pour s'adapter à l'ensemble des paramètres réellement utilisés dans l'évaluation (population analysée, nombre de QTL, part de variance expliquée, informativité des marqueurs).

Le premier article présente donc une étude ayant pour but d'établir à l'aide de simulations des références adaptées au programme français et permettant d'interpréter les résultats issus du terrain. Cet article a également été l'occasion de valider l'intérêt des pratiques mises en place dans le programme SAM et de leurs évolutions possibles.

Le second article évalue à partir de premières données réelles les gains observés sur un lot d'animaux ayant été testés sur descendance. L'interprétation de cet article ne prend son sens qu'accompagné des résultats du premier article.

3.2 Article 1

Article paru dans Genetic Selection Evolution
GUILLAUME F., FRITZ S., BOICHARD D., DRUET T. 2008 . Estimation by simulation of the efficiency of the French marker-assisted selection program in dairy cattle. Genet. Sel. Evol. 40 :91-102

Genet. Sel. Evol. 40 (2008) 91–102
© INRA, EDP Sciences, 2008
DOI: 10.1051/gse:2007036

Available online at:
www.gse-journal.org

Original article

Estimation by simulation of the efficiency of the French marker-assisted selection program in dairy cattle
(Open Access publication)

François GUILLAUME[1,2*], Sébastien FRITZ[3], Didier BOICHARD[1], Tom DRUET[1]

[1] INRA, UR337 Station de génétique quantitative et appliquée, 78350 Jouy-en-Josas, France
[2] Institut de l'élevage, 149 rue de Bercy, 75595 Paris Cedex 12, France
[3] Union nationale des coopératives d'élevage et d'insémination animale, 149 rue de Bercy, 75595 Paris Cedex 12, France

(Received 19 January 2007; accepted 3 September 2007)

Abstract – The efficiency of the French marker-assisted selection (MAS) was estimated by a simulation study. The data files of two different time periods were used: April 2004 and 2006. The simulation method used the structure of the existing French MAS: same pedigree, same marker genotypes and same animals with records.The program simulated breeding values and new records based on this existing structure and knowledge on the QTL used in MAS (variance and frequency). Reliabilities of genetic values of young animals (less than one year old) obtained with and without marker information were compared to assess the efficiency of MAS for evaluation of milk, fat and protein yields and fat and protein contents. Mean gains of reliability ranged from 0.015 to 0.094 and from 0.038 to 0.114 in 2004 and 2006, respectively. The larger number of animals genotyped and the use of a new set of genetic markers can explain the improvement of MAS reliability from 2004 to 2006. This improvement was also observed by analysis of information content for young candidates. The gain of MAS reliability with respect to classical selection was larger for sons of sires with genotyped progeny daughters with records. Finally, it was shown that when superiority of MAS over classical selection was estimated with daughter yield deviations obtained after progeny test instead of true breeding values, the gain was underestimated.

marker-assisted selection / simulation / efficiency / dairy cattle

* Corresponding author: francois.guillaume@jouy.inra.fr

1. INTRODUCTION

Marker-assisted selection (MAS) is expected to be particularly valuable for dairy cattle breeding [2,6]. Indeed, several conditions in which MAS improves the efficiency of classical selection are met: most traits of interest are sex-limited, generation interval is long and progeny-test is a long and costly step. Furthermore, MAS can increase the reliability of breeding values [7]. This would be particularly beneficial for bull dams, which are often selected on pedigree information only [2] or for functional traits, with a low heritability, that are gaining emphasis in breeding goals. Therefore, since the end of 2000, a MAS program has been implemented in France. Breeding companies joined this program in order to improve their selection efficiency. However, since MAS programs are recent and relatively rare, little is known about their efficiency. Indeed, the progeny testing step is relatively long and a comparison of breeding values predicted by MAS before and after progeny testing can be done only more than four years after first MAS predictions. In addition, the number of progeny tested bulls remains limited to estimate MAS efficiency and to draw conclusions. Finally, the true breeding values are unknown and this adds some sampling error. Simulation studies offer the possibility to increase the number of animals and to repeat the analysis, to know the true breeding values and to have direct answers. Different simulation studies [6, 8, 12] have already proven the efficiency of MAS for predicting breeding values. However, simulation studies are often based on simple hypotheses. Thanks to the information accumulated in the French MAS program since 2000, it is now possible to make more realistic assumptions regarding the population structure, the marker informativity, the number of genotyped animals, the number of animals with records and the precision of these records, *etc.* Variances of the QTL used are also better known because they have been estimated recently on a large data sample [3]. The objective of this study was to estimate by simulation the efficiency of the French MAS evaluation for two different time periods.

2. MATERIAL AND METHODS

2.1. French MAS data

Data sets used for French MAS evaluation of April 2004 and 2006 were used in this study. Two different time periods were studied to observe the evolution of the efficiency of MAS. Indeed, the efficiency of MAS should be improved in 2006 because more families were genotyped, dams of young animals were more often genotyped and some new microsatellite markers were used.

Table I. Population structure of the French MAS program in April 2004 and 2006.

	April 2004	April 2006
Animals in pedigree	34 318	55 336
Animals with records	23 137	38 859
Genotyped animals	16 629	27 551
Male candidates	1180	1689
Sires[1]	72	79
Dams[1]	793	1130
Genotyped dams[1]	486	887
Maternal grand-sires[1]	70	114
Sires[1] with more than 20 genotyped progeny daughters	11	12
Progeny tested sire families with 30 sons or more	47	64

[1] Of male candidates.

Three files were used at each evaluation: the pedigree file, the markers file containing the probabilities of transmission for each QTL and the data file.

The pedigree used in the French MAS includes different types of animals. First, candidates are young males or females aged from 1 month to 1 year of age. These animals can be chosen to be parents in the next generation. Males can be selected for progeny testing while females can be used as bull dams. The purpose of MAS is to improve the prediction of breeding values of these candidates, which are therefore genotyped. It is also advised to genotype dams of candidates in order to follow QTL transmission as accurately as possible. Families of progeny tested bulls or groups of progeny daughters were genotyped in order to estimate QTL effects of old bulls or younger bulls (sire of candidates), respectively. Thanks to the genotyped animals, the genotypes of some other animals (*e.g.* sires) were reconstructed. In addition, the pedigree file contained parents over two generations of all these animals. Table I indicates the number of candidates (with their sires and dams), genotyped dams, number of genotyped progeny tested bulls or progeny daughter families.

Animals were genotyped for 43 and 45 microsatellite markers before and after first of January 2005, respectively. These markers are used to follow the transmission of 14 QTL regions [1]. Seven of these QTL affecting milk production or composition traits were used in this study. Two to five microsatellite markers are available for each QTL. These were used to estimate probability of identity-by-descent (pid) matrices using a method similar to that of Wang *et al.* [15] extended to the use of multiple markers as in Pong-Wong *et al.* [10].

Finally, phenotypic records were twice the daughter yield deviations (DYD) for males and yield deviations (YD) for females computed for milk, fat and protein yields and fat and protein percentages, pooled from the first three lactations jointly as in VanRaden and Wiggans [14]. These records were obtained from the official genetic evaluation of April 2004 [11]. Respective weights were estimated as in VanRaden and Wiggans [14] with a correction for the number of cows in each herd. DYD of sires were obtained by using only records of daughters not included in the pedigree file. These phenotypic records were replaced by simulation.

2.2. Simulation

The pedigree file and the file containing pid were exactly the same as in the real MAS program. The structure of the performance file was also kept: the same animals had records and the weights of the records were conserved. Only the records were simulated with the following method. The genetic effect of animal i is computed as

$$g_i = u_i + \sum_{j=1}^{n\_qtl} (v_{ij1} + v_{ij2})$$

where u_i is the polygenic effect of individual i (excluding QTL effects), v_{ij1} and v_{ij2} are allelic effects at QTL j for the paternal and maternal alleles, respectively, and n_qtl is the number of QTL.

For animals without parents, the polygenic effect was sampled from N(0, σ^2_u) while for animals with parents, the polygenic effect was equal to the sum of the mean polygenic effects of the parents and the Mendelian sampling drawn from a normal distribution with the variance adjusted for number of known parents. The polygenic variance (σ^2_u) was defined according to the heritability of the traits and the proportion of genetic variance explained by QTL (Tab. II).

For each QTL j, a biallelic gene with substitution effect α_j was simulated. The estimated percentage of heterozygous sires in the population was used to approximate the allelic frequency in the population. The substitution effect was derived from the simulated QTL variance and the allelic frequencies. The variances used for each QTL for each trait are presented in Table II. These were obtained from Druet *et al.* [3] and from our knowledge of these QTL. For all founder animals, QTL alleles were sampled thanks to the allelic frequencies. Then, the alleles were transmitted to the entire population using the estimated pid. By definition, the pid gives the probability for an offspring to receive the

Table II. Proportion of genetic variance used to simulate QTL effects for dairy traits and polygenic effect (in %).

	Number of the chromosome on which the QTL is located							Polygenic effect	Heritability of the traits
Trait	3	6	7	14	19	20	26		
Milk yield	0	0	5	15	0	10	10	60	0.30
Fat yield	0	5	5	15	5	0	20	50	0.30
Protein yield	0	5	5	10	5	0	10	65	0.30
Fat content	0	5	5	40	0	10	0	40	0.50
Protein content	10	15	0	10	0	15	0	50	0.50

paternal or the maternal allele from its parent. Therefore, these probabilities were used to simulate which QTL allele an offspring had received from its parent. For instance, if the pid was equal to 0.5, the progeny had equal chances to receive the paternal or the maternal allele of its parent while if the paternal pid was equal to 1 then the progeny received the paternal allele of the corresponding parent.

To simulate records, a residual value was sampled from N(0, σ_e^2) where the residual variance is adjusted by the weight from actual phenotypes in the MAS data set. The simulated records were the sum of the genetic and residual values. Additionally, for male candidates, records were simulated with a weight corresponding to the first EBV obtained after progeny testing.

Simulations were repeated 100 times for each trait and both time periods.

2.3. MAS evaluation

The model used in this study was a single trait and multi-QTL model as proposed by Fernando and Grossman [4]:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \sum_{i=1}^{n\_qtl} \mathbf{Z}_{v_i}\mathbf{v}_i + \mathbf{e} \quad (1)$$

where $\mathbf{y}$ is a vector containing records, $\boldsymbol{\beta}$ is a vector of fixed effects (the mean), $\mathbf{u}$ is a vector of random polygenic effects, $\mathbf{v}_i$ is a vector of random gametic effects for QTL i and $\mathbf{e}$ is a vector of random residual terms. $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{Z}_{vi}$ are known design matrices that relate records to fixed, random polygenic and gametic effects, respectively.

Four to five QTL were used for each production trait and the variance components (see Tab. II) were assessed based on a previous study [3].

Table III. Mean information content measured as |1–2p| weighted by QTL variance for each trait for 2004 and 2006 candidates and their parents.

Traits	Mean information content weighted by QTL variance					
	Candidates		Sires		Dams	
	2004	2006	2004	2006	2004	2006
Milk	0.61	0.72	0.36	0.47	0.44	0.54
Fat yield	0.58	0.72	0.33	0.48	0.41	0.54
Protein yield	0.60	0.71	0.34	0.48	0.43	0.55
Fat content	0.70	0.76	0.40	0.49	0.50	0.56
Protein content	0.72	0.75	0.41	0.53	0.52	0.60

3. RESULTS

3.1. Simulated data

The results were obtained for two different sets of candidates (Tab. I). They included males born during the previous AI season, *i.e.* from October to September. The first set was constituted of candidates of year 2004 whereas the second set of candidates of year 2006. Informativity was estimated as $|1 - 2p|$ where p was the probability transmission of a given paternal or maternal QTL allele [2]. When the transmitted allele is known, p is equal to 0 or 1 and $\left|1 - 2p\right|$ is one while when there is no information on which allele was transmitted, p is equal to 0.5 and $\left|1 - 2p\right|$ is zero. So this information content indicates how well the QTL transmission is followed in the population. For each trait, mean information content was computed by weighting the information content of each QTL by the proportion of genetic variance explained by this QTL. This weighted mean information content is presented in Table III for candidates of years 2004 and 2006 and for their sires and dams. For all the traits, information content increased in 2006 with respect to 2004: for candidates, mean information content gains ranged from +0.03 up to +0.14 while for sires they ranged from +0.09 up to +0.15. The gains were comprised between +0.06 and +0.13 for dams.

3.2. Estimation model

Marker-assisted selection was compared to classical selection (model with only a polygenic effect). Accuracies of breeding values (squared correlation R^2 between estimated and true genetic effects) were estimated and are presented in Table IV. For all traits, MAS EBV were more reliable than classical EBV.

Table IV. Reliabilities (R^2) of classical polygenic EBV (POL) and MAS EBV (MAS) for male candidates from 2004 and 2006.

Trait	April 2004			April 2006		
	POL	MAS	Difference	POL	MAS	Difference
Milk yield	0.294	0.327	+0.033	0.313	0.361	+0.048
Fat yield	0.281	0.296	+0.015	0.310	0.373	+0.063
Protein yield	0.254	0.273	+0.019	0.303	0.341	+0.038
Fat content	0.313	0.407	+0.094	0.342	0.453	+0.111
Protein content	0.214	0.301	+0.087	0.342	0.418	+0.076

Table V. Reliabilities of classical polygenic EBV (POL) or marker-assisted EBV (MAS) of candidates of 2004, depending on the status of their sires.

Trait	Sires of candidates without genotyped progeny daughters			Sires of candidates with genotyped progeny daughters		
	POL	MAS	Difference	POL	MAS	Difference
Milk yield	0.266	0.302	+0.036	0.291	0.353	+0.062
Fat yield	0.255	0.263	+0.008	0.277	0.312	+0.035
Protein yield	0.243	0.265	+0.022	0.267	0.307	+0.040
Fat content	0.269	0.384	+0.115	0.304	0.476	+0.172
Protein content	0.200	0.301	+0.101	0.210	0.372	+0.162

In 2004, the gain of reliability ranged from 0.015 for fat yield up to 0.094 for fat content. Gain was relatively limited for yield traits (0.033, 0.015 and 0.019 for milk, fat, and protein yields, respectively) and larger for content traits (0.094 and 0.087 for fat and protein contents, respectively). In 2006, the difference between MAS EBV and classical EBV was larger, especially for yield traits (0.048, 0.063 and 0.038 for milk, fat and protein yields, respectively). Among all 100 replications for 2004, MAS was less efficient than classical selection for eleven and nine replications for fat and protein yields, respectively. In 2006, MAS resulted in lower reliabilities for a single replication for milk yield. For these few negative results, the difference between evaluation methods was close to zero.

In 2004, MAS and classical EBV were also compared with respect to the amount of information available to estimate gametic effects of the sires (Tab. V). Two classes of sires were defined: sires with or without genotyped progeny daughters (at least 20). The improvement of accuracy due to MAS is larger for all traits when a group of progeny daughters is also genotyped. The difference between MAS selection and classical selection when sires of candidates have no genotyped progeny represent only 59, 23, 55, 67 and 63%

Table VI. Correlations between classical polygenic EBV (POL) or marker-assisted EBV (MAS) and simulated DYD of candidates of 2004.

Trait	POL	SAM	Difference			
			Mean	St. Dev.	Minimum	Maximum
Milk yield	0.476	0.502	+0.026	0.013	–0.002	0.074
Fat yield	0.466	0.477	+0.011	0.018	–0.041	0.051
Protein yield	0.441	0.457	+0.016	0.014	–0.021	0.056
Fat content	0.526	0.599	+0.073	0.022	0.022	0.135
Protein content	0.522	0.572	+0.052	0.020	0.013	0.097

of the difference obtained when sires have genotyped progeny for milk, fat and protein yields and fat and protein contents, respectively.

Finally, the comparisons between MAS and classical EBV with simulated DYD (with an accuracy corresponding to first EBV after progeny testing) are shown in Table VI. As expected, MAS EBV are better predictors but the difference between MAS and classical selection varies across replications. The mean correlation gain is equal to 0.026, 0.011, 0.016, 0.073 and 0.052 for milk, fat and protein yield and fat and protein content, respectively. These gains are lower than when comparison is done with true genetic values (comparison on an accuracy scale). The minimum and maximal gains ranged from –0.002 to 0.074, –0.041 to 0.051, –0.021 to 0.056, 0.022 to 0.135 and from 0.013 to 0.097 for milk, fat and protein yields and fat and protein contents, respectively. For some samples, MAS appeared to perform worse than the classical model for fat or protein yield.

4. DISCUSSION

Files involved in the French MAS are increasing on a regular basis as a consequence of continuous addition of new genotyped animals (see Tab. I). Therefore, the MAS evaluation is more demanding in computational terms but the information on QTL is increasing with time. More families are genotyped and QTL transmission is better observed. Both these information improve the estimation of QTL effects and therefore the efficiency of MAS. The increment of genotyped animals is not only due to the continuous application of the MAS program but also to strategic choices decided to improve the French MAS program. For instance, breeding companies genotype dams of candidates more frequently than at the start of the MAS program. At the beginning, neither the dams of sire nor the progeny daughter families were genotyped. During the MAS program, breeding companies were advised to genotype these animals.

The impact of all these decisions is visible in Table III where increasing information can be noted. Some technical changes were also implemented to improve the efficiency of MAS. Some microsatellite markers are no longer used while some more informative markers were integrated in the program. All these elements improved the efficiency of MAS to follow QTL transmission in the population (see Tab. III). The changes in precision of the pid between 2004 and 2006 are important and are consequences of efforts made by breeding companies. Efficiency of MAS can still be improved by the use of denser markers. For instance, if informativity is increased by replacing the microsatellite markers by ten SNP close to the QTL (within 1 cM), the gain of reliability of MAS with respect to classical selection is increased from 43% up to 79% (data not shown). As shown, the gain of efficiency achieved by improving the accuracy of the pid is important but to obtain even larger gains, other MAS strategies must be applied (such as the use of linkage disequilibrium).

Some previous studies showed the advantage of MAS in predicting breeding values [12, 13]. The present study focused on accuracy gain rather than genetic progress gain achievable by MAS; in fact the latter criterion is greatly dependent on the selection strategy whereas accuracy of prediction reflects the methodology efficiency more. In the present study, many conditions were those really applied in the French breeding schemes (pedigree, markers, genotyped animals, *etc.*). Under these conditions, MAS improved the reliability of breeding values but the gain remained limited.

Accuracy improvement appeared larger for content traits than for yield traits. This can be explained by several facts. For content traits, QTL explained in general a larger part of the genetic variation. Part of genetic variance explained by the QTL has a major impact on the efficiency of MAS. Indeed, the gain of reliability achieved by MAS ranked similarly to the part of variance explained by the QTL. However, other parameters influence the efficiency of MAS. For instance, QTL variance is equal for fat yield and protein content but MAS performed better with protein content. Mean information content was higher for content traits. The influence of mean information content can also be seen when comparing the results for yield traits in 2004 and 2006. Efficiency of MAS improved clearly at constant QTL variance thanks to better mean information content. In addition, MAS is more beneficial, at constant part of genetic variance explained by the QTL, when there are fewer QTL (but with larger effects). Indeed, the polygenic model is more appropriate for a situation with many QTL (closer to the infinitesimal model) than with a few QTL. Therefore, the superiority of MAS will be reduced with many small QTL. Finally, QTL effects are estimated more accurately when QTL have larger effects

and when there is less environmental noise. However, for low heritability traits, gains of reliability of MAS are expected to be larger because there is much room for improvement since classical selection performs poorly. In the present study, efficiency of MAS was studied only for heritabilities above 0.30 and no conclusions can be drawn for low heritability traits.

The number of QTL and proportion of total genetic variance explained by them are greater than parameters usually assumed by previous simulation studies [8, 12]. This should enhance MAS efficiency, by reducing the risk that parents are homozygous at all the QTL.

On the contrary to various simulation studies [8, 13], population structure is fairly unbalanced. As shown in Table I, a few sires and maternal grandsires contribute heavily to the population. It is essential to evaluate their gametic effects as accurately as possible. Therefore, it is very important to genotype many animals such as dams and progeny daughters' families. Indeed, the results showed that when sires of candidates have genotyped progeny daughters, MAS was more efficient. This approach has some similarities with the Bottom-up scheme proposed by Mackinnon and Georges [7] and which was shown to increase MAS efficiency. Sires of candidates with genotyped progeny daughters were just a few (11 out of 72 in 2004 and 12 out of 79 in 2006) but generally contributed to a large proportion of candidates (20% in 2004 and 25% in 2006). Since the start of the French MAS, efforts have been made to increase the information available. In 2006, gains of accuracy obtained with MAS were better than in 2004. All the accumulated information improves the French MAS programs.

The study also showed that if the efficiency of MAS is assessed with field data, on DYD for instance, the estimated gain is reduced. Indeed, MAS EBV are better predictors of true genetic values. DYD still contain some errors and MAS EBV do not predict these error terms well.

Although many parameters were estimated on real data, the simulation performed in this study might depart from the underlying biological reality. Therefore, the results presented might over- or under-estimate MAS efficiency. Variance of the QTL was estimated on a large sample independent from the sample used for QTL detection. Still, the variances used might be incorrect. Therefore, the efficiency of MAS was also tested by using under- or over-estimated (by 25%) QTL variances and the differences were marginal: MAS was achieving the same gains. Allelic frequencies or effects might be wrong or the QTL could be multi-allelic. The evaluation model should be robust to these changes and the accuracy of the estimation of QTL effects should not vary much. For instance, the evaluation model does not assume a fixed number of alleles but

rather an infinite number of alleles and could easily handle a multi-allelic QTL. Although reliabilities obtained by MAS might be only slightly affected by different allelic frequencies or multi-allelic QTL, the polygenic model might be more sensitive to the changes and therefore the difference between MAS and the polygenic model might be over- or under-estimated. However, it is difficult to predict if the polygenic model would be penalised under different hypotheses. When more parents are heterozygous (due to multi-allelic QTL or changes in frequencies) and transmission of genetic values departs from the rules used with the polygenic model (half of the breeding value is transmitted), the polygenic model should achieve lower reliabilities.

This is also true when QTL effects are larger because they have a larger influence in the ranking of the animals. On the contrary, the polygenic model performs better with many QTL because this situation is closer to the infinitesimal model.

In this study, QTL were assumed additive but if some QTL had non-additive effects (dominance, epistasis), the impact on the results would be larger since the model would be less robust to it.

Finally, MAS will certainly evolve in the future towards more efficient models using denser marker maps (*e.g.*, Meuwissen *et al.*, [9]) and exploiting linkage disequilibrium. For instance, Hayes *et al.* [5] presented advantages of LD-MAS. With dense maps, small haplotypes around the QTL will be in linkage disequilibrium with the QTL. Therefore, gametic effects will be estimated across families and no longer within each sire family. As a consequence, the effects will be estimated more accurately and less genotyped animals will be required to estimate these effects. In addition, follow-up of transmission of gametic effects will be more precise because information content will improve.

5. CONCLUSIONS

In the French MAS program, accuracy of breeding values of young candidates was shown to be improved thanks to the use of molecular information. The obtained gains of accuracy (in comparison with classical selection) were relatively limited and strongly dependent on the accumulated information in the program. By genotyping more animals (such as dams or progeny daughters of sires of candidates) or using better markers, the efficiency of this program was clearly improved.

Thanks to the development of new genotyping technologies, still improved results are expected with the use of denser marker maps and of linkage disequilibrium.

REFERENCES

[1] Boichard D., Fritz S., Rossignol M.N., Boscher M.Y., Malafosse A., Colleau J.J., Implementation of marker-assisted selection in French dairy cattle, in: Proceedings of the 7th World Congress on Genetics Applied to Livestock Production, 18–23 August 2002, Montpellier, Communication no. 22-03.

[2] Boichard D., Grohs C., Bourgeois F., Cerqueira F., Faugeras R., Neau A., Rupp R., Amigues Y., Boscher M.Y., Levéziel H., Detection of genes influencing economic traits in three French dairy cattle breeds, Genet. Sel. Evol. 35 (2003) 77–101.

[3] Druet T., Fritz S., Boichard D., Colleau J.J., Estimation of genetic parameters for quantitative trait loci for dairy traits in the French Holstein population, J. Dairy Sci. 89 (2006) 4070–4076.

[4] Fernando R.L., Grossman M., Marker assisted selection using best linear unbiased prediction, Genet. Sel. Evol. 21 (1989) 467–477.

[5] Hayes B.J., Chamberlain A.J., Goddard M.E., Use of markers in linkage disequilibrium with QTL in breeding programs, in: Proceedings of the 8th World Congress on Genetics Applied to Livestock Production, 13–18 August 2006, Belo Horizonte, Brazil, Communication no. 30-06.

[6] Kashi Y., Hallerman E., Soller M., Marker-assisted selection of candidate bulls for progeny testing programmes, Anim. Prod. 51 (1990) 63–74.

[7] Mackinnon M.J., Georges M.A.J., Marker-assisted preselection of young dairy sires prior to progeny testing, Livest. Prod. Sci. 54 (1998) 229–250.

[8] Meuwissen T.H.E., Goddard M.E., The use of marker haplotypes in animal breeding schemes, Genet. Sel. Evol. 28 (1996) 161–176.

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

[10] Pong-Wong R., George A.W., Woolliams J.A., Haley C.S., A simple and rapid method for calculating identity-by-descent matrices using multiple markers, Genet. Sel. Evol. 33 (2001) 453–471.

[11] Robert-Granié C., Bonaïti B., Boichard D., Barbat A., Accounting for variance heterogeneity in French dairy cattle genetic evaluation, Livest. Prod. Sci. 60 (1999) 343–357.

[12] Schrooten C., Bovenhuis H., van Arendonk J.A.M., Bijma P., Genetic progress in multiple stage dairy cattle breeding schemes using genetic markers, J. Dairy Sci. 88 (2005) 1569–1581.

[13] Spelman R., Bovenhuis H., Genetic response from marker-assisted selection in an outbred population for differing marker bracket sizes and with two identified quantitative trait loci, Genetics 148 (1998) 1389–1396.

[14] VanRaden P.M., Wiggans G.R., Derivation, calculation, and use of national Animal Model Information, J. Dairy Sci. 74 (1991) 2737–2746.

[15] Wang T., Fernando R.L., van der Beek S., Grossman M., van Arendonk J.A.M., Covariance between relatives for a marked quantitative trait locus, Genet. Sel. Evol. 27 (1995) 251–274.

3.3 Article 2

Article paru dans Journal of Dairy Science
GUILLAUME F., FRITZ S., BOICHARD D., DRUET T. 2008 .Correlations of Marker-Assisted Breeding Values with Progeny Test Breeding Values for 899 French Holstein Bulls. J. Dairy Sci., 91 :2520-2522

J. Dairy Sci. 91:2520–2522
doi:10.3168/jds.2007-0829

Short Communication: Correlations of Marker-Assisted Breeding Values with Progeny-Test Breeding Values for Eight Hundred Ninety-Nine French Holstein Bulls

F. Guillaume,*†[1] S. Fritz,‡ D. Boichard,* and T. Druet*
*INRA, UR337 Station de Génétique Quantitative et Appliquée, F-78350 Jouy en Josas, France
†Institut de l'élevage, 149 rue de Bercy, 75595 Paris Cedex 12, France
‡Union nationale des coopératives d'élevage et d'insémination animale, 149 rue de Bercy, 75595 Paris Cedex 12, France

ABSTRACT

French artificial insemination companies have been running a marker-assisted selection program since 2001 to determine which young bulls should be progeny tested. A first batch of 899 Holstein sires receiving their first proofs based on progeny daughters has been studied. Estimated breeding values with or without marker information were computed based on information available in April 2004, and correlated to daughter yield deviations available in 2007 for production traits. Marker-assisted estimated breeding values presented greater correlations with daughter yield deviations than those calculated using only pedigree index. The average improvement in correlation was 0.043 and ranged from +0.001 for protein yield to +0.103 for fat percentage. This gain was based on the initial and suboptimal conditions of the program and is expected to increase in the coming years because of several improvements implemented since the start of the marker-assisted selection program.
Key words: dairy cattle, marker-assisted selection

Over the last decade, several QTL detection programs have been conducted in dairy cattle (e.g., Georges et al., 1995; Boichard et al., 2003). Such studies revealed the existence of several QTL with large effect on dairy traits such as the gene encoding acylCoA:diacyglycerol acyltransferase (*DGAT1*) on chromosome 14 (Grisart et al., 2002), the ATP-binding cassette, subfamily G, member 2 gene (*ABCG2*) on chromosome 6 (Cohen-Zinder et al., 2005), or the growth hormone receptor (*GHR*) on chromosome 20 (Blott et al., 2003). Use of these QTL in a marker-assisted selection (**MAS**) program has the potential to improve selection efficiency in dairy cattle (Kashi et al., 1990). Since 2001, such a MAS program has been implemented in France in the Holstein, Normande, and Montbéliarde breeds (Boichard et al., 2002). One of the objectives of this program is to help breeding companies select which young bulls should be progeny tested. The accurate genetic value of selected animals can only be calculated after progeny testing (approximately 5 yr after the selection decision has been made). Consequently, efficiency of the MAS program is difficult to prove on real data and can be estimated only after a few years of implementation. In January 2007, the accuracy of production traits' EBV of young bulls born in 2001 was high because of the records of their progeny daughters. These bulls were also genotyped for MAS and can therefore be used to assess the precision of MAS breeding values. The objective of this study was to check, using real data, if breeding values estimated for young animals using MAS were more precise than breeding values obtained at the same age based only on a polygenic model.

Files of the evaluation of April 2004 (the oldest conserved MAS data set) were used in this study. This evaluation used a pedigree of 34,318 animals of which 23,137 had phenotypic records and 16,629 were genotyped. Animals were genotyped for 43 microsatellite markers, which were used to follow the transmission of 14 QTL regions that relied on 2 to 5 evenly spaced microsatellite markers. These QTL were selected on basis of the results of a QTL detection program (Boichard et al., 2003) and confirmed with new sire families (Druet et al., 2006).

Phenotypic records were twice the daughter yield deviations (**DYD**) for males and yield deviations for females computed for milk, fat, and protein yields and fat and protein percentages, pooled from the first 3 lactations jointly as in VanRaden and Wiggans (1991). These records were obtained from the official genetic evaluations of both April 2004 and January 2007 (Robert-Granié et al., 1999). Respective weights were estimated as in VanRaden and Wiggans (1991) with a correction for number of cows in each herd.

The model used in this study was a single-trait and multiple-QTL model as proposed by Fernando and Grossman (1989):

Received November 5, 2007.
Accepted February 8, 2008.
[1]Corresponding author: Francois.Guillaume@jouy.inra.fr

Table 1. Proportions of genetic variance used in the evaluation model for the QTL and polygenic effects for dairy traits and average informativity per QTL and traits (weighted by proportion of genetic variance explained) in the candidate population

Trait	Heritability	Informativity[1]	Chromosome							Polygenic effect, %
			3	6	7	14	19	20	26	
Milk yield	0.30	0.57	0	0	5	15	0	10	10	60
Fat yield	0.30	0.55	0	5	5	15	5	0	20	50
Protein yield	0.30	0.57	0	5	5	10	5	0	10	65
Fat percentage	0.50	0.66	0	5	5	40	0	10	0	40
Protein percentage	0.50	0.67	10	15	0	10	0	15	0	50
Informativity/chromosome[1]			0.71	0.75	0.60	0.68	0.44	0.57	0.42	
Distance between closest marker and QTL location, cM			<1	<1	<1	<1	3	5	5	

[1]Informativity was computed as the average of $|1–2p_{ij}|$, where p_{ij} is the probability for individual i of receiving the paternal QTL of its parent j.

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Zu} + \sum_{i=1}^{n\_qtl} \mathbf{Z}_{v_i}\mathbf{v}_i + \mathbf{e} \quad [1]$$

where $\mathbf{y}$ is a vector containing records, β is a vector of fixed effects (the mean), $\mathbf{u}$ is a vector of random polygenic effects, $\mathbf{v}_i$ is a vector of random gametic effects for QTL i, n_qtl is the number of QTL considered for the trait, and $\mathbf{e}$ is a vector of random residual terms. $\mathbf{X}$, $\mathbf{Z}$, and $\mathbf{Z}_{v_i}$ are known design matrices relating results to fixed, random polygenic, and gametic effects, respectively. Probability of identity-by-descent matrices ($\mathbf{Z}_{v_i}$) were obtained using a method similar to that of Pong-Wong et al. (2001). Between 4 and 5 QTL were used for each production trait, and the variance components (see Table 1) were assessed in a previous study (Druet et al., 2006).

A subset of 899 Holstein progeny-tested sires was used in the validation analysis. Among these genotyped candidates, a range of situations exist regarding knowledge of their ancestors' genotypes. Approximately 3 to 4 generations of sires were genotyped on the paternal path of candidates (as far as 5 generations from the candidate, more than 50% of the male ancestors are genotyped on the paternal side), whereas 0 to 2 generations of parents were genotyped on the maternal side (less than 50% of the female ancestors are genotyped after 2 generations from candidates). These candidates had no daughters with records in April 2004, whereas they had, on average, 76 daughters (SD of 16) with records in 2007. The weighted correlations between their DYD in January 2007 and both polygenic (i.e., pedigree index) and MAS EBV in 2004 (Table 2) were compared. These correlations were obtained for milk, fat, and protein yields and fat and protein percentages.

Correlations between DYD and pedigree index ranged from 0.317 for protein yield to 0.556 for fat content, whereas correlations between DYD and MAS EBV ranged from 0.318 for protein yield to 0.659 for fat content. The gains in correlation obtained with MAS were, on average, around 0.04. They were lower (0.001) for protein yield and greater (0.103) for fat content.

A simulation study performed on the French MAS program (Guillaume et al., 2008) estimated the expected gains of correlation from MAS EBV over pedigree index and also their variation. The present results are in agreement with the results obtained by simulation: correlations are slightly lower for milk yield and protein yield and content, and greater for fat yield and percentage. The simulation study showed that these gains are expected to vary for each crop of bulls and that use of DYD instead of the unknown true genetic values underestimates the gain in accuracy achieved by MAS. Finally, Guillaume et al. (2008) indicated that the gain in accuracy achieved by MAS is clearly larger for young bulls born in 2006 than for animals born during the first years of the program. Indeed, the program is accumulating more information; approximately 10,000 new genotyped animals are integrated into the database each year. These additional genotypes improve the efficiency of the program.

To further increase the efficiency of the program, breeding companies have decided to genotype dams of young bulls and some progeny daughters of sires of

Table 2. Correlations between daughter yield deviations (DYD) of January 2007 and classical or marker-assisted selection (MAS) EBV based on April 2004 evaluation of 899 progeny-tested Holstein bulls for production traits weighted by DYD reliability in 2007

Item	Milk	Fat yield	Protein yield	Fat percentage	Protein percentage
EBV	0.405	0.361	0.317	0.556	0.505
MAS EBV	0.436	0.400	0.318	0.659	0.547
Difference	0.030	0.039	0.001	0.103	0.042

young bulls. This targeted genotyping has been shown to enhance the efficiency of MAS. In the batch of 899 candidates studied, only 66% of the candidates' dams were genotyped and first-crop daughter genotypes of only 6 of 39 sires were available, whereas these numbers are now much larger. Finally, the set of microsatellite markers was also changed in 2005, and the QTL are now followed with more accuracy.

In this study, the greatest increases in terms of correlation were observed for milk content traits. This can be explained by the fact that a larger proportion of the genetic variance was explained by the QTL for these traits. The lowest gain was obtained for protein yield, in which QTL explained only 35% of the genetic variation. Gains in correlation are ordered in the same way as proportion of genetic variance explained by QTL, so that the proportion of genetic variance explained by QTL should be large enough to obtain improvement of accuracy. Furthermore, content traits were influenced by QTL for which average informativity weighted by proportion of variance explained by each QTL was greater (Table 1). This is partly due to the low informativity achieved with the markers for QTL on *Bos taurus* autosomes 19 and 26, which influenced yield traits. To resolve this problem, new microsatellite markers were selected in 2005.

In the present data set, EBV were available only for progeny-tested bulls. Candidates not selected for progeny testing were not included in the study. These animals correspond to those candidates that had poor MAS EBV. A study including all candidates would better assess the efficiency of MAS. Unfortunately, precise genetic values based on progeny daughters are not available for unselected candidates.

In the coming years, efficiency of MAS is expected to improve because our knowledge of the genome is increasing. Methods using genetic markers in animal selection will certainly change. Indeed, future methods will use markers closer to the QTL or use the mutations directly responsible for QTL variation. Linkage disequilibrium between markers and QTL can be used in these conditions. Finally, because of the ability to genotype animals at many more markers at a reasonable cost, genomic selection (Visscher and Haley, 1998; Meuwissen et al., 2001) will be possible.

ACKNOWLEDGMENTS

The French Ministry of Agriculture is acknowledged for financial support through a CASDAR grant, and INRA and UNCEIA are acknowledged for providing access to their data.

REFERENCES

Blott, S., J. J. Kim, S. Moisio, A. Schmidt-Küntzel, A. Cornet, P. Berzi, N. Cambiaso, C. Ford, B. Grisart, D. Johnson, L. Karim, P. Simon, R. Snell, R. Spelman, J. Wong, J. Vilkki, M. Georges, F. Farnir, and W. Coppieters. 2003. Molecular dissection of a quantitative trait locus: A phenylalanine-to-tyrosine substitution in the transmembrane domain of the bovine growth hormone receptor is associated with a major effect on milk yield and composition. Genetics 163:253–266.

Boichard, D., S. Fritz, M. N. Rossignol, M. Y. Boscher, A. Malafosse, and J. J. Colleau. 2002. Implementation of marker-assisted selection in French dairy cattle. Commun. no. 22–03 in Proc. 7th World Congr. Genet. Appl. Livest. Prod., Montpellier, France. J. M. Elsen and V. Ducrocq, ed.

Boichard, D., C. Grohs, F. Bourgeois, F. Cerqueira, R. Faugeras, A. Neau, R. Rupp, Y. Amigues, M. Y. Boscher, and H. Levéziel. 2003. Detection of genes influencing economic traits in three French dairy cattle breeds. Genet. Sel. Evol. 35:77–101.

Cohen-Zinder, M., E. Seroussi, D. M. Larkin, J. J. Loor, A. Evertsvan der Wind, J. H. Lee, J. K. Drackley, M. R. Band, A. G. Hernandez, M. Shani, H. A. Lewin, J. I. Weller, and M. Ron. 2005. Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. Genome Res. 15:936–944.

Druet, T., S. Fritz, D. Boichard, and J. J. Colleau. 2006. Estimation of genetic parameters for quantitative trait loci for dairy traits in the French Holstein population. J. Dairy Sci. 89:4070–4076.

Fernando, R. L., and M. Grossman. 1989. Marker assisted selection using best linear unbiased prediction. Genet. Sel. Evol. 21:467–477.

Georges, M., D. Nielsen, M. Mackinnon, A. Mishra, R. Okimoto, A. T. Pasquino, S. Sargeant, A. Sorensen, M. R. Steele, X. Zhao, J. E. Womack, and I. Hoeschele. 1995. Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. Genetics 139:907–920.

Grisart, B., W. Coppieters, F. Farnir, L. Karim, C. Ford, P. Berzi, N. Cambisano, M. Mni, S. Reid, P. Simon, R. Spelman, M. Georges, and R. Snell. 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., S. Fritz, D. Boichard, and T. Druet. 2008. Estimation by simulation of the efficiency of the French marker assisted selection program in dairy cattle. Genet. Sel. Evol. 40:91–102.

Kashi, Y., E. Hallerman, and M. Soller. 1990. Marker-assisted selection of candidate bulls for progeny testing programmes. Anim. Prod. 51:63–74.

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

Pong-Wong, R., A. W. George, J. A. Woolliams, and C. S. Haley. 2001. A simple and rapid method for calculating identity-by-descent matrices using multiple markers. Genet. Sel. Evol. 33:453–471.

Robert-Granié, C., B. Bonaïti, D. Boichard, and A. Barbat. 1999. Accounting for variance heterogeneity in French dairy cattle genetic evaluation. Livest. Prod. Sci. 60:343–357.

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

Visscher, P. M., and C. S. Haley. 1998. Strategies for marker assisted selection in pig breeding programmes. Commun. no. 23–503–510 in Proc. 6th World Congr. Genet. Appl. Livest. Prod., Armidale, Australia. L. Piper, ed.

3.4 Conclusion

Ces deux articles ont permis de confirmer que pour des candidats au testage, les index SAM étaient en moyenne de meilleurs prédicteurs des valeurs génétiques estimées après testage sur descendance que les index polygéniques classiques. Les gains observés restent modérés mais cohérents avec ce qui pouvait être attendu.

Au delà de ce résultat, il a été possible de bien montrer les contraintes que représentait la prise en compte des réalités pratiques de tout programme de sélection assistée par marqueurs. Ainsi, il est très intéressant de bien voir que la collecte continue de génotypes bénéficie à la qualité des évaluations et que par voie de conséquence, la première validation d'un programme tel que le programme de SAM français reflète plus un gain minimal que le gain à long terme ou même en routine du système dans son ensemble.

Les gains moindres observés sur le caractère le plus sélectionné (la quantité de matière protéique), peuvent susciter quelques interrogations. La quantité de matière protéique est un caractère pour lequel il ne semble exister aucun QTL d'effet très fort, ce qui ne permet pas de gains importants grâce à la SAM.

Deux points majeurs sont mis en évidence, l'effet de l'informativité des marqueurs sur les gains observés et les bénéfices du typage des filles de testage. Comme indiqué par Fernando et Grossman (1989), dès lors qu'une phase gamétique ne peut être attribuable avec certitude, les probabilités de transmission du modèle 2.2 tendent vers 0.5 et on obtient alors un modèle identique à un modèle polygénique. Le modèle ne peut plus prédire l'aléa de méiose et il redevient équivalent à un modèle polygénique. Dans des populations internationales, telle que la Holstein, la majeure partie des ascendants des candidats ne sont pas génotypés, de fait, l'informativité des marqueurs de ces animaux ne peut être élevée dans le cadre d'une évaluation en équilibre de liaison.

Les deux articles précédents se sont focalisés sur la qualité de l'outil d'évaluation par le biais de sa précision. D'autres approches, plus pratiques, pour valider l'utilité de l'outil sont possibles (Guillaume *et al.*, 2006; Boichard *et al.*, 2006). Par exemple, des réductions d'effectif de taureaux testés allant de 5% à 40% auraient été envisageables dans le cadre d'une sélection mono-caractère, sans que celà n'affecte le progrès génétique. De plus, l'index SAM était capable de distinguer dans plus de 65 % des cas, le meilleur individu d'un couple de pleins frères. Ces résultats bien que moins évidents à appréhender permettent au sein d'un programme de sélection de réaliser des économies substantielles qui permettent d'amortir le surcoût dû au génotypage.

L'accumulation de données durant le temps du programme a permis une amélioration de la connaissance des paramètres des QTL ainsi que la mise en évidence de QTL jusqu'alors non détectés (Guillaume *et al.*, 2007b). Cette évolution en termes de connaissances sur les QTL s'observe également à l'échelle mondiale avec un nombre de QTL détectés bien plus important que ce qu'il était au début du programme SAM (Hu *et al.*, 2007). Une mise à jour des QTL utilisés

s'avérait donc indispensable.

La distance existant entre les marqueurs utilisés dans le programme rend également complexe le bon suivi de la transmission des QTL. Comme souligné dans les deux articles, l'augmentation de l'informativité aux QTL est une voie importante de progrès de la précision de l'évaluation. Ce gain peut être rendu possible par la densification en marqueurs de régions QTL. Cette dernière présuppose néanmoins que la localisation de QTL soit suffisamment fine, ce qui est rendu possible notamment par des études de cartographie fine que nous allons présenter dans la prochaine partie.

Chapitre 4

Travaux de cartographie fine

4.1 Introduction

La détection de QTL était une condition préliminaire à la mise en œuvre du programme de sélection assistée par marqueurs de deuxième génération envisagée dans le cadre de cette thèse. En effet, la faible précision de la localisation des QTL limitait considérablement l'efficacité de la SAM de première génération. D'une part, la sélection était restreinte à la composante intra-famille, d'autre part les pertes de charge étaient importantes du fait du suivi médiocre des QTL entre individus distants de plus de 3 générations. Pour viser un suivi quasi systématique des QTL dans un pedigree complet et pour inférer l'information QTL au niveau populationnel par déséquilibre de liaison, il est impératif de connaître la localisation des QTL dans un intervalle de l'ordre du cM. C'est l'objectif des deux études de cartographie fine présentées dans ce chapitre.

À partir de 2005, l'INRA, comme d'autres organismes, a conduit une approche visant l'utilisation de marqueurs SNP à haute densité. Plus de 30 000 marqueurs ont été développés in silico et un projet pilote a été conduit en collaboration avec le CNG pour l'analyse de 1536 marqueurs principalement sur le chromosome 3. Cette approche permet de mettre en œuvre de nouvelles méthodes d'analyse des résultats et de tirer profit de l'information maternelle du dispositif expérimental par déséquilibre de liaison. Cette approche (Druet *et al.*, 2008) a montré que la précision de cartographie était nettement améliorée, avec un intervalle de confiance de l'ordre de 4 cM pour un QTL de fertilité (donc assez difficile à cartographier) en dépit d'un dispositif de taille relativement réduite.

Le projet CartoFine financé par l'ANR et ApisGene prévoyait donc de génotyper un vaste dispositif de 3200 taureaux avec une puce pangénomique dédiée. En 2007, ce projet a été réorienté en abandonnant la puce dédiée interne. L'INRA a rejoint le consortium nord–américain, a fourni ses marqueurs et a bénéficié d'une puce de haute densité (54 000 SNP) à un tarif attractif. Ce

choix s'est révélé très judicieux car cette puce Illumina, devenue la référence internationale, est utilisée par la majorité des équipes dans le monde. Elle est excellente techniquement en termes de couverture du génome (un marqueur tous les 45 kb, pas d'intervalle inférieur à 22 kb et peu d'intervalles supérieurs à 70 kb) et d'informativité élevée (la fréquence moyenne de l'allèle rare est supérieure à 20 % dans la plupart des races Bos taurus). Après un important travail international et en particulier de l'équipe d'André Eggen, la grande majorité des marqueurs sont localisés et ordonnés sur le génome avec une grande précision. À l'aide de ce dispositif, un vaste programme de cartographie fine est conduit pour analyser 25 caractères pour lesquels les taureaux disposent d'index sur descendance.

Objectifs La première étude (Guillaume *et al.*, 2007b) reflète les efforts réalisés jusqu'en 2007 avec une approche "classique" – Un premier QTL primo-localisé est cartographié plus finement principalement en densifiant la région avec des marqueurs microsatellites supplémentaires. Un effort est également réalisé sur la définition du phénotype, pour qu'il colle davantage à l'effet du QTL, de façon à augmenter la part de variance du QTL dans la variance génétique totale, ce qui améliore automatiquement la précision de localisation (Ytournel *et al.*, 2008). Enfin, la population analysée est accrue (plus de 2000 individus), de façon à augmenter le nombre de méïoses informatives, l'autre grand facteur de la précision de la localisation. Malgré ces efforts importants, la précision reste encore limitée et n'atteint pas la précision souhaitée.

Le second article présente les résultats de cartographie fine de QTL issus du programme CartoFine, pour cinq caractères laitiers dans les trois races laitières du projet. La quantité et la précision de localisation des QTL détectés illustre bien le bond technologique permis par l'utilisation des puces pangénomiques de SNP. Ces premiers résultats permettent également d'identifier les nouveaux facteurs limitants des analyses de cartographie fine à haut débit.

4.2 Article 3

Article paru dans Animal Genetics
GUILLAUME F., GAUTIER M., BEN JEMAA S.,FRITZ S.,EGGEN A.,BOICHARD D. and DRUET T.,2007. Refinement of two female fertility QTL using alternative phenotypes in French Holstein dairy cattle, Animal Genetics 38 (1), 7274.

doi:10.1111/j.1365-2052.2006.01542.x

Refinement of two female fertility QTL using alternative phenotypes in French Holstein dairy cattle

F. Guillaume*, M. Gautier†, S. Ben Jemaa†, S. Fritz‡, A. Eggen†, D. Boichard* and T. Druet*

*INRA, UR337, Station de Génétique Quantitative et Appliquée, Jouy-en-Josas F-78350, France. †INRA, UR339, Laboratoire de Génétique Biochimique et de Cytogénétique, Jouy-en-Josas F-78350, France. ‡Union Nationale des Coopératives D'élevage et D'insémination Animale, Paris 75595, France

Summary

Two quantitative trait loci (QTL) affecting female fertility were mapped in French dairy cattle. Phenotypes were non-return rates at 28, 56, 90 and 282 days after insemination. On chromosome 3, a QTL was significant at 1% for non-return rate at 90 days, suggesting that it affects early fertility events. An analysis of *SLC35A3*, which causes complex vertebral malformation, excluded this gene from the QTL interval. On chromosome 7, a QTL was almost significant ($P = 0.05$) for non-return rate at 282 days. This QTL was associated with abortion and stillbirth problems. Use of appropriate phenotypes appeared important for fine-mapping QTL associated with fertility.

Keywords female fertility, non-return rate, quantitative trait loci.

Despite its economical importance, female fertility has been decreasing in the Holstein breed. Several countries have implemented a genetic evaluation for female fertility and have included it in their breeding objectives. Heritability of fertility, however, is very low and the available phenotypes lack accuracy and give little information on the exact timing of events. Therefore, the efficiency of selection remains limited but can be enhanced with marker-assisted selection if quantitative trait loci (QTL) with large effects are detected and characterized.

Previous studies by our group (Boichard *et al.* 2003; Fritz 2003) detected QTL for fertility measured as success/failure of each insemination of the daughters of a bull in a French Holstein population. This study continued analysis of two QTL located on chromosomes 3 and 7 using new markers and new families. Non-return rates (NRR) at 28, 56, 90 and 282 days after artificial insemination (AI) were used to assess whether early or late fertility events were involved in each QTL. NRR282 is equivalent to the fertility measure used in previously mentioned studies.

Re-insemination was used as an indicator of insemination failure to compute NRR. Insemination failure could result from non-fertilization, early embryo mortality (before 17 days), late embryo mortality (from 18 to 42 days) and foetal mortality (after 42 days). In the first two situations, cyclicity is not affected and oestrus is expected 20–23 days after insemination. Many (about 28%) of the early returns to oestrus, however, are not detected because of silent oestrus or poor heat detection, and therefore, re-insemination occurs later. In both cases, insemination failure is, therefore, observed later. Due to a possible delay between return-to-oestrus and its observation, NRRs over different periods suggest but do not identify without fail the cause of insemination failure. NRR28 could be related to early fertility events associated with good heat detection. Returns observed between 28 and 56 or 90 days represent either early failure not detected before (non-fertilization or early embryo mortality) or later failure (late embryo mortality). After 90 days, observed returns to oestrus should mainly be associated with foetal mortality.

The analysis for QTL on BTA3 included 2103 AI bulls distributed in 26 families, whereas the analysis for QTL on BTA7 involved 21 families and 1785 AI bulls. AI bulls were progeny tested with 85 daughters on average. Bulls were genotyped for 16 and 29 microsatellite markers chosen from existing linkage maps of BTA3 and BTA7 respectively (Gautier *et al.* 2003; Ihara *et al.* 2004). Phenotypes were de-regressed estimated breeding values (EBV) for NRR of daughters measured at 28, 56, 90 and 282 days. EBV were obtained from the French evaluation system, adapted for the different traits.

Half-sib linear regression (Knott *et al.* 1996) was performed for QTL detection with the QTLMAP software presented by Boichard *et al.* (2003) with the following model:

$$y_{ij} = s_i + (2p_{ij} - 1)a_i + e_{ij}$$

Address for correspondence

F. Guillaume, Station de Génétique Quantitative et Appliquée, INRA, UR337, Jouy-en-Josas 78350, France.
E-mail: francois.guillaume@jouy.inra.fr

Accepted for publication 8 October 2006

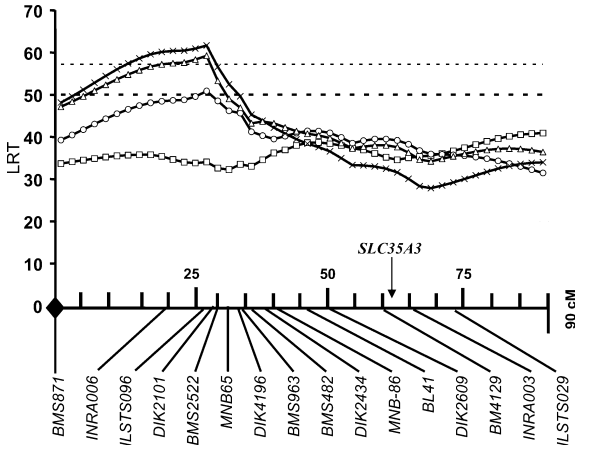


Figure 1 Likelihood curves along BTA3 for non-return rates at different dates: NRR28 (—□—), NRR56 (—○—), NRR90 (—X—) and NRR282 (—△—). Significance thresholds for all traits at 5% and at 1% are presented as bold and thin dotted lines respectively.

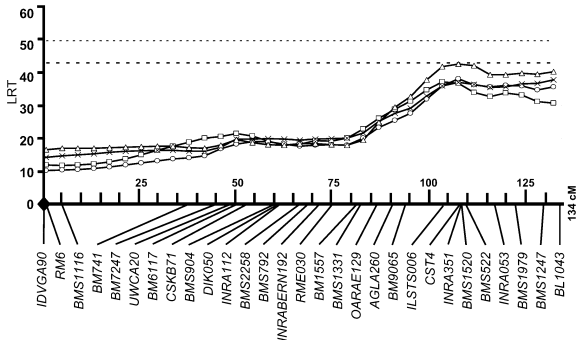


Figure 2 Likelihood curves along BTA7 for different non-return rates: NRR28 (—□—), NRR56 (—○—), NRR90 (—X—) and NRR282 (—△—). Significance thresholds at 5% and 1% are presented as bold and thin dotted lines respectively.

where y_{ij} is the de-regressed EBV of son j of sire i, s_i is the effect of sire i, p_{ij} is the probability of inheriting the first allele from sire i for son j, a_i is half of the substitution effect of the QTL carried by the sire i and e_{ij} is the residual. Residual variance was assumed to be heterogeneous and accounted for the amount of information in the progeny. Chromosome-wide significance thresholds were estimated with 30 000 within-family permutations (Churchill & Doerge 1994).

A QTL with a maximum likelihood ratio test (LRT) peak located on BTA3 at position 26 cM was significant for NRR56 ($P < 0.05$), NRR90 ($P < 0.01$) and NRR282 ($P < 0.01$), whereas no significant QTL were found for NRR28 (Fig. 1). Six sires were estimated to be heterozygous for the QTL at the maximum LRT position and the average substitution effect was 3% of NRR90. Individual sire profiles and results of a two-QTL analysis showed no evidence for a second QTL.

Likelihood ratio test curves of NRR90 and NRR282 were very similar along BTA3. Additional returns to oestrus observed at 282 (associated with foetal mortality) did not improve the efficiency of QTL detection. The same sires contributed the most to the LRT curve for NRR56, NRR90 and NRR282. A single QTL might be responsible for early fertility problems or late embryo mortality, although the LRT was higher for NRR56 and NRR90 than for NRR28. An effect occurring after 90 days was excluded.

The *SLC35A3* gene that carries a mutation causing the complex vertebral malformation (CVM) syndrome in Holstein (Thomsen *et al.* 2006) was not associated with this fertility QTL on BTA3 because (1) five of the six sires segregating for the fertility QTL were not CVM-carriers, (2) *SLC35A3* did not map within the 95% confidence interval of the QTL and (3) the mutation in *SLC35A3* acts after 90 days of gestation while the QTL was shown to act earlier.

A QTL on BTA7 was detected by Boichard *et al.* (2003) in a study that included nine Holstein, three Normande and two Montbéliarde families. In the present study, 12 Holstein families were added to the analysis and only Holstein families were considered. New LRT curves were close to significance ($P = 0.05$) only for NRR282 at position 107 cM and not for the earlier fertility measures (Fig. 2). The differences in results from this study and the previous study are probably due to differences in the families and variation in breeding values. The QTL was not segregating in any of the 12 new families. However, LRT was larger for NRR282 than for the other NRR, indicating that information after 90 days is important for this QTL. Other analyses (data not shown) showed that a significant QTL was detected for stillbirth at the same location. Both observations suggest that the same QTL would affect foetus survival and stillbirth.

In conclusion, we suggest that the QTL on BTA3 is acting on early fertility events and is not the *SLC35A3* gene while the QTL on BTA7 is acting on foetal mortality and stillbirth. This information is useful for the identification of candidate genes within the QTL regions and suggests appropriate phenotypes for further fine-mapping studies.

References

Boichard D., Grohs C., Bourgeois F., Cerqueira F., Faugeras R., Neau A., Rupp R., Amigues Y., Boscher M.Y. & Levéziel H. (2003) Detection of genes influencing economic traits in three French dairy cattle breeds. *Genetics Selection Evolution* **35**, 77–101.

Churchill G.A. & Doerge R.W. (1994) Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963–71.

Fritz S. (2003) *La sélection assistée par marqueurs chez les bovins laitiers validation des paramètres génétiques*. Master thesis, Institut National Agronomique, Paris-Grignon.

Gautier M., Hayes H., Bonsdorff T. & Eggen A. (2003) Development of a comprehensive comparative radiation hybrid map of bovine chromosome 7 (BTA7) vs. human chromosomes 1 (HSA1), 5 (HSA5) and 19 (HSA19). *Cytogenetic Genome Research* **102**, 25–31.

Ihara N., Takasuga A., Mizoshita K. *et al.* (2004) A comprehensive genetic map of the cattle genome based on 3802 microsatellites. *Genome Research* **14,** 1987–98.

Knott S.A., Elsen J.M. & Haley C.S. (1996) Methods for multiple marker mapping of quantitative trait loci in half-sib populations. *Theoretical and Applied Genetics* **93,** 71–80.

Thomsen B., Horn, P. Panitz F., Bendixen E., Petersen A.H., Holm L.E., Nielsen V.H., Agerholm J.S., Arnbjerg J. & Bendixen C. (2006). A missense mutation in the bovine *SLC35A3* gene, encoding a UDP-N-acetylglucosamine transporter, causes complex vertebral malformation. *Genome Research* **16,** 97–105.

4.3 Article 4

Article soumis à BMC Genomics

Whole genome scan IBD analysis for production traits of three French dairy.breeds using a dense SNP chip.

François Guillaume [1,2§], Sebastien Fritz[4], Joaquim Tarrès[2], Aurélia Baur[4], Tom Druet[3], André Eggen[2], Mekki Boussaha[2], Mathieu Gautier[2], Ivo Gut[5], Didier Boichard[2]

[1]Institut de l'élevage, 149 rue de Bercy, 75595 Paris Cedex 12, France

[2]INRA, UMR1313 Génétique Animale et Biologie Intégrative, F-78352 Jouy en Josas

[3]Unit of Animal Genomics, Faculty of Veterinary Medicine and Centre for Biomedical Integrative Genoproteomics, University of Liège (B43), Belgium.

[4]Union nationale des coopératives d'élevage et d'insémination animale, 149 rue de Bercy, 75595 Paris Cedex 12, France

[5]CEA/Institut de Génomique—Centre National de Génotypage, F-91057 Evry, France

[§]Corresponding author

Email addresses:

FG : Francois.guillaume@jouy.inra.fr

Abstract

Background

The aim of this study is to detect and fine map QTL for production traits in a large granddaughter design including 2837 artificial insemination bulls in the Holstein, Montbéliarde, and Normande cattle breeds. Each bull is characterized by the performances of 100 daughters on average for milk, fat, and protein yields, as well as fat and protein contents. Bulls have been genotyped for 54,000 SNPs the the 50k Illumina Beadchip. Data were analysis by accounting for both linkage and linkage disequilibrium with haplotypes of 10 markers. With such a high marker density, this design is aimed to provide many fine mapped QTL with accurate estimates of location and effects and to compare results across the three different breeds analysed separately.

Results

In total, 305 significant regions of various sizes were found across breeds, traits and chromosomes. The number of regions was rather homogeneous across traits, from 49 for fat content to 69 for milk yield. Chromosomes 1, 2, 3, 5, 6, 14, 19, and 20 appeared to be particularly rich in QTLs. Due to a higher power, more QTLs (171) were found in Holstein breed. 157 regions were punctual significant spots. In most situations, the confidence interval based on lod drop-off was very narrow. In the other situations, a complex determinism with several linked QTLs is suspected Variance components ranged from 1 to 37%, with a great majority of values below 5%. An important result is that a large proportion of QTLs is not shared across breeds.

Conclusions

Numerous QTL have been detected and fine mapped, paving the way for both marker-assisted selection and further characterization of the underlying genes and mutations. In many situations, It is interesting to note that with such results, classical marker-assisted selection is an efficient alternative to genomic selection or could be combined with genomic selection. Different QTLs across breeds suggest that a common genomic selection is probably of limited efficiency.

Background

After the first pioneering study of Georges et al [1], large genome wide QTL detection programmes have been carried out in dairy cattle in the last 15 years [1,2,3], nearly all based on the so-called granddaughter design [4]. These studies lead to the discovery of numerous QTL, initially for production traits and more recently for functional traits.

These first studies were followed by a large international effort in fine mapping in order to identify the causative mutation underlying some of these QTL, with spectacular results, however limited in number [5,6,7]. These studies were always targeted to a chromosomal region and, in most cases, to a single breed. These fine-mapping studies relied upon a common strategy with a higher density of markers, sometimes on a larger population and, more recently, on the joint use of linkage and linkage disequilibrium (LDLA) information [8]. These studies were carried out independently of each other in different populations and for different markers. Only in few situations, designs were combined at the raw data level [9] or in meta-analyses [3]. Reviews present some concordant results but the limited accuracy of location estimates prevents to infer that QTLs are identical.

With the availability of high-density SNP arrays covering the whole genome at the cost of the previous low-density genome scan with microsatellites, a one-step procedure could be used combining QTL detection and fine-mapping over the whole genome. Moreover, if the same chip is used, a common set of markers is genotyped in different studies and in different populations, making it easier to compare results and to combine data.

In this paper, we report the findings of a genome wide fine-mapping analysis for three different breeds and 5 dairy traits.

Results

Because of the high density of markers, the use of Linkage Desequilibrium, and the multiplicity of dam origins, likelihood curves were extremely irregular, with a limited redundancy of information between points for distances larger than 5 cM. This pattern is illustrated by Figure 2, corresponding

to milk yield and chromosome 1 in Montbeliarde breed. Consequently, many QTLs were located with a high accuracy. The distribution of the maximum LRT value over the chromosome under H0 was found to be close to a χ^2 distribution with 1.85 degrees of freedom for chromosome 16. We chose a conservative approach by assuming a χ^2 distribution with 2 degrees of freedom and by reporting putative QTLs with LRT values over 10, corresponding to chromosomewise significance of less than 0.006 (1.85 df) or 0.007 (2 df). Nevertheless, a large number of QTLs was detected for each trait and in each breed (Table 1). In total, across breeds, traits and chromosomes, 305 regions of various sizes were found with one-QTL analyses. The 435 individual profiles for each trait x breed x chromosome combination is provided in supplementary material. The number of regions was rather homogeneous across traits, from 49 for fat content to 69 for milk yield. Significant regions were found on all chromosomes, although the contribution of some of them (17, 21, 24, 26, 27, 28) was limited to 3 or 4 significant results. Chromosomes 1, 2, 3, 5, 6, 14, 19, and 20 appeared to be particularly rich in QTLs. As expected, more QTLs (171) were found in Holstein breed, due to the higher power of the design, whereas the numbers of QTLs found in Normande (64) and in Montbeliarde breeds (70) were similar. The distribution of the 305 significant LRT values is presented in table 2. About one third is comprised between 10 and 12, about one third between 12 and 16, 40 high values from 16 to 20, and 42 very high values from 20 to 838. The highest LRT values were observed for chromosome 14 for several traits and in the three breeds, chromosome 6 and protein content for the three breeds, chromosome 20 and 5 for fat and protein contents in Holstein, chromosome 19 for yields in Montbeliarde breed. The significant regions corresponded to different situations. About half of the regions (157) were punctual spots with few windows with significant values. 46 regions were rather narrow, from 1 to 4 cM, 73 were wider but still limited (5 to 15 cM) whereas 29 were larger, from 16 to 30 cM, and sometimes even larger up to 93 cM for protein content on chromosome 20. This width corresponds to the chromosome length with a high LRT value and should not be considered as a confidence interval, as it does not reflect the LRT curve. In most situations, the confidence interval, defined on the basis of the lod drop-off, was very

short. However, in some regions, several peaks were observed and a two-QTL analysis was necessary to conclude. For most QTLs, the estimated variance was rather small, from 1 to 6% of the total genetic variance (Table 3). Only 56 QTLs explained more than 6% of the genetic variance and only 9 more than 10%. The largest values were observed for fat content on the proximal end of chromosome 14 and ranged from 17% in Normande breed to 36% in Montbeliarde breed. The QTL for protein content on chromosome 6 explained 15% in Montbeliarde breed and only 8 and 6% in Normande and Holstein, respectively. In Montbeliarde breed again, the QTL for fat yield on chromosome 2 explained 13% of the genetic variance, whereas it was not significant in the other breeds. Lastly, in the three breeds, a QTL affecting fat content was found on chromosome 5 around position 149 cM, explaining 13% of the genetic variance in Holstein, whereas it explained 8 and 7% in Normande and Montbeliarde. In spite of the low individual values, the sum of the QTL variances for each trait was high, ranging from 55% for protein yield in Normande breed to 115% for protein content in Holstein. At least some of these high values, close to or higher than 100%, suggest that some QTL could be false positive and/or that some variance components are still overestimated. Some regions appeared to be particularly wide. For chromosome 14, the QTL is probably explained by DGAT1. Its very strong effect is clearly seen even at large distances. However, a two-QTL analysis does not reveal a second QTL. Chromosome 20 appeared to be more complex with two QTLs affecting protein content at position 50 and 63 cM. Due to linkage disequilibrium information, these two QTLs could be clearly distinguished, in spite of their rather small distance, and could be accurately located, each in a less than 1 cM interval. Similarly, the 2-QTL analysis confirmed the presence of 2 QTL for milk yield on chromosome 9 in Holstein breed, at positions 87 and 127 cM, although the maximum LRT value of the one-QTL analysis was more reasonable (14.6). Chromosome 6 was found to be particularly complex with at least 3 QTLs for protein content, one of each being the casein locus.

An important result is the low proportion of common QTLs found in the different breeds. The 22 common QTLs (for a total of 49 region x breed combinations) are reported in table 5. In most cases,

they involve only two breeds and, in majority, the Holstein and Montbeliarde breeds. In six cases corresponding to 2 regions, they involve the three breeds, on chromosome 14 for fat and protein contents, near position 130 cM on chromosome 6 (caseins) for milk, protein yield and protein content, and near position 149 cM on chromosome 5 for fat content.

Genetic correlations are strong and positive between milk, fat and protein yields, positive between fat and protein contents, and moderately negative between milk yield and fat and protein contents. Some of these genetic correlations could be found at the QTL level. It is particularly true for milk and protein yields which have very similar profiles, whereas milk and fat yields are less dependent. For instance, in Holstein, the same peaks were observed for milk and protein on chromosome 1 (position 225), 13 (10), 23 (70), for fat and protein yields on chromosome 4 (position 66), 11 (30), 25 (19), and for the three traits on chromosome 2 (position 170), 5 (140), 7 (96), 8 (144), 10 (30), 13 (116). On chromosome 26 at position 43, the same peak was observed for the three traits but it was highly significant only for fat yield because it started from a much higher basal level. The same pattern was observed in Montbeliarde breed, for instance on chromosome 1 (position 98), 2 (180), 7 (70), 10 (94), 14 (1), 19 (several positions), and 20 (126). The agreement was even stronger in Normande breed and should help to refine the location and to fine map the gene of interest (e.g. on chromosomes 1, 3, 7, 15, 17, 21, 22 or 23). For milk, fat and protein contents, results were less concordant, except for well known QTLs on chromosomes 14 and 20.

Discussion

Use of dense SNP markers map and LDLA methodology improves dramatically the power and resolution of the conventional granddaughter design. It makes it possible to use all the information not only from paternal but although from maternal origin. The high density of markers allows to account for linkage disequilibrium and therefore takes advantage of the historical recombinations in the population. It extracts much more information from the same design whereas the conventional approach relies only upon segregations within sires heterozygous at the QTL. In practice, 15 to 30 haplotype effects were estimated at each analyse, making each record fully informative. This

explains why more much QTLs are detected than with a design similar in size but with a low density marker information.

More powerful, the design provides less biased estimates of variance components. As a result, most variance component estimates are smaller than 6% of the total genetic variance. Due to the limited detection power, few QTLs explaining less than 2% of genetic variance are detected. As a consequence, only a left-censored L-shape QTL distribution [10] is observed. Nevertheless, in spite of these reasonable values, many QTL variances are likely to be overestimated, as their sum appears to be very high, and higher than 100% in three breed x trait combinations out of 15.

Significance thresholds were computed with chromosome 16, assuming homogeneous values over chromosomes. These simulations suggest that the maximum LRT value over the chromosome under H0 follows a χ^2 distribution with 1.85 degrees of freedom. This is in agreement with Grignola et al [11] who reported that this distribution is a mixture of two χ^2 distributions with 1 and 2 degrees of freedom. In this study, we reported QTLs with LRT values higher than 10, corresponding to a chromosome-wise critical p-value close to 0.7% assuming 2 degrees of freedom. With a total of 435 chromosomal analyses (3 breeds, 5 traits, and 29 chromosomes), 3 to 4 false positive results are expected over a total of 305 reported results. Of course, this 10 threshold is a little arbitrary but it is believed to be conservative. Assuming a more liberal threshold of 9 would have led to increase the number of regions by more than 100.

Not only power was increased but also resolution. Linkage disequilibrium information at two positions becomes rapidly independent when their genetic distance increases. This is not the case with full LRT which includes linkage information. However, such a design makes it possible to distinguish two QTLs distant by less than 10 cM, which would be completely impossible with traditional microsatellite designs. For instance, Gautier et al [12] presented a comprehensive study of fat yield on chromosome 26 but did not succeed to fine map QTLs with such a high resolution. In the present study, a short region is targeted in Holstein breed, including SORCS1, a candidate gene already mentioned by Gautier et al. Nevertheless, the definition of confidence interval is not trivial.

Indeed, even with a 10-marker window, the likelihood function is not smooth and prevents to use methods based on second derivatives such that of Piepho [13]. For the same reason, lod drop-off seems to be optimistic. In many situations where the pattern is simple with one likely QTL, the peak is very clear and the confidence interval seems to be restricted to a very short region. Even when both linkage and linkage disequilibrium informations combine to detect a QTL resulting in a large region with a high LRT, the curve is sharp enough and points to the QTL. For instance, a large region of 20 cM of chromosome 14 presents a very high LRT value and is detected to carry a very strong QTL but DGAT1, the gene responsible for this QTL, is clearly located in the right marker-window, with a LRT difference of more than 35 with the flanking windows. But some regions appeared to be much more complex. The two-QTL analysis of chromosome 20 revealed two QTL separated by 13 cM, each of them being accurately located in intervals of less than 1 cM. The situation is not always so clear. For instance, the profile of chromosome 19 in Montbeliarde, or of chromosome 6 in Holstein appears to be very complex with an unknown number of QTLs.

The comparison with previous studies is not easy to perform, because of the multiplicity of results both in this paper and in the literature and of the uncertainty about the locations. Khatkar et al (2004) reviewed most results and plotted them per 30 cM intervals over the 29 autosomes. A cattle QTL data base available on-line (http://www.animalgenome.org/cgi-bin/QTLdb/BT/qtraitology?class_ID=1003) reviews QTL data with the most likely location, an estimate of confidence interval, and rarely variance component [14]. Compared with Khatkar's review, our study on three breeds reports more QTLs and only a limited proportion is shared with these studies. For milk yield, we report 41 QTLs whereas 37 are described by Khatkar et al [3], and 17 only could be the same. Corresponding figures are 47, 38 and 17 for protein yield, 43, 58 and 29 for protein content, 47, 31 and 17 for fat yield, and 38, 18 and 10 for fat content. Of course, strong QTLs detected on chromosomes 6, 14, 20 are confirmed. Similarly, some QTL found repeatedly in the distal part of chromosome 1, on chromosome 3, 7, 9, 10, 13, 26, and 29 are found in at least one

breed in our study. But our study confirms a general observation that any new study confirms only few well known strong QTLs and detects a large proportion of new and specific QTLs.

Even within our study, the concordance of results across breeds is limited. This finding could be interpreted by the large number of genes potentially involved in the genetic determinism of these traits. As breeds were created from a limited number of founders and strongly selected, some genes segregating in a population could be fixed in another one, resulting in a large proportion of population-specific segregating genes. This finding is very important both for QTL characterization and marker-assisted or genomic selection. It is commonly assumed that QTLs are the same across breeds and that linkage disequilibrium is maintained only at short distances, providing a fast and efficient method for fine mapping of these QTLs. If segregating QTLs are not always the same across populations, it means that combined analyses are less informative that initially assumed. An alternative would be to select populations where the same QTLs still segregates. Another way could be to combine QTL information from segregating populations and selective sweeps from populations where the QTL is fixed. Nevertheless, the present analysis shows that many QTLs are finely mapped within population with the design commonly developed for genomic selection which is also a good starting point for QTL characterization. As several reference populations are currently developed for genomic selection and are rapidly growing, most of them with the same chip, a huge design is potentially available for QTL mapping, at least in the largest breeds.

For marker-assisted or genomic selection, this finding implies that not only marker-QTL phases are not the same but informative QTLs could be different. This clearly limits the possibility of across-breeds evaluation with a common reference training population. It also makes more complex the use of a common reduced set of markers to be used across breeds.

Conclusions

In spite of its limited size (compared to the potentially available reference populations for genomic selection), this design is much more powerful than the previous designs based on low or medium density microsatellites, due to the large number of SNPs combined with LDLA analysis. Many

QTL are fine mapped in spite of their limited genetic variance, opening strong opportunities for their characterisation. Larger designs, however, are required for the most complex regions to disentangle clusters of linked QTLs. These results could be used in genomic selection but also in marker-assisted selection which uses the biological knowledge about the genetic determinism of the trait.

Materials and methods

Phenotypic data

The QTL experiment was a granddaughter design including 2837 bulls from the three main French dairy cattle breeds (1575 Holstein, 661 Normande, and 601 Montbeliarde bulls, respectively). Bulls were distributed in 69 sire families whereas dams were little related. This design is favourable for Linkage Disequilibrium and Linkage Analysis (LDLA) and provides the robustness of linkage analysis and the resolution of LD analysis. All bulls were progeny tested with phenotypes of 100 daughters on average. For each trait, each bull was given a phenotype defined as the so-called daughter-yield deviation [15], i.e. the average performance of the daughters, adjusted for environmental factors and breeding value of their dam. These phenotypes, as well as their individual residual variances ($1/w_i$), were obtained from the French genetic evaluation system. Five dairy production traits were analysed: milk, fat, and protein yields, and fat and protein contents.

Marker Genotypes

Bulls were genotyped with the Bovine SNP50 BeadChip from Illumina. Genotypes at 54001 SNP were obtained. These SNPs are known to provide a good coverage of the genome with an average spacing of 46 kbases, most intervals comprised between 22 and 80 kbases, and an average minor allele frequency (MAF) around 20% for most Bos Taurus breeds. SNP that were not mapped on BTAU3.0 assembly were discarded. Within each breed, SNP showing a MAF lower than 5 % were also discarded. Finally, the number of markers used for analysis reached 40757, 38885, and 39004 in Holstein, Normande, and Montbeliarde breeds, respectively. Marker density remained high with 15 SNP/Mb on average. Using the family structure and the high marker density, haplotypes were

reconstructed with the method described by Druet et al [16]. This method allows a fast and efficient haplotype reconstruction and more than 95% of the SNP were attributed to a phase.

Statistical model

Data were analysed based on a LD-LA model [8] implemented in an AI-REML software developed and already used by Druet et al [16]. Each breed was analysed separately. The pedigree included bulls, sires and dams**.** IBD probabilities among founder animals were first computed according to Meuwissen et al [8] using haplotypes of 10 markers around the tested position. A clustering step was then used to define a reduced number of haplotype groups. For the remaining animals, probabilities of identity by descent (IBD) were computed for genotyped animals at every tested position according to Pong-Wong et al [17].

The tested model was :

$$\mathbf{y} = \mathbf{1}\,\mu + \mathbf{Z}\,\mathbf{u} + \mathbf{Z_h}\,\mathbf{h} + \mathbf{e}$$

where **y** is the vector of phenotypes, μ the mean, **u** the vector of polygenic effect, **h** is the vector of random QTL effects corresponding to founder's haplotype clusters [16] and **e** a residual vector zero mean and heterogeneous variance $\mathbf{W}^{-1}\sigma_e^2$.

For every second marker along the genome a likelihood ratio test was computed by comparing the models with and without QTL effects [16]. As theoretical thresholds are not clearly defined, empirical thresholds were derived from simulations under H0. Chromosome 16 was used for these simulations and position of every second marker was tested. Then for each of 6000 simulations, the maximum LRT value over the chromosome was retained. The empirical distribution under H0 was found to be similar to a χ^2 distribution with 1.85 degrees of freedom (Figure 1), with the same mean, variance, and 5% percentile**.** Thresholds were defined assuming a χ^2 distribution with 2 degrees of freedom. Thresholds were assumed to be the same for the 29 chromosomes. Genome-wise significance thresholds were obtained by accounting for 29 chromosomes.

Authors' contributions

F Guillaume conducted the analyses and drafted the manuscript.
S Fritz managed the data base and contributed to the analyses.
A Baur, J Tarres contributed to the analyses.
T Druet developed the haplotyping and LDLA software.
M Gautier constructed the genetic map
I Gut and xxx were in charge of the genotyping work
A Eggen supervised the molecular part of the project and provided the DNA samples.
D Boichard supervised the study and drafted the manuscript.

Acknowledgements

The Cartofine project was funded by the French National Research Agency (ANR) and by

ApisGene.

References

1. Georges M, Nielsen D, Mackinnon M, Mishra A, Okimoto R, Pasquino AT, Sargeant LS, Sorensen A, Steele MR, Zhao X, Womack JE, Hoeschele I: **Mapping Quantitative Trait Loci Controlling Milk Productio in Dairy Cattle by Exploiting Progeny Testing.** Genetics 1995, 139(2):907
2. Boichard D, Grohs C, Bourgeois F, Cerqueira F, Faugeras R, Neau A, Rupp R, Amigues Y, Boscher MY,Levéziel H: **Detection of genes influencing economic traits in three French dairy cattle breeds**. Genet Sel Evol 2003, 35:77.
3. Khatkar MS, Thomson PC, Tammen I, Raadsma HW: **Quantitative trait loci mapping in dairy cattle: review and meta-analysis.** Genet Sel Evol 2004, 36(2):163.
4. Weller, J. I., Kashi, Y. et Soller, M. (1990). **Power of daughter and granddaughter designs for determining linkage between marker loci and quantitative trait loci in dairy cattle**. J Dairy Sci, 73(9):2525–2537.
5. Grisart B, Coppieters W, Farnir F, Karim L, Ford C, Berzi P, Cambisano N, Mni M, Reid S, Simon P, Spelma R, Georges M, Snell R: **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 2002, 12(2):222.

6. Blott S, Kim JJ, Moisio S, Schmidt-K untzel A, Cornet A, Berzi P, Cambisano N, Ford C, Grisart B, Johnson D, Karim L, Simon P, Snell R, Spelman R, Wong J, Vilkki J, Georges M, Farnir F, Coppieters W: **Molecular dissection of a quantitative trait locus: a phenylalanine-to-tyrosine substitution in the transmembrane domain of the bovine growth hormone receptor is associated with a major effect on milk yield and composition.** Genetics 2003, 163:253
7. Cohen-Zinder M, Seroussi E, Larkin DM, Loor JJ, van der Wind AE, Lee JH, Drackley JK, Band MR, Hernandez AG, Shani M, Lewin HA, Weller JI, Ron M: **Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 a affecting milk yield and composition in Holstein cattle**. Genome Res 2005, 15(7):936
8. Meuwissen TH, Goddard ME: **Prediction of identity by descent probabilities from marker-haplotypes**. Genet Sel Evol 2001, 33(6):605.
9. Bennewitz J, Reinsch N, Grohs C, Levéziel H, Malafosse A, Thomsen H, Xu N, Looft C, Kuhn C, Brockmann GA, Schwerin M, Weimann C, Hiendleder S, Erhardt G, Medjugorac I, Russ I, Forster M, Brenig B, Reinhardt F, Reents R, Averdunk G, Bl umel J, Boichard D, Kalm E: **Combined analysis of data from two granddaughter designs: A simple strategy for QTL confirmation and increasing experimental power in dairy cattle**. Genet Sel Evol 2003, 35(3):319
10. Hayes B, Goddard ME: **The distribution of the effects of genes affecting quantitative traits in livestock**. Genet Sel Evol 2001, 33(3):209.
11. Grignola FE, Hoeschele I, Zhang Q, Thaller G: **Mapping quantitative trait loci in outcross populations via residual maximum likelihood. II. A simulation study**. Genet Sel Evol 1996, 28:491-504.
12. Gautier, M., Barcelona, R. R., Fritz, S., Grohs, C., Druet, T., Boichard, D., Eggen, A. et Meuwissen, T. H. E. (2006). **Fine mapping and physical characterization of two linked**

quantitative trait loci affecting milk fat yield in dairy cattle on bta26. Genetics, 172(1):425-436.

13. Piepho, H. P. (2001). **A quick method for computing approximate thresholds for quantitative trait loci detection**. Genetics, 157(1):425–432.
14. Hu, Z.-L., Fritz, E. R. et Reecy, J. M. (2007). **Animalqtldb : a livestock qtl database tool set for positional qtl information mining and beyond**. Nucl. Acids Res., 35(suppl_1):D604–609.
15. VanRaden PM, Wiggans GR: **Derivation, calculation, and use of national animal model information**. J Dairy Sci 1991, 74(8):2737.
16. Druet T., Fritz S., Boussaha M., Ben-jemaa S., Guillaume F., Derbala D., Zelenika D., Lechner D., Charon C., Boichard D., Gut I. G., Eggen A., Gautier M. . 2008. **Fine mapping of quantitative trait loci affecting female fertility in dairy cattle on bta03 using a dense single-nucleotide polymorphism map**. Genetics, 178(4) :2227-2235.
17. Pong-Wong R, George A, Woolliams JA, Haley CS: **A simple and rapid method for calculating identity-by-descent matrices using multiple markers**. Genet Sel Evol 2001, 33(33):453

Figures

Figure 1 - Empirical distribution of maximum LRT value over chromosome 16 (6000 simulations), compared to the theoretical χ^2 distribution with 1.85 degrees of freedom (in green).

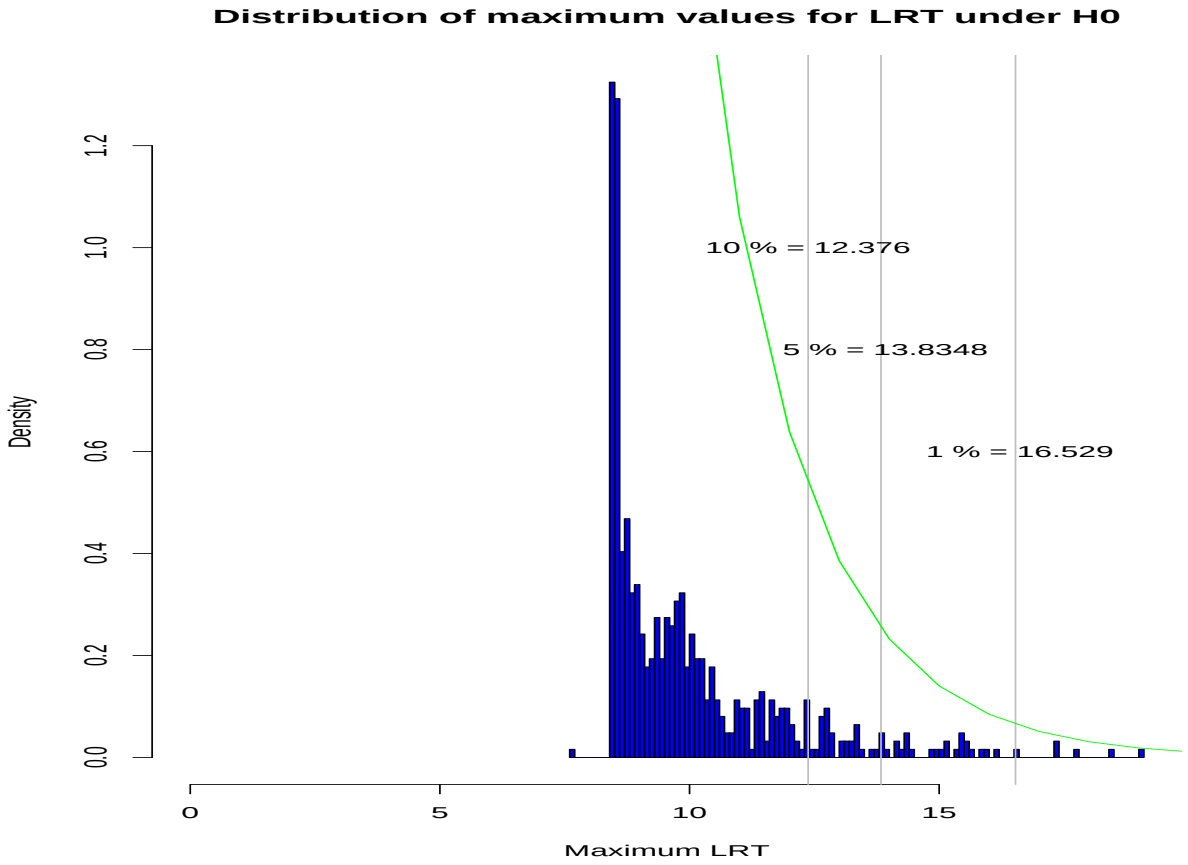
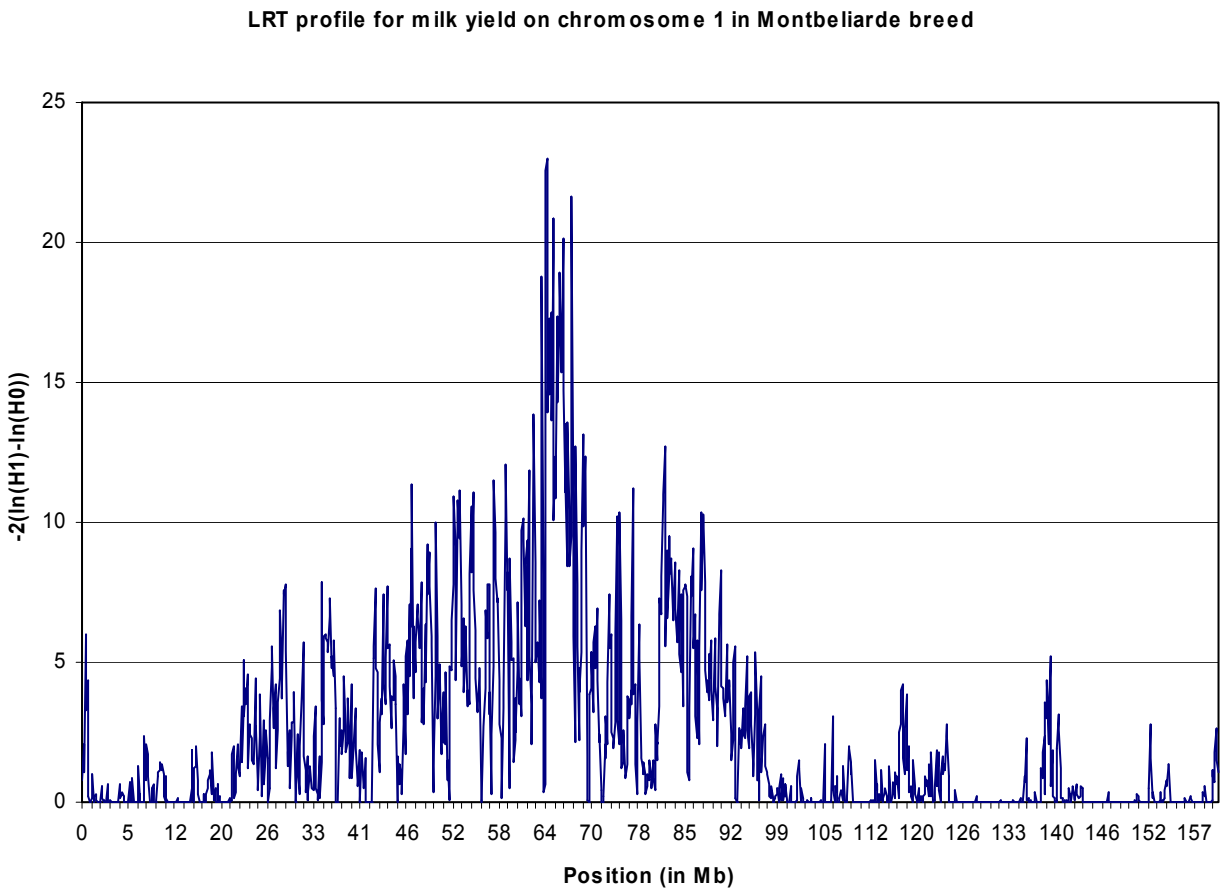


Figure 2 - LRT profile for milk yield on chromosome 1 in Montbeliarde breed.



Tables

Table 1 - List of 305 most significant regions.

A – Montbeliarde breed

(1) MY = milk yield, FY = fat yield, PY = protein yield, F% = fat percent, P% = protein percent

(2) P = punctual, otherwise in cM (Note that these values are not confidence intervals)

Trait (1)	Chromosome	Marker #	Distance (cM)	LRT	Part of genetic variance	Size of significant region (2)
MY	1	1016	96,74	22,99	0,082	3
MY	2	102	7,58	10,25	0,047	P
MY	3	1426	147,67	12,17	0,036	8
MY	4	956	93,80	15,02	0,046	20
MY	4	1112	109,11	17,77	0,055	10
MY	4	1448	139,78	15,60	0,056	P
MY	6	1338	137,54	11,47	0,078	8
MY	6	1652	164,43	10,22	0,038	3
MY	7	776	69,91	12,32	0,043	P
MY	13	412	43,71	12,36	0,055	P
MY	14	14	0,39	13,97	0,032	P
MY	19	584	56,19	23,56	0,057	3
MY	19	700	68,40	22,29	0,057	8
MY	20	1128	106,71	15,73	0,049	2
FY	1	1070	100,87	15,51	0,068	4
FY	2	212	19,27	14,17	0,133	P
FY	2	1696	178,85	11,63	0,049	2
FY	3	900	102,23	16,98	0,048	8
FY	3	1292	137,42	12,04	0,053	P
FY	3	1492	153,94	11,36	0,039	P
FY	3	1628	164,33	11,87	0,032	P
FY	4	894	88,66	13,54	0,080	10
FY	4	1120	109,85	14,93	0,042	P
FY	5	346	37,39	11,70	0,052	5
FY	5	1236	149,43	15,44	0,101	P
FY	19	584	56,19	18,93	0,044	2
FY	19	738	71,23	27,44	0,084	7
FY	19	886	83,95	23,06	0,046	P
FY	20	1116	105,55	10,24	0,034	P
PY	1	1070	100,87	20,02	0,057	9
PY	2	102	7,58	10,51	0,041	P
PY	2	1696	178,85	14,77	0,048	P
PY	3	1478	152,34	13,01	0,045	P
PY	4	894	88,66	16,20	0,093	10
PY	4	1112	109,11	19,42	0,053	3
PY	5	1292	158,50	13,60	0,070	P
PY	6	1160	112,84	14,89	0,046	25
PY	8	266	28,85	10,03	0,076	P
PY	10	966	94,10	13,12	0,067	P
PY	19	584	56,19	26,03	0,067	30
PY	20	1174	110,70	14,13	0,064	6
PY	25	684	59,71	11,04	0,040	P

F%	1	790	75,95	11,95	0,048	P
F%	3	744	83,32	10,54	0,066	P
F%	3	1858	183,32	10,25	0,049	P
F%	4	1356	131,39	15,44	0,078	P
F%	5	400	43,99	10,48	0,054	P
F%	5	942	116,95	13,84	0,043	5
F%	5	1142	139,25	14,02	0,082	15
F%	9	754	82,09	12,28	0,051	3
F%	14	16	1,02	138,90	0,360	23
F%	15	730	71,92	12,97	0,090	P
F%	20	442	41,50	15,31	0,065	P
F%	20	630	62,06	13,43	0,044	3
F%	28	530	51,49	12,48	0,049	3
P%	1	684	66,73	10,08	0,046	P
P%	1	1396	135,54	10,40	0,051	P
P%	2	1906	199,79	11,62	0,044	P
P%	6	1254	127,22	60,53	0,148	56
P%	7	838	80,09	10,19	0,037	P
P%	8	1325	131,41	10,99	0,039	P
P%	10	162	16,42	14,74	0,036	P
P%	11	922	81,54	14,55	0,055	P
P%	15	692	66,54	11,94	0,042	P
P%	20	442	41,50	10,79	0,040	P
P%	20	706	68,37	19,70	0,048	18
P%	21	500	46,48	20,71	0,073	10
P%	21	642	62,42	22,11	0,094	9
P%	21	762	74,99	18,97	0,070	3
P%	21	942	92,45	16,91	0,048	3

B –Normande breed

Trait	Chromosome	Marker #	Distance (cM)	LRT	Part of genetic variance	Size of region
MY	1	2021	196,50	16,22	0,068	P
MY	1	2263	219,25	11,58	0,062	P
MY	4	1224	119,02	11,23	0,037	4
MY	5	118	9,15	14,56	0,062	P
MY	5	671	79,29	11,29	0,043	2
MY	5	1444	178,08	11,01	0,075	P
MY	7	1186	115,95	13,17	0,039	P
MY	13	658	65,74	10,29	0,030	P
MY	13	958	95,20	17,61	0,063	15
MY	15	298	32,93	11,63	0,045	P
MY	16	176	17,35	12,26	0,061	P
MY	17	981	95,28	23,12	0,054	P
MY	22	418	37,32	13,42	0,055	P
MY	23	194	20,29	11,17	0,075	4
MY	23	469	45,36	10,36	0,038	P
MY	29	42	4,50	13,68	0,051	P
MY	29	192	19,13	11,21	0,050	8
FY	1	443	44,46	12,43	0,101	P
FY	1	2021	196,50	14,57	0,064	15
FY	3	346	41,47	14,39	0,043	3
FY	5	19	1,35	10,59	0,042	P

FY	8	1696	169,38	15,06	0,055	3
FY	9	1474	155,33	10,63	0,040	P
FY	13	958	95,20	13,51	0,066	P
FY	14	1100	104,86	13,13	0,053	P
FY	17	983	95,35	14,92	0,056	P
FY	23	194	20,29	10,57	0,076	P
FY	27	62	6,32	10,15	0,032	P
FY	28	138	12,02	12,24	0,046	P
PY	1	2263	219,25	10,60	0,065	P
PY	3	326	38,03	14,99	0,060	2
PY	5	73	5,68	12,71	0,066	P
PY	5	1444	178,08	10,86	0,076	P
PY	6	1316	133,91	10,38	0,027	P
PY	13	958	95,20	14,08	0,051	6
PY	15	298	32,93	15,35	0,051	P
PY	17	981	95,28	24,95	0,053	P
PY	27	66	6,87	13,18	0,060	P
PY	29	42	4,50	11,35	0,047	P
F%	4	1500	143,61	17,03	0,051	2
F%	5	378	41,01	11,98	0,038	P
F%	5	1208	149,07	15,83	0,081	5
F%	5	1299	163,52	12,39	0,030	P
F%	6	986	91,75	16,39	0,075	3
F%	6	1326	135,28	18,98	0,066	25
F%	10	908	86,29	11,02	0,066	P
F%	11	1542	148,50	15,85	0,057	2
F%	11	1672	160,82	12,05	0,029	P
F%	13	658	65,74	10,15	0,026	5
F%	14	14	0,67	70,00	0,174	P
F%	23	252	24,71	13,66	0,048	P
F%	23	692	67,44	10,29	0,046	P
P%	1	2310	222,60	14,35	0,039	7
P%	3	220	23,97	10,75	0,040	P
P%	4	830	83,41	12,99	0,037	P
P%	6	1294	131,86	36,42	0,079	25
P%	10	714	70,58	11,43	0,042	10
P%	10	910	86,53	14,26	0,067	P
P%	10	1000	97,20	11,56	0,048	P
P%	14	14	0,67	12,63	0,047	P
P%	19	474	45,85	15,27	0,052	1
P%	19	692	68,01	11,38	0,064	P
P%	19	833	80,45	13,71	0,051	3
P%	26	362	38,00	12,70	0,034	2

C – Holstein breed

Trait	Chromosome	Marker #	Distance (cM)	LRT	Part of genetic variance	Size of region
MY	1	814	75,83	15,73	0,054	P
MY	1	914	83,51	15,27	0,029	P
MY	1	2383	222,60	15,35	0,028	2
MY	2	1596	168,21	16,90	0,034	7
MY	2	1740	178,96	12,76	0,029	P
MY	2	1894	193,90	11,65	0,018	P

MY	3	754	80,82	12,62	0,029	4
MY	5	1076	123,52	10,52	0,021	P
MY	5	1238	142,32	21,38	0,019	40
MY	6	780	68,14	13,63	0,026	2
MY	6	1286	116,66	12,49	0,012	P
MY	6	1380	127,43	11,68	0,016	P
MY	7	1022	94,73	16,57	0,016	6
MY	8	879	84,93	11,36	0,017	P
MY	8	1472	142,46	14,43	0,039	P
MY	9	64	4,38	11,98	0,018	4
MY	9	754	79,37	12,07	0,024	P
MY	9	852	87,53	14,64	0,046	P
MY	9	950	95,82	12,28	0,018	P
MY	9	1084	112,08	10,72	0,022	P
MY	9	1214	126,83	11,24	0,011	P
MY	10	358	30,85	12,75	0,017	P
MY	10	430	39,51	13,42	0,017	P
MY	13	110	9,49	16,12	0,036	P
MY	14	14	0,42	183,72	0,101	21
MY	14	642	49,76	14,20	0,027	12
MY	14	874	73,70	10,75	0,013	P
MY	14	1080	92,04	19,80	0,032	6
MY	15	540	49,56	11,75	0,025	P
MY	15	756	68,38	10,96	0,044	3
MY	15	856	80,12	10,17	0,025	P
MY	19	556	51,85	11,60	0,022	P
MY	19	750	69,75	10,66	0,021	2
MY	19	880	80,16	13,54	0,026	P
MY	20	524	50,57	15,89	0,023	13
MY	23	291	27,53	13,91	0,015	6
MY	23	736	69,77	24,01	0,033	6
MY	24	144	10,47	11,18	0,016	5
FY	2	780	68,89	17,68	0,021	17
FY	2	1110	104,95	12,30	0,035	4
FY	2	1594	168,07	25,22	0,035	16
FY	4	676	65,60	16,84	0,062	P
FY	5	208	16,21	11,41	0,028	P
FY	5	912	108,01	11,71	0,030	P
FY	5	1232	141,91	19,52	0,026	15
FY	6	726	64,42	15,68	0,025	P
FY	6	952	82,83	18,37	0,021	10
FY	6	1198	106,95	12,51	0,033	P
FY	6	1342	124,09	12,71	0,026	P
FY	7	1022	94,73	24,33	0,035	27
FY	7	1316	123,05	11,25	0,042	13
FY	10	364	31,25	13,81	0,020	P
FY	11	400	29,45	10,87	0,017	P
FY	12	232	20,49	13,30	0,023	P
FY	12	352	34,50	20,12	0,022	11
FY	12	566	61,31	12,55	0,021	3
FY	12	1158	113,78	15,46	0,025	8
FY	13	634	62,29	11,11	0,019	P
FY	14	16	0,85	169,17	0,072	17
FY	15	748	67,01	10,60	0,053	P
FY	18	288	23,50	20,79	0,024	6
FY	18	474	40,00	13,07	0,022	3

FY	18	614	53,98	12,59	0,014	7
FY	19	828	76,41	12,14	0,027	2
FY	20	458	41,50	18,09	0,025	5
FY	20	512	48,15	18,68	0,038	P
FY	24	780	73,02	10,70	0,029	6
FY	25	229	18,72	11,51	0,022	6
FY	26	449	42,95	19,46	0,017	10
FY	28	242	21,27	11,56	0,021	P
PY	1	814	75,83	11,52	0,047	P
PY	1	1039	95,62	11,98	0,024	4
PY	1	2413	224,71	17,91	0,018	8
PY	2	1494	153,80	12,34	0,021	P
PY	2	1596	168,21	26,74	0,038	11
PY	3	1192	126,42	10,85	0,017	P
PY	3	1438	146,57	11,39	0,015	7
PY	3	1644	162,67	16,11	0,020	10
PY	3	1874	182,06	11,77	0,023	P
PY	4	44	6,61	11,14	0,021	P
PY	4	688	67,08	12,87	0,019	8
PY	5	304	26,26	11,27	0,020	9
PY	5	1238	142,32	21,83	0,021	10
PY	5	1504	172,33	16,28	0,025	10
PY	6	862	76,04	10,37	0,027	P
PY	6	1382	127,52	16,15	0,025	16
PY	7	1022	94,73	22,10	0,025	22
PY	8	1472	142,46	17,44	0,042	12
PY	9	876	89,54	17,83	0,028	11
PY	9	1070	110,52	11,65	0,018	P
PY	9	1214	126,83	13,75	0,012	2
PY	10	358	30,85	19,66	0,019	23
PY	11	390	28,86	11,65	0,018	P
PY	11	1724	158,45	14,16	0,030	2
PY	13	110	9,49	18,03	0,035	P
PY	14	14	0,42	43,74	0,030	10
PY	18	16	1,01	10,06	0,015	P
PY	18	318	25,76	12,15	0,018	15
PY	18	946	83,23	13,17	0,015	7
PY	19	248	25,00	17,85	0,025	3
PY	19	884	80,66	13,23	0,030	16
PY	22	714	67,42	10,11	0,017	P
PY	22	908	84,88	10,47	0,033	P
PY	23	310	29,48	15,59	0,026	5
PY	23	736	69,77	15,62	0,031	8
PY	24	144	10,47	11,99	0,017	6
PY	24	256	21,24	10,15	0,013	P
PY	25	229	18,72	10,84	0,020	P
PY	26	731	68,36	16,31	0,024	P
F%	2	41	2,03	10,89	0,018	P
F%	2	1572	166,00	14,90	0,023	P
F%	3	384	43,23	10,19	0,017	P
F%	5	1296	149,27	35,43	0,038	28
F%	5	1614	181,29	13,69	0,132	P
F%	6	752	66,65	14,55	0,040	10
F%	6	1102	98,39	10,53	0,016	3
F%	11	1284	114,23	12,78	0,028	P
F%	11	1522	137,90	14,08	0,014	P

F%	12	566	61,31	12,26	0,016	P
F%	12	1070	105,97	18,54	0,024	9
F%	14	16	0,85	839,55	0,242	29
F%	14	833	70,31	12,68	0,017	P
F%	14	1080	92,04	30,94	0,038	1
F%	15	1136	105,07	10,46	0,024	P
F%	16	946	93,15	12,83	0,076	P
F%	18	894	78,28	10,31	0,034	P
F%	19	846	77,54	21,07	0,016	3
F%	20	514	48,42	59,21	0,079	66
F%	22	776	72,89	13,00	0,027	10
F%	23	291	27,53	12,33	0,021	P
F%	26	451	43,21	10,84	0,016	P
F%	27	584	54,87	12,81	0,023	15
P%	1	898	82,26	11,32	0,027	6
P%	1	1665	158,25	10,28	0,014	P
P%	2	1704	176,49	11,13	0,021	P
P%	3	242	25,15	17,72	0,029	10
P%	3	440	49,18	22,06	0,028	22
P%	3	1012	108,69	11,85	0,029	4
P%	5	498	52,86	26,79	0,026	20
P%	5	1306	151,07	11,74	0,017	P
P%	6	472	40,88	14,57	0,045	5
P%	6	704	62,75	22,35	0,054	36
P%	6	1076	94,96	25,18	0,041	12
P%	6	1422	133,42	48,07	0,058	44
P%	7	1298	121,38	13,33	0,027	5
P%	8	80	7,50	11,83	0,018	P
P%	10	1106	103,45	10,07	0,018	P
P%	11	458	34,03	11,25	0,045	P
P%	11	1158	102,57	11,23	0,028	3
P%	11	1524	138,05	16,38	0,013	6
P%	12	1048	103,08	16,90	0,024	6
P%	13	684	66,28	14,98	0,026	P
P%	14	16	0,85	157,93	0,066	23
P%	14	1096	93,59	38,33	0,064	27
P%	15	408	40,07	19,73	0,022	3
P%	15	1090	100,62	12,51	0,020	P
P%	16	20	1,18	23,15	0,035	5
P%	16	170	16,52	15,87	0,028	5
P%	16	443	45,69	10,70	0,015	P
P%	16	702	70,65	10,32	0,022	P
P%	18	184	13,66	10,66	0,022	P
P%	18	298	24,40	10,37	0,031	P
P%	19	854	78,44	10,33	0,019	P
P%	20	46	3,96	19,32	0,031	3
P%	20	524	50,57	110,33	0,083	93
P%	22	764	71,90	10,90	0,020	P
P%	23	742	70,21	11,97	0,021	P
P%	25	202	16,69	12,03	0,020	7
P%	25	656	55,36	12,17	0,021	7
P%	27	590	55,42	11,58	0,016	P
P%	29	734	69,06	21,01	0,016	23

Table 2 - Distribution of significant LRT values (for each significant region, the highest LRT value is retained)

Class of LRT	Number of values
10-12	109
12-14	68
14-16	46
16-18	24
18-20	16
20-22	0
22-24	21
24-838	21

Table 3 - Distribution of proportion of genetic variance explained by the QTLs

Proportion of Genetic Variance	Traits					Total
	Milk	Fat	Protein	Fat content	Protein content	Total
1 – 2 %	14	4	17	8	11	54
2 – 3%	15	18	15	9	17	74
3 – 4 %	12	10	7	6	10	45
4 – 5 %	9	10	7	7	12	45
5 – 6 %	9	6	6	4	6	31
6 – 10 %	9	8	10	11	9	47
10 – 37 %	1	3		4	1	9
Total	69	59	62	49	66	305

Table 4 - Sum of estimated Proportions of genetic variance explained by the QTLs

Breed	Traits				
	Milk yield	Fat yield	Protein yield	Fat content	Protein content
Montbeliarde	0,73	0,90	0,76	1,08	0,87
Normande	0,93	0,67	0,55	0,79	0,60
Holstein	1,02	0,93	0,92	0,97	1,16

4.4 Conclusion

Le nombre de QTL détectés lors du projet CartoFine s'est avéré beaucoup plus élevé qu'attendu. Il remet largement en cause les idées précédentes supposant que quelques QTL (ceux pris en compte dans le programme SAM1) expliquaient plus de la moitié de la variabilité génétique. Ce résultat essentiel renforce a posteriori le choix d'une analyse pangénomique ignorant les résultats antérieurs, alors que la stratégie antérieure était plutôt l'analyse fine de QTL primo-localisés – stratégie assez lourde à mettre en œuvre pour des résultats relativement réduits comme l'illustre le premier article de ce chapitre.

Ces résultats ouvrent bien sûr la voie à la sélection assistée par marqueurs de seconde génération dans chacune des trois races participant au programme, y compris les races Normande et Montbéliarde pour lesquelles la population génotypée est encore réduite.

Une question complexe, partiellement résolue, est la multiplicité des tests due au grand nombre de marqueurs. Ceci est d'autant plus vrai que la prise en compte du déséquilibre de liaison diminue sensiblement la corrélation entre tests basés sur des marqueurs proches. Les seuils ont été calculés à partir d'un chromosome de longueur moyenne, le chromosome 16, par simulation sous H0, en considérant la valeur maximum observée du rapport de vraisemblance sur l'ensemble du chromosome. Il est souvent dit que la distribution est intermédiaire entre deux χ^2 à un et deux degrés de liberté. Nous avons effectivement observé un χ^2 dont le nombre de degrés de liberté est de 1.85. Faute de pouvoir obtenir des résultats propres à chaque chromosome, nous avons utilisé les mêmes seuils pour l'ensemble des chromosomes.

Un résultat important de cette étude, et une surprise, est la relativement faible concordance des QTL entre races. Si l'on retrouve bien certains QTL communs et fréquemment décrits dans la littérature, la majorité des résultats sont différents entre les trois races. Bien sûr, il est possible que les dispositifs Normands et Montbéliards n'aient pas la taille suffisante pour permettre une localisation fine de QTL observés en race Holstein mais, par ailleurs, le nombre de QTL détectés dans ces deux races et non observés en Holstein est relativement important. Il est possible également que l'on ait des effets d'échantillonnage et que les dispositifs ne soient pas totalement représentatifs des populations. Toutefois, nous privilégions davantage l'interprétation de QTL en ségrégation réellement différents. Autrement dit, lors de la constitution des races, l'échantillonnage de QTL n'aurait été que partiel dans chaque race et qu'ainsi de nombreux QTL auraient été fixés. Autre possibilité, mais sans doute moins probable, la fixation pourrait être postérieure à la constitution des races et liée à la sélection. Ces interprétations renforcent l'intérêt d'une analyse globale et multiraciale des QTL en ségrégation et des signatures de sélection des QTL fixés dans certaines populations.

Ce constat de QTL souvent différents entre races suggère que la stratégie utilisée lors de la SAM1 présentait un risque important d'utiliser des QTL ne ségrégeant finalement pas dans certaines races. Dans les faits l'ensemble des QTL utilisés pour la race Holstein a été confirmé par le projet. Ceci n'est pas toujours le cas pour les races Normande et Montbéliarde, pour lesquelles

quelques QTL, détectés en race Holstein et utilisés dans la SAM, ne sont pas confirmés dans la présente étude. Ce constat nous conduit à définir des choix de QTL pour la SAM dépendant des races. Dans les conditions actuelles avec un génotypage pangénomique à haute densité, ce n'est pas une contrainte excessive. Cela pourrait le devenir si l'objectif était de réduire le coût du génotypage avec une puce réduite qu'il serait difficile de conserver unique, pour l'ensemble des races et l'ensemble des caractères. L'intérêt de cette approche, a priori très attractive et souvent mentionnée au moins pour un premier tri, serait un peu diminué.

Les estimations de variance expliquée par chaque QTL sont très inférieures à celles estimées précédemment (Boichard *et al.*, 2003; Druet *et al.*, 2006) et utilisées dans le cadre du programme SAM de première génération (Guillaume *et al.*, 2008a). Ainsi, pour le QTL affectant la quantité de lait présent sur le BTA20, proche du gène GHR (Blott *et al.*, 2003), la part de variance génétique initialement estimée par Boichard *et al.* (2003) et utilisée dans le programme était de 10%, tandis qu'elle n'était plus que de 4.1% dans l'étude de Druet *et al.* (2006) et 2.3 % dans la présente étude. Ces variations dans les estimations sont surtout dues aux éléments suivants :

- le premier dispositif était réduit en termes de nombre de familles, ce qui tendait à surestimer les variances ;
- dans la seconde étude de Druet *et al.*, seule l'information paternelle était utilisée, ce qui rend l'échantillon utile de petite taille : la variance due au QTL n'était estimée que par la fraction de pères hétérozygotes et l'effet de substitution intra père hétérozygote ;
- dans la présente étude, la variance due au QTL est estimée à partir des effets et des fréquences de l'ensemble des haplotypes rencontrés dans la population, y compris la population maternelle, son estimation est donc beaucoup plus robuste.

Enfin, une autre raison, plus biologique, peut expliquer ces baisses de variance observées. En effet, il apparaît fréquemment qu'une région initialement large soit complexe et contienne plusieurs QTL. Resserrer la position conduit à focaliser sur un seul des QTL, n'expliquant qu'une partie de la variabilité de la région. Des analyses multi-QTL, permettraient (sous réserve d'une résolution suffisante) de tester l'existence d'autres QTL dans la région (par exemple celle de GHR) et d'affiner l'estimation des parts de variance aux QTL. Ces analyses n'ont pas encore été mises en œuvre de façon systématique faute de temps.

Les conséquences de cette multiplicité de QTL, chacun n'expliquant qu'une faible part de variance, conduit à en intégrer un grand nombre dans la SAM, de façon à disposer du maximum d'efficacité. Ce grand nombre de QTL considérés pour l'évaluation invite, quant à lui, à conserver un modèle simple (tout au moins permettant une évaluation rapide) , nous présentons dans le prochain chapitre la solution adoptée en France pour le programme d'évaluation assistée par marqueurs de seconde génération .

Chapitre 5

Utilisation d'haplotypes pour la sélection assistée par marqueurs

5.1 Introduction

L'utilisation de marqueurs microsatellites assez distants des QTL a limité le gain de précision apporté par l'information moléculaire dans le programme de sélection assistée par marqueurs de première génération. La disponibilité de puces SNP pangénomiques bovines à coût abordable a non seulement permis une localisation fine des QTL mais également l'intégration du déséquilibre de liaison dans le modèle d'évaluation.

Dans un contexte général où d'autres équipes développaient des méthodes de sélection génomique, la stratégie choisie en France a été la sélection assistée par marqueurs. Ce choix est justifié par les arguments suivants :

- le modèle SAM se base sur des QTL connus et il est donc supposé robuste ;
- il est efficace même pour des populations de taille moyenne comme la Normande et la Montbéliarde qui ne disposent pas de grandes populations de référence comme la Holstein ;
- il constitue une évolution logique par rapport au programme antérieur depuis 2001, d'où une meilleure maîtrise de sa mise en œuvre et ses propriétés ;
- il est relativement peu exigent en temps de calcul, sachant qu'il doit être opérationnel pour plusieurs races et de nombreux caractères, avec des traitements fréquents.

Le modèle d'évaluation retenu dans cette optique comprend donc une composante polygénique et une somme d'effets haplotypiques aux QTL.

$$y_i = u_i + \sum_{j=1}^{N_{QTL}} (h_{ij}^p + h_{ij}^m) + e_i \tag{5.1}$$

où la valeur génétique g_i de l'individu i est la somme d'un effet polygénique u_i, et des effets haplotypiques paternels h_{ij}^p et maternels h_{ij}^m pour les différents QTL.

Les effets haplotypiques de deux animaux quelconques dans la population sont considérés identiques si les haplotypes observés à un QTL sont identiques par état (identity by state, IBS). De ce fait, on considère l'existence d'un fort déséquilibre de liaison entre haplotype et allèle au QTL. Cette hypothèse se justifie par la taille efficace réduite de la population évaluée, la longueur (moins de 0.7 cM en moyenne) des haplotypes utilisés et le polymorphisme important permis par les haplotypes. L'avantage de ce modèle est de considérer un nombre réduit d'effets à calculer pour chaque QTL permettant une évaluation rapide. De plus, le calcul de probabilités de transmission n'est plus nécessaire : on les suppose égales à un si deux haplotypes sont identiques par état et à zéro sinon .

Le modèle 5.1 peut être vu comme une évolution du modèle 2.1 où les effets gamétiques seraient remplacés par des effets haplotypiques, la logique sous-jacente étant que l'on peut tracer sans erreur l'origine de l'allèle au QTL, jusqu'à un fondateur commun à tous les individus portant l'haplotype observé. Un autre parallèle peut être fait, avec le modèle de sélection génomique 2.9 : en effet le modèle utilisé peut être vu comme un modèle de sélection génomique dans lequel un nombre modéré (le nombre de QTL) de fragments chromosomiques est évalué. Les autres fragments étant considérés nombreux et chacun de variance faible, ils sont modélisés par le biais d'un terme polygénique. La variance de chaque effet haplotypique est supposée connue, car estimée préalablement lors de la détection de QTL. Ainsi, les similitudes existant entre les différents modèles d'évaluation assistée par marqueurs laisse présager d'une certaine similitude entre résultats d'évaluation.

Objectifs Cet article a pour but de vérifier la qualité de ces hypothèses à l'aide de simulations et de données réelles, et de mesurer la précision des prédictions. La qualité du calcul du coefficient de détermination des évaluations sera également analysée, afin d'évaluer les meilleurs indicateurs de fiabilité disponibles.

5.2 Article 5

Article soumis à Genetic Selection Evolution

Efficiency of Marker Assisted Selection with dense Single Nucleotide Polymorphism maps in French Dairy Cattle.

François. Guillaume [1,2*§], Didier Boichard[2], Tom Druet[3] and S. Fritz[4],

[1]Institut de l'élevage, 149 rue de Bercy, 75595 Paris Cedex 12, France

[2]INRA, UMR1313 Génétique Animale et Biologie Intégrative, F-78352 Jouy en Josas

[3]Unit of Animal Genomics, Faculty of Veterinary Medicine and Centre for Biomedical Integrative Genoproteomics, University of Liège (B43), Belgium.

[4]Union nationale des coopératives d'élevage et d'insémination animale, 149 rue de Bercy, 75595 Paris Cedex 12, France

[§]Corresponding author

Email addresses:

FG : Francois.guillaume@jouy.inra.fr

Abstract

Background

Use of high throughput SNP genotyping provides new opportunities to enhance classical marker-assisted selection, as an alternative to genomic selection. Indeed, dense markers improve the estimation of QTL transmission probability from parents to progeny as well as the estimation of identity-by-descent probability for QTL between founders. Therefore it should increase estimated breeding value reliability. This study based on a real data set in dairy cattle was aimed at quantifying the efficiency of this approach.

Methods

A population of 1592 Holstein sires were genotyped for 54000 SNP. The subset of the 468 youngest bulls were used for validation. Polygenic pedigree index and marker-assisted prediction (MA-EBV) were compared to the results after progeny testing. Both simulation and real data were used. In simulations, the population structure was kept and true marker haplotypes were used to generate biallelic QTL and flanking marker information. For the evaluation, haplotypes of 2 to 10 markers were used.

Results

MA-EBV outperformed classical polygenic EBV. The greatest increases of reliabilities were obtained for haplotypes of 4 SNP in simulations or 6-8 SNP in real data. Real data results were in agreement with their expectation from simulations. Estimation of reliability from the inverse of the coefficient matrix appeared to be inadequate and is a critical point which deserves further investigation.

Conclusions

This classical two-step approach based on QTL fine-mapping before evaluation appeared to be efficient and could be an attractive alternative to genomic selection, at a low computing cost and with biological interpretation of results.

Background

A large marker-assisted programme (MAS) has been carried out since 2001 by eight French breeding companies, in collaboration with INRA and Labogena. This first generation MAS programme used a limited number (45) of microsatellite markers to follow 14 QTL regions [1]. Linkage equilibrium was assumed at the population level and only within-family information was used on breeding value prediction. More than 70 000 animals were genotyped over 8 years, generating a large database used to confirm the QTL and accurately estimate their variance components [2]. Efficiency of MAS was assessed for dairy traits [3,4].

Denser marker maps are now available and new technologies offer the opportunity to genotype animals for many markers spanning the entire genome. According to Meuwissen et al. [5], genomic selection has become very popular and is shown to be efficient to predict breeding values. However, this genomic information could also be used to enhance marker-assisted selection efficiency. Indeed, since many QTL are finely mapped, QTL transmissions could be accurately followed across generations and linkage disequilibrium could be used to link information across families. MAS accounts only for identified QTL in contrast with genomic selection, but it could make a better use of each QTL. We believe that both approaches could be unified if MAS uses many QTL and if genomic selection is based on marker haplotypes and uses prior information on known QTL.

The aim of this study was to assess the efficiency of a marker-assisted selection when dense marker information was used, both by simulation and use of real data. In addition, the properties of several reliability indicators were investigated.

Methods

French MAS Data

A population of 1592 Holstein sires has been genotyped with the Illumina 54 k SNP chip. This population was divided into two parts: a training sample of 1124 bulls, born between 1998 and 2002 and with a progeny-test evaluation available before 2007; and a validation sample of 468 candidate bulls with their first progeny-test evaluation available in 2008. The pedigree was extended to three generations of ancestors, resulting in a population of 6219 individuals.

For prediction, phenotypic information was that available in 2004, i.e. when the candidates were selected based only on pedigree information. For validation, the information available in 2008 was used. The phenotypic information was twice the daughter yield deviations (DYD, VanRaden and Wiggans, 1991 [6]) for milk yield, protein percentage and somatic cell counts. DYD used phenotypic data from the first three lactations of the daughters. DYD and their corresponding weights were obtained from the French national evaluation of April 2004 [7]. The DYD of candidates was also obtained from the June 2008 evaluation.

Simulation

In the simulation step, the pedigree and the performance weights were conserved, whereas the true records and the genotypes were simulated as follows.

QTL parameters were those estimated on the true data. Between 19 and 32 QTL were used for each trait and the variance components are shown in table I.

The genetic effect of animal i is computed as :

$$g_i = u_i + \sum_{j=1}^{n\_qtl} (v_{ij1} + v_{ij2}) \quad [1]$$

where u_i is the polygenic effect of individual i (excluding QTL effect), v_{ij1} and v_{ij2} are allelic effects at QTL j for the paternal and maternal alleles, respectively, and n_qtl is the number of QTL.

For animals with unknown parents, the polygenic effect was sampled from $N(0, \sigma_u^2)$ while for animals with known parents, the polygenic effect was the sum of the mean polygenic effects of the parents and the Mendelian sampling drawn from a normal distribution with the variance adjusted for the number of known parents. The polygenic variance (σ_u^2) was defined according to heritability of the traits and the proportion of genetic variance explained by QTL (Table I).

Each QTL was assumed to be biallelic and the frequency was estimated from the percentage of heterozygous sires in the actual MAS population. Its substitution effect was then derived from the simulated QTL variance and the allelic frequency [8]. In order to generate a marker and QTL structure reflecting the real situation, the following two-step procedure was used.

In a first step, the maternal phases of the 1592 bulls for 2631 SNP on chromosome BTA1 were predicted by DualPHASE [9]. This information was then used in the simulations. Each QTL was defined by a given real SNP selected randomly only on the basis of its allelic frequency. Each QTL region was defined by 11 loci, i.e. the QTL SNP itself and 5 SNP upstream and downstream.

In a second step, genotypes were assigned in the pedigree. For each founder and each QTL, two haplotypes were sampled from the collection of haplotypes. Then haplotypes were transmitted from parent to progeny following Mendelian rules, allowing recombination according to the genetic distance between markers.

The records were simulated as the sum of the genetic effect and a residual with zero mean and variance depending on the weight.

In addition, records of male candidates were simulated according to their DYD weight in 2008. Simulations were repeated 100 times for each trait.

Polygenic and MAS evaluation

Two models were used in this study. The first one was a polygenic model, assuming no QTL. In the second model, a regression on an identical-by-state (IBS) haplotype groups was added :

$$\mathbf{y}=\mathbf{X\beta}+\mathbf{Zu}+\sum_{i=1}^{n\_qtl}\mathbf{Z}_{h_i}h_i+\mathbf{e} \quad [2]$$

where **y** is a vector containing records, **β** is a vector of fixed effects (the mean), **u** is a vector of random polygenic effects, $\mathbf{h}_i$ is a vector of random haplotype effects for QTL i and **e** is a vector of random residual terms. **X**, **Z** and $\mathbf{Z}_{h_i}$ are known design matrices relating records to fixed, random polygenic and haplotypic effects, respectively. Different sizes of haplotypes were tested with 2, 4, 6, 8, and 10 markers. The central SNP of the haplotype, which was assumed to be the QTL, was disregarded. All variances were assumed to be known, although it has been observed they have limited impact on breeding value prediction (data not shown).

Estimation of efficiency of MAS

Since the efficiency of the haplotype-based model relies on the capacity of haplotypes to estimate QTL effects properly, we investigated the relations between haplotypes and QTL. The linkage disequilibrium between QTL and haplotypes was measured on the simulated data set according to equation 3 [10]:

$$r^2(h,q)=\left(\sum_{i=1}^{n}\frac{D_i^2}{p_i}\Big/ q_1q_2\right) \quad [3]$$

where $D_i^2 = (pq_i - p_i q_1)^2$ is the disequilibrium between haplotype i and allele 1 of the QTL, pq_i is the frequency of haplotype i carrying allele 1 at the qtl , p_i is the frequency of haplotype i and q_1, q_2 are the frequencies of alleles 1 and 2 at the QTL, respectively.

On real data sets, true EBV are not known and accuracy is generally approximated based on the terms of the inverse of the MME, whereas with simulations, reliabilities of EBV can be obtained from true genetic values. Therefore, in the simulations we compared the estimated and true reliabilities to assess the goodness of approximation. We furthermore tested an approximation of reliability based on the correlation between MA-EBV and DYD after progeny testing, as in Van Raden [11].

$$R^2 = \frac{corr(DYD_{2008}, EBV_{2004})^2}{E(R^2(EBV_{2008}))} \quad [4]$$

Finally, since the best genetic indicator available after progeny testing is DYD,
we compared the correlation really obtained between 2004 EBV and 2008 DYD for the candidates to the corresponding figures obtained by the simulations.

Results

Simulated data

A relatively small number of haplotypes was observed in the population (Table II), so that the number of haplotype effects to estimate remained small. For instance, 39 haplotypes were observed on average with 10-SNP haplotypes, to compare with 2^{10} theoretically possible haplotypes. Linkage disequilibrium between QTL and haplotypes increased asymptotically with haplotype length (from 0.331 to 0.427) but it hardly increased (< 0.01) for haplotypes larger than four markers. The observed correlations between estimated haplotype effects and simulated gene effects among non-candidates were high, ranging from 0.83 to 0.92. Haplotypes of four SNP gave

the most accurate estimates for the three traits, whereas haplotypes of two SNP gave the less accurate estimates.

Among candidates, despite the higher linkage disequilibrium (Table III) which was roughly two times higher than in the whole population, and an overrepresentation of the most frequent haplotypes, the accuracy of the haplotype effect estimation was a bit lower than among non-candidate populations.

MA-EBV outperforms classical EBV for traits under studies, whatever their heritability. Reliabilities ranged from 0.47 to 0.61 for MA-EBV and from 0.30 to 0.35 with classical EBV. From all tested haplotype lengths, haplotypes of four SNP resulted in the most accurate EBV for all traits (Table IV).

The reliability obtained from the inverse of the mixed model equations varied much less (ranging from 60.8 to 63.1) than the true reliability (ranging from 47.5 to 54.9), and tended to favour short haplotype models because fewer haplotype effects were estimated.

Although a bit underestimated, reliabilities computed based on the correlation between DYD and EBV, better fitted the true reliabilities.

Field data

The correlations between MA-EBV and DYD after progeny testing were always higher than correlations between DYD and polygenic EBV as expected (Table V). Haplotypes of six SNP gave the best correlation for milk yield and somatic cell count, whereas the haplotypes of eight SNP gave the best result for protein content. In all cases, haplotypes of two SNP gave the worst results. More variations in the results of different haplotype lengths were observed with real data as compared to simulation data. Based on simulations, the correlations obtained from haplotypes of different lengths were quite homogeneous with a maximum difference of 0.033 for milk yield,

0.022 for protein content, and 0.039 for somatic cell count. The corresponding correlations obtained from real data were more heterogeneous, with maximum differences of 0.073, 0.055 and 0.052. Compared to the polygenic model, the highest gains in correlation were 0.199 for protein percentage and haplotypes of eight SNP, 0.197 for milk yield with haplotypes of six SNP and 0.102 for somatic cell count with haplotypes of six SNP. The observed gains due to MA-EBV were greater than their expectation based on simulations for protein percentage and milk yield whereas they were lower than their expectation for SCS.

Discussion

Compared to a similar study using Fernando-Grossman's model and a low density micro-satellite marker map [3,4], gains in terms of reliability were much higher for protein content and milk yield, although the number of genotyped individuals considered was 5 times smaller and the proportion of total genetic variance was also smaller. These findings emphasize the greater potential gain allowed by the use of high density SNP maps even with a smaller number of genotyped animals.

Accuracies obtained were higher than in previous simulations where LD was obtained by a simulation of random drift [6,12]. This result is due to the higher density of markers used and the quantity of phenotypes considered.

Here, we used both a real pedigree structure and true founder haplotypes making it possible to fix QTL position and frequency, in contrast with other simulation studies [6,13]. The marker density used here (on average 12 SNP per cM) was fairly high compared to most simulation studies, but is a realistic density with the current technologies.

Similar gains on field data have been reported by Van Raden et al. [11] with a genomic selection method considering the whole SNP data to compute MA-EBV.

This two-stage approach gave quite good results at low computational costs and with a limited training population as compared to genomic selection.

Regarding the number of QTL and their contribution to total genetic variance, the parameters changed drastically compared to our former validation study. All the QTL previously used were kept but their contribution to genetic variance decreased based on new estimations of their contribution with a LD-LA model. Numerous new QTL were detected and added to the model, nevertheless they generally explained a small proportion of total genetic variance (<2%). This was particularly the case with SCS where only one QTL explained more than 2 % of total genetic variance. This particular situation, i.e. numerous QTL explaining a small part of total genetic variance, may explain the lower results obtained with SCS.

The main idea in using linkage disequilibrium in the MAS model is that markers can be a proxy to QTL alleles. In these simulation studies, the correlation observed between true breeding values and their estimation by use of markers were close to what could be expected from a gene-assisted selection (GAS) model (data not shown), meaning that the haplotype effect fits the gene effect properly. On the contrary to other simulations studies, the results tend to show that haplotypes of four SNP give better results than smaller haplotypes due to the higher linkage disequilibrium obtained. For haplotypes longer than four SNP, estimation errors of the haplotype effect tend to decrease the benefits from higher linkage disequilibrium. Field data indicate that optimal length of haplotypes tends to be longer than with simulations (6 or 8 SNP). A possible explanation is the number of QTL alleles, assumed to be 2 in simulations. It is possible that a larger number of marker haplotypes better describes the diversity of QTL alleles.

Linkage disequilibrium is expected to decrease over generations and limits the prediction accuracy. In the simulation study, this was not observed because the regions under study were short (0.8 cM on average) and the population was restricted to a limited number of generations with genotypes. Furthermore, due to the limited number of sires in the candidate population, the frequency of the most common haplotypes tended to increase and led to a better prediction of the candidates. Anyway, validation studies (even based on a candidate population) are to be taken with caution, because candidates are not a fully representative subset of the learning population.

The theoretical reliability derived from the inverse of the mixed model equations overestimated the true reliability (table IV). In fact this reliability would be correct if the haplotype–QTL disequilibrium was complete, which was not the case. This was especially clear with haplotypes shorter than 4 SNP: a smaller number of haplotypes gave less opportunity to address QTL alleles. Furthermore, theoretical reliabilities decreased with haplotype length and, therefore with the number of haplotypes. This trend, however, was not confirmed by the true reliability computed from simulations. More work is needed to properly estimate marker-assisted selection reliability.

Moreover, the theoretical reliability appears to be very sensitive to the prior values of QTL variances used in the evaluation. As an illustration, Table VI shows that average theoretical reliabilities ranged from 0.56 to 0.70, when QTL variances decreased or increased by 25% in the evaluation model. The corresponding figures observed on correlation between DYD and MA-EBV were much more stable around 0.60 (Table VI). It can be concluded that prior variance parameters have little effect on the realised accuracy of breeding value prediction, whereas it strongly affects the estimation of reliability.

Finally, it should be noted that MAS efficiency could be improved by customizing some parameters to each QTL. For instance, in the routine MAS evaluation, haplotypes are defined based on the analysis of heterozygous sires in the population. In fact, for most QTL the observation of haplotypes and substitution effect among sires is relevant for the definition of haplotypes. This strategy may help to discard uninformative SNP and define more sensible haplotypes to get better results (Table VII).

Conclusions

This preliminary study gave encouraging results, and supports the idea that even with a simple model, use of a dense SNP marker map allows for a drastic improvement in terms of reliability. The results obtained were similar to those achievable with genomic selection but were much less intensive in terms of computing. Reliabilities obtained within the range of expected values should still be improved.

Competing interests

None declared

Authors' contributions

FG performed the analysis and draft of the manuscript; DB revised the manuscript and supervised the analysis; TD revised the manuscript and conceived most of the software used in this study ; SF performed the analysis and gave helpful advise for the manuscript .

Acknowledgements

The authors gratefully acknowledge UNCEIA, LABOGENA and INRA for the access to the MAS program data. FG acknowledge the French Ministry of Agriculture for financial support through a CASDAR grant.

References

1. Boichard D, Fritz S, Rossignol MN, Boscher MY, Malafosse A, Colleau JJ: **Implementation of marker-assisted selection in French dairy cattle**. *In Proceedings of the 7th World Congress on Genetics Applied to Livestock Production, August 19-23 2002, Montpellier,* Edited by Elsen JM, Ducrocq V, 2002, 22-03.

2. Druet T, Fritz S, Boichard D, Colleau JJ: **Estimation of Genetic Parameters for Quantitative Trait Loci for Dairy Traits in the French Holstein Population**. *J Dairy Sci* 2006, **89**: 4070-4076.

3. Guillaume F, Fritz S, Boichard D, Druet T: **Estimation by Simulation of the Efficiency of the French Marker Assisted Selection Program in Dairy Cattle**. *Genet Sel Evol* 2008, **40**: 91-102.

4. Guillaume F, Fritz S, Boichard D, Druet T: **Correlations of marker-assisted breeding values with progeny-test breeding values for eight hundred ninety-nine French Holstein bulls.** *J Dairy Sci* 2008, **91**:2520–2522.

5. Meuwissen THE. Hayes BJ. Goddard ME: **Prediction of total genetic value using genome-wide dense marker maps.** *Genetics* 2001, **157**, 1819-1829.

6. Van Raden. PM, Wiggans GR: **Derivation. calculation. use of national Animal Model Information.** *J Dairy Sci* 1991, **74**, 2737-2746.

7. Robert-Granié C, Bonaïti B, Boichard D, Barbat A: **Accounting for variance heterogeneity in French dairy cattle genetic evaluation.** *Livest Prod Sci* 1999, **60**: 343-357.

8. Falconer DS, Mackay TFS: *An Introduction to Quantitative Genetics.* Longman Group, Essex, UK; 1996.

9. Druet T, Coppieters W, Farnir F, Georges M: **Large scale haplotype reconstruction program using pedigree information a Hidden Markov Model.** *In Proceedings of the XX International Congress of Genetics, Berlin, 12-17 July 2008*.

10. Hayes B, Chamberlain AJ, Goddard ME: **Use of markers in linkage disequilibrium with QTL in breeding programs**. *In Proceedings of the 8th World Congress on Genetics Applied to Livestock Production, Belo Horizonte, August 13-18 2006.* 30-06.

11. VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, Schenkel FS: **Reliability of genomic predictions for North American Holstein bulls** *J Dairy Sci* 2009, **92**: 16-24.

12. Calus MPH, Meuwissen THE, de Roos APW, Veerkamp RF: **Accuracy of Genomic Selection Using Different Methods to Define Haplotypes**. *Genetics* 2008, **178**, 553-561.

13. Solberg TR, Sonesson AK, Woolliams JA, Meuwissen THE: **Genomic selection using different marker types density.** *In Proceedings of the 8th World Congress on Genetics Applied to Livestock Production, Belo Horizonte, August 13-18 2006.* 22–13.

Tables

Table 1 - Parameters used for simulations and MA-Evaluations

Traits	Heritability	Genetic variance explained by QTL	Nb QTL	Number of QTL explaining more than 2% of total genetic variance	Total number of SNP used in the haplotypes
Protein %	0.50	0.41	19	5	73
Milk yield	0.30	0.46	26	6	103
SCS	0.15	0.46	32	1	128

Table 2 - Description of haplotypes in the population

Haplotype length	Mean number of haplotypes	r^2	Observed accuracy
2	4.0	0,331	0,834
4	8.2	0,405	0,922
6	18.2	0,412	0,896
8	27.6	0,421	0,881
10	39.0	0,427	0,876

Figures were averaged over the three traits under study for 100 simulations each. Linkage disequilibrium between QTL allele and haplotypes (r^2) were computed according to equation 2. Accuracy is the correlation between estimated haplotype effect and simulated gene effect.

Table 3 - Haplotype effect accuracy in the candidate population

	Observed accuracy among candidates			Estimated LD (r^2) among candidates		
Nb SNP	Prot %	Milk	SCS	Prot %	Milk	SCS
2	0.767	0.784	0.786	0.763	0.764	0.763
4	0.844	0.871	0.875	0.937	0.938	0.937
6	0.788	0.821	0.824	0.953	0.953	0.953
8	0.757	0.795	0.796	0.974	0.974	0.975
10	0.747	0.788	0.796	0.989	0.989	0.989

Figures were averaged over 100 simulations.

Table 4 - Comparison of reliability obtained over 100 simulations in the candidate population.

	Protein %			Milk yield			SCS		
model	True R^2	DYD	MME R^2	True R^2	DYD	MME R^2	True R^2	DYD	MME R^2
polygenic	35.6	34.4	48.0	33.0	31.0	39.0	30.0	29.8	36.6
2 SNP	47.6	46.1	63.0	54.9	51.4	63.1	54.9	54.4	62.0
4 SNP	50.4	49.1	62.8	60.3	56.4	62.9	61.2	60.7	61.7
6 SNP	48.7	47.1	62.5	57.9	54.1	62.6	58.2	57.8	61.2
8 SNP	47.7	46.3	62.3	56.6	52.9	62.3	56.7	56.3	60.9
10 SNP	47.5	46.0	62.2	56.2	52.5	62.2	56.2	55.9	60.8

True R^2 were obtained by squaring the correlation between true breeding value and estimated breeding value. DYD refers to reliability assessed according to equation 4. Finally, MME R^2 were derived from the inverse of the mixed model equations.

Table 5 - Correlations between polygenic and MAS EBV based on 2004 information and DYD of 2008 for the 468 candidates

	Protein %		Milk yield		SCS	
	expected	observed	expected	observed	expected	observed
Polygenic	0,572	0,523	0,543	0,400	0,518	0,479
2	0,662	0,649	0,699	0,542	0,700	0,529
4	0,683	0,690	0,732	0,591	0,739	0,576
6	0,669	0,706	0,717	0,597	0,721	0,581
8	0,663	0,722	0,709	0,574	0,712	0,570
10	0,661	0,707	0,706	0,554	0,709	0,567

Correlations are weighted by 2008 DYD weight. Expected correlations rely on the average correlation over 100 simulations. Observed correlations were obtained based on field data

Table 6 - Comparison of theoretical reliabilities and correlation between MAS EBV and DYD when QTL variance vary.

	Theoretical reliabilities			Observed Correlation		
	%prot	Milk yield	SCS	%prot	Milk yield	SCS
-25%	0,564	0,557	0,527	0,706	0,601	0,584
-10%	0,598	0,599	0,572	0,706	0,599	0,583
0	0,621	0,627	0,602	0,706	0,597	0,581
+10%	0,645	0,655	0,633	0,704	0,594	0,580
+25%	0,680	0,699	0,680	0,702	0,588	0,576

Figures were obtained based on the candidate population for a 6 SNP model.

Table 7 - Highest correlations for the 468 candidates between polygenic and MAS EBV based on 2004 information and DYD of 2008, weighted by 2008 DYD weight.

	Simulation			Field data		
			Correlation			Correlation
	Pol	MAS	Gain	Pol	MAS	Gain
Protein %	0.572 (0.06)	0.683(0.04)	0.111	0.523	0.737	0.214
Milk yield	0.543 (0.06)	0.732(0.04)	0.189	0.406	0.624	0.218
SCS	0.518 (0.06)	0.739(0.05)	0.221	0.479	0.608	0.129

5.3 Conclusion

Les résultats obtenus confirment les gains substantiels permis par l'utilisation de cartes denses. Du fait de problème de récupération de typages, les effectifs de taureaux de validation a été légèrement plus faible qu'initialement prévu et un effectif plus conséquent aurait été appréciable, de nouvelles validations devraient permettre de conforter ce résultat sur un échantillon plus important. De même, de nouvelles études sur un échantillon de données plus important devraient permettre de confirmer l'existence de certains QTL, actuellement non intégrés dans les modèles d'évaluation, ainsi que d'affiner les parts de variance attribuées à chaque QTL. On peut donc considérer que les gains de précision présentés puissent encore être améliorés.

Le choix de poser un modèle n'utilisant plus l'analyse de liaison, n' a pas été discuté dans le précédent article. L'utilisation d'un modèle Fernando-Grossman avec des marqueurs SNP a été testée à l'aide de simulations (Guillaume *et al.*, 2007a) , dans un cadre identique à celui testé dans le premier article de cette thèse (population candidate de 2006), mais en substituant l'information des marqueurs microsattelites par 10 SNP dans un intervalle de 1 cM. Les CD attendus apportaient déja une amélioration (cf. Table 5.1) par rapport à ceux obtenus avec des microsattelites de la SAM. Ces résultats ont été confirmés (Guillaume *et al.*, 2008b), sur la population candidate du dernier article, mais en ne considérant que 4 QTL pour l'évaluation du lait. Ces résultats auraient pu être amélioré par l'intégration du déséquilibre de liaison au niveau des effets gamétiques fondateurs. Le manque de temps, nous a malheureusement contraint à restreindre les modèles d'évaluation testés, aux modèles simples et faciles à mettre en place. De plus, les corrélations obtenues avec le modèle SAM2 étant nettement supérieures, ce dernier modèle à donc été rapidement privilégié. Une comparaison de modèles plus exhaustive (incluant notamment des modèles de type sélection génomique) sera mise en œuvre à courts termes dans le cadre du projet AMASGEN.

Tout comme les deux premiers articles de cette thèse, les résultats se basent sur un effectif restreint. Dans la pratique les volumes de données croissant régulièrement (1500 nouveaux génotypages par mois) , les gains de précision devront s'en trouvé améliorés. Le comportement des méthodes d'évaluation génomique lorsque le volume de données à traiter augmente est un point essentiel qui pèse énormément dans l'intérêt de chaque méthode. Si les résultats reportés dans la littérature jusqu'à présent semblent indiquer des gains de précision de l'ordre de ce qui est observé dans cette étude, il sera intéressant de vérifier si cet état de fait demeure vrai avec l'accumulation de donnés génotypiques.

Une difficulté majeure rencontrée dans cet exercice de validation est la mise en place de critères fiables d'évaluation de la précision des valeurs génétiques. Or, en fonction de la méthode de validation utilisée et de la population de référence utilisée, de grande variations peuvent être observées (Legarra *et al.*, 2008). La validation telle que nous l'avons menée dans cette étude se calque sur une situation pratique et répond bien aux questions des unités de sélection lors du choix des taureaux à mettre en testage . Un biais est néanmoins possible puisque les animaux de validation sont très fortement apparentés aux animaux ayant servi à la détection de QTL. Dans

la mesure où les index SAM2 ne seront pas utilisés que pour le choix des animaux à mettre en testage et que des animaux moins apparentés à ceux du dispositif CartoFine devront être évalués, la méthode de validation de l'outil SAM2 devra être complétée ou adaptée. Un travail important d'inventaire et d'analyses des différents outils possibles reste à mettre en œuvre pour permettre une comparaison juste des différentes méthodes d'évaluation.

	Polygénique	SAM	SAM SNP
Lait	0,313	0,361	0,391
MG	0,310	0,373	0,420
MP	0,303	0,340	0,374
TB	0,342	0,453	0,508
TP	0,342	0,418	0,455

TABLE 5.1 – CD moyens des index obtenus selon 3 modèles différents (polygénique, SAM microsattelites, SAM SNP (10SNP-1cM)

Chapitre 6

Discussion générale

Depuis une décennie, les avancées de la génomique ont ouvert des perspectives nouvelles dans la connaissance du génome et tout particulièrement pour le monde de la sélection. Nous présenterons premièrement les impacts des typages à haut débit sur la détection de QTL ainsi que les voies de progrès dans ce type d'analyses. Nous discuterons dans un second volet l'évolution des outils de sélection assistée par marqueurs pour finir sur les impacts que pourront avoir ces outils sur la filière de sélection.

6.1 Cartographie de QTL

Tout d'abord, les résultats de détection de QTL ont radicalement changé en l'espace de quelques années grâce à l'apport des données de typage à haut débit (cf. figure 6.1). Les différences observées sont dues d'une part à l'information marqueurs beaucoup plus abondante et d'autre part aux méthodes de détection de QTL, dont les hypothèses sous-jacentes, avec l'apport du déséquilibre de liaison, réduisent l'auto-corrélation entre tests et fournissent une bien meilleure résolution (Tarres *et al.*, 2009). Avant la disponibilité de puces de SNP, la détection de QTL suivait un processus en deux étapes : une première couverture du génome avec des marqueurs à basse densité permettait une primolocalisation des QTL dans des régions larges (généralement quelques dizaines de cM pour des QTL d'effet moyen) ; cette primolocalisation était suivie d'une étape de cartographie fine par densification en marqueurs afin d'affiner la localisation du QTL, si possible dans un intervalle suffisamment restreint pour orienter efficacement le choix de gènes candidats. Avec les puces de haute densité, les deux phases sont maintenant confondues puisque la primolocalisation est d'emblée très précise. Rançon du succès, le nombre de QTL est maintenant très élevé, demandant de nouvelles stratégies pour l'identification des gènes et mutations causales, par exemple par capture de séquence et séquençage massif. En attendant, des stratégies sont

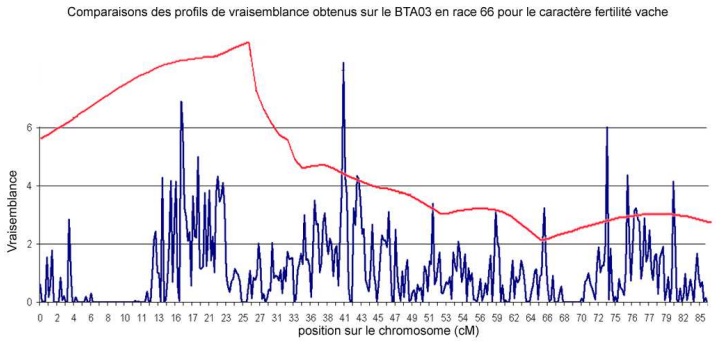


FIGURE 6.1 – Profils de détection de QTL de fertilité sur le BTA03 en race prim'Holstein obtenus : par régression intra père à l'aide de quelques des microsatellites (en rouge)(Guillaume *et al.*, 2007b) ; ou par analyse LD-LA avec des SNP dans le cadre de cartofine (en bleu)

nécessaires pour limiter le nombre de faux positifs et exploiter au mieux l'ensemble des régions QTL détectées.

D'un point de vue pratique, les volumes de données à traiter et les temps de calculs deviennent des facteurs très limitants. À titre d'illustration, l'ensemble des détections LDLA du projet CartoFine représente plus d'un an de calcul. Paradoxalement, la réduction du volume de données à traiter devient donc primordiale. Ainsi, lors de la préparation des données de génotypage, il est d'usage de supprimer les marqueurs trop peu informatifs ou n'étant pas en équilibre de Hardy-Weinberg. Les impacts que peuvent avoir ces pratiques sur les résultats n'ont pas été analysés. Lors de la détection de QTL, afin de réduire le nombre de tests à effectuer, un marqueur sur deux a été testé, faisant l'hypothèse que du fait de l'utilisation de probabilités IBD sur 10 marqueurs, des tests consécutifs sont très corrélés. En pratique, on constate que le profil de la statistique de vraisemblance est très erratique, le choix de ne traiter qu'un marqueur sur deux est donc discutable. De façon plus générale, l'ensemble des simplifications faites lors des analyses devront être revisitées pour en mesurer l'importance réelle.

Enfin, la question des règles de décision est un point critique découlant des deux points précédents. Les hypothèses testées généralement sont l'existence d'un QTL contre l'absence de QTL, ce test est réalisé point par point. Les seuils de test calculés n'ont normalement de sens que lorsqu'ils sont appliqués au maximum de vraisemblance sur chaque chromosome, ce qui implique qu'une analyse complète du génome ne devrait permettre de détecter qu'un QTL par chromosome et par caractère. Bien sûr, l'excellente résolution permet généralement de distinguer plusieurs QTL dès lors que leur distance est suffisante. Il n'en demeure pas moins qu'une analyse multiQTL est nécessaire. L'approche adoptée pour le projet CartoFine a été d'intégrer les effets de plusieurs QTL par chromosome lorsque le le profil de vraisemblance le suggérait. Elle n'a donc été conduite que dans un nombre limité de situations où elle s'est montrée particulièrement efficace, cartographiant finement et distinguant aisément des QTL distant de moins de 10 cM. . Cette démarche bien qu'apparemment efficace n'est pas rigoureuse, il conviendrait de mettre en

place une stratégie de détection multiQTL systématique si l'on voulait détecter et cartographier finement le maximum de QTL, mais cette démarche rallongerait d'autant le temps nécessaire pour effectuer l'analyse complète du génome et, dans les conditions de cette thèse, n'aurait pas été réaliste. Un dernier aspect du problème des règles de décision vient de l'absence de méthode de calcul des seuils de significativité réellement satisfaisante. Le choix d'une loi de χ^2 à 2 degrés de liberté sous H0 est discutable même si elle constitue sans doute une bonne approximation. Le recours à des simulations entraîne de gros problèmes de calculs (à la fois de convergence et de temps) tandis que les méthodes approchées (Piepho, 2001) supposent des profils à variations continues et sont peu adaptées aux profils hachés observés. D'important progrès sont donc encore nécessaires afin de permettre une meilleure utilisation des résultats de détection de QTL issus d'analyses LDLA.

Les puces denses de SNP constituent un immense progrès dans la détection de QTL. Même si les méthodes pour en tirer le maximum d'information ont été adaptées et ont fourni de nombreux résultats nouveaux et originaux, il faut sans doute les revisiter pour en tirer le meilleur parti et les adapter au haut débit, pour traiter rigoureusement les questions des seuils de rejet, de la multiplicité des tests, de la stabilité numérique et des temps de calculs.

6.2 Bilan du premier programme SAM

Le premier programme de sélection assistée par marqueurs souffre maintenant de la comparaison avec les programmes mis en place aujourd'hui, il est donc important de souligner rétrospectivement son intérêt. En effet, au début des années 2000, ce programme était une prouesse technique et organisationnelle, ce qui a limité le nombre d'exemples réellement développés et de grande ampleur. Bien sûr, l'efficacité était relativement réduite, en accord avec les études préalables (Ruane et Colleau, 1995). Mais elle était suffisante chez les bovins laitiers pour rentabiliser l'effort consenti. En effet, le coût du programme est très élevé (40 000 euros par taureau testé) de sorte que le surcoût de la SAM ne représente que 3 à 5% du programme complet. Les résultats de validation obtenus confirment que la SAM, telle qu'elle a été mise en œuvre, est plus précise qu'une évaluation polygénique. Le programme permettant d'éliminer 15% des jeunes taureaux avant testage sans perte de progrès génétique, il est financièrement rentable. Ce résultat peut cependant être considéré comme une exception car dans la majorité des filières, un programme de ce type n'aurait pas été rentable.

La population génotypée dans le cadre du programme a permis d'orienter les travaux de cartographie fine de QTL en fournissant des populations de grande taille et des informations de primo-localisation très abondantes. Un programme de sélection assistée par marqueurs ne se résume pas à une simple évaluation. Un travail important de logistique (organisation de la collecte des génotypes, identification des animaux à génotyper, stockage du matériel biologique) est nécessaire. Un tel dispositif existe en France depuis 2001, ce qui explique que le passage à l'évaluation assistée par marqueurs de seconde génération s'est faite sans discontinuité. De même,

la filière a intégré de longue date les changements nécessaires dans les programmes de sélection, même si le choc culturel de l'arrêt du testage est nouveau pour la majorité des acteurs.

6.3 Évaluation SAM2 et évolutions possibles

Les résultats de validation de la SAM2 indiquent clairement un gain de précision et ce à partir des premiers effectifs disponibles qui restent modérés. Ces gains semblent atteindre des niveaux équivalents à ceux cités dans la littérature, avec d'autres méthodes d'évaluation, même si la comparaison reste actuellement sujette à caution du fait de l'hétérogénéité des méthodes et de l'absence de règles de validation internationalement reconnues. Une force du modèle de sélection assistée par marqueurs de deuxième génération est de reposer sur des connaissances éprouvées. Les parts de variance des QTL précédemment identifiés ont été revues à la baisse. Mais l'ajout d'un grand nombre de nouveaux QTL au modèle SAM 2 permet d'expliquer plus de la moitié de la variance génétique dans chaque population. Bien sûr, une minorité de ces nouveaux QTL peuvent être des faux positifs mais l'accumulation de nouveaux génotypes et phénotypes permettra de les valider. Par ailleurs, l'extension progressive du dispositif génotypé avec une puce pangénomique permettra d'en découvrir d'autres.

Un modèle SAM aura toujours une efficacité bornée par la proportion de la variance expliquée par les QTL pris en compte. Un avantage supposé de la sélection génomique est son aptitude à prendre en compte l'ensemble de la variabilité génétique, y compris celle due aux petits QTL. Il conviendrait de valider cette propriété très attractive mais de plus en plus discutée. En admettant qu'elle soit exacte, une alternative serait d'intégrer au modèle d'évaluation SAM un terme génétique résiduel prédit par une approche de type sélection génomique pour augmenter la part de variance génétique prise en compte. L'idée serait d'allier la connaissance a priori des QTL à l'évaluation a posteriori des effets génomiques. L'avantage de cette méthode serait une prise en compte immédiate de toute nouvelle information intégrée au système, son inconvénient éventuel serait une augmentation des temps de calcul. Il reste à démontrer l'efficacité de cette approche, ce qui revient à questionner l'efficacité de la sélection génomique pour utiliser les QTL difficiles à détecter et cartographier. L'avenir des modèles d'évaluation moléculaire est donc dans la mise au point de modèles capables de traiter de façon optimale des volumes de données toujours plus importants en un temps acceptable reposant sur des modèles et paramètres robustes.

6.4 Évolutions des programmes de sélection

Ce travail s'est focalisé sur un outil qu'est l'évaluation génétique assistée par marqueurs. Discutons maintenant l'utilisation de cet outil et son impact en sélection. Nous allons ainsi rappeler le principe d'un schéma de sélection bovin laitier et décrire l'intégration possible des

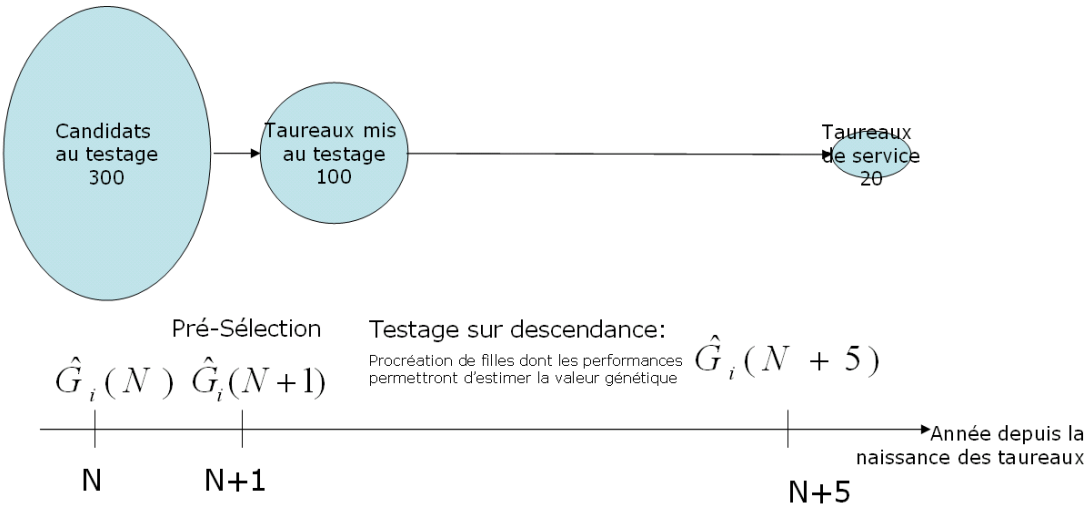


FIGURE 6.2 – Représentation d'un schéma de testage type

évaluations génétiques assistées par marqueurs dans ces schémas. Un schéma de sélection recrute des candidats mâles à partir des meilleurs parents, en choisit un nombre fixe pour l'étape la plus lourde du testage sur descendance (cf. Figure 6.2) et ne commercialise que les meilleurs issus de cette épreuve. Le coût du testage et la pression de sélection en sortie rend ce programme très coûteux, ce qui impose une diffusion massive des reproducteurs sélectionnés.

Hors information moléculaire, le choix des animaux testés est réalisé sur la base des index polygéniques sur ascendance disponibles, dont les Coefficient de Détermination (CD), sont proches de 0,3 pour la majorité des caractères et peuvent varier de 0,15 à 0,40. Après 3 à 4 ans de testage, temps nécessaire à la procréation et le phénotypage des descendants, les taureaux reçoivent des évaluations génétiques avec des niveaux de CD de l'ordre de 0,8 pour les caractères héritables à 0,5 pour les caractères faiblement héritables. L'augmentation de l'intervalle de génération est justifiée par le gain de CD permis par le testage.

Dans le cadre de la sélection assistée par marqueurs de première génération, les évaluations SAM apportent des gains de CD modérés. La SAM de première génération permet une pré-sélection des candidats au testage et augmente l'intensité de sélection dans la phase précoce du programme. Cette nouvelle étape permet d'étendre la population de candidats (et de mères de candidats) tout en diminuant le nombre de taureaux mis effectivement en testage. À ce stade, le testage sur descendance reste nécessaire pour atteindre la fiabilité souhaitée avant une forte diffusion des reproducteurs testés. Par ailleurs, le coût du reproducteur commercialisable n'est pas sensiblement diminué, ce qui impose donc toujours une forte diffusion et donc une précision élevée.

La SAM de deuxième génération, induit deux évolutions majeures :

– une grande amélioration de la fiabilité des index des candidats dès leur plus jeune âge ;
– une hausse du coût de typage.

Comparativement au programme SAM de première génération, les coûts individuels de typage ont fortement augmenté (le coût d'un typage ayant été multiplié en gros par 3). Ces hausses de prix ont été consenties au regard de la meilleure précision des évaluations obtenues, mais elles ne se justifient que si elles induisent de profondes mutations des pratiques des schémas de sélection. Avec les niveaux de CD atteints grâce à la SAM2, le testage sur descendance n'apporte plus qu'un gain de précision marginal, ne justifiant plus son coût ni l'allongement de l'intervalle de générations.

En reprenant la formule du progrès génétique annuel 1.2 et en faisant varier l'intervalle de génération de 2,5 ans (situation de diffusion de taureaux sans testage) à 7 ans (situation de diffusion de taureaux après testage), on observe que pour maintenir un progrès génétique constant, malgré l'allongement de l'intervalle de génération, il faudrait que les CD des index avant testage soient inférieurs à 0,4 et dépassent la valeur de 0,8 après testage. Dans les cas où le CD des index dépassent 0,5 avant testage, l'allongement de l'intervalle de générations de 2,5 ans entraîne une baisse de progrès génétique annuel. Ainsi, dans la plupart des cas, le temps nécessaire au testage sur descendance pénalise le progrès génétique réalisable. Maintenir le testage sur descendance entraîne donc des contraintes importantes (financières et temporelles) et n'est pas optimal.

Dans le cadre de caractères ayant des CD assez faibles (telle que la fertilité) un testage sur descendance peut demeurer intéressant. Dans ce cas, il faut être en mesure d'atteindre après testage un CD assez élevé, ce qui est d'autant plus difficile que l'héritabilité est faible. L'effectif de descendantes par père doit ainsi être nettement plus élevé que ce qui est pratiqué actuellement (200 filles serait un minimum) ce qui tend à diminuer (de moitié au moins) le nombre de taureaux testés à effort de testage constant. Cette réduction éventuelle des taureaux de testage doit tout de même être réalisée sans pour autant limiter l'intensité de sélection, il est alors prudent d'élargir au maximum la population de candidats génotypés afin d'être sûr de retenir les meilleurs candidats pour le testage. C'est un risque supplémentaire pour le maintien de la variabilité génétique, un objectif déjà difficile à gérer actuellement.

Globalement, peu d'arguments légitiment donc un maintien du testage sur descendance et la diffusion de taureaux évalués à l'aide d'informations moléculaires dès leur puberté semble la pratique entraînant à la fois le meilleur progrès génétique et une réduction substantielle des coûts des schémas de sélection. Selon les options choisies, des gains de 30 % à 80 % de progrès génétique sur l'objectif de sélection semblent envisageables.

L'optimisation des schémas de sélection a souvent été réduite à l'optimisation du choix des taureaux à mettre en testage. Cette étape entraînant d'une part les coûts les plus importants et garantissant d'autre part, le retour sur investissement des entreprises de sélection du fait du nombre réduit de taureaux testés puis diffusables. Néanmoins, d'autres facteurs d'optimisation sont à considérer.

La réduction des intervalles de générations peut encore permettre un gain de progrès génétique. Ainsi, l'utilisation de génisses comme mère à taureaux réduit l'intervalle de génération de 1 ans, celle de taureaux non testés comme pères à taureaux, une réduction de 2,5 ans. Cette voie est limitée ne serait-ce que d'un point de vue physiologique (gestation de 9 mois, maturitée sexuelle : 15 mois) (Schaeffer, 2006) et des solutions telles que la velogenetic (Georges et Massey, 1991) ou la "whizzo génétique" (Haley et Visscher, 1998) consistant à collecter des ovocytes dès la naissance de génisses puis les féconder in-vitro, doivent encore montrer leur faisabilité technique et surtout économique et éthique.

Une évolution importante est l'augmentation de la précision des index femelles, ce qui est une autre révolution : en effet, il était de règle que seuls les mâles testés pouvaient disposer d'index précis. Compte tenu de la taille de la population femelle, une pression de sélection intense et précise devient possible, augmentant encore le progrès génétique.

L'augmentation de la pression de sélection est quant à elle un autre levier important d'amélioration des schémas de sélection. Cette augmentation doit toutefois passer par une augmentation de la taille de population candidate et non par une réduction des animaux diffusés. La non réduction du nombre de taureaux diffusés, voire son augmentation, s'explique d'une part par une meilleure gestion du risque de surestimation d'un individu (les CD des évaluations assistées par marqueurs demeurant plus faibles que ceux d'un testage sur descendance), une meilleure gestion de la variabilité génétique, une meilleure satisfaction d'objectifs de sélection adaptés à des besoins variés et enfin par une contrainte pratique qui est de répondre à la demande de semences bovines qui est de l'ordre de 4 millions de doses par an.

L'augmentation de la taille de la population candidate peut être réalisée par une diminution des niveaux d'index polygénique requis pour intégrer le schéma de sélection, l'idée n'étant pas de diminuer la valeur génétique moyenne des mères à taureaux, mais de donner une chance à des animaux exceptionnels issus de familles moyennes, non retenues dans l'approche actuelle. Une seconde voie est la multiplication des élites grâce au biotechnologies de la reproduction, telles que la transplantation embryonnaire, ce point pallie une limite importante des schémas bovins laitiers : la faible prolificité des bovins. En effet, avec un veau par an et par vache dans le meilleur des cas, il faut en espérance au moins deux ans pour obtenir un mâle, avec une chance sur deux que ce mâle ait un aléa de méiose positif, le temps nécessaire à l'obtention d'un descendant interessant est rapidement limitant. Avec un transfert embryonnaire, le nombre d'embryons produits est plus de l'ordre de 5 à 6, ce qui doit permettre de procréer suffisamment d'individus pour que l'un d'entre eux bénéficie d'un aléa de méiose favorable.

L'ensemble de ces mesures va dans le sens d'une meilleure gestion de la variabilité génétique, car cette question, malheureusement trop peu intégrée dans les choix et orientations des schémas de sélection est à la base même du progrès génétique à longs termes.

Enfin, il convient d'insister sur le fait que non seulement la précision des index est augmentée mais qu'elle devient relativement homogène entre caractères, pour toutes les catégories d'individus.

En conséquence, le progrès génétique réalisé est davantage proportionnel au poids économique de chacun des caractères. Ceci est un avantage considérable pour la gestion durable de la sélection, en particulier pour les caractères fonctionnels souvent peu héritables et génétiquement opposés aux caractères de production.

6.5 Application de la sélection assistée par marqueurs aux races à petits effectifs et autres espèces

Ce travail a été focalisé sur un cas particulier (celui des bovins laitiers) qui est un cas idéal pour l'application des méthodes d'évaluation présentées ici. L'application de ces méthodes à d'autres races et autres espèces animales peut aboutir à des résultats moins convaincants du fait d'effectifs trop faibles ou de structures de populations moins favorables.

Différents éléments sont à considérer pour juger de la pertinence de la mise en place d'outils de sélection assistée par marqueurs :

- La quantité et la qualité des informations non moléculaires disponibles. En effet, les outils génomiques permettent une meilleure valorisation de ces informations, mais ne se substituent pas à elles . Si la SAM est peu dépendante de la mesure du phénotype chez le candidat, elle n'est efficace que si la prédiction génomique repose sur une population phénotypée et génotypée, régulièrement renouvelée. La situation est toujours beaucoup plus favorable quand les phénotypes préexistent. La situation est exceptionnelle dans les grandes races bovines laitières avec un grand nombre de taureaux évalués sur descendance. Dans bien des cas, les généalogies sont inconnues de sorte que seules les performances propres sont utilisables. Dans ces conditions, la population de référence est caractérisée par des phénotypes d'héritabilité variable et parfois faible, ce qui réduit l'informativité du dispositif et augmente donc sa taille souhaitable. Ainsi, le nombre de phénotypes nécessaires augmente lorsque l'héritabilité du caractère diminue et lorsque la précision désirée augmente (Goddard, 2008) . Pour un caractère d'héritabilité 0,2, Hayes *et al.* (2009) estiment que l'effectif nécessaire passe de 2000 à 9000 pour augmenter la précision de l'index génomique de 0,5 et 0,7. Pour des caractères héritables, les efforts de phénotypage restent acceptables et permettent assez facilement d'atteindre des précisions supérieures à celles usuellement obtenues, ainsi dans la filière bovine allaitante par exemple, de nombreux caractères courant d'évaluation pourraient bénéficier de l'apport d'informations moléculaires.
- La rentabilité de ces outils, l'amélioration de la précision des index se doit d'être valorisée par la filière. Le point crucial est le coût du génotypage par rapport au coût de la sélection et à la valorisation d'un reproducteur. Les bovins laitiers « bénéficient » d'une situation paradoxale où la sélection classique est très coûteuse. Il en serait de même pour le cheval. Mais on conçoit aisément que la situation sera plus difficile dans de nombreuses filières. En toute rigueur, le coût du programme incluant le génotypage devrait être comparé au progrès génétique qu'il apporte dans l'ensemble de la population. Ainsi, dans les espèces

à organisation pyramidale, la rentabilité devrait être facile à obtenir. Cela reste théorique car très souvent, la valeur ajoutée est largement partagée par l'ensemble de la filière et des consommateurs et ne bénéficie guère au sélectionneur. En pratique donc, la SAM sera sans doute difficile à intégrer dans de nombreuses filières.

6.6 Conclusion

Ce travail de thèse s'est déroulé durant une période de changements aussi profonds que rapides dans le monde de la génétique. En l'espace de quelques années, les potentialités des outils à disposition des généticiens ont littéralement explosé, bouleversant ainsi la nature des questions et enjeux de la sélection.

À son commencement, cette thèse avait pour but d'étudier l'efficacité du programme de sélection assistée par marqueurs. Aujourd'hui, même si de nombreux travaux sont nécessaires pour tirer le meilleur parti des informations, peu de doutes subsistent concernant le gain de précision permis par l'intégration de l'information moléculaire à haut débit. Les enjeux d'une utilisation massive, pour tout type d'animaux dans les différentes filières relèvent davantage de critères économiques.

Bibliographie

ASHWELL, M. S., TASSELL, C. P. V. et SONSTEGARD, T. S. (2001). A genome scan to identify quantitative trait loci affecting economically important traits in a us holstein population. *J Dairy Sci*, 84(11):2535–2542.

BENNEWITZ, J., REINSCH, N., GUIARD, V., FRITZ, S., THOMSEN, H., LOOFT, C., KÜHN, C., SCHWERIN, M., WEIMANN, C., ERHARDT, G., REINHARDT, F., REENTS, R., BOICHARD, D. et KALM, E. (2004a). Multiple quantitative trait loci mapping with cofactors and application of alternative variants of the false discovery rate in an enlarged granddaughter design. *Genetics*, 168(2):1019–1027.

BENNEWITZ, J., REINSCH, N., PAUL, S., LOOFT, C., KAUPE, B., WEIMANN, C., ERHARDT, G., THALLER, G., KUHN, C., SCHWERIN, M., THOMSEN, H., REINHARDT, F., REENTS, R. et KALM, E. (2004b). The dgat1 k232a mutation is not solely responsible for the milk production quantitative trait locus on the bovine chromosome 14. *J. Dairy Sci.*, 87(2):431–442.

BENNEWITZ, J., REINSCH, N., REINHARDT, F., LIU, Z. et KALM, E. (2004c). Top down preselection using marker assisted estimates of breeding values in dairy cattle. *Journal of Animal Breeding and Genetics*, 121(5):307–318.

BLOTT, S., KIM, J.-J., MOISIO, S., SCHMIDT-KÜNTZEL, A., CORNET, A., BERZI, P., CAMBISANO, N., FORD, C., GRISART, B., JOHNSON, D., KARIM, L., SIMON, P., SNELL, R., SPELMAN, R., WONG, J., VILKKI, J., GEORGES, M., FARNIR, F. et COPPIETERS, W. (2003). Molecular dissection of a quantitative trait locus : a phenylalanine-to-tyrosine substitution in the transmembrane domain of the bovine growth hormone receptor is associated with a major effect on milk yield and composition. *Genetics*, 163(1):253–266.

BOICHARD, D., FRITZ, S., ROSSIGNOL, M. N., BOSCHER, M. Y., MALAFOSSE, A. et COLLEAU, J. J. (2002). Implementation of marker-assisted selection in french dairy cattle. *in Proc. 7th World Cong. Genet. Appl. Livest. Prod., Montpellier, France.*, Electronic communication 22-03.

BOICHARD, D., FRITZ., S., ROSSIGNOL, M.-N., GUILLAUME, F., COLLEAU, J.-J. et DRUET, T. (2006). Implementation of marker selection : practical lessons from dairy cattle. *Proc 8th WCGALP*, pages 22–11.

BOICHARD, D., GROHS, C., BOURGEOIS, F., CERQUEIRA, F., FAUGERAS, R., NEAU, A., RUPP, R., AMIGUES, Y., BOSCHER, M. Y. et LEVÉZIEL, H. (2003). Detection of genes influencing economic traits in three french dairy cattle breeds. *Genet Sel Evol*, 35(1):77–101.

BOLARD, M. et BOICHARD, D. (2002). Use of maternal information for qtl detection in a (grand) daughter design. *Genet Sel Evol*, 34(3):335–352.

BULMER, M. G. (1971). The effect of selection on genetic variability. *Am. Nature*, 105:201–211.

CHURCHILL, G. A. et DOERGE, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics*, 138(3):963–971.

COHEN-ZINDER, M., SEROUSSI, E., LARKIN, D. M., LOOR, J. J., van der WIND, A. E., LEE, J.-H., DRACKLEY, J. K., BAND, M. R., HERNANDEZ, A. G., SHANI, M., LEWIN, H. A., WELLER, J. I. et RON, M. (2005). Identification of a missense mutation in the bovine abcg2 gene with a major effect on the qtl on chromosome 6 affecting milk yield and composition in holstein cattle. *Genome Res*, 15(7):936–944.

de ROOS, A. P. W., SCHROOTEN, C., MULLAART, E., CALUS, M. P. L. et VEERKAMP, R. F. (2007). Breeding value estimation for fat percentage using dense markers on bos taurus autosome 14. *J Dairy Sci*, 90(10):4821–4829.

DRUET, T., FRITZ, S., BOICHARD, D. et COLLEAU, J. J. (2006). Estimation of genetic parameters for quantitative trait loci for dairy traits in the french holstein population. *J Dairy Sci*, 89(10): 4070–4076.

DRUET, T., FRITZ, S., BOUSSAHA, M., BEN-JEMAA, S., GUILLAUME, F., DERBALA, D., ZELENIKA, D., LECHNER, D., CHARON, C., BOICHARD, D., GUT, I. G., EGGEN, A. et GAUTIER, M. (2008). Fine mapping of quantitative trait loci affecting female fertility in dairy cattle on bta03 using a dense single-nucleotide polymorphism map. *Genetics*, 178(4):2227–2235.

FERNANDO, R. L. et GROSSMAN, M. (1989). Marker assisted selection using best linear unbiased prediction. *Genet Sel Evol*, 21:467–477.

FISHER, R. (1918). The correlation between relatives on the supposition of mendelian inheritance trans. *Soc. Edinb*, 52:399.

GAUTIER, M., BARCELONA, R. R., FRITZ, S., GROHS, C., DRUET, T., BOICHARD, D., EGGEN, A. et MEUWISSEN, T. H. E. (2006). Fine mapping and physical characterization of two linked quantitative trait loci affecting milk fat yield in dairy cattle on bta26. *Genetics*, 172(1):425–436.

GAUTIER, M., CAPITAN, A., FRITZ, S., EGGEN, A., BOICHARD, D. et DRUET, T. (2007). Characterization of the dgat1 k232a and variable number of tandem repeat polymorphisms in french dairy cattle. *J Dairy Sci*, 90(6):2980–2988.

GEORGES, M. et MASSEY, J. (1991). Velogenetics, or the synergistic use of marker assisted selection and germ-line manipulation. *Theriogenology*, 35(1):151–159.

GEORGES, M., NIELSEN, D., MACKINNON, M., MISHRA, A., OKIMOTO, R., PASQUINO, A. T., SARGEANT, L. S., SORENSEN, A., STEELE, M. R., ZHAO, X., WOMACK, J. E. et HOESCHELE, I. (1995). Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. *Genetics*, 139(2):907–920.

GIBSON, J. (1994). Short-term gain at the expense of long-term response with selection of identified loci. Proc. 5th World Cong. Genet. Appl. Livest. Prod. 21:201–204.

GILMOUR, A. (1999). *ASREML Reference Manual*. NSW Agriculture.

GODDARD, M. (1992). A mixed model for analyses of data on multiple genetic markers. *Theoretical and Applied Genetics*, 83(6):878–886.

GODDARD, M. (2008). Genomic selection : Prediction of accuracy and maximisation of long terme response. *Genetica*, 136:245–257.

GRISART, B., COPPIETERS, W., FARNIR, F., KARIM, L., FORD, C., BERZI, P., CAMBISANO, N., MNI, M., REID, S., SIMON, P., SPELMAN, R., GEORGES, M. et 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(2):222–231.

GUILLAUME, F., FRITZ., S., BOICHARD, D. et DRUET, T. (2006). Application of mas in french dairy cattle : first results. *10th QTL-MAS Workshop Salzburg*.

GUILLAUME, F., FRITZ, S., BOICHARD, D. et DRUET, T. (2008a). Estimation by simulation of the efficiency of mas in the french marker-assisted selection program in dairy cattle. *Genet Sel Evol*, 40(1):91–102.

GUILLAUME, F., FRITZ, S., BOICHARD, D. et T.DRUET (2007a). Use of snp for marker assisted selection in french dairy cattle. *Proc EAAP 2007*, session 18 communication 6.

GUILLAUME, F., GAUTIER, M., JEMAA, S. B., FRITZ, S., EGGEN, A., BOICHARD, D. et DRUET, T. (2007b). Refinement of two female fertility qtl using alternative phenotypes in french holstein dairy cattle. *Anim Genet*, 38(1):72–74.

GUILLAUME, F., TARRÈS, J., BOICHARD, D., T.DRUET et FRITZ, S. (2008b). Accuracies of different types of mas-ebv in the french mas program. *Proc EAAP 2008*, session 04 communication 10.

HABIER, D., FERNANDO, R. L. et DEKKERS, J. C. M. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, 177(4):2389–2397.

HALEY, C. S., KNOTT, S. A. et ELSEN, J. M. (1994). Mapping quantitative trait loci in crosses between outbred lines using least squares. *Genetics*, 136(3):1195–1207.

HALEY, C. S. et VISSCHER, P. M. (1998). Strategies to utilize marker-quantitative trait loci associations. *J. Dairy Sci.*, 81(Suppl_2):85–97.

HAYES, B. et GODDARD, M. E. (2001). The distribution of the effects of genes affecting quantitative traits in livestock. *Genet Sel Evol*, 33(3):209–229.

HAYES, B. J., VISSCHER, P. M. et GODDARD, M. E. (2009). Increased accuracy of artificial selection by using the realized relationship matrix. *Genetics Research*, 91(01):47–60.

HAZEL, L. N. (1943). The genetic basis for constructing selection indexes. *Genetics*, 28(6):476–490.

HEATH, S. (1997). Markov chain monte carlo segregation and linkage analysis for oligogenic models. *The American Journal of Human Genetics*, 61(3):748–760.

HENDERSON, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31(2):423–447.

HEYEN, D. W., WELLER, J. I., RON, M., BAND, M., BEEVER, J. E., FELDMESSER, E., DA, Y., WIGGANS, G. R., VANRADEN, P. M. et LEWIN, H. A. (1999). A genome scan for qtl influencing milk production and health traits in dairy cattle. *Physiol Genomics*, 1(3):165–175.

HU, Z.-L., FRITZ, E. R. et REECY, J. M. (2007). Animalqtldb : a livestock qtl database tool set for positional qtl information mining and beyond. *Nucl. Acids Res.*, 35(suppl_1):D604–609.

KASHI, Y., HALLERMAN, E. et SOLLER, M. (1990). Marker-assisted selection of candidate bulls for progeny testing programmes. *Anim. Prod*, 51:63–74.

KHATKAR, M. S., THOMSON, P. C., TAMMEN, I. et RAADSMA, H. W. (2004). Quantitative trait loci mapping in dairy cattle : review and meta-analysis. *Genet Sel Evol*, 36(2):163–190.

KNOTT, S. (1994). Prediction of the power of detection of marker-quantitative trait locus linkages using analysis of variance. *Theoretical and Applied Genetics*, 89(2):318–322.

LANDE, R. et THOMPSON, R. (1990). Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics*, 124(3):743–756.

LARZUL, C., MANFREDI, E. et ELSEN, J. (1997). Potential gain from including major gene information in breeding value estimation. *Genet Sel Evol*, 29(2):161–184.

LEGARRA, A., ROBERT-GRANIE, C., MANFREDI, E. et ELSEN, J.-M. (2008). Performance of genomic selection in mice. *Genetics*, 180:611–618.

MACKINNON, M. et GEORGES, M. (1998). Marker-assisted preselection of young dairy sires prior to progeny-testing. *Livestock Production Science*, 54(3):229–250.

MEUWISSEN, T. et GODDARD, M. (1999). Marker assisted estimation of breeding values when marker information is missing on many animals. *Genet. Sel. Evol.*, 31:375–394.

MEUWISSEN, T., SOLBERG, T., SHEPHERD, R. et WOOLLIAMS, J. (2009). A fast algorithm for bayesb type of prediction of genome-wide estimates of genetic value. *Genetics Selection Evolution*, 41(1):2.

MEUWISSEN, T. H. et ARENDONK, J. A. V. (1992). Potential improvements in rate of genetic gain from marker-assisted selection in dairy cattle breeding schemes. *J Dairy Sci*, 75(6):1651–1659.

MEUWISSEN, T. H. et GODDARD, M. E. (1996). The use of marker haplotypes in animal breeding schemes. *Genet Sel Evol*, 28(2):161–176.

MEUWISSEN, T. H. et GODDARD, M. E. (2001). Prediction of identity by descent probabilities from marker-haplotypes. *Genet Sel Evol*, 33(6):605–634.

MEUWISSEN, T. H., HAYES, B. J. et GODDARD, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829.

PIEPHO, H. P. (2001). A quick method for computing approximate thresholds for quantitative trait loci detection. *Genetics*, 157(1):425–432.

PONG-WONG, R., GEORGE, A., WOOLLIAMS, J. A. et HALEY, C. S. (2001). A simple and rapid method for calculating identity-by-descent matrices using multiple markers. *Genet Sel Evol*, 33(33):453–471.

RON, M. et WELLER, J. I. (2007). From qtl to qtn identification in livestock–winning by points rather than knock-out : a review. *Anim Genet*, 38(5):429–439.

RUANE, J. et COLLEAU, J. J. (1995). Marker assisted selection for genetic improvement of animal populations when a single qtl is marked. *Genet Res*, 66(1):71–83.

RUANE, J. et COLLEAU, J. J. (1996). Marker-assisted selection for a sex-limited character in a nucleus breeding population. *J Dairy Sci*, 79(9):1666–1678.

SCHAEFFER, L. (2006). Strategy for applying genome-wide selection in dairy cattle. *Journal of Animal Breeding and Genetics*, 123(4):218–223.

SCHROOTEN, C., BOVENHUIS, H., van ARENDONK, J. A. M. et BIJMA, P. (2005). Genetic progress in multistage dairy cattle breeding schemes using genetic markers. *J Dairy Sci*, 88(4):1569–1581.

SOLBERG, T. R., SONESSON, A. K., WOOLLIAMS, J. A. et MEUWISSEN, T. H. E. (2008). Genomic selection using different marker types and densities. *J Anim Sci*, 86(10):2447–2454.

SPELMAN, R. et BOVENHUIS, H. (1998). Genetic response from marker assisted selection in an outbred population for differing marker bracket sizes and with two identified quantitative trait loci. *Genetics*, 148(3):1389–1396.

SPELMAN, R. et GARRICK, D. (1997). Utilisation of marker assisted selection in a commercial dairy cow population. *Livestock Production Science*, 47(2):139 – 147.

SPELMAN, R. J., GARRICK, D. J. et van ARENDONK, J. A. M. (1999). Utilisation of genetic variation by marker assisted selection in commercial dairy cattle populations. *Livestock Production Science*, 59(1):51 – 60.

SPELMAN, R. J. et van ARENDONK, J. A. (1997). Effect of inaccurate parameter estimates on genetic response to marker-assisted selection in an outbred population. *J Dairy Sci*, 80(12):3399–3410.

STELLA, A., LOHUIS, M. M., PAGNACCO, G. et JANSEN, G. B. (2002). Strategies for continual application of marker-assisted selection in an open nucleus population. *J Dairy Sci*, 85(9):2358–2367.

TARRES, J., GUILLAUME, F. et FRITZ, S. (2009). A strategy for qtl fine-mapping using a dense snp map. *BMC Proceedings*, 3(Suppl 1):S3.

THOMSEN, H., REINSCH, N., XU, N., LOOFT, C., GRUPE, S., KUHN, C., BROCKMANN, G. A., SCHWERIN, M., LEYHE-HORN, B., HIENDLEDER, S., ERHARDT, G., MEDJUGORAC, I., RUSS, I., FORSTER, M., BRENIG, B., REINHARDT, F., REENTS, R., BLUMEL, J., AVERDUNK, G. et KALM, E. (2001). Comparison of estimated breeding values, daughter yield deviations and de-regressed proofs within a whole genome scan for qtl. *Journal of Animal Breeding and Genetics*, 118(6):357–370.

VANRADEN, P. M. et WIGGANS, G. R. (1991). Derivation, calculation, and use of national animal model information. *J Dairy Sci*, 74(8):2737–2746.

VERRIER, E. (2001). Marker assisted selection for the improvement of two antagonistic traits under mixed inheritance. *Genet Sel Evol*, 33(1):17–38.

VILLUMSEN, T., JANSS, L. et LUND, M. (2009). The importance of haplotype length and heritability using genomic selection in dairy cattle. *Journal of Animal Breeding and Genetics*, 126(1):3–13.

VISSCHER, P. M. et HALEY, C. S. (1998). Strategies for marker assisted selection in pig breeding programmes. *in Proc. 6th World Cong. Genet. Appl. Livest. Prod., Armidale, Australia.*, 23:503–510.

WANG, T., FERNANDO, R., van der BEEK, S., GROSSMAN, M. et van ARENDONK, J. (1995). Covariance between relatives for a marked quantitative trait locus. *Genet Sel Evol*, 27:251–274.

WELLER, J. I., KASHI, Y. et SOLLER, M. (1990). Power of daughter and granddaughter designs for determining linkage between marker loci and quantitative trait loci in dairy cattle. *J Dairy Sci*, 73(9):2525–2537.

YTOURNEL, F., GILBERT, H. et BOICHARD, D. (2008). Comment affiner la localisation d'un qtl ? *INRA Prod. Anim.*, 21(2):147–158.

Liste des acronymes et abréviations

CD Coefficient de Détermination.

cM centiMorgan.

DL déséquilibre de liaison.

IBD identity by descent.

IBS identity by state.

LD-LA linkage disequilibrium and linkage analysis.

QTL quantitative trait locus.

SAM sélection assistée par marqueurs.

SNP Single Nucleotide Polymorphism.

Publications scientifiques

- GUILLAUME F., GAUTIER M., BEN JEMAA S.,FRITZ S.,EGGEN A.,BOICHARD D. and DRUET T.,2007. Refinement of two female fertility QTL using alternative phenotypes in French Holstein dairy cattle, Animal Genetics 38 (1), 7274.
- GUILLAUME F., FRITZ S., BOICHARD D., DRUET T. 2008 . Estimation by simulation of the efficiency of the French marker-assisted selection program in dairy cattle. Genet. Sel. Evol. 40 :91-102
- GUILLAUME F., FRITZ S., BOICHARD D., DRUET T. 2008 .Correlations of Marker-Assisted Breeding Values with Progeny Test Breeding Values for 899 French Holstein Bulls. J. Dairy Sci.
- DRUET, T., FRITZ, S., BOUSSAHA, M., BEN-JEMAA, S., GUILLAUME, F., DERBALA, D., ZELENIKA, D., LECHNER, D., CHARON, C., BOICHARD, D., GUT, I. G., EGGEN, A. et GAUTIER, M. . 2008. Fine mapping of quantitative trait loci affecting female fertility in dairy cattle on bta03 using a dense single-nucleotide polymorphism map. Genetics, 178(4) :2227-2235.
- BEN JEMAA S, DRUET T., GUILLAUME F., FRITZ S., EGGEN A., GAUTIER M. 2008 Detection of QTL affecting female fertility traits in French dairy cattle. Journal of Animal Breeding and Genetics 125, 280-288.
- TARRÈS, J., GUILLAUME, F. et FRITZ, S. 2009. A strategy for QTL fine-mapping using a dense SNP map. BMC Proceedings, 3(Suppl 1) :S3.

Colloques internationaux

- GUILLAUME F., FRITZ S., BOICHARD D. et DRUET T. 2006. Application of mas in french dairy cattle : first results. 10th QTL-MAS Workshop Salzburg.
- BOICHARD D., FRITZ S., ROSSIGNOL M.N., GUILLAUME F., COLLEAU J.J., et DRUET T., 2006. Implementation of marker selection : practical lessons from dairy cattle. Proc 8th WCGALP, 22-11
- GUILLAUME F., FRITZ S., BOICHARD D., et DRUET T., 2007 . Use of SNP for marker assisted selection in French dairy cattle in EAAP Dublin 2007 session18 communication n°6.
- GUILLAUME F., BOICHARD D.,TARRRÈS J., DRUET T. et FRITZ S., 2008 . Accuracies of different types of MAS-EBV in the French MAS program in EAAP Vilnius 2008 session 04 communication n°10.
- DUCROCQ V., FRITZ S., GUILLAUME F. et BOICHARD D. 2009 . French report on the use of genomic evaluation in Interbull bulletin n° 39

Autres colloques

- GUILLAUME F., BEN-JEMAA S., FRITZ S., DRUET T., GAUTIER M., 2005. Développement d'indicateurs alternatifs de la fertilité femelle chez les bovins laitiers. Applications à la cartographie fine de QTL. 12ème journées Rencontres Recherches Ruminants, 158.
- BARBAT A., BONAITI B., GUILLAUME F. , DRUET T. , COLLEAU J.J., BOICHARD D.,2005, Bilan phénotypique de la fertilité à l'insémination artificielle dans les trois principales races laitières françaises. 12ème journées Rencontres Recherches Ruminants
- BASSO B., FRITZ S., DRUET T., GUILLAUME F., ROSSIGNOL M.N., AMIGUES Y., GABRIEL R., SELLEM E., SALAS-CORTES L., HUMBLOT P., DRUART X. ,2005, Estimation de paramètres génétiques et détection de QTL liés à des caractères de fertilité mâle, de production de semence et de qualité de la semence chez le taureau laitier. 12ème journées Rencontres Recherches Ruminants
- LE MEZEC P., GUILLAUME F., MATTALIA S., MOUREAUX S., PACCARD P.,PONSARD C., FRERET S. . 2006 . Fertilité des vaches laitières :vers une description fine des phénotypes, Poster, séminaire Agenae 2006
- EGGEN A., GAUTIER M., BOUSSAHA M. FRITZ S., BEN JEMAA S., GUILLAUME F., TARRES J., MALAFOSSE A. BOICHARD D., 2007, Cartographie fine de QTL en bovins laitiers, Poster, Séminaire Agenae (15 16 octobre) Dourdan France .
- FRITZ S., DRUET T. , GUILLAUME F, BOSCHER M.Y. , EGGEN A. GAUTIER M, COLLEAU J.J., BOICHARD D . 2007 . Bilan du programme de Sélection Assistée par Marqueurs dans les trois principales races bovines laitières françaises et perspectives d'évolution. 14ème journées Rencontres Recherches Ruminants
- BOICHARD D., GUILLAUME F., TARRES J., BAUR A., EGGEN A., DRUET T. , FRITZ S. 2008. Sélection assistée par marqueurs. séminaire Agenae 2008
- FRITZ S., GUILLAUME F., TARRÈS J., BAUR A., BOUSSAHA M., BOSCHER M.Y., JOURNAUX L., MALAFOSSE A., GAUTIER M., COLLEAU J.J., EGGEN A., BOICHARD D. 2008. Utilisation des résultats de cartographie fine de QTL en sélection chez les bovins laitiers . 15ème journées Rencontres Recherches Ruminants
- GUILLAUME F. ,FRITZ S., DUCROCQ V., BOICHARD D. 2008. Utilisation de marqueurs SNP pour l'évaluation génétique des bovins laitiers. Journées du département de génétique animale, Lacanau 2008

Liste des tableaux

Table des figures