Modèles de prédiction pour l'évaluation génomique des bovins laitiers français : application aux races Holstein et Montbéliarde

Carine Colombani

Université de Toulouse

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :
Institut National Polytechnique de Toulouse (INP Toulouse)

Discipline ou spécialité :

Pathologie, Toxicologie, Génétique et Nutrition

Présentée et soutenue par :
Carine COLOMBANI

le : mardi 16 octobre 2012

Titre :

Modèles de prédiction pour l'évaluation génomique des bovins laitiers français : application aux races Holstein et Montbéliarde

---

Ecole doctorale :
Sciences Ecologiques, Vétérinaires, Agronomiques et Bioingénieries (SEVAB)

Unité de recherche :
UR631 INRA-SAGA

Directeur(s) de Thèse :

Christèle ROBERT-GRANIÉ

Rapporteurs :

Pascale LE ROY
Jean-Michel ROGER

Membre(s) du jury :
Florence PHOCAS
Zulma VITEZICA
Robert SABATIER

Modèles de prédiction pour l'évaluation génomique des bovins laitiers français

Applications aux races Holstein et Montbéliarde

Carine COLOMBANI

Directeur de thèse : Christèle ROBERT-GRANIÉ

Laboratoires d'accueil : UR631 INRA-SAGA (Station d'Amélioration Génétique des Animaux)

Membres du jury :

- Pascale Le Roy (rapporteur)
- Jean-Michel Roger (rapporteur)
- Florence Phocas
- Zulma Vitezica
- Robert Sabatier
- Christèle Robert-Granié (encadrante)

Date de soutenance : 16 octobre 2012
Lieu : INRA Auzeville

À mon père,
Pour son écoute et sa soif de savoir

Remerciements

Voici venu l'un des moments les plus émouvants de ma thèse : il est temps pour moi de penser à toutes les personnes qui ont accompagné ma vie professionnelle et personnelle de ces dernières années.

Le financement de mon travail de thèse a été assuré par le département de génétique animale (DGA) et l'ANR via le projet AMASGEN. À ce titre, je tiens à remercier Didier Boichard et Denis Milan, ancien et actuel chefs du département GA et Vincent Ducrocq, coordinateur du projet AMASGEN.

J'exprime mes profonds remerciements à ma directrice de thèse, Christèle Robert-Granié, qui a su me laisser la liberté nécessaire à l'accomplissement de mes travaux, tout en y gardant un œil critique et avisé. Merci également de m'avoir accueillie dans l'unité SAGA.

Ma vive reconnaissance va à Pascale Leroy et Jean-Michel Roger pour avoir accepté d'être les rapporteurs de ma thèse, à Zulma Vitezica et Florence Phocas, pour en avoir été les examinatrices et à Robert Sabatier, pour la présidence de mon jury de thèse et son intérêt dans mon travail par sa participation à mon comité de thèse. Je leur suis à tous reconnaissante d'avoir fait de ma soutenance une épreuve aussi agréable et formatrice.

Ma gratitude va également aux autres membres de mon comité de thèse, Vincent Ducrocq, François Guillaume, Bertrand Servin et Andrés Legarra pour leur contribution et les discussions stimulantes qui en ont découlées.

Ces quelques années à la SAGA se sont déroulées dans un environnement très amical, et je voudrais donc adresser un grand merci collectif à toutes les personnes travaillant dans l'enceinte de la SAGA. Merci à Nancy, Line, Valérie et Dounia pour leur aide chaleureuse, à Carine pour sa disponibilité et aux « Informaticiens » pour leur efficacité. Un grand merci à Virginie qui a rendu mes premiers mois de thèse moins effrayants en m'accueillant dans son bureau et en m'offrant toute sa gentillesse. Pour leur bienveillance, merci à Loys, Eduardo, Jean-Michel et Francis.

Je pense bien sûr à tous les membres du projet AMASGEN, qui ont toujours su montrer de l'intérêt dans mon travail et qui ont répondu à mes sollicitations lorsque le besoin s'en faisait sentir. Un merci particulier à Pascal et Andrés, pour leur aide précieuse et leurs conseils avisés. Merci à tous de votre bonne humeur, je n'imagine pas de meilleure ambiance dans un groupe de travail.

Il m’est impossible d’oublier mes nouveaux collègues d’Ingenomix qui ont dû supporter une doctorante de dernière année avec tout ce que ça représente de stress et de doutes. Merci pour votre soutien et votre patience. Un grand merci à Sébastien pour ses encouragements et sa disponibilité.

D’un point de vue plus personnel, je tiens à remercier notre petit groupe de cantine, de pauses café et de soirées : ma colocataire de bureau Cécile, pour son écoute et son aide au quotidien, Benjamin, qui n’a jamais eu peur de nos soirées « filles » et qui nous manque déjà beaucoup, Momo et Guillaume B, pour leur bonne humeur et leur amitié. Merci à Aurélie qui m’a toujours comprise, pour les « vidages de sac » de début de soirée, et son indéfectible soutien : ne reste pas trop loin trop longtemps. Un merci tout aussi spécial à Julie, pour les fous rires, les bavardages interminables et nos incomparables soirées mousse !

Je ne sais comment exprimer mon immense gratitude, et ma profonde amitié aux filles de la mafia de la SAGA (merci Eduardo, ce petit nom nous restera). Chloé, Mila et Charlotte, votre compagnie quotidienne a rempli ma thèse de souvenirs heureux. Je n’oublierai jamais nos pauses café, nos escapades du midi et nos soirées filles. Je suis très heureuse de voir que même aujourd’hui, avec les kilomètres qui nous séparent, on fait encore partie d’un même clan.

Ces remerciements ne sauraient être complets si je n’y incluais ma famille pour l’aide morale et la motivation qu’ils m’ont apportées pour achever ce travail et tous mes proches pour leur patience, leur confiance et leurs encouragements tout au long de ces années de thèse. Merci à ma sœur Charlène pour son écoute pendant nos débriefings quasi-quotidiens et pour m'avoir donné la passion des voyages. Nos évasions ont été des bouffées d'oxygène plus que nécessaires ! Merci de me garder une place de choix dans ta vie malgré les kilomètres et mon indisponibilité chronique de ces derniers temps. Aujourd’hui, comme tous les jours, mes pensées se tournent vers mon père, qui a su me transmettre le goût d’apprendre, et qui me manque, mais m’accompagne au quotidien.

Mille tendres mercis, enfin, à Jean-François, qui aura eu à subir à travers moi les désagréments de la vie de doctorante et dont la patience et l’aide précieuse auront contribué à me donner l’énergie nécessaire pour y mettre un point final.

Résumé

L'évolution rapide des techniques de séquençage et de génotypage soulève de nouveaux défis dans le développement des méthodes de sélection pour les animaux d'élevage. Par comparaison de séquences, il est à présent possible d'identifier des sites polymorphes dans chaque espèce afin de baliser le génome par des marqueurs moléculaires appelés SNP (Single Nucleotide Polymorphism). Les méthodes de sélection des animaux à partir de cette information moléculaire nécessitent une représentation complète des effets génétiques. Meuwissen *et al.* (2001) ont introduit le concept de sélection génomique en proposant de prédire simultanément tous les effets des régions marquées puis de construire un index "génomique" en sommant les effets de chaque région. Le challenge dans l'évaluation génomique est de disposer de la meilleure méthode de prédiction afin d'obtenir des valeurs génétiques précises pour une sélection efficace des animaux candidats.

L'objectif général de cette thèse est d'explorer et d'évaluer de nouvelles approches génomiques capables de prédire des dizaines de milliers d'effets génétiques, sur la base des phénotypes de centaines d'individus. Elle s'inscrit dans le cadre du projet ANR AMASGEN dont le but est d'étendre la sélection assistée par marqueurs, utilisée jusqu'à lors chez les bovins laitiers français, et de développer une méthode de prédiction performante.

Pour cela, un panel varié de méthodes est exploré en estimant leurs capacités prédictives. Les méthodes de régression PLS (Partial Least Squares) et sparse PLS, ainsi que des approches bayésiennes (LASSO bayésien et BayesCπ) sont comparées à deux méthodes usuelles en amélioration génétique : le BLUP basé sur l'information pedigree et le BLUP génomique basé sur l'information des SNP. Ces méthodologies fournissent des modèles de prédiction efficaces même lorsque le nombre d'observations est très inférieur au nombre de variables. Elles reposent sur la théorie des modèles linéaires mixtes gaussiens ou des méthodes de sélection de variables, résumant l'information massive des SNP par la construction de nouvelles variables. Les données étudiées dans le cadre de ce travail proviennent de deux races de bovins laitiers français (1 172 taureaux de race Montbéliarde et 3 940 taureaux de race Holstein) génotypés sur environ 40 000 marqueurs SNP polymorphes.

Toutes les méthodes génomiques testées ici produisent des évaluations plus précises que la méthode basée sur la seule information pedigree. On observe un léger avantage prédictif des méthodes bayésiennes sur certains caractères mais elles sont cependant trop exigeantes en temps de calcul pour être appliquées en routine dans un schéma de sélection génomique. L'avantage des méthodes de sélection de variables est de pouvoir faire face au nombre toujours plus important de données SNP. De plus, elles sont capables de mettre en évidence des ensembles réduits de marqueurs, identifiés sur la base de leurs effets estimés, c'est-à-dire ayant un impact important sur les caractères étudiés. Il serait donc possible de développer une méthode de prédiction des valeurs génomiques sur la base de QTL détectés par ces approches.

Abstract

The rapid evolution in sequencing and genotyping raises new challenges in the development of methods of selection for livestock. By sequence comparison, it is now possible to identify polymorphic regions in each species to mark the genome with molecular markers called SNPs (Single Nucleotide Polymorphism). Methods of selection of animals from genomic information require the representation of the molecular genetic effects. Meuwissen et al. (2001) introduced the concept of genomic selection by predicting simultaneously all the effects of the markers. Then a genomic index is built summing the effects of each region. The challenge in genomic evaluation is to find the best prediction method to obtain accurate genetic values of candidates.

The overall objective of this thesis is to explore and evaluate new genomic approaches to predict tens of thousands of genetic effects, based on the phenotypes of hundreds of individuals. It is part of the ANR project AMASGEN whose aim is to extend the marker-assisted selection, used in French dairy cattle, and to develop an accurate method of prediction.

A panel of methods is explored by estimating their predictive abilities. The PLS (Partial Least Squares) and sparse PLS regressions and Bayesian approaches (Bayesian LASSO and BayesCπ) are compared with two current methods in genetic improvement: the BLUP based on pedigree information and the genomic BLUP based on SNP markers. These methodologies are effective even when the number of observations is smaller than the number of variables. They are based on the theory of Gaussian linear mixed models or methods of variable selection, summarizing the massive information of SNP by new variables. The datasets come from two French dairy cattle breeds (1172 Montbéliarde bulls and 3940 Holstein bulls) genotyped with 40 000 polymorphic SNPs.

All genomic methods give more accurate estimates than the method based on pedigree information only. There is a slight predictive advantage of Bayesian methods on some traits but they are still too demanding in computation time to be applied routinely in a genomic selection scheme. The advantage of variable selection methods is to cope with the increasing number of SNP data. In addition, they are able to extract reduced sets of markers based of their estimated effects, that is to say, with a significant impact on the trait studied. It would be possible to develop a method to predict genomic values on the basis of QTL detected by these approaches.

Table des matières

Introduction

L'amélioration des animaux d'élevage est une préoccupation majeure des éleveurs qui cherchent à sélectionner les meilleurs reproducteurs afin d'obtenir les descendants les plus performants et les mieux adaptés aux conditions d'élevage d'aujourd'hui et de demain. La génétique quantitative permet de modéliser la sélection en s'intéressant à des caractères quantitatifs, c'est-à-dire de variation continue et qui sont supposés polygéniques. La performance *P* d'un animal s'exprime comme la somme d'un effet génétique, noté *G,* et d'un effet environnemental noté *E* (*P* = *G* + *E*), afin d'estimer l'influence respective de l'hérédité et du milieu. Ce modèle est appelé modèle polygénique additif et est à la base de la génétique quantitative. Il repose sur deux hypothèses fortes :

- L'additivité des effets du génotype et des effets environnementaux, qui implique que les effets génétiques ne varient pas quand les conditions du milieu changent ;
- Et l'indépendance entre génotype et milieu, c'est-à-dire qu'on peut supposer que les conditions d'élevage et les génotypes sont associés au hasard.

Les effets d'environnement ou effets du milieu E regroupent ce qui peut potentiellement avoir une influence sur la performance d'un animal comme les caractéristiques de son élevage, son âge ou l'effet de son troupeau. Certains effets du milieu comme les erreurs de mesure ou certaines conditions particulières de mesure, ne sont pas identifiables et constituent une part de l'erreur du modèle, de même que l'ensemble des effets sur la performance qui sont inexpliqués. L'effet du génotype *G* se décompose en la somme d'une valeur génétique additive A correspondant à la somme des effets moyens des gènes, et une partie non additive réunissant les effets de dominance et d'épistasie, c'est-à-dire les interactions entre gènes. Le but de la sélection est d'améliorer la valeur additive A moyenne au sein d'une population en sélectionnant les candidats selon leurs performances propres et/ou celles de leurs apparentés. La part d'origine génétique additive dans la performance d'un animal sur un caractère et qui est représentée par la variabilité des phénotypes de ce caractère s'exprime par le terme d'héritabilité h^2. Il permet de

prédire la réponse à la sélection des descendants. Plus un caractère est héritable et plus il sera facile à sélectionner.

En France comme dans beaucoup de pays, il y a encore quelques années, l'amélioration génétique des bovins laitiers reposait sur le testage sur descendance. La valeur génétique (ou **EBV** pour Estimated Breeding Value) d'un potentiel reproducteur était prédite à partir de ses données généalogiques et des performances de ses ascendants, collatéraux et descendants. En effet, en espérance, la valeur génétique additive d'un animal est due pour moitié à la mère de l'individu et pour moitié à son père. Le testage sur descendance était réalisé sur un ensemble de jeunes taureaux prometteurs issus d'accouplements raisonnés et qui entraient en station afin d'en contrôler la croissance, l'efficacité alimentaire, la conformation et les fonctions sexuelles. À deux ans, ces taureaux étaient mis à la reproduction afin d'obtenir au bout de trois ans, entre 50 et 120 filles en première lactation sur lesquelles étaient menées des analyses de la composition du lait, de la production laitière et des contrôles morphologiques ou de fertilité. Ainsi, il fallait attendre que le taureau à évaluer soit âgé de 5 à 6 ans pour avoir un nombre suffisant de filles et obtenir une valeur génétique précise. Les taureaux obtenant les valeurs génétiques les plus élevées étaient ensuite diffusés à grande échelle. Il faut cependant noter que le coût du testage sur descendance est relativement élevé (environ 50 000€ par taureau testé) d'autant plus qu'on ne retient en moyenne qu'un taureau testé sur dix. De plus, les caractères étudiés, tels que la production laitière et ses composantes, induisent des intervalles de génération généralement longs car mesurés sur les femelles.

Dans les années 1990, pour réduire les coûts de la sélection, on a cherché à utiliser l'information moléculaire sous forme de marqueurs **ADN** (Acide DésoxyriboNucléique), pouvant être disponibles sur n'importe quel individu dès la naissance et permettant d'en prédire la valeur génétique. La prise en compte de l'information moléculaire est particulièrement avantageuse sur des caractères difficilement mesurables, des caractères qui nécessitent la mort de l'animal (qualité de la viande) ou mesurables sur les femelles uniquement (production laitière) et les caractères peu héritables ou exprimés tardivement. En 2001, la Sélection Assistée par Marqueurs de première génération (**SAM1**) s'est imposée dans l'évaluation française des bovins laitiers dans les trois principales races : Holstein, Normande et

Montbéliarde. Elle repose sur l'utilisation de quelques **QTL** (Quantitative Trait Loci) importants : ce sont des régions du génome connus pour avoir un effet sensible sur les caractères d'intérêt. L'évolution des techniques de génotypage avec notamment le développement des puces à **SNP** (Single Nucleotide Polymorphism) a donné lieu à la création de plusieurs programmes de recherche dont le but premier était de rechercher les QTL (tels que le programme Qualvigène en 2003 sur les qualités de carcasse et de viande sur les trois principales races bovines allaitantes françaises ou le programme PhénoFinLait à partir de 2008 sur la composition fine des laits des espèces bovines, caprines et ovines). A partir de 2005, le projet Cartofine dont le principal objectif était la cartographie fine de QTL chez les bovins laitiers français, a permis de caractériser une population de taureaux de référence, génotypés et phénotypés, pour les trois principales races bovines laitières françaises. Ces mêmes populations de référence devaient permettre de tester de nouvelles méthodes de prédiction des valeurs génétiques des animaux en exploitant les informations moléculaires. En 2008, l'arrivée des puces SNP de moyen débit (puces 54 000 SNP ou « puces 54k ») a permis de développer une méthode SAM de deuxième génération (**SAM2**), reposant sur un nombre de QTL plus élevé et cartographiés plus finement et sur une population de référence plus étendue qu'en SAM1. Il était attendu que cette nouvelle méthode améliore substantiellement la précision des résultats des évaluations génétiques.

Le concept de sélection génomique a été introduit par Meuwissen *et al.* (2001) afin de considérer un ensemble de marqueurs couvrant tout le génome et ainsi, de ne plus se limiter à un nombre réduit de QTL. L'idée principale de la sélection génomique est que chaque marqueur est susceptible d'être associé à un QTL et chaque QTL est en déséquilibre de liaison avec certains marqueurs. Les effets estimés de tous les marqueurs sont sommés afin d'obtenir la valeur génétique de l'individu. Ainsi tous les QTL sont considérés simultanément et il n'est pas nécessaire de procéder à une étape de détection de QTL. La théorie décrite par Meuwissen *et al.* (2001) n'était cependant pas applicable à l'époque en raison des coûts de génotypage élevés et du peu de marqueurs disponibles. Le développement des technologies de séquençage a permis la mise à disposition de milliers de SNP et le développement en pratique de méthodes de sélection génomique.

L'objectif est donc de trouver des méthodes de régression capables de prédire la performance d'un animal *i* (variable réponse y_i) à partir de son génotype représenté sous forme de marqueurs SNP_j (variables explicatives x_{ij}). Le principal challenge statistique soulevé par l'utilisation des données SNP est que le nombre d'observations disponibles *n* (quelques centaines voire quelques milliers d'individus) est limité par les coûts de génotypage et est donc très inférieur au nombre de variables explicatives *p* (quelques milliers voire des dizaines de milliers de marqueurs).

Le projet ANR **AMASGEN** (Approches Méthodologiques et Application de la Sélection GENomique, 2009-2011) coordonné par V. Ducrocq (INRA-GABI, équipe G2B, Jouy-en-Josas) a pour but de comparer des approches statistiques capables de répondre au problème $p>>n$ afin de développer une méthode robuste d'évaluation génomique applicable aux bovins laitiers français. Financé pour trois ans par ApisGene et l'Agence Nationale pour la Recherche (**ANR**), le projet réunit des équipes du département INRA de Génétique Animale (Unités de Recherche GABI et SAGA), de l'UNCEIA et de l'Institut de l'Élevage. Il s'articule autour de cinq grandes tâches dont les idées principales sont :

- faire l'inventaire et évaluer un ensemble de méthodes prometteuses pour l'évaluation génomique des bovins laitiers afin d'obtenir des équations de prédiction fiables (tâche 1) ;
- en étudier les propriétés et les conditions d'application (tâche 2) ;
- développer et mesurer le gain apporté par une méthode de prédiction combinant SAM2 et évaluation génomique (tâche 3) ;
- créer une puce de présélection génomique et évaluer l'apport du génotypage femelle (tâche 4) ;
- et enfin, quantifier l'impact d'une présélection génomique des taureaux sur les évaluations nationales et internationales (tâche 5).

Mon projet de thèse s'inscrit principalement dans les tâches 1 et 2 du projet ANR AMASGEN : comparaison et mise en application des méthodes de prédiction génomique. Les trois premiers chapitres de ma thèse sont consacrés :

- à la définition des notions liées à l'évaluation génomique des reproducteurs et à l'inventaire des travaux de sélection génomique chez les bovins laitiers ;

- à la présentation des méthodes statistiques étudiées et mises en œuvre au cours de ma thèse ;
- et enfin, à la description des données utilisées dans ce travail.

Les deux chapitres suivants présentent les différents travaux menés sur les données bovines laitières françaises :

- Un chapitre est consacré aux méthodes de réduction de dimensions et de sélection de variables, en particulier la régression PLS (Partial Least Squares) et la sparse PLS. L'apport de ces méthodes par rapport aux approches courantes de sélection génomique y est discuté.
- Le chapitre suivant présente mes travaux sur les méthodes bayésiennes BayesCπ et LASSO bayésien. L'application de ces approches sur les données bovines laitières Holstein et Montbéliarde y est exposée, au regard des méthodes développées dans le chapitre précédent et des méthodes classiques telles que le BLUP génomique et le BLUP sur pedigree. De plus, nous avons cherché à savoir si ces différentes méthodes étaient capables de repérer des régions chromosomiques importantes pour les caractères d'intérêt et si leurs résultats étaient comparables aux méthodes de détection de QTL classiques.

Le dernier chapitre est une discussion générale des travaux réalisés dans cette thèse, des dernières avancées en évaluation génomique et des perspectives d'amélioration et d'application de la sélection génomique en France.

Chapitre 1 Introduction à la sélection génomique chez les bovins laitiers

1.1 Du modèle polygénique à la sélection génomique

1.1.1 Le modèle polygénique

La sélection dans toutes les espèces d'élevage a d'abord été menée sur les caractéristiques phénotypiques des animaux mesurées par l'éleveur ou des pointeurs, selon des critères plus ou moins subjectifs. Dans les années 1950, la sélection a bénéficié des nouvelles techniques de reproduction comme l'insémination artificielle (IA), le contrôle des performances associé à l'identification des animaux et à la collecte des informations généalogiques (ascendance, collatéraux, descendance) des animaux candidats à la sélection. De plus, les progrès de l'informatique et des statistiques pour la génétique ont conduit, au XX^{ème} siècle, à d'importantes améliorations avec les apports de Fisher (1918) avec le modèle polygénique, puis de Hazel (1943), avec la théorie des index de sélection.

La sélection des animaux, pour un caractère donné, se base alors sur une évaluation de la valeur génétique transmissible des animaux, corrigée des effets non génétiques. L'évaluation génétique d'un reproducteur peut être définie comme la prédiction de sa valeur génétique à partir des performances mesurées sur lui-même et/ou sur des individus apparentés. Elle est considérée comme un outil primordial d'aide à la sélection puisqu'elle fournit le critère optimal pour réaliser le choix des reproducteurs. Elle permet également de mesurer *a posteriori* l'efficacité des programmes de sélection.

L'évaluation génétique des animaux s'effectue à partir des phénotypes (ou performances) observés. Le phénotype (P) d'un animal se décompose traditionnellement de la manière suivante : P = E + G où E est l'effet de l'environnement (ou milieu) sur la performance de l'animal et G est l'effet du génotype de l'animal. Chaque composante du modèle est décomposée dans le but de représenter les effets des gènes et des conditions d'élevage de la manière la plus précise possible. Génotype et environnement sont supposés agir de façon additive.

Lorsque ce n'est pas le cas, on parle d'interaction génotype-milieu et les choses se compliquent. Le génotype (G) peut être décomposé en G = A + D + I où :

- A est la valeur génétique additive de l'animal, c'est à dire la somme des effets moyens des gènes régissant l'expression du caractère ;
- D est la valeur de dominance, c'est-à-dire la somme des effets d'interaction entre les allèles de chaque locus impliqué dans l'expression du caractère ;
- I est la valeur d'épistasie, c'est-à-dire la somme des effets d'interaction entre allèles de loci différents.

Lors de la méiose, un seul allèle par locus est conservé dans chaque gamète. De ce fait, la valeur de dominance n'est pas transmissible à la descendance. En revanche, une partie des effets d'épistasie peut être maintenue dans le gamète (principalement les effets d'interactions impliquant des loci proches sur un même chromosome) car le taux de recombinaison entre ces loci est faible.

Toutefois, la part du génotype d'un animal transmis à sa descendance est principalement liée à la valeur génétique additive (A), dont la moitié sera en espérance transmise à chaque descendant. Les reproducteurs sont donc choisis sur leur valeur génétique additive. En l'absence d'identification des effets de l'ensemble des gènes régissant la plupart des caractères d'intérêt zootechnique, l'évaluation génétique des animaux repose généralement sur le modèle polygénique infinitésimal : un caractère quantitatif est gouverné par un très grand nombre de loci ayant des effets individuels faibles et indépendants. Par application du théorème de limite centrale, la somme des effets moyens de ces gènes suit une distribution normale de moyenne nulle et de variance σ_a^2, appelée variance génétique additive.

L'environnement (E) regroupe 3 types d'effets de milieu :

- Ceux qu'on a identifiés et quantifiés (dont on connait *a priori* l'influence sur les caractères étudiés). Chez les bovins laitiers par exemple, la production laitière dépend de la durée de la lactation. Elle est standardisée pour une durée de 305 jours.
- Ceux qu'on a identifiés mais dont on ne connait pas l'influence exacte sur les caractères étudiés. Ils doivent alors être pris en compte et estimés dans le modèle d'analyse. On sait par exemple que le mois et l'année de vêlage d'une vache, son âge au vêlage et son troupeau sont des facteurs influençant sa production laitière.

- Ceux qu'on ne maîtrise pas du tout. Ils constituent l'erreur du modèle qu'on cherchera à minimiser.

Aussi précise que soit la modélisation des effets génétiques ou environnementaux, une part de la performance reste toujours inexpliquée. Tous les facteurs non pris en compte constituent alors l'erreur du modèle. Elle est par définition la résultante de l'action d'un grand nombre de facteurs ayant des effets individuels faibles et peut donc être considérée comme la réalisation d'une variable aléatoire obtenue par tirage dans une population d'effets selon une distribution donnée. On suppose que les erreurs sont normalement et indépendamment distribuées.

Les modèles statistiques de la génétique quantitative sont des modèles mixtes à effets fixes et à effets aléatoires. Les effets de milieu (troupeau, année, saison, rang de vêlage, *etc*.) sont généralement considérés comme fixes alors que les valeurs génétiques sont supposées aléatoires. En conséquence, chaque performance d'un animal est décomposée à l'aide d'un modèle d'analyse avec au minimum 3 composantes additives : $y_i = x_i\beta + u_i + e_i$ où y_i est la performance de l'animal *i*, $x_i\beta$ est la somme des effets de milieu modélisés auxquels est soumise la performance et u_i est la valeur génétique additive de l'animal *i*. La distribution du vecteur **u** des effets u_i est normale de moyenne nulle et de variance $\mathbf{A}\sigma_u^2$ où **A** est la matrice de parenté de la population (dont l'élément (*k*,*l*) est égal au coefficient de parenté entre les individus *k* et *l*). L'erreur du modèle e_i a une distribution normale, indépendante de celle de **u** et d'espérance nulle. Les erreurs sont, de plus, supposées indépendantes les unes des autres. Habituellement, elles sont supposées de même variance σ_e^2.

Le modèle linéaire mixte gaussien (combinant à la fois des effets fixes et des effets aléatoires) est particulièrement indiqué pour la description des données générées dans le cadre de programmes d'amélioration génétique chez les grandes espèces domestiques. Le fait de considérer les deux types d'effets simultanément permet, à l'aide d'une estimation adéquate, de corriger les estimées d'un effet pour tous les autres effets inclus dans le modèle, qu'ils soient fixes ou aléatoires. Dans le cas des modèles mixtes, cette méthode statistique est le BLUP (Meilleure prédiction linéaire non biaisée ; Henderson, 1973). Elle possède des propriétés intéressantes

dont l'absence de biais dans l'estimation des effets aléatoires à condition que le modèle d'analyse soit correct.

Sous forme matricielle, le modèle linéaire mixte gaussien à un seul effet aléatoire **a** (en plus de la résiduelle) se définit de la manière suivante :

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e}$$

où **y** est le vecteur des observations y_i, **β** est le vecteur des différents effets fixes considérés dans le modèle, **u** est le vecteur des effets aléatoires, **e** est le vecteur aléatoire des erreurs et **X** et **Z** sont des matrices dites « d'incidence » qui relient les observations aux effets fixes et aléatoires qui les ont influencés et qui, contrairement aux effets **β** et **u,** sont connues. Les espérances de **u** et **e** sont supposées connues (E[**u**] = E[**e**] = **0**) ainsi que les matrices de variance-covariance des effets aléatoires :

$$\text{Var}(\mathbf{u}) = \mathbf{G} = \mathbf{A}\sigma_u^2 \text{ ; Var}(\mathbf{e}) = \mathbf{R} = \mathbf{I}\sigma_e^2 \text{ et Cov}(\mathbf{u,e}) = \mathbf{0}$$

On a alors Var(**y**) = **V** = **ZGZ'**+**R** et par construction E[**y**] = **Xβ** où **β** est un vecteur d'inconnues.

Henderson *et al.* (1959) ont montré que les estimations des effets fixes et les prédictions des valeurs génétiques (effets aléatoires) peuvent s'obtenir comme solutions du système suivant :

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} + \frac{\sigma_e^2}{\sigma_u^2}\mathbf{A^{-1}} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix}$$

Ce système d'équations est appelé système des équations du modèle mixte.

La prédiction de la valeur génétique d'un individu à l'aide du BLUP-modèle animal prend en compte, par l'intermédiaire de la matrice de parenté **A,** l'ensemble des performances mesurées sur tous les animaux apparentés. Les valeurs génétiques de tous les animaux sont prédites quelle que soit la quantité d'information disponible pour chaque animal, en résolvant les équations du modèle mixte.

Le BLUP-modèle animal constitue actuellement la référence internationale, grâce à ses propriétés théoriques et son aspect unificateur, facilitant ainsi les comparaisons et les échanges internationaux. Il assure une évaluation individuelle de tous les animaux, qu'ils soient issus d'une ou plusieurs générations, qu'ils aient ou non des données et quelle que soit la structure des apparentements en jeu.

Cette méthode a permis un gain génétique certain ces dernières années dans de nombreuses filières. Cependant, elle peut être complexe et coûteuse, en particulier quand le phénotype ne peut pas être mesuré sur le candidat à la sélection comme chez les taureaux laitiers. Ce dispositif de sélection demande de longues années d'observation et de mesures sur descendance pour identifier les animaux candidats à la reproduction. De plus, même quand le caractère est observable, la sélection a une efficacité limitée quand le phénotype dépend essentiellement du milieu, c'est-à-dire quand l'héritabilité est faible. C'est avec le séquençage de l'ADN que de nouvelles perspectives pour la sélection se sont ouvertes.

La situation chez les bovins laitiers. La sélection des bovins laitiers dans le monde entier repose sur un socle que l'on pensait immuable : depuis les années 1970, dans tous les pays développés, les bases de données nationales centralisent les mesures phénotypiques de production, de fertilité, de morphologie, de résistance à certaines infections (mammites), de longévité, de facilité de naissance ou de vêlage. Par ailleurs, le très fort pouvoir de diffusion des taureaux permis par l'insémination artificielle (parfois plus de 100 000 filles pour un taureau) requiert un testage sur descendance avant d'entreprendre une diffusion massive des tous meilleurs taureaux. Dans ce contexte, il est capital d'estimer le plus précisément possible la valeur génétique additive des taureaux sur tous les caractères d'intérêt, d'où un recours à des modèles et des méthodes statistiques sophistiqués. Malgré un intervalle de génération élevé, le testage sur descendance s'est révélé très efficace avec des progrès génétiques annuels importants, en particulier sur la production (supérieurs à 100 kg/an/vache de lait en race Holstein). En 2007, le séquençage complet du génome bovin a permis d'avoir accès à un polymorphisme important autorisant la création de puces à SNP de densité élevée. La disponibilité de ces puces SNP en 2008 a ouvert ainsi de nouvelles perspectives dans le domaine de la sélection animale.

1.1.2 Les fondements de la génomique

L'utilisation des marqueurs ADN pour la sélection animale présente l'avantage d'augmenter la précision des valeurs génétiques (ou EBV : Estimated Breeding Values) notamment sur les jeunes animaux et sur des caractères difficilement

mesurables (Meuwissen et Goddard, 1996 ; Dekkers, 2004). Cela pourrait conduire à un plus fort gain génétique pour un intervalle de génération réduit (Schaeffer, 2006).

Les marqueurs génétiques. Après les espèces modèles comme la drosophile et la plante *Arabidopsis Thaliana*, les animaux d'élevage ont fait l'objet du séquençage intégral de leur génome. Depuis le milieu des années 1990, la génomique appliquée aux espèces d'élevage a permis de détecter des marqueurs moléculaires de régions chromosomiques ayant un effet quantitatif significatif sur les caractères d'intérêt appelés **QTL** (Quantitative Trait Loci). Un marqueur génétique peut être défini comme un endroit du génome (locus) se présentant sous différentes formes (allèles) dans une population. Ils constituent les outils élémentaires pour « baliser » le génome et sont aussi de véritables traceurs de l'information génétique entre générations. Les deux grands types de marqueurs utilisés en génomique sont les microsatellites et les SNP.

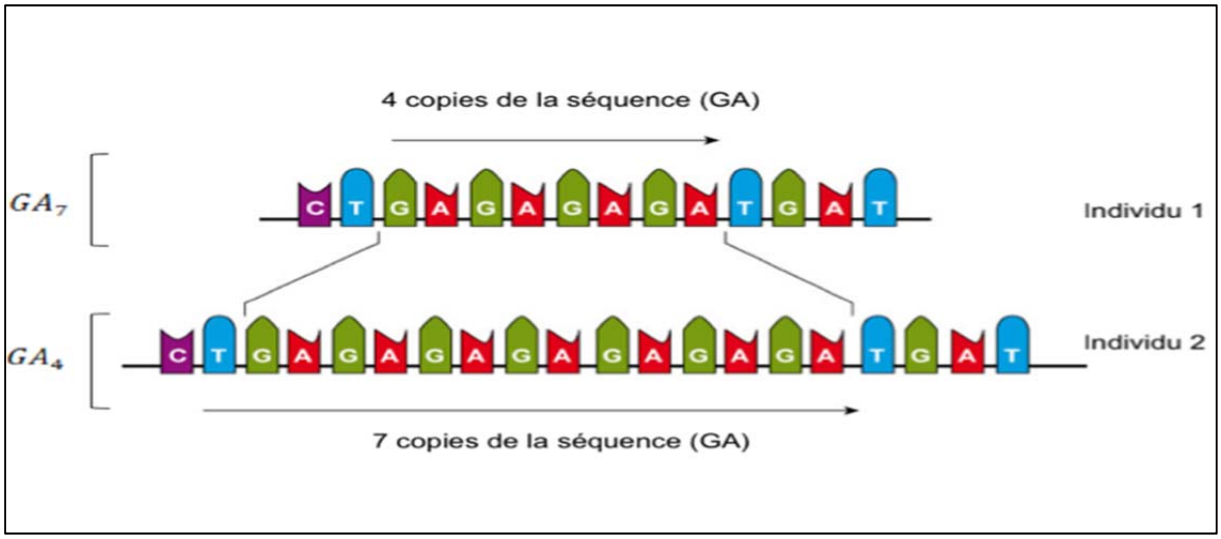


Figure 1.1 : Représentation d'un marqueur microsatellite ; répétition du motif GA

Les marqueurs satellites (figure 1.1) sont mis en évidence par la répétition de séquence d'un faible nombre de bases (de 2 à 5 paires de bases). Ils ont l'avantage d'être très polymorphes donc très informatifs.

Les marqueurs SNP (figure 1.2) correspondent quant à eux au polymorphisme d'une seule paire de bases (A, G, T, C). Contrairement aux marqueurs microsatellites, ils sont très abondants (1 SNP toutes les 100 à 1 000 bases soit

plusieurs millions de SNP dans une population donnée) et peuvent être présents dans les gènes. Ils sont le plus souvent bi-alléliques (a, A) c'est-à-dire qu'ils sont constitués de l'allèle ancestral initial et de l'allèle muté ce qui les rend peu informatifs individuellement : cela est compensé par leur grand nombre. Pour un locus donné, un individu sera donc porteur d'un des deux génotypes homozygotes (aa ou AA) ou du génotype hétérozygote (aA ou Aa indistinctement).

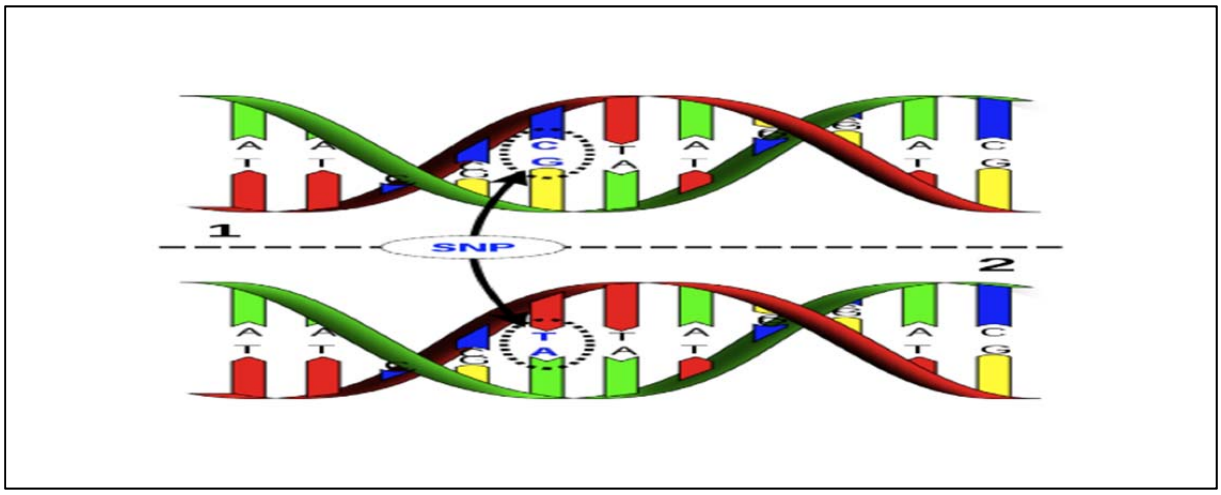


Figure 1.2 : Représentation d'un marqueur SNP ; la molécule d'ADN de l'individu 1 diffère de celle de l'individu 2 par un seul nucléotide (C/T)

Le déséquilibre de liaison. Les marqueurs SNP, qui se répartissent sur l'ensemble du génome à intervalle moyen de l'ordre de 40−50 kb (pour des puces d'environ 50 000 marqueurs), permettent de localiser les QTL plus précisément en utilisant les analyses d'association. Ces analyses exploitent le phénomène de déséquilibre de liaison (**LD** pour linkage disequilibrium), association non aléatoire d'allèles en plusieurs loci dans la population des chromosomes (ou gamètes) présents dans la population. Un déséquilibre de liaison est généralement dû à l'existence d'une liaison entre deux loci, et il est d'autant plus probable que ces deux loci sont plus proches l'un de l'autre. À l'inverse, deux loci éloignés (notamment sur deux chromosomes distincts) tendent à être en équilibre de liaison. Les études d'association entre un marqueur et un phénotype supposent donc que ce marqueur est en déséquilibre de liaison avec un QTL dont le polymorphisme cause une partie de la variabilité du caractère.

On peut illustrer le LD en considérant deux marqueurs SNP A et B sur un même chromosome, avec A possédant les allèles (A_1, A_2) et B les allèles (B_1, B_2). Quatre haplotypes sont possibles : A_1B_1, A_1B_2, A_2B_1 et A_2B_2. Si chaque allèle a une fréquence de 0,5 alors chaque haplotype devrait avoir une fréquence de 0,25. Ainsi, toute déviation à 0,25 de cette fréquence haplotypique signe un déséquilibre de liaison.

Pour chaque région les animaux ayant hérité des mêmes allèles aux SNP ont une très grande probabilité d'avoir hérité en même temps du même allèle au QTL. On est donc capable d'extrapoler le génotype d'un individu à un QTL en disposant de l'information génotypique de ses marqueurs génétiques.

1.1.3 La sélection assistée par marqueurs (SAM)

La SAM de première génération (SAM1). L'évaluation SAM de première génération a été mise en place en 2001 dans les trois principales races bovines laitières (Holstein, Normande et Montbéliarde) (Boichard *et al.*, 2002). La différence d'une évaluation SAM par rapport à une évaluation polygénique classique est dans l'information utilisée : on ajoute à l'effet polygénique, estimé à partir de l'information des performances d'un individu et de ses apparentés, la somme des effets des allèles aux QTL que cet individu porte.

Un premier modèle a été proposé par Fernando et Grossman (1989) et utilise le déséquilibre de liaison intra famille. 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 statistique sous-jacent est le suivant :

$$y_i = u_i + \sum_{j=1}^{N_{QTL}} (v_{ij}^{P} + v_{ij}^{m}) + e_i$$

où u_i est l'effet polygénique de l'individu *i*, v_{ij}^{p} et v_{ij}^{m} sont les effets paternels et maternels de l'allèle au QTL *j* de l'individu *i* et $e_i$est un terme résiduel. Les N_{QTL} QTL introduits dans le modèle sont suivis par des marqueurs microsatellites.

Fernando et Grossman (1989) montrent également que la méthodologie BLUP est applicable à ce modèle. 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. Ainsi, la SAM1 n'est efficace qu'intra-famille et permet, par exemple, de choisir entre plein-frères et pleines sœurs. Elle était utilisée pour le tri des jeunes candidats (futures mères à taureaux ou veaux avant le testage) au sein d'une même famille, ce qui présentait un grand intérêt notamment dans la voie mâle puisque 95% des taureaux testés sur descendance sont issus de transplantation embryonnaire (Colleau *et al.,* 1998).

En France, la SAM1 repose sur l'identification de 14 QTL suivis par 43 marqueurs microsatellites pour 7 caractères évalués. Parmi les arguments développés en faveur de la sélection sur index SAM, la réduction possible de 8% à 33% selon les caractères, du nombre de taureaux à tester pour obtenir le même progrès génétique qu'en l'absence d'informations moléculaires, a été mise en avant (Boichard *et al.*, 2002). Dans un article de bilan, Fritz *et al.* (2007) montrent que l'application de la SAM en bovins laitiers en France entre 2001 et 2007 a permis d'augmenter la précision des index des jeunes taureaux : par exemple, pour la quantité de lait, la précision de l'index, mesuré par le Coefficient de Détermination (**CD**) a augmenté de 0,33 à 0,44.

Cependant, les marqueurs microsatellites induisent une localisation des QTL imprécise car leur intervalle de position s'étend de 5 cM à 20 cM. Ils sont donc peu abondants et non uniformément répartis sur le génome. Ils sont aussi très difficiles à analyser à haut débit avec un coût associé élevé (d'environ 1,5€ par marqueur) et ne permettent de suivre que les QTL ayant un impact majeur. Le développement de puces SNP bovines de 54 000 SNP (principalement la puce « Illumina BovineSNP50™ beadchip ») à faible coût (moins d'un centime par marqueur) et couvrant l'ensemble du génome a considérablement accéléré l'évolution du programme, permettant à la fois d'affiner 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 SAM de deuxième génération (Guillaume *et al.,* 2008), utilisant des haplotypes de SNP en déséquilibre de liaison avec les principaux QTL localisés finement.

La SAM de deuxième génération (SAM2). Les marqueurs SNP ont permis de localiser plus finement les QTL sur des petits segments de chromosome contenant

une dizaine de gènes au maximum. Les estimations des effets des QTL reposent sur le déséquilibre de liaison entre QTL et marqueurs sur l'ensemble de la population et non plus au sein d'une famille. Le nombre de QTL intégré dans la SAM est passé de 14 à une trentaine voire une cinquantaine de QTL selon le caractère évalué. Ces QTL sont marqués par des haplotypes de 4 à 6 marqueurs SNP. Chaque QTL n'explique qu'une faible part de variance ce qui conduit à en intégrer un grand nombre dans la SAM2. Ces QTL expliquent plus de la moitié de la variance génétique d'une population donnée. Il est donc important de conserver un modèle simple afin de maintenir la rapidité des évaluations pour que le modèle SAM2 gagne à la fois en efficacité mais aussi en simplicité. Le modèle d'évaluation retenu dans cette optique comprend donc une composante polygénique et une somme d'effets haplotypiques aux QTL :

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

où h_{ij}^{p} et h_{ij}^{m} sont les effets haplotypiques paternels et maternels pour le QTL *j* de l'individu *i*.

Le choix de la France (Boichard et *al.*, 2002, 2003) d'adopter dans un premier temps une stratégie de sélection assistée par marqueurs plutôt qu'une sélection génomique s'explique par le fait que le modèle SAM est robuste car il est basé sur des QTL connus. De plus, il reste efficace sur des populations de taille moyenne comme la race Montbéliarde et la race Normande. Ainsi, en juin 2010, les dispositifs de testage classiques en race Montbéliarde, Normande et Holstein ont laissé la place à la diffusion de taureaux sélectionnés sur index génomiques estimés à partir des évaluations SAM2.

1.1.4 La sélection génomique

Définition de la sélection génomique. Lande et Thompson (1990) conceptualisent la sélection génomique en proposant d'effectuer la sélection des reproducteurs potentiels sur la base d'un score moléculaire. 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 « gros » QTL 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 (figure 1.3).

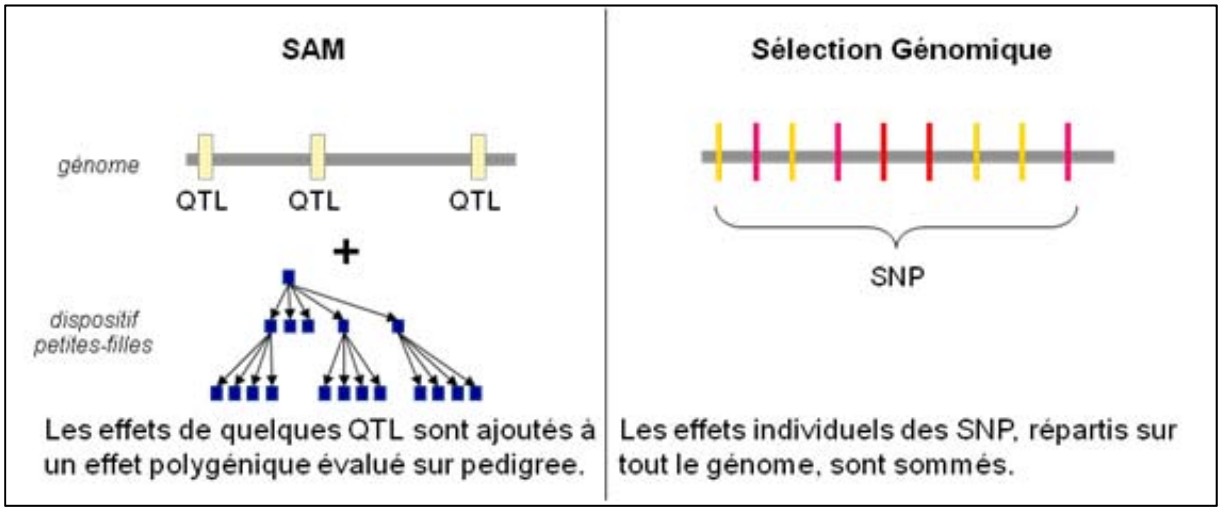


Figure 1.3 : Différence entre sélection assistée par marqueurs et sélection génomique

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). Trois idées sont alors sous-jacentes :

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 ;
3. on cherche les marqueurs les plus prédictifs, ce ne sont pas forcément les plus proches des QTL.

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) montrent 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 a l'inconvénient d'introduire une proportion croissante de faux QTL réduisant l'efficacité de l'approche.

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 0,5, la parenté vraie peut varier théoriquement de 0 à 1. De plus, la sélection génomique permettrait de résoudre le problème de surestimation des effets QTL due à la prise en compte simultanée de tous les effets aux marqueurs.

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 d'un modèle de prédiction construit sur une population ressources (population de référence) d'animaux génotypés et phénotypés. Une fois la relation entre les génotypes aux marqueurs et les phénotypes (donc le modèle de prédiction) établie, il est alors possible de prédire un index de sélection génomique pour des jeunes animaux (sans phénotypes), uniquement sur la base de leur génotype aux marqueurs SNP.

La population de référence. La population de référence est composée de taureaux d'insémination artificielle génotypés et évalués sur descendance donc avec des phénotypes. En pratique, la population de référence est partagée en deux sous-ensembles : une population d'apprentissage et une population de validation (figure 1.4).

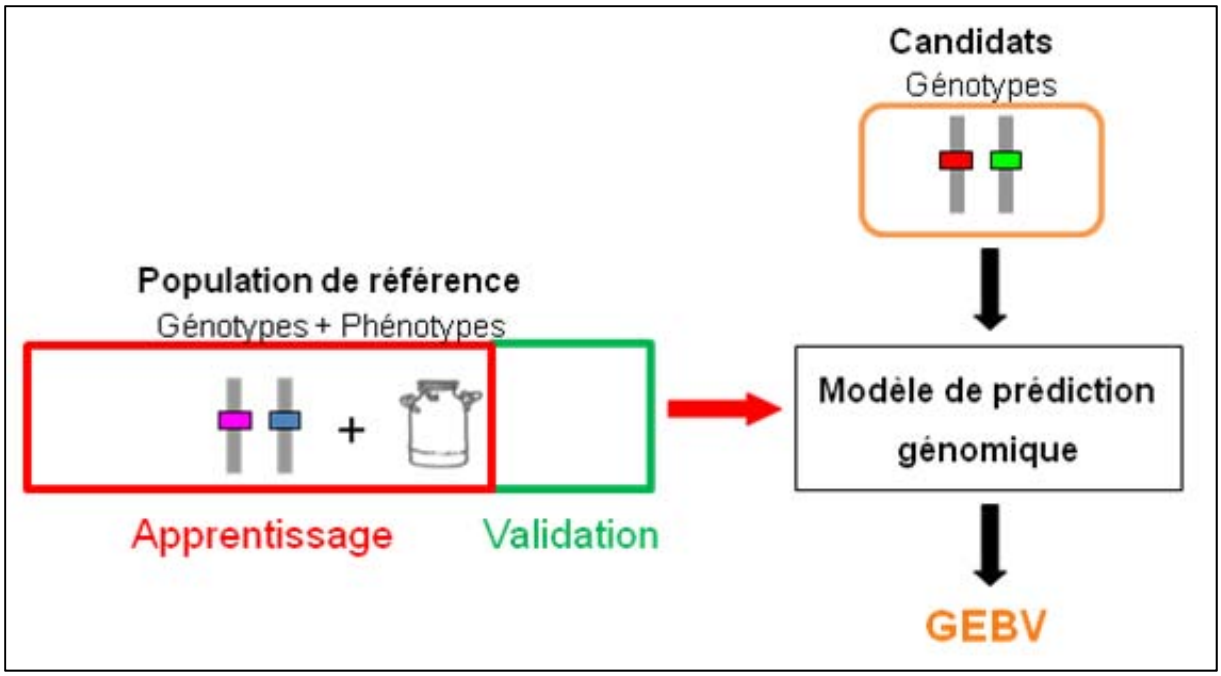


Figure 1.4 : Les animaux candidats sont sélectionnés selon leurs valeurs génomiques (GEBV) à partir du modèle de prédiction établi sur la population de référence

Les différentes méthodes de sélection génomique choisies sont appliquées sur la population d'apprentissage afin d'obtenir des équations de prédiction basées sur l'estimation des effets des SNP. Ensuite, on vérifie la qualité prédictive de ces modèles en les appliquant à la population de validation afin de mesurer l'écart entre les vraies valeurs génétiques des animaux de la population de validation et les valeurs génomiques prédites (**GEBV**). Le but est de construire des modèles capables d'évaluer de façon fiable un ensemble d'animaux candidats à la sélection dont on ne connait que les génotypes. Pour prédire une valeur génomique pour les animaux qui ne sont pas dans la population de référence, les effets des allèles aux SNP que ces animaux portent sont sommés sur l'ensemble du génome : $\boldsymbol{GEBV} = \sum_{j=1}^{p} \boldsymbol{X_j}\, \hat{g}_j$ où *p* est le nombre de SNP, $\boldsymbol{X_j}$ est la matrice d'incidence associant l'effet du marqueur SNP *j* aux animaux candidats et $\hat{g}_j$ est l'effet estimé du SNP *j*.

Il est important que la population de référence soit d'une taille suffisamment élevée afin de permettre des prédictions précises. Goddard et Hayes (2009) mettent l'accent sur le nombre élevé d'animaux génotypés nécessaires à de bonnes prédictions génomiques. Il est clairement mis en évidence que la population de

référence doit être d'autant plus grande que l'héritabilité du caractère d'intérêt est faible. De Roos (2011) suggère quant à lui qu'une petite population de référence peut être utilisée si la densité des marqueurs disponibles est assez élevée. La taille minimale de la population de référence doit être de 5 000 animaux pour atteindre une précision minimale de 0,3 d'après Goddard et Hayes (figure 1.5). En pratique, il est difficile d'atteindre un nombre aussi élevé d'animaux phénotypés et génotypés. On estime alors que la population de référence doit être au moins de l'ordre de 1 000 individus afin de rendre compte de façon fiable de la relation entre génotype et phénotype.

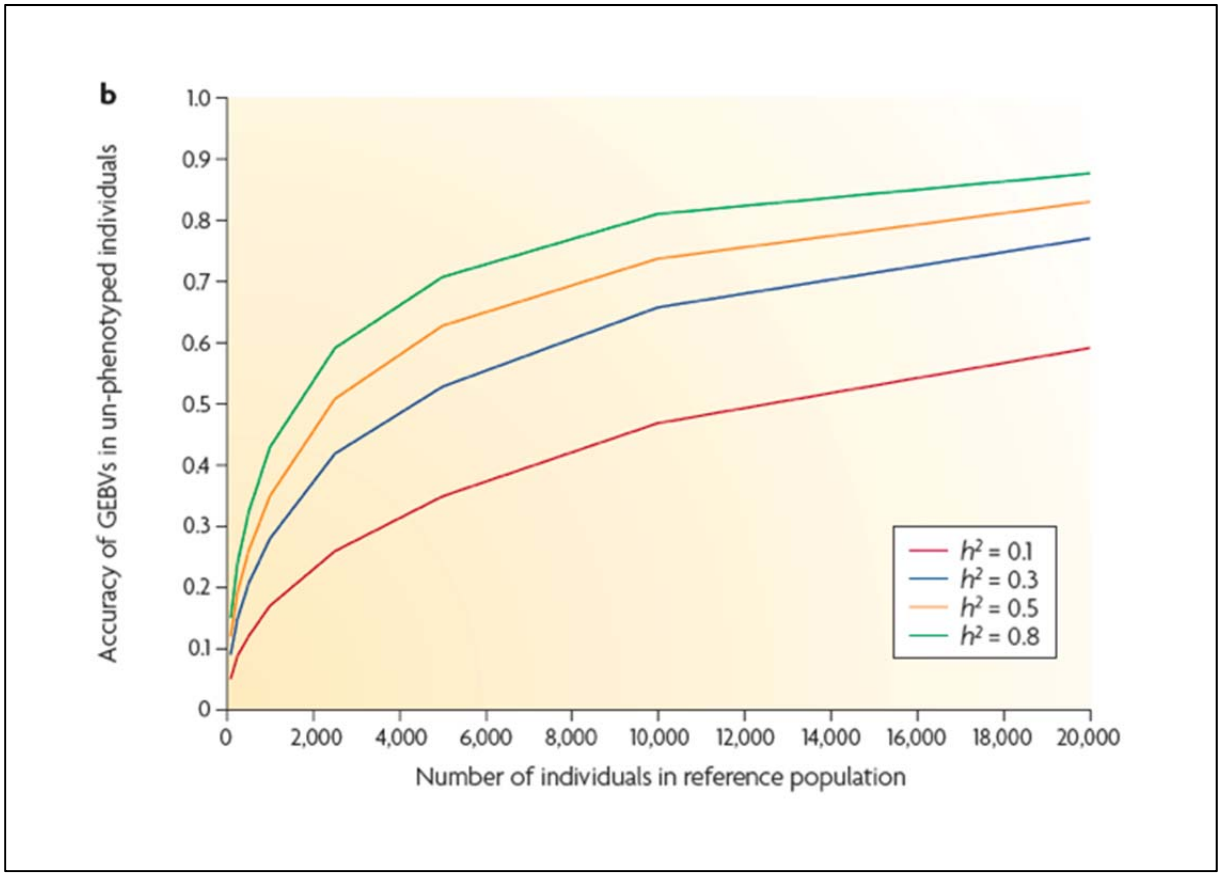


Figure 1.5 : Fiabilité des GEBV d'animaux non phénotypés pour une taille croissante de la population de référence, selon l'héritabilité du caractère (Goddard et Hayes, 2009)

Le volume de données disponibles permettant d'agrandir la population atteindra un plafond quand tous les taureaux testés seront génotypés. Les sources nouvelles d'augmentation seront les taureaux nouvellement évalués sur descendance, les taureaux provenant d'échanges internationaux et très

probablement les femelles. On peut également envisager de combiner l'information de plusieurs races (quand cela est judicieux) pour augmenter la taille de la population de référence. Cependant des études de faisabilité sont nécessaires dans un tel cas.

Les données phénotypiques. Chez les bovins laitiers, le dispositif sur lequel la sélection génomique a débuté est un dispositif petites-filles (Weller *et al*., 1990) constitué de familles de taureaux demi-frères de père. Ces taureaux sont évalués sur descendance (environ sur une centaine de filles), ce qui leur confère un phénotype très particulier, la performance moyenne de leurs filles (VanRaden et Wiggans, 1991). Ce phénotype est équivalent à une performance propre pour un caractère d'héritabilité égale à la précision de l'index.

Dans le cadre de la sélection génomique, deux types de phénotypes ont été utilisés, selon le sexe de l'individu : d'une part, pour les mâles, les **DYD** (Daughter Yield Deviation) définis comme la moyenne des performances de ses filles, corrigées pour l'ensemble des facteurs inclus dans le modèle d'indexation officielle et de la moitié de la valeur génétique de leurs mères (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. Les DYD et YD sont des sous-produits des évaluations génétiques officielles, basées sur le modèle polygénique et utilisant les performances des individus et l'information pedigree. 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 préconisés (Thomsen *et al.,* 2001). En effet, les index utilisent déjà une information génétique que l'on voudrait capturer à travers les SNP. Les **EDC** (Effective Daughter Contribution) rendent compte de la précision que l'on peut associer aux pseudo-phénotypes (DYD, YD ou index) calculés (Fikse et Banos, 2001). Un EDC est alors associé à chaque animal ; il repose à la fois sur le nombre de filles phénotypées disponibles pour cet animal et sur l'héritabilité du caractère.

1.2 Les phases de préparation des données de génotypage

1.2.1 Contrôle de qualité des données génomiques

Les technologies utilisées pour le génotypage des SNP ont été développées de façon importante mais comme toutes données, les données de génotypage peuvent présenter des erreurs ou des valeurs manquantes. Il est alors indispensable de procéder à un ensemble de contrôles sur les données afin que cela ne perturbe pas les estimations des valeurs génomiques.

L'équilibre d'Hardy-Weinberg. La théorie de l'équilibre d'Hardy-Weinberg (noté HWE) a été proposée indépendamment par Hardy (1908) et Weinberg (1908). Elle stipule que les fréquences des allèles et du génotype d'un locus restent constantes de générations en générations (d'où la notion d'équilibre) si les hypothèses suivantes sont respectées :

– La population est de taille infinie (ou une population de taille assez importante pour que la loi des grands nombres s'applique).

– Les espèces étudiées sont diploïdes et à reproduction sexuée.

– Il n'y a pas de migration.

– Il n'y a pas de sélection.

– Il n'y a pas de mutation.

– Le régime de reproduction est panmictique (les gamètes s'associent au hasard, ou les couples se forment aléatoirement)

– Les fréquences alléliques des mâles et des femelles sont identiques.

Pour représenter une situation d'équilibre d'Hardy-Weinberg, prenons le cas d'un SNP A possédant 2 allèles, notés A_1 et A_2, de fréquences respectives p et q avec $p + q = 1$. Les fréquences génotypiques doivent être : $f(A_1A_1) = p^2$, $f(A_1A_2) = 2pq$ et $f(A_2A_2) = q^2$. Il existe plusieurs tests permettant de voir si on dévie de l'équilibre d'Hardy-Weinberg. Le plus simple est le test de Pearson (plus connu sous le nom de "test du khi2") dont la distribution sous l'hypothèse nulle suit asymptotiquement une loi de χ^2. Le test de Pearson n'est pas optimal lorsque la fréquence d'un des génotypes présents est faible. Dans ces conditions, il est préférable d'utiliser un test exact de Fisher. On trouve facilement, dans la littérature, d'autres tests exacts comme ceux de Wigginton *et al.* (2005) ou Guo et Thompson (1992). Il est important

d'écarter des analyses génomiques les SNP qui ne répondent pas aux conditions d'équilibre d'Hardy-Weinberg pour éviter des marqueurs techniquement difficiles à typer.

Call Freq et Call Rate. On vient de voir que le test de HWE permet de détecter certaines erreurs de génotypage. Cependant, il ne permet pas de les corriger ou d'imputer les génotypes manquants, dus à une mauvaise séparation entre les « clusters » prédéfinis de la puce, qui permettent d'affilier chaque individu à un génotype. En général, un premier filtrage consiste à supprimer les SNP dont le pourcentage d'individus génotypés avec succès est inférieur à un seuil généralement fixé autour de 80% (il s'agit du call freq).

De même, un individu dont le pourcentage de génotypes manquants sur l'ensemble des marqueurs est trop important (en général supérieur à 2%) est supprimé des futures analyses (il s'agit du call rate). Le call rate d'un individu *i* représente le taux de données manquantes sur son génotype et est calculé de la façon suivante. Soit $ntyp_i$ le nombre d'allèles renseignés pour l'animal *i, nsnp* le nombre de SNP disponibles sur la puce et snp_nontyp_i le nombre de SNP non typés pour l'animal *i* alors le call rate de l'animal *i* s'écrit :

$$cr_i = \frac{ntyp_i}{2(nsnp - snp_{nontyp_i})}$$

La fréquence de l'allèle mineur. La fréquence moyenne de l'allèle rare de chaque SNP est supérieure à 20% chez la plupart des races *bos taurus*. Cependant de nombreux SNP sont très peu polymorphes et ne sont donc pas suffisamment informatifs pour apporter un gain de précision aux modèles prédictifs. De nombreux auteurs imposent un seuil minimal pour la fréquence de l'allèle mineur afin d'éliminer ces SNP. Les valeurs les plus couramment utilisées sont 1%, 3% et 5%.

Ces différentes phases d'élimination des données génotypiques « à problème » sont très importantes car ces marqueurs pourraient avoir un impact négatif sur le phasage ou l'imputation des génotypes manquants. Les fichiers de pedigree sont aussi soigneusement vérifiés. Les erreurs liées à l'enregistrement des pedigrees sont, en général, de l'ordre de 3%. Les inversions de prélèvements d'ADN

ou d'enregistrement des animaux font que le pourcentage d'incompatibilité des données génomiques peut atteindre 5%. Il faut donc supprimer de la population de référence ces animaux car s'ils présentent des phénotypes extrêmes, cela pourrait influer sur la qualité des modèles établis et donc sur les prédictions génomiques des animaux candidats.

1.2.2 Imputation des génotypes manquants

Après ces différentes étapes de contrôle de la qualité des données, il peut encore rester des individus avec des données incomplètes. Restreindre les analyses aux seules données complètes pourrait porter la population de référence à une taille trop réduite pour mener à une bonne modélisation génomique. Pour remplacer les génotypes manquants, une solution basique serait d'utiliser la moyenne des génotypes observés ou le génotype le plus probable. Mais cela modifierait le déséquilibre de liaison avec les marqueurs proches et conduirait ainsi à des biais ou à une perte de puissance du modèle. L'idée est plutôt de remplacer les génotypes manquants par une valeur prédite basée sur les génotypes observés aux SNP voisins. En général, les méthodes existantes, par maximum de vraisemblance ou bayésiennes, permettent d'affecter une valeur aux génotypes manquants et de reconstruire les phases simultanément.

Il y a encore peu de temps, les méthodes les plus populaires d'imputation de données étaient dans les logiciels PHASE (Stephens *et al.*, 2001), fastPHASE (Scheet et Stephens, 2006) et IMPUTE (Marchini *et al.*, 2006). D'autres stratégies existent comme celles basées sur des méthodes de classification où le génotype manquant est copié des autres individus qui ont les mêmes génotypes aux marqueurs voisins, ou des méthodes de régression comme celle de Souverein *et al.* (2006) qui modélisent les génotypes manquants comme une fonction de génotypes d'autres marqueurs et de phénotypes dans une régression logistique polytomique.

De multiples méthodes et logiciels permettant d'inférer les phases haplotypiques se sont développés autour de ces trois types d'approches. Le logiciel le plus connu, PHASE (Stephens *et al.*, 2001), utilise une approche bayésienne où la probabilité *a priori* est calculée à partir de la théorie de la coalescence. Ce logiciel donne des estimations précises (Marchini *et al.*, 2006) mais il est très couteux en temps de calcul. Plus récemment, les logiciels fastPHASE (Scheet et Stephens,

2006) et BEAGLE (Browning et Browning, 2007) ont été créés à partir des chaînes de Markov cachées (**HMM**) et de l'algorithme **EM** (Expectation-Maximization ; Dempster *et al.*, 1977). Ils donnent une précision aussi bonne que PHASE tout en étant plus rapides. Browning (2008) propose une revue des différentes méthodes existantes sur ce sujet.

Généralement, en population animale, les individus sont apparentés et les génotypes des pères et des descendants sont disponibles. Dans ce contexte, les méthodes de phasage citées précédemment ne sont pas optimales. Il existe des méthodes qui prennent en compte la partie transmission d'haplotypes entre parents et descendants et qui, de plus, utilisent le déséquilibre de liaison. On trouve parmi celles-ci DualPHASE et DagPHASE (Druet et Georges, 2010) qui sont basées sur fastPHASE et BEAGLE. Enfin, l'étude comparative de Browning (2008) montre la supériorité des logiciels fastPHASE et BEAGLE à la fois sur le taux d'erreur dans l'imputation de données génomiques manquantes mais aussi sur leur rapidité d'exécution.

1.3 Modélisation statistique de la sélection génomique

Plusieurs méthodes ont été développées pour estimer les effets des SNP dans un contexte de prédiction génomique. Le modèle statistique général est équivalent au modèle mixte présenté précédemment, et s'écrit :

$$\mathbf{y} = \mu\mathbf{1_n} + \mathbf{Xg} + \mathbf{e}$$

avec **y** le vecteur des phénotypes (DYD) pour le caractère étudié des *n* individus de l'ensemble d'apprentissage, μ la moyenne générale du caractère étudié sur la population d'apprentissage, **X** la matrice des génotypes des *p* marqueurs SNP sur les *n* individus, **g** le vecteur des effets aléatoires des SNP d'éléments g_j et tels que $\mathbf{g} \sim \mathrm{N}(\mathbf{0}, \sigma_g^2)$, et **e** le vecteur des effets résiduels tel que $\mathbf{e} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}\sigma_e^2)$. La valeur génétique d'un individu *i* est estimée par $\widehat{u_i} = \mathbf{x_i}\hat{\mathbf{g}}$, où $\mathbf{x_i}$ est le vecteur ligne qui contient le génotype aux marqueurs de l'individu *i*, et $\hat{\mathbf{g}}$ est l'ensemble des effets estimés des marqueurs.

Dans l'article de Meuwissen *et al.* (2001), plusieurs méthodologies ont été mises au point pour l'estimation des effets SNP : une méthode de régression des moindres carrés, une méthode de type BLUP dans laquelle on considère que les

effets SNP sont des effets aléatoires issus d'une loi normale de variance σ_g^2 et deux méthodes bayésiennes appelées BayesA et BayesB.

1.3.1 La régression des moindres carrés

Cette première approche ne suppose aucune distribution sur les effets des SNP. Les effets SNP sont considérés comme des effets fixes dans une approche classique des moindres carrés. Meuwissen *et al.* (2001) proposent une approche en trois étapes :

- Les effets individuels de chaque SNP *j* sont estimés un à un par régression simple selon le modèle $\mathbf{y} = \mu\mathbf{1_n} + \mathbf{X_j}\, g_j + \mathbf{e}$;

- Les *m* (*m* < *n*) marqueurs obtenant un effet estimé supérieur à un seuil fixé sont sélectionnés ;

- Enfin, les effets de ces *m* marqueurs sont réestimés dans un modèle de régression multiple tel que : $\mathbf{y} = \mu\, \mathbf{1_n} + \sum_{j=1}^{m} \mathbf{X_j}\, g_j + \mathbf{e}$.

La régression des moindres carrés présente l'inconvénient de reposer sur une présélection des SNP selon leur effet estimé individuellement les uns des autres : cela multiplie les analyses et conduit à une surestimation des effets aux QTL. Nous verrons dans le chapitre suivant que la régression Ridge apporte une solution à cette question.

1.3.2 Le BLUP génomique (GBLUP)

Le modèle BLUP proposé par Meuwissen *et al.* (2001) est très proche du BLUP sur pedigree présenté au début de ce chapitre. Dans le BLUP génomique, les effets des marqueurs sont supposés aléatoires de loi normale avec une variance homogène pour tous les marqueurs. Goddard (2009) montre que ce BLUP génomique est équivalent à un modèle BLUP traditionnel en remplaçant la matrice de parenté **A** calculée à partir de l'information pedigree par une matrice de parenté génomique. Les éléments de cette matrice génomique appelée « matrice **G** » correspondent aux relations génomiques entre les individus et rendent compte de la similarité entre leurs génotypes. En supposant que l'espérance des observations **y** (ici DYD) est égale à *μ,* le modèle polygénique basé sur l'information pedigree

s'écrit : $\mathbf{y} = \mu\mathbf{1_n} + \mathbf{Zu} + \mathbf{e}$. Les équations du modèle mixte d'Henderson permettant d'obtenir les prédictions des valeurs génétiques $\hat{u}$ s'écrivent :

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

Le BLUP appliqué à des données génomiques est souvent appelé **GBLUP**, pour « Genomic BLUP » et conduit au système d'équations suivant :

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

À partir de l'expression du GBLUP ci-dessus, la valeur génétique d'un individu *i* est égale à $\hat{u}_i = \sum_{\mathbf{j}} \mathbf{x_{ij}}\,\hat{\mathbf{g}}_{\mathbf{j}}$. Il est alors possible de transformer le modèle initial en un modèle totalement équivalent de façon à obtenir directement les mêmes valeurs génétiques individuelles $\hat{u}_i$ (Habier *et al.,* 2007 ; VanRaden, 2008 ; Goddard, 2009 ; Hayes *et al.,* 2009b). Le système d'équations du BLUP s'écrit alors de la manière suivante :

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

L'utilisation de l'appellation « GBLUP » pour cette expression se justifie pleinement. Différents auteurs (VanRaden, 2008 ; Goddard, 2009) montrent que, pour un codage particulier (centré sur 0) des SNP, il est possible de construire une matrice dite « de parenté génomique », $\mathbf{G} = \frac{\mathbf{XX'}}{2\sum p_j q_j}$ analogue à la matrice de parenté **A** et jouant le même rôle dans les équations BLUP. Le terme p_j représente la fréquence de l'allèle de référence du SNP *j* et $q_j = 1 - p_j$. Les éléments de la matrice **G** mesurent la proportion moyenne d'allèles partagés par deux individus pour tous les SNP, pondérés par leur fréquence : le partage d'allèles plus rares est plus indicatif de la parenté. À condition de la calculer avec une quantité suffisante de marqueurs, cette « parenté génomique » est plus précise que celle basée sur le pedigree, car cette dernière est basée sur une généalogie qui sera tôt ou tard incomplète et elle ne prend pas en compte les écarts à la théorie dus à la liaison entre loci et à la taille réelle (finie) du génome.

Misztal *et al.* (2009) proposent de modifier la matrice de parenté **A** par une matrice **H** qui prend en compte à la fois l'information des relations de parenté classiques et l'information génomique : $\mathbf{H} = \mathbf{A} + \mathbf{A_\Delta}$ où $\mathbf{A_\Delta}$ contient les écarts entre les coefficients de parenté attendus et observés. Si on intègre à la fois les animaux non génotypés (notés 1) et les animaux génotypés (notés 2) alors la matrice de parenté A s'écrit :

$$\mathbf{A} = \begin{bmatrix} \mathbf{A_1^1} & \mathbf{A_2^1} \\ \mathbf{A_1^2} & \mathbf{A_2^2} \end{bmatrix}.$$

La matrice $\mathbf{A_1^1}$ contient donc les relations de parenté entre animaux non génotypés, les matrices $\mathbf{A_2^1}$ et $\mathbf{A_1^2}$ les relations de parenté entre animaux non génotypés et génotypés et la matrice $\mathbf{A_2^2}$ les relations de parenté entre animaux génotypés c'est-à-dire la matrice **G**.

La matrice **H** peut donc s'écrire :

$$\mathbf{H} = \begin{bmatrix} \mathbf{A_1^1} & \mathbf{A_2^1} \\ \mathbf{A_1^2} & \mathbf{G} \end{bmatrix} = \mathbf{A} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G} - \mathbf{A_2^2} \end{bmatrix} \text{ d'où } \mathbf{A_\Delta} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G} - \mathbf{A_2^2} \end{bmatrix}.$$

Pour éviter des multiplications de matrices trop complexes, Legarra *et al.* (2009) présentent plusieurs écritures de la matrice **H** en partitionnant les animaux en plusieurs groupes :

- les ancêtres non génotypés avec valeur génétique ;
- les animaux génotypés avec valeur génétique ;
- et les descendants non génotypés mais avec valeur génétique.

Cela permet de prendre facilement en compte tous les animaux disponibles, qu'ils soient génotypés ou pas.

1.3.3 Les approches bayésiennes.

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

Meuwissen *et al.* (2001) modélisent les données sur deux niveaux :

- Au niveau des observations, le modèle correspond à une régression aléatoire simple des effets SNP telle que : $\mathbf{y} = \mu\mathbf{1_n} + \mathbf{X_j g_j} + \mathbf{e}$ où la seule distribution *a priori* informative est celle de $\mathbf{g_j}$, soit la distribution des effets des marqueurs. Elle est telle que $\mathbf{g_j} \sim \mathrm{N}(\mathbf{0}, \sigma^2_{g_j})$: on suppose donc bien des variances hétérogènes sur les effets des marqueurs.
- Au niveau de la modélisation des variances des effets des marqueurs, Meuwissen *et al.* (2001) proposent des distributions *a priori* différentes entre les méthodes BayesA et BayesB. On suppose que les effets des SNP proviennent d'une distribution normale avec une variance spécifique, différente d'un SNP à l'autre, de manière à ce que l'ordre de grandeur potentiel des effets des marqueurs soit variable. Gianola *et al.* (2009) démontrent qu'en fait, cela revient à postuler pour les effets des SNP une distribution *t* multivariée de faible degré de liberté. Cette distribution ressemble à une distribution normale « écrasée » avec des queues plus épaisses : de « gros » effets de SNP deviennent possibles contrairement au GBLUP

BayesA. Dans la méthode BayesA, Meuwissen *et al.* (2001) stipulent ainsi que les effets des SNP proviennent d'une distribution normale avec une variance spécifique d'un SNP à l'autre. Les variances sont modélisées selon une loi de χ^2 inverse : la plupart des marqueurs a un petit effet et quelques-uns ont un gros effet. La loi *a priori* des variances des effets SNP s'écrit :

$$P\left(\sigma^2_{g_j}\right) \sim \chi^{-2}(v, S)$$

avec *S* le paramètre d'échelle et *v* le nombre de degrés de liberté. Cela a l'avantage, si on considère une distribution normale des données, de conduire à une loi conditionnelle *a posteriori* de χ^2 inverse également :

$$P\left(\sigma^2_{g_j} | \, \mathbf{g_j}\right) \sim \chi^{-2}(v + n_j, S + \mathbf{g_j'g_j})$$

La loi *a posteriori* combine à la fois l'information apportée par les données et par la loi *a priori* supposée. Un échantillonnage de Gibbs est utilisé pour estimer les effets et leurs variances (voir chapitre 2).

BayesB. La méthode BayesB a l'avantage de ne sélectionner que les marqueurs utiles c'est-à-dire ceux dont l'effet sur le caractère étudié est significatif. Dans la méthode BayesB, Meuwissen *et al.* (2001) proposent un modèle dans lequel une proportion π (arbitrairement fixé à 0,95) des marqueurs a un effet nul ce qui évite un bruit de fond. La distribution *a priori* des variances des effets aux marqueurs s'écrit alors :

$$\begin{cases} \sigma^2_{g_j} = 0 \; avec \; une \; probabilté \; \pi \\ P\left(\sigma^2_{g_j}\right) \sim \chi^{-2}(v,S) avec \; une \; probabilté \; 1-\pi \end{cases}$$

L'échantillonnage de Gibbs ne peut pas être utilisé pour estimer les effets et les variances du modèle BayesB en raison de la probabilité élevée sur certains marqueurs d'être de variance nulle. On utilisera donc un algorithme de Metropolis-Hastings qui permet l'estimation simultanée de $\sigma^2_{g_j}$ et de $\mathbf{g_j}$.

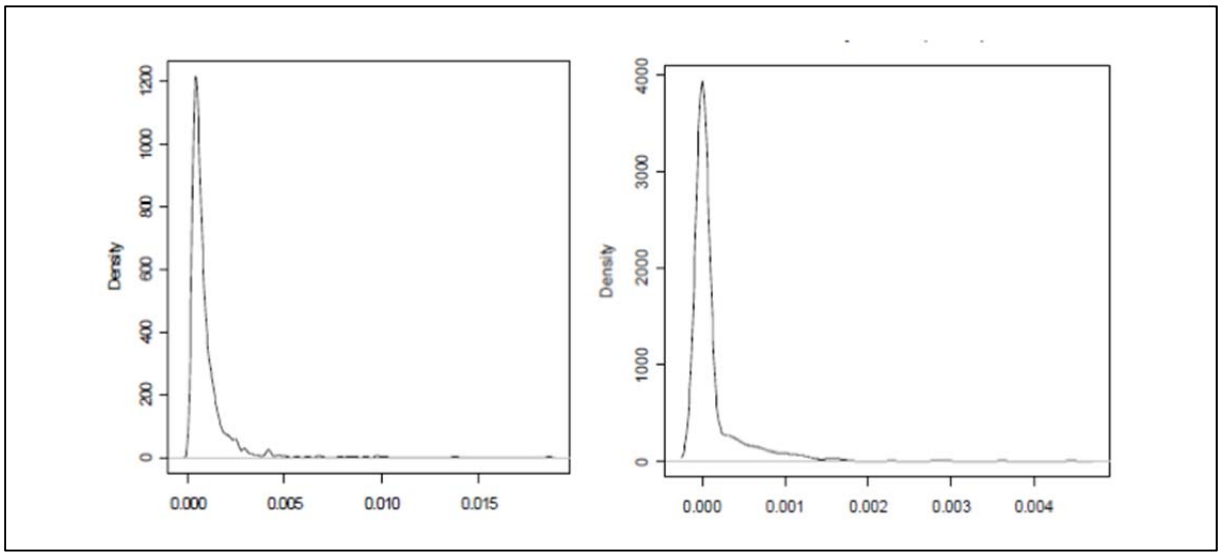


Figure 1.6 : Distributions a priori des variances des effets des marqueurs des méthodes BayesA et BayesB. (Hayes, 2011)

La figure 1.6, extraite de Hayes (2011) représente les différences entre les distributions *a priori* des méthodes BayesA et Bayes B. On remarque que dans la méthode BayesA, une grande quantité de marqueurs a une très faible variance alors que dans la méthode BayesB, une grande quantité a une variance nulle. De plus, dans les deux méthodes, une petite quantité de marqueurs a une variance

importante, 10 fois plus grande pour BayesA (de 0,015 à 0,020) que pour BayesB (de 0,002 à 0,004).

Sur la base des résultats de Meuwissen *et al.* (2001) et de nombreux travaux ultérieurs, la méthode BayesB est souvent considérée comme la référence en termes d'efficacité de prédiction génomique, mais elle est extrêmement coûteuse en temps de calcul. Cependant, Meuwissen *et al.* (2009) proposent une alternative à la méthode BayesB qui repose sur un algorithme rapide. Nous allons voir dans le paragraphe suivant que les méthodes bayésiennes se sont imposées comme les méthodes de référence pour la sélection génomique des bovins laitiers.

1.4 Application de la sélection génomique chez les bovins laitiers

Meuwissen *et al.* (2001) évaluent leurs méthodes sur un jeu de données simulées, composé d'un génome de 10 chromosomes de 100 cM chacun avec un marqueur tous les 1 cM. Les marqueurs sont introduits dans les modèles sous forme d'haplotypes. La population d'apprentissage réunit 2 200 animaux et la population de validation est composée des 2 000 descendants des animaux de la population d'apprentissage. Les capacités prédictives des différentes méthodes sont estimées en calculant la corrélation (ρ) et le coefficient (pente b) de régression entre les vraies valeurs génétiques et les valeurs prédites sur les animaux de la population de validation. Une valeur proche de 1 pour la pente de régression signifie que les valeurs prédites sont non biaisées. Le tableau 1.1 présente les résultats obtenus avec les méthodes GBLUP, la régression des moindres carrés et les approches BayesA et BayesB.

Tableau 1.1 : Capacités prédictives (corrélation ρ et pente de régression b entre valeurs génétiques vraies et valeurs prédites) des méthodes des moindres carrés, GBLUP, BayesA et BayesB sur données simulées (Meuwissen et al. 2001)

	ρ	***b***
Moindres carrés	0,318	0,285
GBLUP	0,732	0,896
BayesA	0,798	0,827
BayesB	0,848	0,946

La méthode des moindres carrés est la moins efficace car elle surestime les effets aux QTL (Moser *et al*., 2009). L'approche BayesB est la plus précise à la fois en termes de corrélation et de pente de régression. Il reste cependant un faible écart entre la pente de régression obtenue par les méthodes bayésiennes et 1 probablement dû à l'hypothèse d'une distribution *a priori* χ^{-2} pour BayesA et BayesB différente de la distribution simulée des variances. Goddard et Hayes (2009) comparent la corrélation de 0,85 annoncée par Meuwissen *et al*. (2001) à des résultats obtenus sur données réelles par VanRaden *et al.* (2009), Legarra *et al*. (2008) et González-Recio *et al*. (2009). VanRaden *et al*. (2009) produisent une corrélation moyenne sur plusieurs caractères de 0,71 à partir d'une population de référence de plus de 3 500 taureaux. Pour comparaison, la précision des estimations des valeurs génétiques des veaux, à la naissance, en se basant sur la moyenne de leurs parents, n'est que de 0,5 environ. Sur des données de souris, Legarra *et al*. (2008) démontrent l'avantage d'une évaluation génomique sur l'utilisation seule des données de pedigree : sur une moyenne de 4 caractères, le gain de précision d'une évaluation génomique par rapport à une évaluation sur pedigree inter-familles, est d'environ 0,13. Ces derniers résultats rejoignent la conclusion de González-Recio *et al*. (2009) sur des données issues de populations de poulets.

Les corrélations affichées par Meuwissen *et al*. (2001) sont du même ordre de grandeur que celles attendues entre prédictions de la valeur génétique après un testage sur descendance et valeurs génétiques vraies. Cependant, ces résultats sont issus de simulations qui postulent une homogénéité de l'information tant moléculaire que généalogique. De plus le caractère simulé présente une héritabilité modérée à élevée (h^2 = 0,3) alors qu'un des enjeux de la sélection génomique est aussi de progresser sur les caractères à faible héritabilité (h^2 < 0,1). Il était donc important de tester ces différentes méthodes sur des données réelles.

Les capacités prédictives des différentes méthodes sont très souvent jaugées en calculant, dans la population de validation, la corrélation entre les valeurs génétiques ou les DYD prédits et les DYD observés rétrospectivement ou les index sur descendance. La première étape a été de montrer l'avantage des index génomiques sur les index sur ascendance, calculés comme la moyenne des valeurs génétiques des parents (**PA** pour Parent Average). Des études ont montré la supériorité des évaluations génomiques (VanRaden, 2008) ou de la sélection

assistée par marqueurs, en France (Boichard *et al.,* 2002) sur les index sur ascendance : les corrélations entre valeurs prédites et observations sont meilleures que les index PA.

De nombreux auteurs ont appliqué les premières méthodes d'évaluation génomique décrites par Meuwissen *et al.* (2001) ou leurs méthodes dérivées sur des données réelles. Les approches BayesA et BayesB obtiennent des résultats souvent similaires ou légèrement supérieurs au GBLUP, pour la race bovine australienne Holstein-Friesian (+0,02 à +0,07 de gain de corrélation entre valeurs prédites et observées), par exemple (Hayes *et al.,* 2009a) et en Nouvelle-Zélande (+2% de gain de corrélation ; Harris *et al.,* 2009). Cependant, la méthode GBLUP est moins exigeante en temps de calcul que la méthode BayesA (Moser *et al.*, 2009 ; Solberg *et al*., 2009). Gredler *et al.* (2009) démontrent la supériorité de la méthode BayesB, en termes de précision des estimations génomiques, sur une méthode BayesA modifiée pour intégrer un effet polygénique (Hayes, 2009).

Ainsi, même si la méthode BayesB semble légèrement plus efficace que la méthode BayesA, de nombreuses études montrent que la méthode BayesB n'est pas tellement plus performante en termes de précision des estimations génomiques qu'un modèle GBLUP (Habier *et al*., 2010 ; Luan *et al*., 2009 ; Gredler *et al*., 2010). Les différences entre les meilleures méthodes sont généralement faibles et le plus souvent inférieures à 3%.

Nous verrons dans le chapitre suivant, que d'autres méthodes bayésiennes existent (BayesCπ, LASSO bayésien) et peuvent être efficacement applicables dans un contexte de sélection génomique des reproducteurs. Nous nous intéresserons également aux méthodes de sélection de variables et nous nous appliquerons à comparer chacune de ces méthodes entre elles, dans le but d'approfondir nos connaissances sur les approches d'évaluation génomique.

Chapitre 2 Modélisation et propriétés des approches statistiques

D'un point de vue statistique, l'évaluation génomique des reproducteurs utilise l'information génomique et/ou de parenté d'un ensemble d'animaux génotypés (variables explicatives) pour expliquer les performances observées (variable réponse) afin de construire le meilleur modèle prédictif. Puis, une fois le meilleur modèle établi, il permet de prédire les valeurs génomiques d'un ensemble d'individus candidats sur lesquels seuls les génotypes sont connus. Travailler avec des données génomiques (les marqueurs SNP) représente un challenge statistique compte tenu d'un nombre de variables explicatives (quelques dizaines de milliers de SNP) très supérieur au nombre d'observations (de plusieurs centaines à quelques milliers). Dans ce contexte, la régression linéaire multiple simple et la régression sur les moindres carrés de Meuwissen *et al.* (2001, voir chapitre 1) ne sont pas applicables en raison de la multicolinéarité (c'est-à-dire du caractère redondant) des variables explicatives. Cette multicolinéarité des variables (les marqueurs SNP) est due au déséquilibre de liaison entre les marqueurs. Si le nombre de facteurs est trop important, le modèle obtenu à partir de ces méthodes répondra parfaitement aux données sur lesquelles il a été construit mais ne prédira pas de façon précise de nouvelles données : c'est ce qu'on appelle le sur-apprentissage.

Les méthodes de sélection de variables (comme le LASSO bayésien, les méthodologies BayesCπ et Elastic Net) et de réduction de dimensions (telles que la régression PLS et la sparse PLS) permettent de construire des équations de prédiction à partir d'un nombre restreint d'individus. Néanmoins, la plupart de ces méthodes reste peu répandue dans le cadre de l'évaluation génomique des animaux domestiques tels que les bovins laitiers. Au cours de ma thèse, j'ai exploré et adapté les méthodes PLS et sparse PLS et appliqué les approches BayesCπ et LASSO bayésien afin de réaliser des évaluations génomiques. J'ai ensuite comparé les résultats obtenus aux méthodes d'évaluation génomique classiquement utilisées telles que le BLUP sur pedigree et le GBLUP (voir chapitre 1). Ce chapitre présente les spécificités statistiques de chaque approche (propriétés, modélisation et paramétrage statistiques) : les méthodes PLS et sparse PLS seront développées

plus particulièrement car elles ont fait l'objet d'une étude plus poussée que les méthodes bayésiennes dans mon travail de thèse.

Dans la plupart des méthodes décrites dans ce chapitre, le modèle statistique général est un modèle de régression multiple aléatoire, qui peut s'écrire sous la forme $\mathbf{y} = \mu\mathbf{1_n} + \mathbf{Xg} + \mathbf{e}$ où :

- **y** représente le vecteur des pseudo-performances des *n* individus de l'ensemble d'apprentissage ;
- μ est la moyenne globale de ce caractère sur la population d'apprentissage ;
- **X** est la matrice *nxp* des variables explicatives, vecteurs des génotypes des *p* marqueurs SNP sur les *n* individus ;
- **g** est un vecteur aléatoire contenant les coefficients de régression (les effets des SNP) d'éléments g_j et tels que $g_j \sim N(0, \sigma_g^2)$;
- et **e** est un vecteur aléatoire des effets résiduels tel que $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$.

Chaque colonne de la matrice **X** correspond à une variable explicative c'est-à-dire un SNP, codé selon si l'individu considéré est homozygote pour l'allèle 1 (codage 0), homozygote pour l'allèle 2 (codage 2) ou hétérozygote (codage 1) ; ce point est détaillé dans le chapitre 3. Le but est donc de prédire **y** à partir de la matrice **X**.

2.1 Les méthodes de régression pénalisée

Dans le cadre des méthodes classiques de régression pénalisée, les effets des SNP sont considérés comme des effets fixes du modèle (contrairement au modèle présenté ci-dessus). Les méthodes de sélection de variables dites de « régression pénalisée » ont l'avantage par rapport à la régression des moindres carrés de sélectionner les variables les plus importantes tout en estimant leurs effets et d'éliminer ou régresser vers zéro les autres variables. La régression Ridge (Hoerl et Kennard, 1970 ; Frank et Friedman, 1993) et la méthode **LASSO** (Least Absolute Shrinkage and Selection Operator, introduite par Tibshirani en 1996) sont deux méthodes de régression pénalisée : elles font tendre vers 0 certains effets SNP afin d'éliminer du modèle de régression les marqueurs dont l'effet est trop faible. La méthodologie Elastic Net (Zou et Hastie, 2003) allie la régression Ridge et la méthode LASSO afin de s'adapter au mieux aux données étudiées.

2.1.1 La régression Ridge

La différence entre régression Ridge et LASSO vient de la définition de la pénalisation. Une pénalité de type norme l_2, conduit à la régression Ridge. Les coefficients de régression g_j sont définis de la façon suivante :

$$\hat{\mathbf{g}}_{\boldsymbol{Ridge}} = argmin \left\{ \sum_{i=1}^{n} \left(y_i - \mu - \sum_{j=1}^{p} x_{ij} g_j \right)^2 + \lambda \sum_{j=1}^{p} g_j^2 \right\}$$

Le paramètre λ contrôle l'intensité de la pénalisation : plus λ est grand et plus les coefficients de régression sont corrigés et tendent vers 0 et les uns vers les autres. Au contraire quand λ est proche de 0, on tend vers la solution des moindres carrés. La pénalisation permet d'éviter que deux variables fortement négativement corrélées (donc avec des coefficients de régression opposés), s'annulent. Tous les effets g_j sont non nuls.

Le principal problème posé par l'utilisation de cette méthode est que le paramètre λ est généralement choisi de façon arbitraire. En pratique, on est amené à calculer la solution de la régression Ridge pour un grand nombre de valeurs du paramètre λ, afin de choisir la valeur qui convient. Le calcul de l'estimateur des coefficients de régression nécessite une inversion matricielle, la résolution du problème pour une large grille de λ peut rapidement s'avérer coûteuse numériquement. De plus, lorsqu'une faible valeur est choisie pour λ, le nombre de facteurs intervenant dans le modèle de régression est important et la solution du modèle peut ne pas être unique. Les auteurs Xu (2003) et Whittaker (2000) proposent des méthodes pour choisir ce paramètre. Xu (2003) conclut que l'utilisation de cette méthode de régression n'est pas adaptée car elle suppose que tous les marqueurs ont des effets de même variance, tout au long du génome alors que certains effets peuvent être négligeables. Cependant, cela n'a pas d'impact sur l'estimation des valeurs génomiques car les effets sont au final sommés sur beaucoup de marqueurs.

2.1.2 La méthode LASSO

La méthode LASSO (Tibshirani, 1996) repose sur une pénalisation de la valeur absolue des coefficients de régression de type norme l_1. Les coefficients de régression g_j sont tels que :

$$\hat{\mathbf{g}}_{\boldsymbol{LASSO}} = argmin \left\{ \sum_{i=1}^{n} \left(y_i - \mu - \sum_{j=1}^{p} x_{ij} g_j \right)^2 + \lambda \sum_{j=1}^{p} |g_j| \right\}$$

D'une part, les coefficients sont régressés vers 0 : l'introduction d'un biais entraîne une réduction de la variance. D'autre part, certains coefficients sont mis à zéro : par conséquent, l'estimation et la sélection de variables sont effectuées simultanément. Seules les variables ayant un effet significatif sur la variable réponse sont intégrées dans le modèle, ce qui correspond à l'hypothèse d'un nombre réduit de QTL avec un effet fort sur le caractère étudié (Hayes et Goddard, 2001). Si le paramètre λ est fixé à une très petite valeur, donc pour une sélection de variables très faible, alors les estimateurs LASSO des coefficients de régression correspondent aux estimateurs des moindres carrés (Hastie *et al.* 2001).

Dans le cas où $p > n$ (nombre de variables supérieur aux nombres d'observations), le LASSO sélectionne au mieux n variables : cela est dû à la nature convexe du problème d'optimisation. De plus, dans le cas où $n > p$ et en présence de variables fortement corrélées, la capacité prédictive du LASSO apparait inférieure à celle de la régression Ridge (Hastie *et al.*, 2001).

Un des inconvénients majeurs de cette méthode est qu'elle tend à sélectionner de manière aléatoire une seule variable parmi un ensemble de variables très corrélées. De plus, comme pour la régression Ridge, le paramètre de pénalisation λ est difficile à fixer. Le LASSO est rarement utilisé pour l'évaluation génomique mais des études ont cependant montré son efficacité, proche de celle d'un GBLUP ou de la méthode BayesA (Usai *et al*., 2009). Enfin, Tibshirani (1996) et Fu (1998) comparent la régression Ridge et la méthode LASSO sans réussir à montrer la supériorité de l'une ou de l'autre. Cependant, pour traiter des ensembles de données de taille de plus en plus importantes, la méthode de sélection de variables LASSO est beaucoup plus prometteuse.

2.1.3 L'Elastic Net

L'Elastic Net (**EN**, Zou et Hastie, 2003) repose sur une pénalité qui est une combinaison linéaire de celles de la régression Ridge et du LASSO. L'algorithme Elastic Net nécessite donc l'introduction d'un second paramètre noté α, compris entre 0 et 1, représentant la part de la pénalité LASSO et celle de la régression Ridge à prendre en compte dans le modèle :

$$\hat{\mathbf{g}}_{\boldsymbol{EN}} = argmin\left\{\sum_{i=1}^{n}\left(y_i - \mu - \sum_{j=1}^{p} x_{ij}g_j\right)^2 + \lambda\left((1-\alpha)\sum_{j=1}^{p} g_j^2 + \alpha\sum_{j=1}^{p}|g_j|\right)\right\}$$

Si α = 1, on obtient un modèle LASSO complet et si α = 0, on a un modèle de régression Ridge. De fait, l'Elastic Net a l'avantage de pouvoir palier certains comportements limitant du LASSO et de la régression Ridge. Il permet de choisir un sous-ensemble de variables pertinentes parmi un grand ensemble de variables y compris lorsque celles-ci sont très corrélées entre elles. L'utilisation de l'Elastic Net permet également de sélectionner toutes les variables alors que le LASSO sélectionne au plus *n* variables, soit le nombre d'observations du modèle.

Zou et Hastie (2003) montrent qu'en présence d'un ensemble de variables explicatives très corrélées, la régression Ridge assigne le même coefficient à toutes les variables de l'ensemble (« effet de groupe »), en les gardant toutes dans le modèle. Le LASSO, quant à lui, ne retient qu'une seule variable. Ainsi, l'Elastic Net procure un bon compromis entre ces deux cas extrêmes : dans le cadre d'une évaluation génomique, cette méthode sera préférée plutôt qu'un LASSO ou une régression Ridge.

Nous verrons à la fin de ce chapitre que l'étalonnage de l'Elastic Net peut être délicat et très dépendant de l'ensemble de données considéré.

2.2 Les méthodes de réduction de dimensions

La régression multiple peut être utilisée sur un nombre très important de variables explicatives tout en restant limité au nombre d'observations. De plus, même si le nombre de variables explicatives est élevé, il se peut qu'un ensemble restreint de facteurs sous-jacents (variables latentes) explique à lui seul la majorité de la variable réponse. Les méthodes de réduction de dimensions permettent

d'établir des équations de prédiction à partir de ces variables latentes $\boldsymbol{\xi_h}$, h=1,…,H, (où H est inférieur à p, le nombre total de variables explicatives) qui sont des combinaisons linéaires des données originelles X_j. Cela implique une réduction de la complexité du modèle et en facilite la compréhension. Les méthodes de réduction de dimensions diffèrent dans la façon de construire les combinaisons linéaires. On suppose que la matrice **X** est centrée réduite.

2.2.1 La régression sur composantes principales (RCP)

La régression sur composantes principales est une méthode de régression de la variable réponse **y** sur H combinaisons linéaires de variables explicatives issues d'une analyse en composantes principales. Dans l'analyse en composantes principales, **X** s'écrit sous la forme d'une décomposition en valeurs singulières (**SVD** pour Singular Value Decomposition) telle que **X** = **VΔS'** avec **V'V** = **S'S**= **I** et **Δ** est une matrice diagonale dont les éléments diagonaux sont les valeurs singulières. Les vecteurs singuliers (les colonnes de **V**) sont utilisés dans la RCP car leur orthogonalité élimine le problème de multicolinéarité de **X**. La RCP se décompose en trois étapes :

- La première étape consiste à effectuer une analyse en composantes principales sur l'ensemble des variables explicatives (**X**). Les composantes principales (notées v_j avec j =1,…,p) sont construites de façon à expliquer la plus grande variabilité de l'espace des variables explicatives possible. Cependant, rien ne garantit que ces composantes soient pertinentes pour expliquer les observations **y**.
- La deuxième étape consiste à régresser la variable réponse **y** sur ces composantes par un modèle de régression linéaire simple afin de sélectionner les H composantes principales v_j qui sont les plus corrélées avec la variable réponse.
- Enfin, **y** est régressé sur les H variables latentes $\boldsymbol{\xi_h} = \boldsymbol{X}v_h$ afin d'obtenir les coefficients de régression du modèle linéaire.

La première composante principale v_1 est telle que $\boldsymbol{\xi_1} = \boldsymbol{X}v_1$: elle a la plus grande variance parmi toutes les combinaisons linéaires des colonnes de **X** (Hastie *et al.*, 2001). Ainsi, la dernière composante principale a la plus petite variance.

2.2.2 La régression PLS

Historiquement, la régression PLS est née de l'algorithme **NIPALS** (Nonlinear estimation by Iterative Partial Least Squares) développé par Wold (1966) pour l'analyse en composantes principales. Cet algorithme permet de réaliser une analyse en composantes principales sur des données incomplètes, sans avoir à supprimer ni à estimer les données manquantes. L'acronyme PLS signifie à la fois :

- « Partial Least Squares » car le calcul des coefficients de régression est basé sur une régression des moindres carrés ;
- et « Projection to Latent Structures » car la prédiction repose sur l'extraction, parmi les variables explicatives, d'un ensemble de facteurs orthogonaux appelés variables latentes.

La régression PLS peut être considérée comme une généralisation de la régression multiple et de la régression sur composantes principales. Cette méthode connait un très grand succès dans le domaine de la chimie, particulièrement dans les applications concernant des données de chromatographie ou de spectrographie mais peut s'appliquer à de nombreux domaines (Tenenhaus, 1998) comme l'économie, la bioinformatique, le traitement des données, l'imagerie médicale et la génomique.

Le but de la régression PLS est de prédire **y** à partir de **X** et de décrire leurs similarités. Elle est particulièrement bien adaptée au cas où le nombre d'individus est très inférieur au nombre de variables explicatives et en présence de variables explicatives hautement corrélées.

La régression PLS repose sur la construction de variables latentes, qui sont des combinaisons des variables initiales, associées à des poids appelés vecteurs *loadings*. Dans mon travail de thèse, j'ai cherché à expliciter un seul caractère à la fois, bien que la régression PLS permette d'étudier des variables réponse multivariées. Ainsi, les variables latentes $\boldsymbol{\xi_h}$, h=1,…, H, sont construites seulement à partir des variables explicatives X_j de manière à maximiser, à chaque étape h, la covariance entre **y** et la variable latente $\boldsymbol{\xi_h}$. Il a été montré que la régression PLS recherche des variables latentes qui ont de fortes variances et de fortes corrélations avec la variable réponse (Frank et Freidman, 1993).

On suppose que le vecteur **y** est centré et que chaque X_j est centré et réduit. L'algorithme PLS débute par la construction d'une première variable latente telle que :

$$\boldsymbol{\xi_1} = \boldsymbol{\omega_{11}} X_1 + \ldots + \boldsymbol{\omega_{1p}} X_p$$

où les vecteurs *loadings* $\boldsymbol{\omega_{1j}}$ sont tels que $\boldsymbol{\omega_{1j}} = \frac{\mathrm{cov}(\mathrm{X_j},\mathbf{y})}{\sqrt{\sum_{j=1}^{p} \mathrm{cov}^2(\mathrm{X_j},\mathbf{y})}}$.

Puis, si le pouvoir explicatif de la régression de **y** sur $\boldsymbol{\xi_1}$ est trop faible, on recommence sur le vecteur des résidus $\mathbf{y_1}$ de la régression de **y** sur $\boldsymbol{\xi_1}$. Cette deuxième variable latente sera alors combinaison linéaire des résidus X_{1j} des régressions des variables X_j sur la composante $\boldsymbol{\xi_1}$. L'itération se poursuit jusqu'à atteindre le nombre de dimensions (nombre de variables latentes) *H* optimal. L'algorithme PLS peut se résumer sous la forme du problème d'optimisation suivant :

$$\max_{\|\mathbf{u_h}\|} cov\,(\mathbf{X_{h-1}}\mathbf{u_h}, \mathbf{y_{h-1}})$$

avec $\boldsymbol{\xi_1} = \mathbf{X_{h-1}}\mathbf{u_h}$. Le vecteur $\mathbf{u_h}$ est le vecteur *loading* associé à la variable latente $\boldsymbol{\xi_h}$, et où $\mathbf{X_h}$ et $\mathbf{y_h}$ sont les matrices et vecteurs résidus des régressions de $\mathbf{X_{h-1}}$ et $\mathbf{y_{h-1}}$ sur $\boldsymbol{\xi_h}$ à chaque itération *h*. L'algorithme est initialisé à $\mathbf{X_0} = \mathbf{X}$ et $\mathbf{y_0} = \mathbf{y}$.

La principale originalité de la régression PLS est qu'elle conserve l'asymétrie des liens entre les variables explicatives et la variable réponse (Abdi, 2007). Si le nombre de variables latentes extraites correspond au rang de la matrice **X** alors la régression PLS est équivalente à la régression multiple. L'avantage de la régression PLS sur la RCP est qu'elle prend en compte l'utilité individuelle des variables latentes pour prédire **y** alors que les composantes principales ne sont calculées qu'à partir des variables initiales **X**. En effet, la RCP est basée sur la décomposition de **X'X** alors que la régression PLS repose sur la décomposition en valeurs singulières de **X'y**.

Nous verrons à la fin de ce chapitre, que le nombre de variables latentes introduites dans le modèle de régression final peut être déterminé par des techniques de validation croisée.

La régression PLS peut également être utilisée comme méthode de sélection de variables en éliminant du modèle PLS, une fois entièrement construit, les marqueurs dont les effets estimés sont trop faibles (**PLS-VIP**, Lê Cao et Le Gall,

2011). Un critère permettant de rendre compte de cet effet est le coefficient **VIP** (Variable Importance in the Projection), décrit dans l'ouvrage de Tenenhaus (1998). Ce coefficient permet de juger de l'importance d'une variable explicative dans la construction du modèle de régression PLS. Le coefficient VIP de la variable *j* dans un modèle PLS à *H* composantes, est donné par :

$$VIP_{Hj} = \sqrt{\frac{p}{\sum_{h=1}^{H} cor^2(\mathbf{y}, \boldsymbol{\xi}_\mathbf{h})}} \sum_{h=1}^{H} cor^2(\mathbf{y}, \boldsymbol{\xi}_\mathbf{h}) \omega_{hj}^2$$

avec $\sum_{j=1}^{p} VIP_{Hj}^2 = p$. La contribution d'une variable X_j à la construction de la composante $\boldsymbol{\xi}_\mathbf{h}$ est mesurée par le poids ω_{hj}. Pour mesurer la contribution de la variable X_j sur la variable réponse **y** à travers la variable latente $\boldsymbol{\xi}_\mathbf{h}$, il faut prendre en compte le pouvoir explicatif de la variable latente $\boldsymbol{\xi}_\mathbf{h}$. Pour cela, il faut évaluer la redondance de l'information de $\boldsymbol{\xi}_\mathbf{h}$ sur **y** par le terme $cor^2(\mathbf{y}, \boldsymbol{\xi}_\mathbf{h})$. Les variables les plus importantes sont celles ayant un VIP au carré supérieur à 1 car $\sum_{j=1}^{p} VIP_{Hj}^2 = p$.

Dans l'étude des données bovines laitières présentée dans ce manuscrit, les coefficients VIP sont utilisés pour rendre compte de l'importance des effets des SNP sur le caractère étudié. Pour la sélection de variables, une approche d'estimation des effets SNP et de sélection en une étape sera utilisée : la sparse PLS.

2.2.3 La sparse PLS

La régression PLS n'est pas adaptée pour éliminer les variables qui ont un effet négligeable sur la variable réponse : toutes les variables explicatives de départ contribuent à la construction des variables latentes. Huang *et al.* (2004) proposent une méthode PLS pénalisée qui impose une valeur seuil à l'estimateur PLS final. Cependant, cela ne conduit pas nécessairement à des combinaisons linéaires faisant intervenir un nombre réduit de variables latentes. La sparse PLS a pour but d'imposer une « sparsité », c'est-à-dire une sélection de variables, directement dans l'étape de réduction de dimension. Chun et Keles (2010) ont introduit une version de la sparse PLS pour l'analyse de données de biopuces. Une pénalisation l_1, de type LASSO, est utilisée pour sélectionner les variables explicatives les plus pertinentes à intégrer dans chaque variable latente. Ces nouvelles variables sont ensuite

introduites dans un modèle de régression PLS final ce qui garantit l'orthogonalité des composantes. Chun et Keles (2010) comparent leur modélisation de la sparse PLS aux capacités prédictives des méthodologies RCP, Elastic Net et PLS sur des petits jeux de données simulées (un nombre d'observations *n* égal à 40 et un nombre de variables explicatives *p* égal à 80). L'erreur quadratique moyenne est calculée pour chacune des méthodes par validation croisée. Les méthodes de sélection de variables (sparse PLS et EN) obtiennent une erreur jusqu'à 30 fois moins importante que la régression PLS, et 17 fois moins importante que la RCP. Les meilleurs résultats sont obtenus en utilisant la sparse PLS et l'Elastic Net. Sur un deuxième jeu de données simulées (n = 100 et p = 5 000) en présence de covariables corrélées, la sparse PLS est comparée à la régression PLS, la RCP et la régression Ridge. À nouveau, la méthode de sélection de variables domine les autres méthodes. Cette étude permet de mettre en avant l'avantage des méthodes de sélection de variables sur les méthodes de régression basée sur l'ensemble des données (Chun et Keles, 2010).

La sparse PLS introduite par Lê Cao *et al.* (2008) a été développée pour l'analyse des données génomiques, protéomiques, métabolomiques et phénotypiques de grande dimension pour apporter aux biologistes une aide à la compréhension de leurs données. En effet, l'interprétation biologique est favorisée par la sélection, et donc la mise en avant des variables les plus pertinentes. Elle a été développée à la fois dans un cadre de régression et d'analyse canonique pour l'intégration de deux tableaux de données biologiques et la sélection de variables. Les premiers résultats ont été obtenus sur simulations et sur données réelles (données d'expression de gènes et transcriptomiques) dans le but de comparer les performances prédictives entre régressions PLS et sparse PLS. La sparse PLS augmente la capacité prédictive du modèle (taux de vrais positifs) et permet de mettre en avant certaines variables liées au problème biologique sous-jacent. Le paramétrage de la sparse PLS et notamment le choix du nombre de SNP à sélectionner repose sur une expertise biologique difficilement transposable à tous les domaines d'application. Nous verrons cependant, à la fin de ce chapitre, comment ces deux paramètres peuvent être estimés. La sparse PLS est une méthode rapide mais qui, dans l'étude de Lê Cao *et al.* (2008), n'a été testée que sur un nombre réduit d'échantillons (40 observations pour les données simulées et de 4 à 43 observations pour les trois jeux de données réelles testés).

La sparse PLS a également été appliquée à des études multivariées. Elle repose sur l'algorithme PLS-SVD (Lorber *et al*., 1987) qui propose de décomposer la matrice **X'Y** en vecteurs singuliers à la première étape de la régression PLS afin d'obtenir directement les premiers vecteurs *loadings* $\mathbf{u_1}$ et $\mathbf{v_1}$ de **X** et **Y**. La sparse PLS ajoute à chaque étape de la PLS, une pénalisation sur les vecteurs *loadings* construits. Cette pénalisation est la même que celle utilisée par Shen et Huang (2008) dans la sparse ACP.

Au cours de ce travail, j'ai simplifié l'algorithme sparse PLS, développé par Lê Cao *et al.* (2008) afin d'éliminer l'étape décomposition SVD qui devient inutile pour notre étude univariée. Le problème d'optimisation à résoudre est donc :

$$\max_{\|\mathbf{u_h}\|=1} cov\left(\mathbf{X_{h-1}u_h}, \mathbf{y_{h-1}}\right) + \mathrm{g}_\lambda\left(\mathbf{u_h}\right)$$

où $\mathrm{g}_\lambda(\mathbf{x}) = sign(x)(|x| - \lambda)_+$ est la fonction de pénalité *soft-thresholding* qui permet de réaliser la sélection de variables.

2.3 Les méthodes bayésiennes

Les méthodes bayésiennes reposent sur un modèle statistique où tous les effets sont considérés comme des effets aléatoires. Dans le cadre de la sélection génomique, on s'attend à obtenir une large majorité de marqueurs avec des effets faibles, voire négligeables et peu de marqueurs avec des effets de grand taille. En effet, la très grande majorité des SNP ne sont pas des mutations causales ou ne sont même pas en très fort déséquilibre de liaison avec elles. Les méthodes bayésiennes permettent d'introduire des distributions *a priori* des effets des SNP plus adaptées à cette idée. Elles sont très largement utilisées dans l'évaluation génomique bovine, comme méthodes de sélection de variables. Elles se distinguent des méthodes fréquentistes par l'utilisation d'une *loi a priori* qui exprime l'information « intuitive » qu'on peut avoir sur les données sans les avoir traitées. La *loi a posteriori*, c'est-à-dire après l'étude des données, exprime la petite part d'incertitude restante et sera à la base des prédictions des observations futures. Ce sont les lois *a priori*, c'est-à-dire les hypothèses faites sur la distribution des SNP qui distinguent les méthodes bayésiennes les unes des autres.

2.3.1 Le LASSO Bayésien

Le LASSO bayésien (Park et Casella, 2008) est l'équivalent bayésien du LASSO décrit au point 2.1.2. Il fait l'hypothèse que les effets des marqueurs suivent une loi de Laplace (ou « double exponentielle »). La loi de Laplace peut être vue comme l'association de deux lois exponentielles, accolées dos à dos : on suppose donc qu'un grand nombre de SNP a un effet pratiquement nul et que très peu ont un effet très important.

Tibshirani (1996) montre que la loi des estimateurs LASSO peut s'écrire :

$$P(g_j|\sigma^2,\lambda) = \frac{\lambda}{2}\exp(-\lambda|g_j|).$$

Il suggère que les estimateurs LASSO peuvent être interprétés comme un mode *a posteriori* d'un modèle dans lequel les paramètres de régression seraient indépendants et identiquement distribués selon une loi *a priori* de Laplace.

Park et Casella (2008) proposent d'utiliser une approche bayésienne complète en faisant l'hypothèse d'une distribution *a priori* des coefficients de régression tels que :

$$P(g_j|\sigma^2,\lambda) = \frac{\lambda}{2\sqrt{\sigma^2}}\exp(-\frac{\lambda|g_j|}{\sqrt{\sigma^2}}).$$

où σ^2 représente la variance des effets résiduels du modèle et la variance du vecteur des effets SNP **g** est $\sigma_g^2 = \frac{2\sigma^2}{\lambda}$. Ils démontrent qu'il est important de conditionner la loi des estimateurs par σ afin de garantir une loi *a posteriori* unique. Les applications du LASSO bayésien à la sélection génomique proposées par de los Campos *et al.* (2009) et Weigel *et al.* (2009) utilisent la même variance σ² pour modéliser à la fois la distribution des effets des SNP et les résidus.

Au cours de ce travail, j'ai choisi d'utiliser le LASSO bayésien général développé par Legarra *et al.* (2011) pour la sélection génomique et qui se rapproche du LASSO classique en divisant la variance σ² en un terme purement résiduel (σ_e^2) et une variance due aux marqueurs (σ_g^2). La loi *a priori* des effets des marqueurs est la même que pour le LASSO classique où les résidus suivent une loi normale multivariée :

$$\mathbf{e}|\mathbf{I}\sigma_e^2 \sim MVN(0, \mathbf{I}\sigma_e^2).$$

Comme dans le LASSO classique, le paramètre *λ* est un paramètre d'échelle : il est utilisé pour définir l'intensité de sélection des SNP. On suppose que la distribution *a priori* de *λ* est une loi uniforme entre 0 et un très grand nombre. Le

LASSO bayésien traite ce paramètre d'intensité comme un hyperparamètre inconnu et génère son échantillonnage en même temps que les autres paramètres du modèle. Dans l'article de Legarra *et al.* (2011) est présentée la méthode pour estimer les paramètres de ce modèle. De los Campos *et al.* (2009) montrent que le LASSO bayésien est proche en termes de précision des prédictions, de la méthode BayesB mais avec une réduction importante de la complexité des calculs. C'est aussi un bon compromis entre le LASSO classique et la régression Ridge. Cependant, le LASSO bayésien régresse les coefficients des marqueurs à effets faibles vers 0 plus rapidement que la régression Ridge ce qui porte à croire que la distribution de Laplace est avantageuse sur la loi Gaussienne. De plus, le nombre de marqueurs pouvant avoir un effet nul n'est pas limité au nombre d'observations comme pour le LASSO classique. Le LASSO bayésien fait donc partie des méthodes qui seront appliquées aux données bovines laitières françaises.

2.3.2 La méthode BayesC

Comme montré dans le chapitre précédent, les méthodes bayésiennes telles que BayesA et BayesB (Meuwissen *et al*., 2001) ont été largement utilisées pour réaliser des évaluations génomiques. Des méthodes apparentées existent, avec des performances similaires, développées dans le but de réduire les temps de calcul et de simplifier les modélisations statistiques.

La méthode BayesC (Kizilkaya *et al*., 2010) diffère du BayesB en supposant la variance associée aux SNP commune à tous les marqueurs. En BayesC, tout comme en BayesB, la probabilité π qu'un SNP ait un effet non nul, est supposée connue. Le modèle est semblable au modèle BayesB mais pour une variance des effets homogène sur tous les loci :

$$\begin{cases} \sigma_g^2 = 0 \; avec \; une \; probabilité \; 1 - \pi \\ P\left(\sigma_g^2\right) \sim \chi^{-2}(v, S) avec \; une \; probabilité \; \pi \end{cases}$$

2.3.3 La méthode BayesCπ

Le principal problème de la méthode BayesC est que la part de SNP ayant un effet non nul, est supposée connue. Avec la méthode BayesA, le paramètre π est égal à 1 ce qui implique que tous les marqueurs ont un effet. Pour la méthode

BayesB, π est strictement inférieur à 1 afin de prendre en compte l'hypothèse que certains SNP peuvent avoir un effet nul mais est fixé de façon arbitraire alors que l'intensité de la sélection de variables est contrôlée par ce paramètre : il devrait donc être estimé à partir des données. Habier *et al*. (2011) proposent de modifier la méthode BayesC en estimant le paramètre π : le paramètre π est supposé inconnu. Ainsi, la distribution *a priori* de π devient uniforme sur [0,1]. La modélisation des effets SNP est la même qu'avec BayesC :

$$\begin{cases} P(g_j|\pi,\sigma_g^2) = 0 \; avec \; une \; probabilité \; 1-\pi \\ P\left(g_j|\pi,\sigma_g^2\right) \sim N\left(0,\sigma_g^2\right) \; où \; P\left(\sigma_g^2\right) \sim \chi^{-2}(v,S) avec \; une \; probabilité \; \pi \end{cases}$$

Les différents paramètres de ce modèle sont estimés par des méthodes **MCMC**, Markov Chain Monte Carlo (Metropolis *et al.,* 1953 ; Robert, 1996) comme proposé par Habier *et al*. (2011). La valeur initiale assignée à la variance des effets des marqueurs est décrite par VanRaden *et al.* (2009). Elle s'écrit en fonction de la variance génétique additive σ_a^2 : $\sigma_g^2 = \frac{\sigma_a^2}{(1-\pi)\sum_{j=1}^{p} 2p_j(1-p_j)}$ où p_j est la fréquence allélique du SNP *j*.

2.4 Estimation et choix des paramètres statistiques

Un panel varié de méthodes a été testé ou implémenté dans le cadre du projet AMASGEN et au cours de mon travail de thèse : une méthode de régression pénalisée (l'Elastic Net), une méthode de réduction de dimensions pure (la régression PLS), deux méthodes bayésiennes de sélection de variables (le LASSO bayésien et le BayesCπ) et une méthode alliant réduction de dimensions et sélection de variables (la sparse PLS). Chacune d'entre elles réunit un nombre de paramètres plus ou moins faciles à fixer. L'étalonnage des méthodes fréquentistes passe par un problème d'optimisation de critères sur validation croisée et l'étalonnage des approches bayésiennes, par des méthodes d'échantillonnage MCMC. L'étalonnage des méthodes Elastic Net, PLS et sparse PLS sera particulièrement développé dans ce chapitre car cela a été un point essentiel de discussion sur l'efficacité de ces méthodes, dans le cadre du projet AMASGEN. Le LASSO bayésien et le BayesCπ ont été appliqués sur les données bovines laitières françaises en utilisant le programme GS3 développé par Legarra *et al.* (2011, http://snp.toulouse.inra.fr/~alegarra). Les paramètres et les distributions *a posteriori*

des méthodes bayésiennes sont récupérés en sortie de ce programme. Les variables μ, g_j et σ_g^2, σ_e^2 et π sont estimées par une méthode MCMC, l'échantillonnage de Gibbs (Geman et Geman, 1984).

Pour modéliser les données, étalonner puis estimer la capacité prédictive d'un modèle de régression statistique, l'idéal serait de disposer de trois jeux de données indépendants : un ensemble d'apprentissage pour estimer les effets des variables explicatives, un ensemble de validation pour étalonner les paramètres du modèle et un ensemble test sur lequel serait évalué le modèle ainsi construit. Cependant, il est très rare de disposer d'un ensemble de données suffisamment grand pour faire un tel partage. La validation croisée est une méthode d'échantillonnage souvent appliquée pour l'estimation de la fiabilité d'un modèle. Il existe trois techniques principales de validation croisée :

- « Test set validation » : cette première méthode, très simple, consiste à diviser l'échantillon de taille *n* en un échantillon d'apprentissage (environ 75% de l'échantillon) et un échantillon de validation (25% de l'échantillon). Le modèle est bâti sur l'échantillon d'apprentissage et validé sur l'échantillon de validation. Ainsi, l'ensemble de validation est utilisé pour choisir le modèle optimal mais aussi pour rendre compte de sa fiabilité. L'erreur est estimée en calculant l'erreur quadratique moyenne.
- « K-fold validation » : l'ensemble des données est divisé *k* fois, puis un des *k* échantillons est sélectionné comme ensemble de validation tandis que les (*k-1*) autres échantillons constituent l'ensemble d'apprentissage. On calcule, comme dans la première méthode, par exemple, l'erreur quadratique moyenne. Puis on répète l'opération en sélectionnant un autre échantillon de validation parmi les (*k-1*) échantillons qui n'ont pas encore été utilisés pour la validation du modèle. L'opération se répète ainsi *k* fois pour qu'au final, chaque sous-échantillon ait été utilisé exactement une fois comme ensemble de validation. Enfin, la moyenne des *k* erreurs quadratiques moyennes est calculée pour estimer l'erreur de prédiction.
- « Leave-one-out validation » : il s'agit d'un cas particulier de la seconde méthode où $k = n$; l'apprentissage est réalisé sur (*n-1*) observations puis la validation du modèle porte sur la $n^{ième}$ observation. Cette opération est répétée *n* fois.

Il existe plusieurs critères pour choisir les paramètres d'un modèle par validation croisée : on peut calculer l'erreur moyenne de prédiction sur tous les échantillons ou chercher à maximiser la corrélation entre valeurs vraies et valeurs prédites.

2.4.1 Paramétrage de l'Elastic Net

Dans les méthodes de régression pénalisée, choisir la meilleure valeur pour le coefficient de pénalité revient à choisir la taille du sous ensemble de données optimale c'est-à-dire celle qui donnera la plus petite erreur de prédiction ou la plus forte corrélation entre les valeurs vraies et les valeurs prédites. On utilise pour cela la première méthode de validation croisée car, dans le contexte de l'évaluation génomique, un ensemble d'apprentissage et un ensemble de validation sont déjà définis. De manière générale, l'ensemble d'apprentissage est composé d'une sous-population d'individus les plus anciens et l'ensemble de validation des animaux les plus jeunes (candidats à la sélection).

L'Elastic Net est contrôlé par deux paramètres : le paramètre d'intensité de sélection λ et le paramètre α qui attribue la part de LASSO et de régression Ridge à introduire dans le modèle EN. Ces deux paramètres sont liés l'un à l'autre et sont donc très difficiles à estimer. Au cours de ce travail, plusieurs valeurs des paramètres α et λ ont été testées selon une grille de recherche établie par Croiseau *et al.,* (2011). L'objectif est de maximiser la corrélation entre les valeurs de **y** observées et les valeurs prédites par les différents modèles testés sur l'ensemble de validation. De cette façon, le modèle est construit spécifiquement pour ces ensembles de données ce qui peut être un avantage par rapport aux autres méthodes. Le paramètre α est fixé par recherche dichotomique entre 0 et 1. On commence par tester le modèle pour α=0,1 et 0,5 en calculant la corrélation entre les valeurs observées dans la population de validation et les valeurs prédites par les différents modèles, bâtis sur l'ensemble d'apprentissage en fixant α à la valeur choisie. Si la meilleure corrélation est obtenue avec une valeur de α égale à 0, alors, à la seconde itération, l'intervalle sera réduit à [0 ; 0,5]. Si la valeur retenue est 1, alors l'intervalle sera réduit à [0,5 ; 1] et si α = 0,5, alors l'intervalle sera [0,25 ; 0,75]. Les itérations s'arrêtent quand la différence entre deux valeurs de α testées est inférieure à 0,02. De plus, pour chaque valeur de α, 500 valeurs de λ sont testées

avec des valeurs comprises entre 0 et la valeur absolue du coefficient de régression le plus élevé d'un modèle de régression non pénalisé. Au final, le modèle Elastic Net qui donne la meilleure corrélation est sélectionné.

2.4.2 Paramétrage de la régression PLS

En utilisant la régression PLS, le nombre de variables latentes intervenant dans le modèle final doit être choisi. Dans l'algorithme PLS, la construction d'une nouvelle variable latente est justifiée si elle améliore la prédiction de **y** pour de nouvelles observations. Plusieurs auteurs, par exemple Denham (2000) et Mevik et Cederkvist (2004) proposent d'utiliser la troisième méthode de validation croisée en calculant à chaque itération *i* l'erreur de prédiction. L'erreur de prédiction moyenne (**MSEP**, Mean Squared Error of Prediction) sur les *n* échantillons construits en enlevant une seule donnée à chaque itération, correspond à la somme des écarts au carré entre la valeur observée sur la donnée écartée et son estimation sur le modèle à *n-1* observations et *h* variables latentes :

$$MSEP = \sum_{i=1}^{n} (y_i - \hat{y}_h\,(-i))^2 \,.$$

Au final, le modèle sélectionné sera celui qui minimise le critère MSE.

Sur la même méthode de validation croisée (« leave-one-out validation ») peuvent être appliqués les critères **RSS** (somme des carrés des résidus du modèle) et **PRESS** (somme de tous les carrés des erreurs de prédiction calculées sur les différents échantillons de validation). La variable latente $\boldsymbol{\xi_h}$ est retenue dans le modèle (Abdi, 2010) si $\sqrt{PRESS_h} \leq 1 - \sqrt{RSS_{h-1}}$.

Une méthode de validation très répandue avec la régression PLS repose sur ces deux critères : le coefficient de corrélation multiple « cross-validé », noté Q^2. Une nouvelle composante $\boldsymbol{\xi_h}$ est significative si Q_h^2 est supérieur à 0,095 ce qui est équivalent au critère précédent. Cependant, les premiers résultats du coefficient de corrélation multiple « cross-corrélé » obtenus sur les données bovines laitières françaises, estime qu'un modèle à une seule variable latente serait le plus ajusté. Mais la capacité prédictive des modèles PLS à une variable latente apparait très faible donc ce critère de validation n'a plus été utilisé par la suite.

Dans notre étude, le nombre de variables latentes est choisi de la même façon que le critère de pénalité de l'Elastic Net c'est-à-dire en maximisant la

corrélation entre les valeurs observées et prédites dans l'ensemble de validation prédéfini. En général, la qualité de prédiction d'un modèle aléatoire n'augmente pas forcément avec le nombre de variables latentes utilisées. Elle atteint un maximum puis diminue. Toutes les valeurs comprises entre 1 et 100 ont été testées mais nous verrons dans le chapitre 4 que le critère de corrélation atteint un plateau bien avant 50 variables latentes puis diminue, ce qui suggère qu'il n'est pas utile de construire des modèles plus complexes. Une autre approche aurait été de choisir le nombre minimal de variables latentes dont les résidus ne sont pas significativement plus grands que ceux du modèle avec l'erreur de prédiction minimale afin de réduire le nombre de variables latentes dans la régression PLS.

2.4.3 Paramétrage de la sparse PLS

Deux paramètres sont à estimer lors de l'utilisation de la sparse PLS : le nombre de variables latentes *H* (comme pour la régression PLS) et le paramètre d'intensité de sélection, défini par Lê Cao *et al.* (2008) comme le nombre de variables à sélectionner par dimension du modèle, noté n_{dim}. Le paramétrage de l'Elastic Net s'est avéré très difficile en raison de la présence de deux paramètres liés l'un à l'autre et qui sont estimés simultanément. Pour plus de simplicité et en suivant la démarche de Lê Cao *et al.* (2008), on suppose que le nombre de variables à sélectionner est le même sur chacune des variables latentes. Dans leur étude, Chun et Keles (2010) n'utilisent pas non plus de paramètres d'intensité de sélection différents sur chaque variable latente pour ne pas multiplier le nombre de paramètres à estimer. Dans notre étude, une grille de valeurs a été choisie arbitrairement : *H* varie de 1 à 100 et 10 valeurs de n_{dim} ont été testées, exprimées selon un pourcentage du nombre de variables explicatives initiales (de 0,2% à 10%).

Le modèle optimal est ensuite choisi par validation croisée « 10-fold » c'est-à-dire qu'un modèle est construit pour chaque combinaison de paramètres sur 10 ensembles d'apprentissage différents. La **RMSEP** (Root Mean Squared Error of Prediction) est calculée pour chaque combinaison de paramètres : $RMSEP = \frac{1}{10}\sum_{k=1}^{10}(\mathbf{y_k} - \hat{\mathbf{y}}_\mathbf{k})^2$ où $\widehat{\boldsymbol{y_k}}$ est le vecteur des valeurs prédites de l'ensemble de validation du $k^{ième}$ tirage pour une combinaison de paramètres testée. Le paramètre n_{dim} est choisi de façon à minimiser la RMSEP moyenne quelque soit la valeur de *H*.

Puis, *H* est fixé de la même façon que pour la régression PLS, c'est-à-dire de manière à ce que la corrélation entre les valeurs prédites par le modèle et les valeurs observées dans l'ensemble de validation (25% des plus jeunes individus de la population totale) soit maximale. La démarche suivie et les différents tests effectués sur les données bovines laitières des races Holstein et Montbéliarde, pour estimer les paramètres de la sparse PLS sont décrits dans le chapitre 4.

2.5 Comparaison des méthodes de sélection génomique

2.5.1 Prise en compte des EDC dans les modèles

Les données phénotypiques utilisées dans les études d'évaluation génomique des bovins laitiers sont transformées en DYD. À chaque DYD, et donc à chaque animal, est associé un poids, appelé EDC. Il représente la fiabilité du DYD : il est fortement lié au nombre de filles de chaque taureau et donc à la quantité d'information disponible. Le fait d'associer un EDC différent pour chaque animal revient à considérer un modèle à variances hétérogènes. En supposant que les observations sont centrées, le modèle de régression s'écrit sous la forme suivante : **y** = **Xg** + **e** où **g** est le vecteur des effets aléatoires des SNP d'éléments g_j et tels que $g_j \sim N(0, \sigma_g^2)$ et **e** est le vecteur des effets résiduels tel que $\mathbf{e} \sim N(0, \mathbf{I}\,\frac{\sigma_e^2}{EDC})$ (VanRaden et Wiggans, 1991). Pour obtenir des variances résiduelles homogènes tout en conservant un modèle équivalent, il est nécessaire de multiplier y_i et la $i^{ème}$ ligne de la matrice d'incidence **X** par $\sqrt{EDC_i}$:

$$\sqrt{EDC}\mathbf{y} = \sqrt{EDC}\mathbf{X}\mathbf{g} + \mathbf{e} \text{ avec } \mathbf{e} \sim N(0, \mathbf{I}\,\sigma_e^2).$$

Le vecteur des observations **y** et la matrice d'incidence des SNP sont affectés par les EDC, quelque soit la méthode de régression étudiée. Nous verrons dans le chapitre 4, que la prise en compte des EDC dans les méthodes d'évaluation génomique a un impact sur la qualité des résultats.

2.5.2 Capacités prédictives des méthodes

Pour comparer les capacités prédictives des différentes méthodes testées, des critères de validation ont été choisis. Le critère le plus répandu dans les études

d'évaluation génomique est la corrélation ρ entre les DYD observés ($\mathbf{DYD_{obs}}$) et prédits ($\widehat{\mathbf{DYD}}$) des animaux de l'ensemble de validation. Ce coefficient est égal au rapport de leur covariance $cov(\mathbf{DYD_{obs}}, \widehat{\mathbf{DYD}})$ et du produit non nul de leurs écarts types $\sigma_{DYD_{obs}}$et $\sigma_{\widehat{DYD}}$:

$$\rho = cor\left(\mathbf{DYD_{obs}}, \widehat{\mathbf{DYD}}\right) = \frac{cov(\mathbf{DYD_{obs}}, \widehat{\mathbf{DYD}})}{\sigma_{DYD_{obs}}\sigma_{\widehat{DYD}}}$$

Le coefficient de corrélation est compris entre -1 et 1. Il est égal à 1 dans le cas où l'une des variables est une fonction affine croissante de l'autre variable et -1 dans le cas où la fonction affine est décroissante. Les valeurs intermédiaires renseignent sur le degré de dépendance linéaire entre les deux variables. Plus le coefficient est proche des valeurs extrêmes -1 et 1, plus la corrélation entre les variables est forte. Une corrélation égale à 0 signifie que les variables sont linéairement indépendantes.
Cette corrélation est calculée par validation croisée car deux ensembles de données différents sont utilisés : un pour établir les équations de prédiction (ensemble d'apprentissage) et le deuxième pour calculer le critère de corrélation (ensemble de validation). Elle donne une appréciation de la précision des évaluations génomiques produites. Les EDC sont pris en compte dans le calcul de la corrélation en utilisant la formule suivante :

$$\rho_{DYD_{obs}\widehat{DYD}} = \frac{\sum_i EDC_i(DYD_{obs,i} - \overline{DYD_{EDC}})(\widehat{DYD_i} - \overline{\widehat{DYD}_{EDC}})}{\sqrt{\sum_i EDC_i(DYD_{obs,i} - \overline{DYD_{EDC}})^2 \sum_i EDC_i(\widehat{DYD_i} - \overline{\widehat{DYD}_{EDC}})^2}}$$

où $\overline{DYD_{EDC}} = \frac{\sum_i EDC_i DYD_{obs,i}}{\sum_i EDC_i}$ et $\overline{\widehat{DYD}_{EDC}} = \frac{\sum_i EDC_i \widehat{DYD_i}}{\sum_i EDC_i}$.

Dans le cadre de coopérations européennes menées par Interbull, un autre critère de validation des estimations génomiques est la pente de la régression des DYD observés des animaux de l'ensemble de validation sur les prédictions génomiques. Nous nous sommes donc aussi intéressés à ce critère de validation. La droite de la régression linéaire des $\mathbf{DYD_{obs}}$sur les $\widehat{\mathbf{DYD}}$ est calculée par la méthode des moindres carrés selon le modèle :

$$\mathbf{DYD_{obs}} = b\widehat{\mathbf{DYD}} + a.$$

Le coefficient *b* représente la pente de la droite de régression et correspond au rapport entre la covariance de $\mathbf{DYD_{obs}}$ et $\widehat{\mathbf{DYD}}$ et la variance de $\widehat{\mathbf{DYD}}$:

$$b = \frac{cov(\mathbf{DYD_{obs}}, \widehat{\mathbf{DYD}})}{\sigma_{\widehat{DYD}}}$$

Dans la régression des $\mathbf{DYD_{obs}}$ sur les $\widehat{\mathbf{DYD}}$, les EDC sont pris en compte de la même façon que dans l'établissement des équations de prédiction. La valeur de la pente doit être la plus proche possible de 1, selon les recommandations du comité Interbull de mars 2011. On peut ainsi construire pour chaque coefficient de régression *b* un intervalle de confiance à 95%, égal à ± 1,96 fois l'erreur standard (quantile au seuil de confiance α=95% de la loi de Student).

2.5.3 Le test de Hotelling-Williams

L'égalité entre les corrélations obtenues à partir des différentes méthodes a été testée en appliquant le test de Hotelling-Williams (Van Sickle, 2003). Il est utilisé pour comparer deux corrélations dépendantes, partageant une variable (dans notre cas, les DYD observés sur la population de validation). L'hypothèse nulle correspond à l'égalité entre les deux corrélations considérées. Si on note $\widehat{\mathbf{DYD_A}}$ le vecteur des DYD prédits à partir de la méthode A, $\widehat{\mathbf{DYD_B}}$ le vecteur des DYD prédits à partir de la méthode B et $\mathbf{DYD_{obs}}$ le vecteur des DYD observés dans la population de validation alors :

$$\rho_{A,obs} = cor(\widehat{\mathbf{DYD_A}}, \mathbf{DYD_{obs}}), \rho_{B,obs} = cor(\widehat{\mathbf{DYD_B}}, \mathbf{DYD_{obs}})$$

et

$$\rho_{A,B} = cor(\widehat{\mathbf{DYD_A}}, \widehat{\mathbf{DYD_B}})$$

La statistique du test de Hotelling-Williams s'écrit :

$$t = \left|\rho_{A,obs} - \rho_{B,obs}\right| \sqrt{\frac{(n-1)(1+\rho_{A,B})}{2\frac{n-1}{n-3}|R| + \bar{r}^2(1-\rho_{A,B})^3}}$$

où $|R| = 1 - \rho_{A,obs}^2 - \rho_{B,obs}^2 - \rho_{A,B}^2 + 2\rho_{A,obs}\rho_{B,obs}\rho_{A,B}$ et $\bar{r} = \frac{\rho_{A,obs}+\rho_{B,obs}}{2}$.

Sous l'hypothèse nulle, la statistique de test suit une loi de Student à *n-3* degrés de liberté. Les corrélations obtenues par les différentes méthodes sont comparées deux à deux, au seuil de significativité de 5%.

Cinq méthodes (Elastic Net, PLS, sparse PLS, BayesCπ et LASSO bayésien) ont donc été choisies pour leurs propriétés statistiques et leur fiabilité dans le cadre du projet AMASGEN. Les résultats des méthodes PLS, sparse PLS, BayesCπ et LASSO bayésien, seront présentés dans la suite de ce manuscrit. Une partie de mon travail de thèse, pour les régressions PLS et sparse PLS, a été d'adapter et de modifier le programme R du package mixOmics (Lê Cao *et al.,* 2009) pour l'application de ces méthodes aux données bovines laitières françaises. Pour les méthodes BayesCπ et LASSO bayésien, j'ai dû apprendre à maitriser le logiciel GS3 (Legarra *et al.,* 2011) et préparer les fichiers d'entrée. Les évaluations BLUP, GBLUP et Elastic Net ont été réalisées par les autres membres du projet AMASGEN. L'efficacité de ces méthodes sera vérifiée en les comparant aux deux méthodes d'évaluation animale les plus répandues : le BLUP sur pedigree et le GBLUP. Le chapitre suivant présente les données issues des deux races de bovins laitiers français sur lesquelles ont été appliquées ces méthodologies.

Chapitre 3 Deux populations de référence de bovins laitiers français : la race Holstein et la race Montbéliarde

3.1 Description des caractères

La sélection génétique vise à choisir, en fonction de leur profil génétique, les meilleurs reproducteurs, c'est à dire ceux susceptibles de produire des descendants réunissant certaines qualités attendues par les éleveurs. L'intérêt principal chez les bovins laitiers porte évidemment sur l'amélioration de la production laitière et de ses composantes (quantité de matières, taux butyreux et protéiques) ainsi que certaines aptitudes fonctionnelles (résistance aux mammites, fertilité, longévité, points de conformité au standard de la race, *etc*.). Ainsi, pour chaque taureau destiné éventuellement à la reproduction, son potentiel génétique sera évalué à partir de ses performances et de celles de ses éventuels descendants, ascendants et collatéraux. Comme cela a été décrit dans le chapitre 1, pour obtenir le phénotype d'un taureau sur des caractères laitiers (donc mesurés sur les femelles), on utilise la moyenne des performances de ses filles, corrigée des effets environnementaux (effets fixes et aléatoires non génétiques) et de la part génétique maternelle, afin de ne conserver que la part correspondant aux variations génétiques dues au taureau. C'est ce qu'on appelle le DYD qui constituera la variable réponse dans les modèles décrits dans ce document. Plus le nombre de filles évaluées pour un même taureau est élevé, et plus la « pseudo-performance » de ce taureau est précise. Un EDC est alors associé à chaque « pseudo-performance » ; cet EDC dépend du nombre de filles considérées dans le calcul du DYD ainsi que des caractéristiques du caractère étudié (voir chapitre 1).

Au cours du projet AMASGEN, un panel d'environ 25 caractères a été récolté et analysé :

- des caractères de morphologie comme par exemple, des mesures relatives à la mamelle, la qualité du bassin et des membres (inclinaison du bassin, aplombs, *etc*.), ou la vitesse de traite ;

- des caractères de production comme la quantité de lait et le taux de matière grasse du lait ;

- des caractères de fertilité (fertilité vache et génisse).

Au cours de ma thèse, je me suis concentrée sur 6 caractères présentant des héritabilités h^2 (part de variabilité phénotypique d'origine génétique) variant de 0,02 pour la fertilité à 0,5 pour les taux butyreux et protéiques. La fertilité des vaches telle que définie par Boichard et Manfredi (1994) repose sur une mesure binaire représentant la réussite ou l'échec d'une unique insémination. Le tableau 3.1 résume les caractéristiques des phénotypes étudiés.

Tableau 3.1 : Caractéristiques des phénotypes étudiés

Caractères	Notation	Unités de mesure	h^2
Quantité de lait	Lait	kg	0,3
Quantité de matière grasse	MG	kg	0,3
Quantité de matière protéique	MP	kg	0,3
Taux butyreux	TB	g/kg	0,5
Taux protéique	TP	g/kg	0,5
Fertilité	Fer	-	0,02

3.2 Composition des populations de référence

3.2.1 Les effectifs

Le projet AMASGEN s'intéresse à l'application de la sélection génomique de taureaux issus de races françaises de bovins laitiers. Dans le cadre de ma thèse, j'ai analysé les données issues des deux plus grandes races bovines laitières françaises en termes d'effectif : la Holstein et la Montbéliarde.

La Prim'Holstein, selon l'appellation française de la race Holstein, est répartie sur l'ensemble du territoire national avec plus de 2,8 millions de vaches, soit plus de 60% de l'effectif total des vaches laitières en France. La collecte nationale de lait destiné à l'industrie laitière repose à 80% sur la production Holstein. La Holstein est aussi la première race laitière au monde (États-Unis, Pologne, Allemagne…).

La Montbéliarde est la race laitière dominante dans l'Est de la France et représente 92% des vaches de Franche-Comté. Elle s'est très bien adaptée à tous les massifs français avec environ 700 000 vaches laitières.

La sélection génomique repose sur l'analyse d'une population de référence (c'est-à-dire des taureaux génotypés et phénotypés) la plus importante possible. Plus cette population sera grande, et plus les équations de prédiction qui en seront dérivées seront précises. L'objectif est de réussir à estimer, à partir des données de la population actuelle de référence, les phénotypes des taureaux des nouvelles générations en se basant exclusivement sur leurs génotypes. Afin de représenter au mieux ce schéma, la population de référence est divisée en deux parties : une population d'apprentissage « **A** » et une population de validation « **V** » constituée des taureaux les plus jeunes.

Pendant la durée du projet AMASGEN, plusieurs jeux de données, de taille de plus en plus importante, ont été créés et mis à notre disposition. En effet, le nombre de taureaux génotypés a augmenté avec le temps et les moyens investis dans les frais de génotypage. Deux jeux de données pour chaque race présentant une augmentation significative de la taille des populations de référence ont été étudiés. Le tableau 3.2 présente les effectifs de ces deux races sur les jeux de données disponibles en janvier 2009 puis en octobre 2009. Pour chaque race, le nombre de taureaux qui composent l'ensemble d'apprentissage a plus que doublé entre ces deux dates.

Tableau 3.2 : Nombres de taureaux génotypés et phénotypés par population étudiée

	Janvier 2009		**Octobre 2009**	
	Holstein	**Montbéliarde**	**Holstein**	**Montbéliarde**
Apprentissage	1 216	451	2 976	950
Validation	540	227	964	222
Total	1 756	678	3 940	1 172

Outre l'accroissement du nombre de taureaux génotypés entre ces deux dates, cette différence s'explique également au travers d'une modification de la stratégie de validation. La différence entre la stratégie utilisée sur l'ensemble « Janvier 2009 » et celle utilisée sur l'ensemble « Octobre 2009 » réside dans le choix de la date utilisée pour calculer les DYD des taureaux d'apprentissage. Le DYD est calculé à partir de l'information phénotypique disponible à une date donnée.

Or, la performance d'un taureau est fonction des phénotypes mais aussi du nombre de ses descendantes. Un taureau sélectionné dans nos populations de référence continue à avoir des filles. Plus la date choisie pour calculer les DYD sera récente et plus le nombre de filles de ce taureau sera donc important. Pour la population d'apprentissage « Janvier 2009 », c'est l'information phénotypique à l'année de naissance du plus jeune taureau de validation (2004) qui est utilisée. Sur l'ensemble « Octobre 2009 », les DYD des taureaux de la population d'apprentissage sont évalués avec l'information disponible en 2009, soit en prenant en compte le nombre de filles de chaque taureau en 2009. Le nombre de filles de chaque taureau de la population d'apprentissage est donc plus important sur l'ensemble « Octobre 2009 » que sur l'ensemble « Janvier 2009 ». Certains taureaux qui ne pouvaient pas entrer dans la population d'apprentissage en janvier 2009 à cause d'un nombre de filles insuffisant ont pu entrer dans la population d'octobre 2009 ce qui explique l'augmentation de la taille de la population de référence entre janvier et octobre 2009.

À noter que ce changement de stratégie a eu lieu en même temps que la mise en place du projet EuroGenomics fin 2009 (Lund *et al.,* 2010). Cinq entreprises européennes, travaillant en étroite collaboration avec des instituts de recherche et impliquées dans la sélection bovine, ont décidé d'unir leurs forces pour améliorer les résultats en sélection génomique. Cette coopération repose essentiellement sur la mise en commun de leurs populations de référence (16 000 taureaux génotypés fin 2009) en race Prim'Holstein. Cette augmentation de la taille de la population de référence améliorera considérablement la fiabilité des index génomiques pour cette race.

Pour partager les populations de référence de chaque race en un ensemble d'apprentissage et un ensemble de validation, une date de naissance minimale pour les taureaux de validation a été fixée arbitrairement pour l'ensemble de janvier 2009. Pour l'ensemble d'octobre 2009, cette date de naissance minimale avait été choisie de manière à avoir les 75% plus vieux taureaux de l'effectif total en ensemble d'apprentissage et les 25% plus jeunes en ensemble de validation. C'est pourquoi, en race Montbéliarde, les taureaux de l'ensemble de validation de janvier 2009 (227 animaux) sont plus nombreux que ceux de l'ensemble d'octobre 2009 (222 animaux).

En premier lieu, l'impact de l'augmentation de la taille de la population d'apprentissage sur la précision des estimations des valeurs génomiques des individus de la population de validation a été analysé. Pour cela, les méthodes BLUP sur information pedigree, GBLUP, Elastic Net, PLS et sparse PLS ont été appliquées sur les ensembles de janvier et d'octobre 2009. La figure 3.1 présente les corrélations pondérées par les EDC entre les phénotypes observés (DYD observés), des taureaux de l'ensemble de validation et les phénotypes prédits de ces mêmes taureaux par les différentes méthodes testées sur deux caractères (Lait et TB), en races Holstein et Montbéliarde.

Le BLUP (noté Pol pour modèle polygénique) est très peu affecté par la différence de taille des populations entre les ensembles de janvier et d'octobre 2009. Mais pour les autres approches, la précision des estimations augmente avec la taille de la population d'apprentissage. En Holstein, la corrélation moyenne sur toutes les méthodes basées sur les marqueurs SNP (GBLUP, Elastic Net, PLS, sparse PLS) passe de 0,42 à 0,54 en Lait et de 0,64 à 0,72 en TB. En Montbéliarde, elle est de 0,34 et 0,44 en janvier 2009 et de 0,39 et 0,56 en octobre 2009, sur le Lait et le TB, respectivement. Ces résultats sont en accord avec l'étude de Hayes *et al*. (2009b) qui montre l'importance d'avoir une population de référence la plus étendue possible et son impact sur la précision des GEBV (voir chapitre 1).

Ainsi, afin d'obtenir des modèles de prédiction les plus précis possible, nous ne présenterons dans la suite de cette thèse que les travaux réalisés sur la population issue des données d'octobre 2009 (population de plus grande taille). Cependant, la stratégie de validation de l'ensemble de janvier 2009 qui consiste à utiliser l'information disponible à la date de naissance des taureaux de validation et pas à une date ultérieure pour les taureaux de la population d'apprentissage, correspond mieux au futur schéma de sélection. On cherche à construire des équations de prédictions capables d'estimer la valeur génétique d'un animal dès sa naissance. Outre le fait que la population d'octobre 2009 est plus étendue, sa stratégie de validation est celle adoptée par les autres membres d'EuroGenomics dans le but d'obtenir les DYD les plus précis possible. Elle est donc finalement la stratégie adoptée dans le cadre du projet AMASGEN.

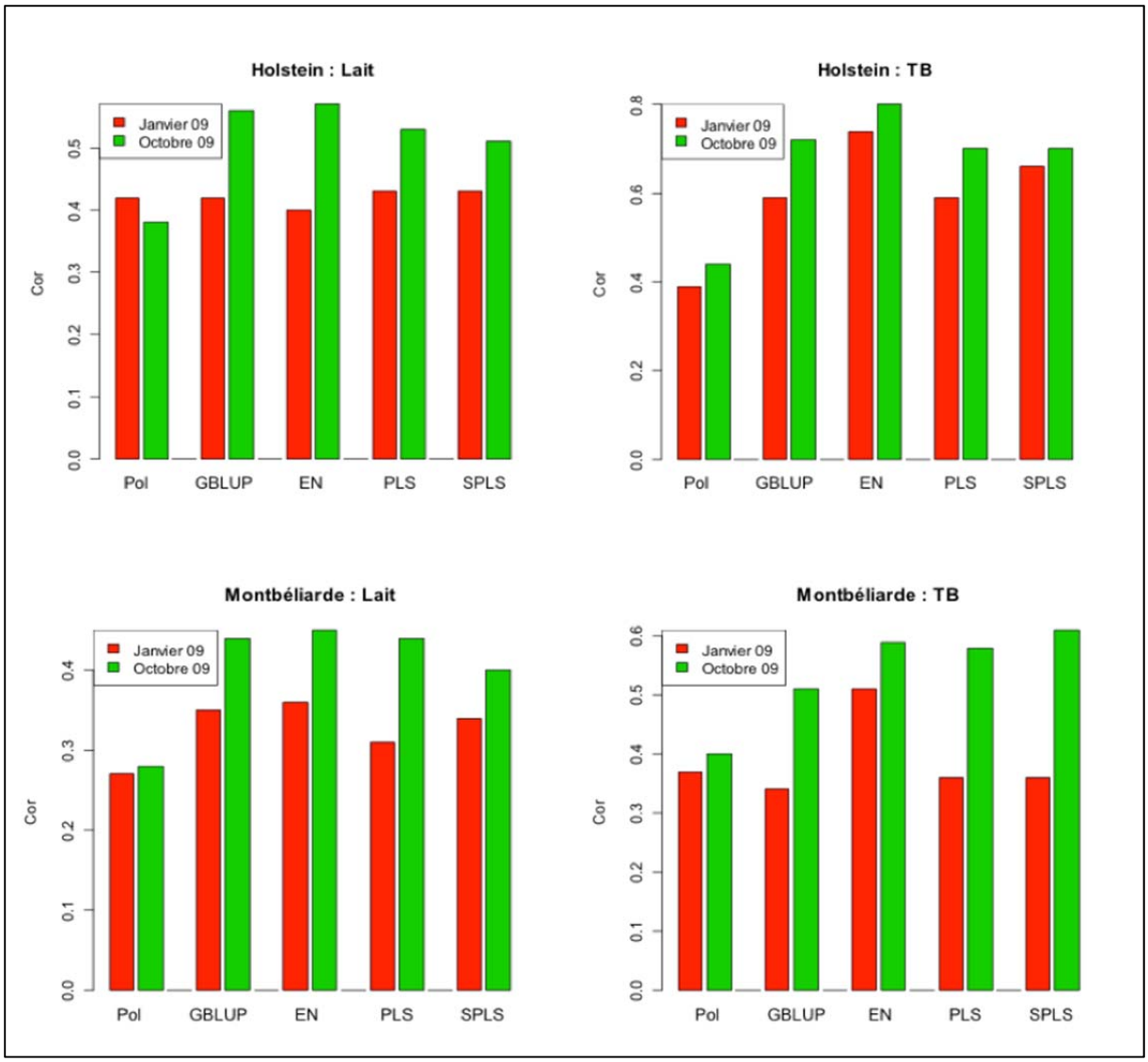


Figure 3.1 : Corrélations pondérées par les EDC entre phénotypes observés et phénotypes prédits dans la population de validation, obtenues par les méthodes BLUP sur pedigree (Pol), GBLUP, Elastic Net (EN), PLS et sparse PLS (SPLS) sur les ensembles de janvier et octobre 2009

3.2.2 Structure de la population de l'ensemble d'octobre 2009

La constitution des populations Holstein et Montbéliarde a débuté avec le projet de recherche de cartographie fine de QTL « Cartofine » basé sur un protocole petites-filles qui rassemblait, en Holstein, 34 familles de pères pour 1 855 taureaux d'insémination artificielle et, en Montbéliarde, 17 familles de pères pour 671 taureaux. Ces taureaux ont été testés sur descendance sur un minimum de 100 filles phénotypées. Ils constituent la base des populations de référence Holstein et Montbéliarde du projet AMASGEN.

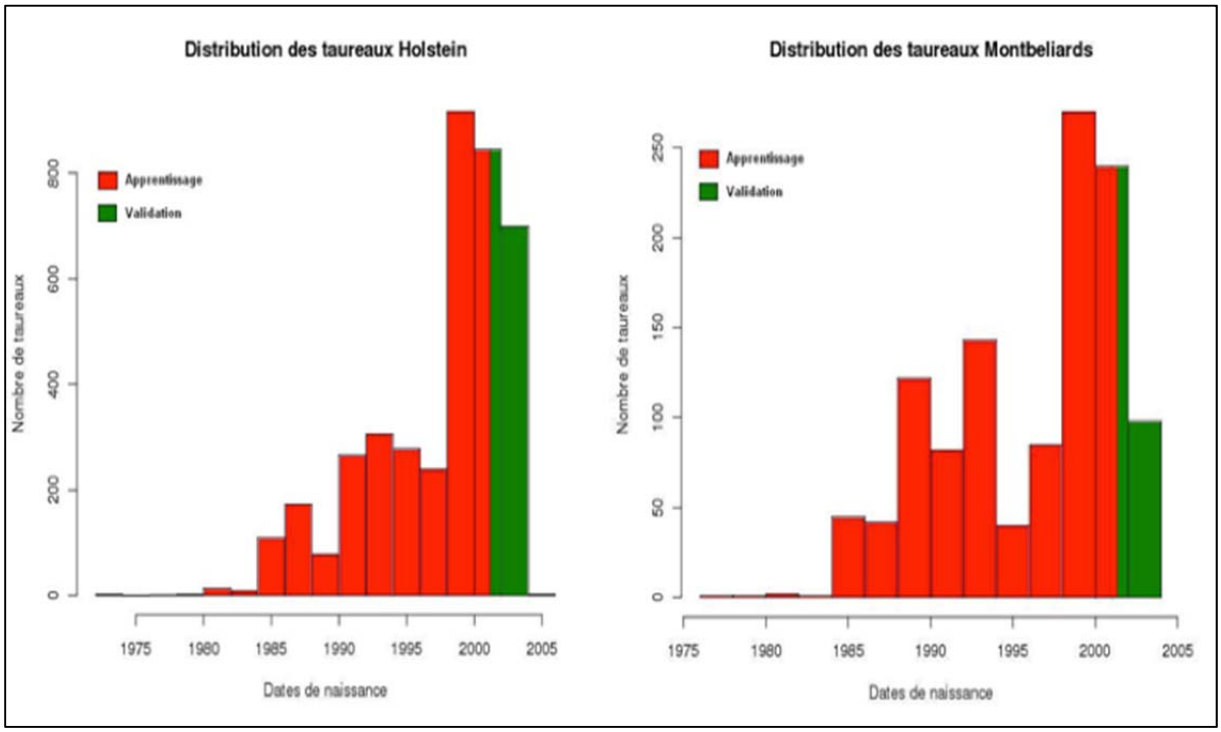


Figure 3.2 : Histogramme des taureaux entre ensembles d'apprentissage (en rouge) et de validation (en vert) selon leur date de naissance, en race Holstein et en race Montbéliarde

La figure 3.2 présente la répartition par année de naissance des taureaux de race Holstein et de race Montbéliarde pour la population de référence d'octobre 2009. Les taureaux de l'ensemble d'apprentissage sont nés avant juin 2002 et les taureaux de validation sont nés entre juin 2002 et 2005. Les deux tiers des taureaux de la population de référence de chaque race sont nés après 1998.

Pour être inclus dans l'analyse, les taureaux des populations du projet AMASGEN doivent avoir été évalués sur un minimum de 20 filles sur le caractère quantité de lait (et donc aussi sur tous les autres caractères de production car le nombre de filles évaluées est le même pour tous les caractères de production). Le tableau 3.3 présente le nombre de filles minimal, maximal et moyen sur les populations d'apprentissage et de validation Holstein. Les caractères de production et le caractère de fertilité ont été mesurés en moyenne sur un nombre de filles très proche afin d'obtenir une précision des données semblable en fertilité et en production laitière. Par exemple, dans la population d'apprentissage le nombre moyen de filles est de 1 662 pour les caractères de production contre 1 654 pour la fertilité.

Tableau 3.3 : Valeurs minimales, maximales et moyennes du nombre de filles par taureau dans la population d'apprentissage et de validation, en Holstein

	Apprentissage			Validation		
Caractères	**minimum**	**maximum**	**moyenne**	**minimum**	**maximum**	**moyenne**
Production	29	118 700	1 662	40	5 378	97
Fertilité	8	125 900	1 654	41	11 460	117

En race Montbéliarde, le nombre de filles est aussi semblable pour l'évaluation des caractères de production et de la fertilité. Le nombre de filles minimum imposé aux taureaux de la population d'apprentissage Montbéliarde est le même qu'en Holstein (20 filles). Les taureaux qui composent la population de validation sont les plus jeunes, donc leurs filles sont moins nombreuses, en race Montbéliarde comme en Holstein. Nous ne présentons que la répartition du nombre de filles en Holstein en raison de la similitude des observations sur les deux races.

Sur la figure 3.3, qui représente la distribution du nombre de filles par taureau sur les caractères de production et de fertilité, dans l'ensemble d'apprentissage et l'ensemble de validation en Holstein, on remarque qu'une grande majorité d'animaux a un nombre de descendantes limité. En effet, dans l'ensemble d'apprentissage, seulement 44 taureaux en production laitière (et 46 en fertilité) ont entre 20 000 et 40 000 filles, 25 taureaux en production laitière (et 23 en fertilité) ont entre 40 001 et 80 000 filles et on compte 5 taureaux avec plus de 80 000 filles sur les caractères de production laitière et 2 en fertilité. Dans l'ensemble de validation, on trouve 2 taureaux ayant entre 500 et 1 000 filles pour tous les caractères ; 3 taureaux et respectivement 6 taureaux avec plus de 1 000 filles en production laitière et fertilité respectivement. Dans la population de référence Holstein comme en Montbéliarde, les taureaux qui ont un nombre très important de filles en production laitière sont, en général, ceux qui ont un grand nombre de filles en fertilité. Cependant, la figure 3.3 ne permet pas de distinguer la répartition des taureaux qui ont un nombre moyen de filles. La figure 3.4 représente la répartition du nombre de filles pour les taureaux Holstein avec moins de 500 filles.

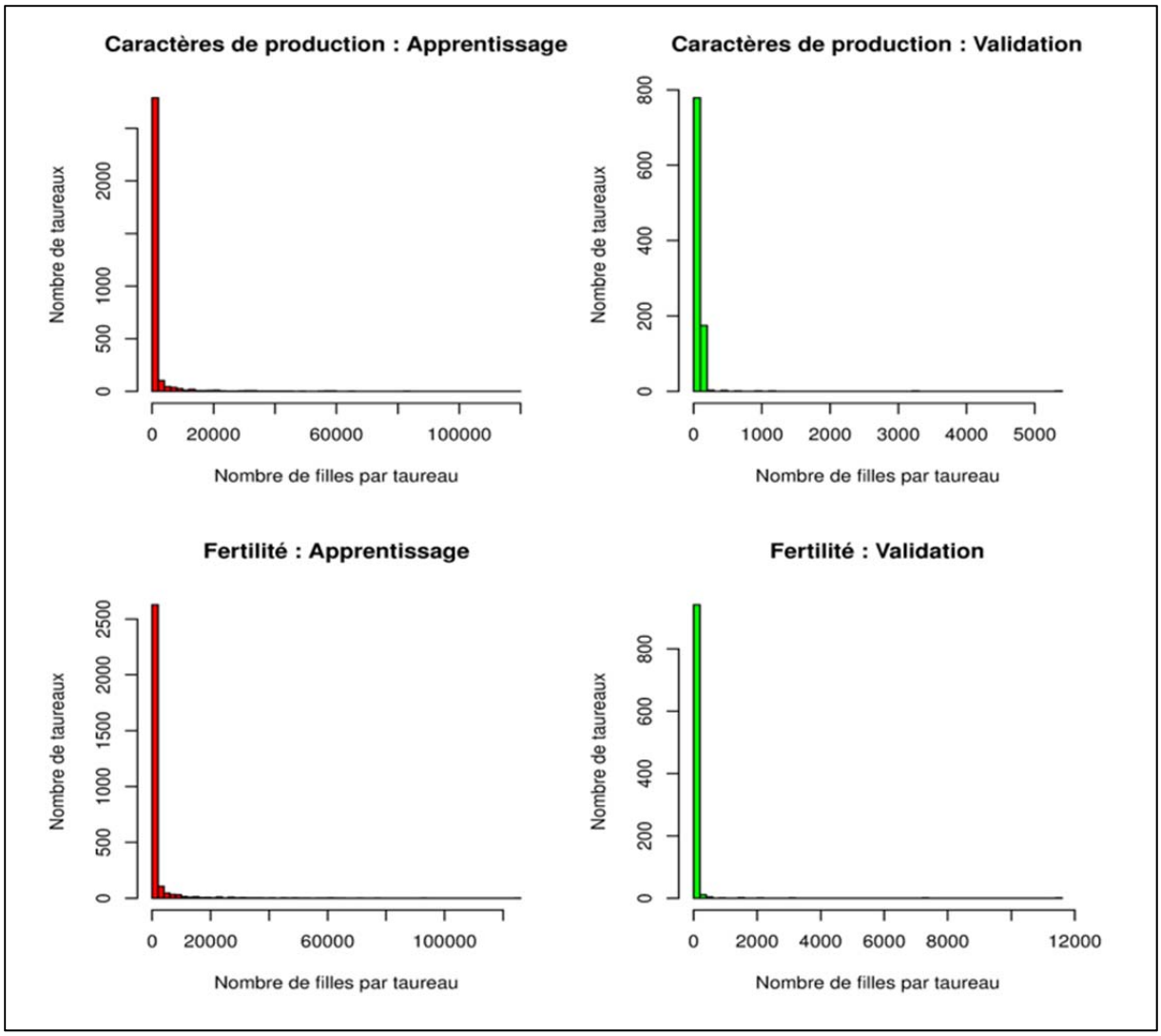


Figure 3.3 : Histogramme du nombre de filles par taureau sur les caractères de production et de fertilité dans l'ensemble d'apprentissage et l'ensemble de validation, en race Holstein

Les distributions de la figure 3.4 sont proches d'une loi normale mais avec une queue de distribution longue. De plus, elles sont très semblables au sein d'une même population entre les caractères de production et le caractère de fertilité. Pour toute population (Apprentissage ou Validation) ou tout caractère considéré, on remarque également que la majorité des taureaux a moins de 200 filles. Donc la moyenne du nombre de filles par taureau devrait être inférieure à 200. Hors, le nombre de filles moyen dans la population d'apprentissage est très élevé sur tous les caractères (entre 1 654 et 1 652). Cela est dû à la présence de quelques taureaux très diffusés et qui ont donc un nombre de filles très nettement supérieur aux autres taureaux.

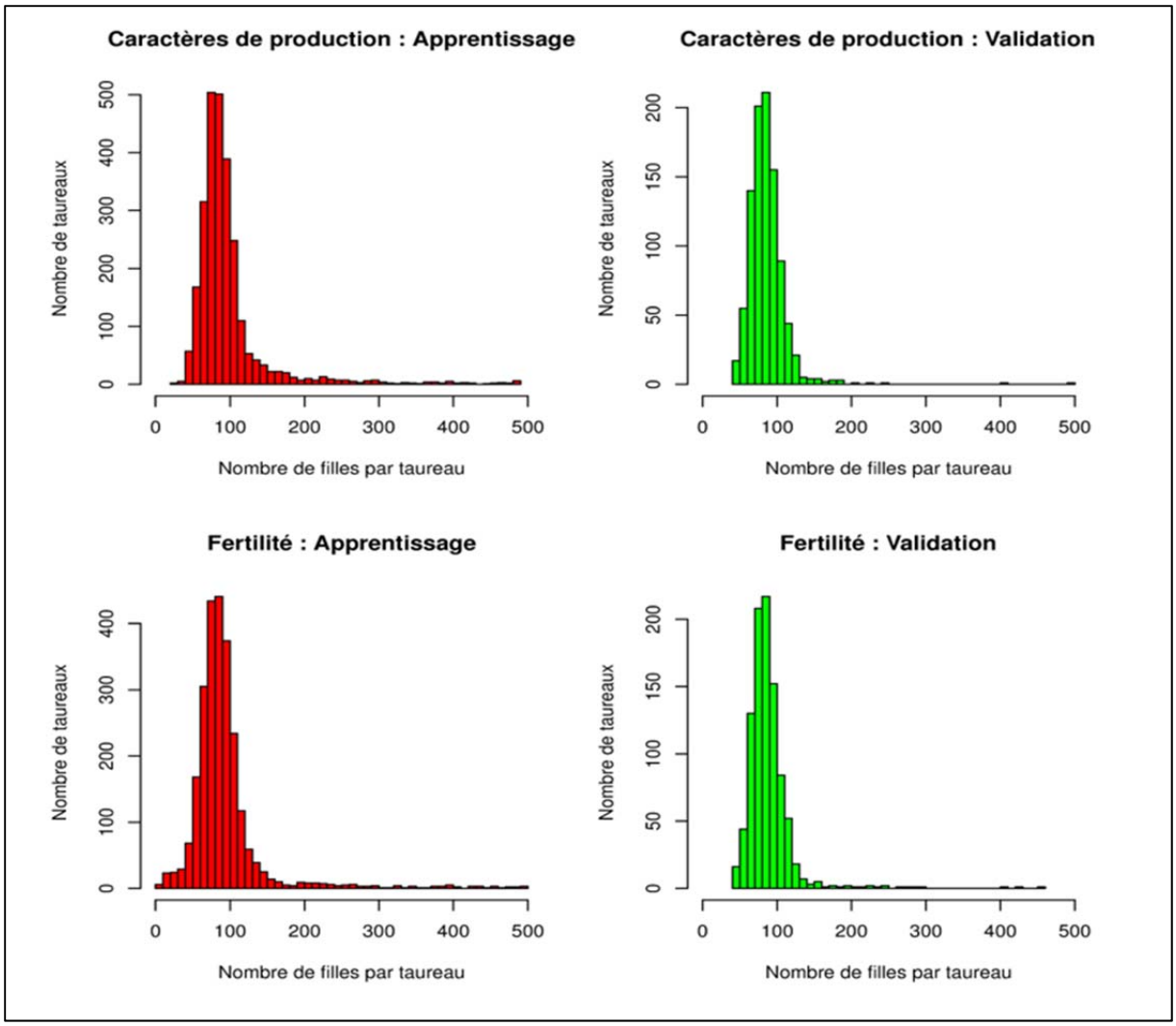


Figure 3.4 : Histogramme du nombre de filles par taureau sur les caractères de production et de fertilité dans l'ensemble d'apprentissage et l'ensemble de validation pour les taureaux avec moins de 500 filles, en race Holstein

Les mêmes observations peuvent être faites pour la population Montbéliarde. Au vu de ces premiers résultats, nous pouvons dire que certains taureaux ont un nombre de filles très important ce qui conduira à des EDC très élevés pour ces taureaux particuliers. Nous étudierons dans le chapitre 4 l'impact de la prise en compte de ces forts EDC sur les méthodes de prédiction génomique testées en Holstein et en Montbéliarde.

3.3 Statistiques descriptives des phénotypes

Notre étude repose sur l'analyse de données phénotypiques (DYD) et génotypiques (SNP) de deux races bovines laitières françaises, sur six caractères (cinq caractères de production laitière et un caractère de fertilité vache). Afin de comparer efficacement les caractères de production entre eux, rappelons que la sélection des reproducteurs a reposé, depuis 1989, sur la valeur pour chaque taureau d'un index combinant les composantes du lait appelé **INEL** (INdex Economique Laitier). Cet index était construit de sorte à privilégier la quantité de matières protéiques. L'INEL est la partie production laitière d'un index global dont l'objectif est d'aider à améliorer la rentabilité de l'exploitation laitière dans son ensemble en tenant compte de l'évolution probable de l'économie. La nouvelle définition de l'INEL, mise en application depuis 2001, pondère chaque composante du lait en fonction de sa valeur économique. On distingue dans le lait trois éléments distincts : la matière grasse (MG), la matière protéique (MP) et le vecteur (vecteur = Lait – [MG + MP]). Les pondérations associées à chacun de ses éléments reposent ainsi sur leurs valeurs économiques mais aussi sur les paramètres génétiques de chaque caractère (héritabilité, corrélations). La définition du nouvel index INEL, adoptée en race Holstein, est la suivante :

$$INEL = 0{,}98\ [\ MP + 0{,}2\ MG + TP + 0{,}5\ TB\ ].$$

En tenant compte des variabilités respectives des différents caractères, leur importance relative devient : 71% pour le MP, 19% pour le MG et 5% pour chacun des taux. Dans l'INEL de la race Montbéliarde, une part plus importante est donnée au taux protéique :

$$INEL = 1{,}055\ [\ MP + 0{,}1\ MG + 3TP + 0{,}5\ TB\ ].$$

L'INEL rend compte de l'importance qu'il a été accordé aux caractères de production qui nous intéressent ici et permet d'avoir ainsi une vision globale de l'effort de sélection réalisé.

Le paragraphe suivant a pour but de décrire en détail les caractères pour chacune des races étudiées afin de faciliter la compréhension de la structure des données traitées dans la suite de ce travail.

3.3.1 Description des DYD en races Holstein et Montbéliarde

Pour chaque caractère étudié, la distribution des valeurs phénotypiques (c'est-à-dire des DYD observés) dans chacune des deux populations (apprentissage et validation) et pour chacune des races a été représentée. La figure 3.5 correspond à la race Holstein et la figure 3.6 à la race Montbéliarde.

La population d'apprentissage (en rouge) est composée des taureaux les plus anciens et la population de validation (en vert) des taureaux les plus jeunes. Du fait de la sélection sur ces caractères, on s'attend à des animaux plus performants (et donc des animaux avec des DYD plus ou moins élevés selon que l'on cherche à augmenter ou baisser la valeur phénotypique des taureaux sur les caractères considérés) dans la population de validation que dans la population d'apprentissage.

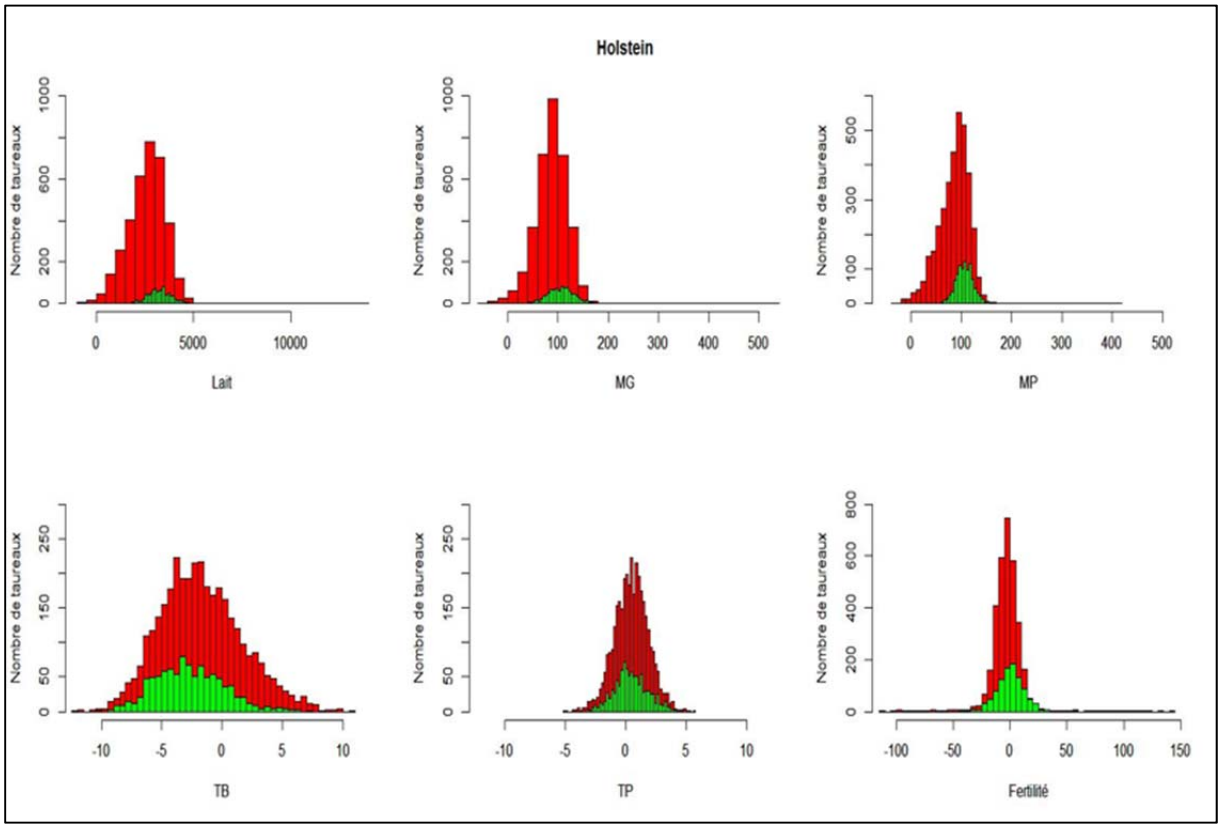


Figure 3.5 : Histogramme des variables phénotypiques (DYD) observées sur chacun des caractères étudiés dans l'ensemble d'apprentissage A (en rouge) et l'ensemble de validation V (en vert), en race Holstein

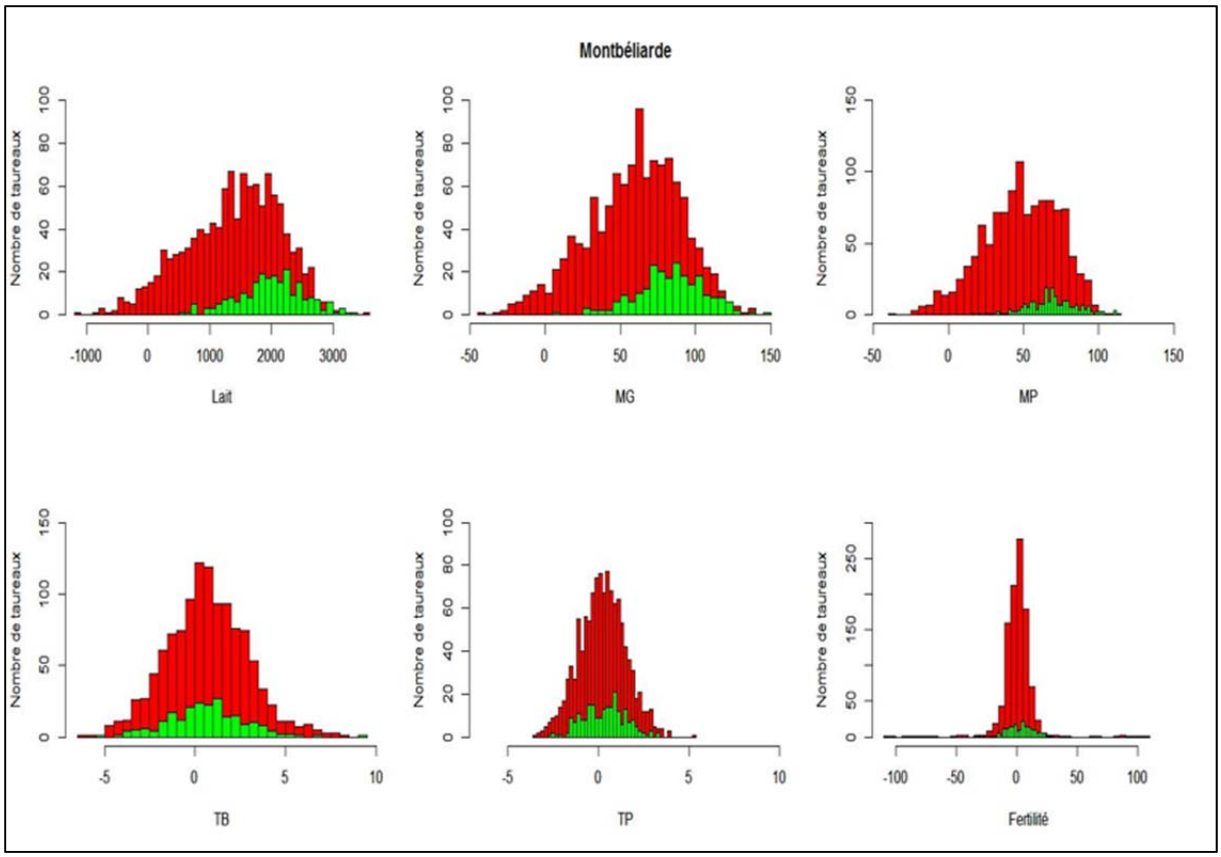


Figure 3.6 : Histogramme des variables phénotypiques (DYD) observées sur chacun des caractères étudiés dans l'ensemble d'apprentissage A (en rouge) et l'ensemble de validation V (en vert), en race Montbéliarde

On remarque sur la figure 3.5 pour la race Holstein, pour chaque caractère un écart de la moyenne de distribution entre population d'apprentissage et de validation que l'on observe également en race Montbéliarde, sur la figure 3.6.

Tableau 3.4 : Statistiques élémentaires des variables phénotypiques (DYD), dans l'ensemble d'apprentissage (A) et l'ensemble de validation (V) en race Holstein : minimum (min), maximum (max), moyenne (μ), écart-type (σ) et héritabilité (h^2)

	Lait		**MG**		**MP**		**TB**		**TP**		**Fer**	
	A	**V**	**A**	**V**	**A**	**V**	**A**	**V**	**A**	**V**	**A**	**V**
min	-968	1 318	-53,3	27,9	-32,9	42,1	-12,1	-10,5	-5,2	-3,2	-103,3	-110,5
max	1 3912	5 356	522,9	190,9	416,9	106,7	10,6	8,1	5,6	0,5	143,8	11,6
μ	2 581	3 257	86,7	105,9	85,4	161,4	-1,8	-2,7	0,5	5,3	-2,9	-0,02
σ	947,4	593,1	31,7	25,5	30,0	16,6	3,4	2,9	1,4	1,4	15,7	14,4
h^2	0,3		0,3		0,3		0,5		0,5		0,02	

Le tableau 3.4 et le tableau 3.5 présentent les valeurs minimales, maximales, moyennes et écart-types des caractères étudiés pour chaque sous-population (apprentissage et validation). Ils permettent d'apprécier plus précisément la différence entre les populations d'apprentissage et de validation en races Holstein et Montbéliarde respectivement.

Tableau 3.5 : Statistiques élémentaires des variables phénotypiques (DYD), dans l'ensemble d'apprentissage (A) et l'ensemble de validation (V) en race Montbéliarde : minimum (min), maximum (max), moyenne (μ), écart-type (σ) et héritabilité (h^2)

	Lait		**MG**		**MP**		**TB**		**TP**		**Fer**	
	A	**V**	**A**	**V**	**A**	**V**	**A**	**V**	**A**	**V**	**A**	**V**
min	-1 152	590	-42,2	9,5	-38,6	17,3	-6,3	-5,5	-3,4	-2,8	-109	-35,7
max	3 560	3 391	138,5	148,8	110,1	111,7	8,1	9,3	5,3	3,2	109,4	29,6
μ	1 390	1 997	59,3	83,5	48	68,8	0,7	0,7	0,2	0,3	0,6	3
σ	760,5	542,6	30,9	22,2	25,1	17,1	2,2	2,1	1,3	1,2	14,4	11,1
h^2	0,3		0,3		0,3		0,5		0,5		0,02	

En Holstein comme en Montbéliarde, la moyenne des caractères quantité de lait et quantités de matières MG et MP (exprimées en kg) a augmenté significativement entre l'ensemble d'apprentissage et l'ensemble de validation : +676kg en Holstein et +607kg en Montbéliarde en lait, +19,2kg et +24,2kg en MG et +76kg et +20,8kg en MP. Les taux (exprimés en g/kg) et la fertilité ont évolué différemment. Il n'y a pas de différences importantes de moyenne entre les taux mesurés sur la population d'apprentissage et les taux mesurés sur la population de validation. On observe, en Holstein, une diminution du taux butyreux moyen entre les animaux de la population d'apprentissage et la population de validation (-0,9g/kg).

La sélection réalisée jusqu'à aujourd'hui pour la production laitière chez les bovins laitiers permet de valider nos commentaires précédents. L'index INEL a été utilisé pour choisir les reproducteurs et donc les animaux qui composent les populations de référence : il est construit de façon à donner plus de poids aux quantités de matières (MG et MP) qu'aux taux (TB et TP). L'effort de sélection s'est aussi porté de façon évidente sur la quantité de lait ce qui explique des valeurs

nettement plus élevées en moyenne sur les taureaux de l'ensemble de validation que sur ceux de l'ensemble d'apprentissage pour le lait et les quantités de matières que pour les taux. En fertilité, le peu de différence entre les valeurs phénotypiques entre la population d'apprentissage et la population de validation en Holstein (+2,87%) et Montbéliarde (+2,4%) s'explique par une faible héritabilité du caractère (2%). L'héritabilité représente la part de variabilité d'un caractère qui est d'origine génétique additive. La fertilité étant un caractère peu héritable, il est donc difficile à sélectionner ce qui s'exprime par une augmentation limitée des moyennes des phénotypes entre les taureaux les plus anciens et les taureaux les plus jeunes.

Au niveau des écart-types phénotypiques (ou DYD) observés, les résultats sont en accord avec ceux obtenus dans la littérature et en particulier par Boichard et Bonaïti (1987). Dans la population d'apprentissage, nous observons des écart-types phénotypiques de 947 pour la quantité de lait, de l'ordre de 3 pour la quantité de matière grasse et la quantité de matière protéique, de 3,4 pour le taux butyreux, 1,4 pour le taux protéique et 15 pour la fertilité en race Holstein. Des résultats légèrement plus faibles sont observés pour chacun des caractères en race Montbéliarde. Enfin, les écart-types observés dans la population de validation sont inférieurs à ceux obtenus dans la population d'apprentissage, principalement en raison d'effectifs moins importants pour les deux races.

3.3.2 Corrélations entre les caractères de production laitière

Comme nous venons de le voir, les pondérations associées à chaque caractère dans l'index de sélection expliquent en partie la différence de progrès génétique que l'on observe entre les caractères. L'intensité de la liaison entre chaque caractère (pouvant être représentée par une corrélation phénotypique) permet de mesurer directement l'impact de la sélection sur tous les caractères de production les uns par rapport aux autres. Les corrélations « phénotypiques » entre les caractères de production laitière (c'est-à-dire les corrélations entre les DYD des caractères) ont donc été calculées. Les résultats sont présentés dans le tableau 3.6 pour la race Holstein et dans le tableau 3.7 pour la race Montbéliarde.

Les taureaux des populations de référence ont été principalement sélectionnés sur la quantité de lait et la matière protéique. En Holstein (tableau 3.4), les différences entre la moyenne des DYD observée dans la population d'apprentissage et la moyenne observée dans la population de validation sont importantes en lait et MP, faibles en MG et négatives en TB. La valeur des corrélations phénotypiques entre caractères confirment ces résultats. Si on prend comme caractère de référence la quantité de lait, on s'attendait bien à obtenir un important gain sur le MP (corrélation phénotypique de 0,91) et plus modéré sur le MG (corrélation phénotypique de 0,65). La corrélation moyenne négative entre lait et TB implique une baisse du taux butyreux quand on sélectionne sur le lait, comme on a pu l'observer précédemment.

Tableau 3.6 : Corrélations phénotypiques entre les cinq caractères de production dans la population de référence en race Holstein

	MG	**MP**	**TB**	**TP**
Lait	0,65	0,91	-0,57	-0,18
MG		0,75	0,25	0,28
MP			-0,35	0,23
TB				0,53

En Montbéliarde, les caractères MG et MP évoluent de la même façon (tableau 3.5). Les taux sont stabilisés quand on compare les DYD obtenus par les animaux de la population d'apprentissage et les animaux de la population de validation. Les corrélations phénotypiques faibles entre la quantité de lait et les taux (-0,2 environ) et entre le MP et les taux (-0,02 en TB et 0,16 en TP) confirment l'impact limité d'une sélection de ces caractères sur les taux. Les quantités de lait et de matières sont fortement et positivement liées (0,86 entre Lait et MG et 0,93 entre Lait et MP) d'où une augmentation significative de ces quantités au cours des générations. Les quantités MG et MP évoluent de la même façon, avec une corrélation de 0,9.

Tableau 3.7 : Corrélations phénotypiques entre les cinq caractères de production dans la population de référence en race Montbéliarde

	MG	**MP**	**TB**	**TP**
Lait	0,86	0,93	-0,22	-0,21
MG		0,90	0,30	0,07
MP			-0,02	0,16
TB				0,53

Les corrélations génétiques présentées dans la littérature en race Holstein et Montbéliarde (Boichard et Bonaïti, 1987 ; Barillet et Bonaïti, 1992) sont similaires aux corrélations phénotypiques calculées sur nos données. Par exemple en race Montbéliarde, sont obtenues 0,9 de corrélation génétique contre 0,93 de corrélation phénotypique entre lait et MP, 0,85 et 0,86 entre lait et MG, -0,3 et -0,22 entre lait et TB, 0,85 et 0,9 entre les quantités de matières et 0,6 et 0,53 entre les taux.

Ainsi, pour améliorer la quantité et la composition du lait grâce à la sélection des reproducteurs, il est nécessaire de tenir compte du lien qu'il existe entre les caractères considérés. Par exemple, une pondération trop importante sur le lait peut induire une diminution de la concentration de matière grasse. Il n'est pas indispensable d'exploiter tout le potentiel génétique des animaux car en sélection, les animaux doivent, pour être considérés comme performants, soit produire plus et mieux, soit produire autant mais à moindre coût. Ainsi, afin de limiter les coûts d'élevage, les caractères qui entrainent l'abattage involontaire des animaux comme une fertilité femelle trop faible, sont devenus importants. Rappelons que la fertilité des vaches est faiblement héritable (h^2 = 0,02) et il est important de pouvoir quantifier l'apport de la sélection génomique sur de tels caractères. Enfin, il a été montré qu'un niveau génétique en production laitière trop élevé implique des performances de reproduction réduites (Grimard *et al*., 2006 ; Ledoux *et al*., 2006), il est donc indispensable de considérer dans le calcul des index en évaluation génétique, un caractère de fertilité femelle. Pour toutes ces raisons et pour avoir un panel varié de caractères, il nous a semblé nécessaire de considérer dans notre étude un caractère de fertilité femelle.

3.4 Définition des données génotypiques : les marqueurs SNP

L'intégration des données génomiques pour la sélection animale repose sur l'étude des SNP. Un SNP est une mutation ponctuelle du génome pouvant être utilisée comme un marqueur à deux allèles car il représente une large majorité de l'ensemble des variations des génomes des êtres vivants. Ces marqueurs sont particulièrement intéressants du fait de leur abondance et des possibilités d'automatisation permettant d'acquérir plusieurs dizaines de milliers de génotypes à la fois. Les SNP sont utilisés comme marqueurs moléculaires, notamment au travers des puces à ADN, afin de repérer la transmission d'un segment chromosomique d'un individu à un autre. Comme évoqué au chapitre 1, la répartition uniforme sur le génome des SNP permet d'en avoir une vue d'ensemble. On suppose que les SNP sont fortement liés aux QTL qui agissent sur les caractères d'intérêt et permettent donc d'en capturer les effets.

Pour être intégré à notre population de référence, un taureau doit avoir un call rate supérieur à 0,95 ce qui garantit une quantité minimale de génotypes manquants. Les taureaux de nos populations de référence ont été génotypés avec la puce à ADN Illumina Bovine SNP50 Beadchip® qui contient 54 001 SNP fortement informatifs pour la plupart des principales races bovines et uniformément distribués sur le génome (un marqueur tous les 50kb en moyenne). Ces SNP ont été validés sur 19 races laitières et allaitantes.

Les SNP étant majoritairement bialléliques, ils se présentent sous la forme de données binaires sous le codage 1 et 2, avec pour chaque individu, un allèle qui vient du père et un allèle qui vient de la mère. On utilise ensuite la fréquence de l'allèle 2 au locus considéré. Ainsi, un individu homozygote pour l'allèle 2 (de génotype 22) est codé 1+1 =2, un individu homozygote pour l'allèle 1 (de génotype 11) est codé 0+0 =0 et un individu hétérozygote (de génotype 12) est codé 1+0 =1.

Une fois les génotypes obtenus pour l'ensemble des taureaux, plusieurs étapes, réalisées au sein de l'équipe G2B de l'UMR INRA-GABI, sont nécessaires pour filtrer, nettoyer et normaliser les données afin d'éliminer les incohérences et les données manquantes :

- une vérification et élimination des incohérences entre parents et descendants en se basant sur les pedigrees ;

- une prédiction des allèles aux marqueurs manquants à l'aide du logiciel DualPhase (Druet et Georges, 2009) qui exploite le déséquilibre de liaison entre marqueurs (voir chapitre 1) ;
- une élimination des animaux présentant des génotypes manquants dans la population de référence (en général, moins d'une dizaine d'animaux).

Ainsi, les ensembles de données phénotypiques et génotypiques des populations de référence Holstein et Montbéliarde ne présentent aucune valeur manquante.

Le tableau 3.8 résume les effectifs d'animaux dans la population de référence et de SNP avant et après les phases de contrôle de qualité des données dans les deux races. Après les phases de contrôle, seuls les SNP suffisamment informatifs sont conservés c'est-à-dire ceux dont la fréquence de l'allèle minoritaire (ou MAF pour Minor Allele Frequency) est supérieure à 3%.

Tableau 3.8 : Effectifs de la population de référence et du nombre de SNP disponibles avant et après les phases de contrôle de qualité des données en races Holstein et Montbéliarde

	Holstein		**Montbéliarde**	
	Avant	**Après**	**Avant**	**Après**
Taureaux	4 471	3 940	1 392	1 172
SNP	48 658	39 738	48 860	38 462

Dans la suite de ce travail, seront présentés (sauf mentionné autrement) les travaux réalisés sur les données après contrôle de qualité soit un ensemble de 3 940 taureaux sur 39 738 SNP en race Holstein et 1 172 taureaux sur 38 462 SNP en race Montbéliarde.

Chapitre 4 Étude comparative des méthodes de régression PLS et sparse PLS

4.1 Introduction

La régression PLS permet de réduire le nombre de variables composant le modèle de prédiction en construisant des combinaisons linéaires des variables explicatives originales, puis en régressant la variable réponse sur ces nouvelles variables. C'est une méthode de réduction de dimensions présentant de bonnes capacités prédictives y compris lorsque le nombre d'individus est limité par rapport au nombre de variables explicatives. Plusieurs études ont appliqué une régression PLS pour calculer les évaluations génomiques des animaux et en étudier les propriétés et les performances, notamment en termes de précision des GEBV.

Solberg *et al.* (2009) comparent la régression PLS à la régression sur composantes principales (RCP) et à la méthode BayesB pour l'estimation des valeurs génomiques : pour cela, ils ont simulé plusieurs jeux de données en faisant varier la densité de marqueurs SNP considérés dans le modèle. Un génome de 10 chromosomes de 100cM a été simulé avec 4 densités de marqueurs différentes (100, 200, 400 et 800 marqueurs par Morgan). Sachant que la taille du génome bovin est de 30M, ce schéma correspondrait à l'étude d'un génome bovin à partir de 3 000 à 24 000 SNP. Cette étude a permis de confronter la régression PLS à la RCP pour ce type de données. Cette étude montre que la corrélation entre les DYD observés et les DYD prédits est moins affectée par l'augmentation de la densité des marqueurs pour la régression PLS et la RCP (+7% de gain de corrélation et +6% pour la régression PLS et la RCP, respectivement quand on passe de 100 marqueurs par Morgan à 800 marqueurs par Morgan) qu'en utilisant la méthodologie BayesB (+17%). Même si la régression PLS s'avère plus précise que la RCP, elles restent toutes deux inférieures à la méthode BayesB en terme de prédiction des valeurs génomiques (avec une corrélation maximale de 0,67 pour la RCP, 0,68 pour la PLS et 0,86 pour BayesB). Cependant, elles sont plus simples à implémenter et plus rapides en temps de calcul (jusqu'à 65 fois moins long qu'avec la méthode BayesB soit, pour 8 080 SNP, 60mn pour la RCP, 3mn pour la régression PLS et

plus de 46h pour BayesB). Coster *et al.* (2010) utilisent eux aussi un jeu de données simulées pour comparer la régression PLS, la méthode BayesB et l'approche **LARS** (Least Angle Regression). Dans leur simulation, ils ont fait varier le nombre de QTL ayant un effet sur le caractère d'intérêt (5%, 25% et 50% de l'ensemble des loci expliquant 90% de la variance génétique) et la distribution de la variance de ces QTL (homogène ou hétérogène sur l'ensemble des QTL considérés dans leur analyse). Ils ont ainsi confronté 6 scenarios simulés. La précision des valeurs génomiques représentée par la corrélation entre TBV et GEBV varie de 0,65 à 0,75 par la méthode LARS, de 0,60 à 0,77 par la méthode bayésienne et de 0,66 à 0,68 par la régression PLS. Les auteurs recommandent la régression PLS quand le nombre de QTL est grand ou inconnu car la corrélation entre GEBV et TBV est moyenne par rapport aux autres méthodes mais reste très stable entre les différents scénarios. La régression PLS est aussi plus rapide : 4s de temps d'exécution contre 211s pour l'approche LARS et 430s pour la méthode BayesB. L'étude de Moser *et al.* (2009) nous permet d'évaluer l'efficacité de la PLS vis-à-vis de 4 autres méthodes (les moindres carrés, la méthode BayesA, le GBLUP et la régression sur **SVM** -Support Vector Machines ou Machines à vecteurs de support) sur un ensemble d'apprentissage de 1 239 taureaux génotypés sur 7 372 SNP. Les auteurs concluent que ces méthodes présentent des performances similaires mais mettent en avant le fait que la régression PLS et le GBLUP sont moins exigeants en temps de calcul : 4s et 22s respectivement contre 3mn, 4mn et 421mn pour les moindres carrés, le GBLUP et BayesA, respectivement.

Les méthodes de réduction de dimension comme la PLS et la RCP ont l'avantage d'être rapides car elles construisent des équations de prédiction sur un nombre réduit de variables latentes. Nous avons choisi d'étudier la régression PLS car, contrairement à la RCP, elle utilise à la fois l'information phénotypique et l'information des marqueurs en cherchant à maximiser le lien entre la variable réponse et les variables explicatives par le biais de la construction de variables latentes. Cependant, la régression PLS peut être pénalisée par l'utilisation de l'ensemble des marqueurs, pouvant être très nombreux, dans la construction de chaque variable latente. En effet, on peut s'attendre à ce que la plupart des marqueurs SNP ne soient pas nécessairement liés au caractère d'intérêt. Ainsi, certains auteurs ont cherché à améliorer le modèle PLS en ajoutant une étape de sélection de variables à leurs algorithmes : la sparse PLS. Chun et Keles (2010) ont

introduit une version de la sparse PLS pour l'analyse de données de biopuces. Une pénalisation l_1 est utilisée pour sélectionner les variables explicatives les plus pertinentes à intégrer dans chaque variable latente. Cette version de la sparse PLS a l'avantage de conserver l'orthogonalité des variables latentes (voir chapitre 2). Cette méthode a été reprise par Long *et al.* (2011) et comparée à la RCP supervisée qui présélectionne un ensemble réduit de SNP en se basant sur la corrélation entre les SNP et le caractère étudié. Les capacités prédictives de ces méthodes ont été testées sur 32 518 SNP et 4 703 taureaux Holstein. Même si la précision des estimations des valeurs génomiques reste légèrement inférieure pour la RCP supervisée et pour la sparse PLS que pour la RCP et la régression PLS, cette étude démontre l'avantage d'intégrer une étape de sélection de variables à des méthodes de réduction de dimensions afin de limiter le coût de calcul dans la prédiction des valeurs génétiques.

La sparse PLS introduite par Lê Cao *et al.* (2008) a été développée pour l'analyse des données génomiques, protéomiques, métabolomiques et phénotypiques de grande dimension. Les premiers résultats ont été obtenus sur simulations et sur données réelles (données d'expression de gènes et transcriptomiques). Ils montrent que la sparse PLS augmente la capacité prédictive du modèle (taux de vrais positifs) et permet de mettre en avant certaines variables pertinentes, liées au problème biologique étudié. Les premiers essais en Holstein sur l'ensemble de janvier 2009 (population d'apprentissage de 1 216 taureaux sur environ 40 000 SNP), réalisé dans le cadre de ma thèse, ont démontré que la sparse PLS telle que décrite par Lê Cao *et al.* (2008), et implémentée via le package R mixOmics (Lê Cao *et al.*, 2009) restait performante sur un nombre élevé d'individus. Cependant, l'étude de la population Holstein d'octobre 2009 (2 976 taureaux) a soulevé des problèmes d'espace mémoire. Une partie de mon travail a donc consisté à adapter le code existant aux données SNP et aux machines de calcul disponibles. Les caractères sont étudiés indépendamment les uns des autres (analyse univariée) : j'ai donc simplifié l'algorithme sparse PLS en éliminant l'étape de décomposition en valeurs singulières et en appliquant une pénalisation l_1 sur les vecteurs *loadings* associés aux SNP seulement et obtenus à partir d'une régression PLS simple au lieu d'une PLS-SVD. Ainsi, l'algorithme sparse PLS ne nécessite plus d'inversion de matrice et n'est donc plus limité par le nombre de SNP ou d'individus considérés.

La littérature a montré l'intérêt des méthodes PLS et sparse PLS mais sans démontrer leur éventuelle supériorité sur les autres méthodes de sélection génomique. Dans ce travail, ces deux méthodes ont été testées sur les données réelles décrites dans le chapitre 3 et comparées aux méthodes couramment utilisées pour l'évaluation des bovins laitiers afin d'étudier leur capacité prédictive.

4.2 Application des méthodes de régression PLS et sparse PLS aux données bovines laitières françaises : la race Holstein

4.2.1 Propriétés statistiques et capacités prédictives des méthodes

Le but de cette étude (article 1) est de tester les capacités prédictives des méthodes PLS et sparse PLS sur les données génomiques bovines en race Holstein. La population de référence est de taille relativement importante, environ 4 000 taureaux, dont 2 976 animaux pour la population d'apprentissage et 964 animaux pour la population de validation. Pour paramétrer les méthodes PLS et sparse PLS, l'erreur de prédiction moyenne sur dix échantillons par validation croisée dans l'ensemble d'apprentissage a été calculée de même que la corrélation pondérée par les EDC entre les DYD prédits par régression PLS et sparse PLS et les DYD observés dans l'ensemble de validation. Après avoir obtenu les modèles optimaux, les deux méthodes ont été comparées en utilisant comme critère la corrélation définie ci-dessus ainsi que la pente de régression des DYD observés sur les DYD prédits.

Le test de Hotelling-Williams (tableau 2 de l'article 1) a mis en évidence l'infériorité significative de la capacité prédictive de la sparse PLS par rapport à la régression PLS, avec une différence moyenne au niveau des corrélations de -0,04 et de -0,09 au niveau des pentes de régression. L'intérêt de cette étude réside également dans l'impact de la prise en compte des poids (EDC) dans la modélisation PLS et sparse PLS. Pour chaque caractère analysé et d'après les résultats du tableau 3 (article 1), le modèle construit sans prendre en compte les EDC dans les approches PLS ou sparse PLS présente une précision très proche du modèle intégrant les EDC. En revanche, sa complexité est très affectée par l'utilisation des EDC avec un nombre de variables latentes (ou dimensions) très supérieur au modèle sans prise en compte des EDC. Afin de mieux comprendre les différences observées

entre les modèles avec et sans EDC, des essais supplémentaires ont été menés en faisant varier la structure de population des ensembles d'apprentissage et de validation : ces résultats sont présentés en seconde partie de ce paragraphe. Dans l'article 1, seuls les modèles les plus simples pour chacun des caractères étudiés sont présentés en n'utilisant pas les EDC dans les équations de prédiction et dans le calcul des corrélations entre valeurs observées et prédites.

Dans ce contexte, les profils des graphes des coefficients VIP, qui représentent les effets de chaque SNP, pour la régression PLS et la sparse PLS ont été confrontés (figure 4 de l'article 1). On remarque que les zones du génome qui sont désignées comme ayant potentiellement un effet fort sur le caractère d'intérêt, sont semblables d'une méthode à l'autre mais que la sparse PLS permet d'accentuer les effets des SNP importants. En effet, une grande partie des effets de SNP est annulée par sparse PLS ce qui permet d'attribuer des effets plus forts aux autres SNP car la somme des coefficients VIP de l'ensemble des SNP au carré est égale au nombre de variables explicatives. Ainsi, pour la sparse PLS, la précision des valeurs génomiques prédites, c'est-à-dire la corrélation entre valeurs observées et valeurs prédites, est légèrement plus faible que pour la régression PLS mais la sélection de variables apportée par la sparse PLS permet d'obtenir des équations de prédiction basées sur un nombre réduit de SNP en mettant en avant les SNP ayant un effet significatif sur le caractère.

Enfin, les capacités prédictives des méthodes PLS et sparse PLS ont été confrontées à deux méthodes couramment utilisées en évaluation génétique des bovins laitiers : le BLUP (qui ne prend en compte que l'information pedigree et les phénotypes) et le GBLUP (qui prend en compte l'information des marqueurs SNP). Le tableau 4 de l'article 1 montre que les méthodes de sélection génomique telles que le GBLUP, la régression PLS et la sparse PLS ont des capacités prédictives significativement plus élevées que le BLUP pour cinq caractères parmi les six étudiés, d'après le test de Hotelling-Williams avec un seuil de significativité de 5%. On observe une supériorité significative du GBLUP sur les méthodes PLS et sparse PLS sur la quantité de lait (+0,04 de corrélation entre PLS et GBLUP) et la fertilité (+0,10 de corrélation entre PLS et GBLUP), mais pas sur les quatre autres caractères.

Ce travail a permis de mettre en évidence que sur les données bovines laitières de la race Holstein, les méthodes utilisant l'information génomique étaient plus précises qu'un BLUP utilisant seulement l'information pedigree. Il n'y a pas de différences significatives entre la régression PLS et la sparse PLS mais la sélection de variables intégrée à la sparse PLS permet d'obtenir des modèles qui mettent en avant plus facilement les SNP qui ont un effet significatif sur les caractères d'intérêt que la PLS.

Cet article a été accepté pour publication dans *Journal of Dairy Science* en décembre 2011.

J. Dairy Sci. 95:2120–2131
http://dx.doi.org/10.3168/jds.2011-4647

A comparison of partial least squares (PLS) and sparse PLS regressions in genomic selection in French dairy cattle

C. Colombani,*[1] P. Croiseau,† S. Fritz,‡ F. Guillaume,§ A. Legarra,* V. Ducrocq,† and C. Robert-Granié*
*INRA, UR631-SAGA, BP 52627, 31326 Castanet-Tolosan Cedex, France
†INRA, UMR1313-GABI, 78352 Jouy en Josas, France
‡UNCEIA, 149 rue de Bercy, 75595 Paris, France
§Institut de l'Élevage, 149 rue de Bercy, 75595 Paris, France

ABSTRACT

Genomic selection involves computing a prediction equation from the estimated effects of a large number of DNA markers based on a limited number of genotyped animals with phenotypes. The number of observations is much smaller than the number of independent variables, and the challenge is to find methods that perform well in this context. Partial least squares regression (PLS) and sparse PLS were used with a reference population of 3,940 genotyped and phenotyped French Holstein bulls and 39,738 polymorphic single nucleotide polymorphism markers. Partial least squares regression reduces the number of variables by projecting independent variables onto latent structures. Sparse PLS combines variable selection and modeling in a one-step procedure. Correlations between observed phenotypes and phenotypes predicted by PLS and sparse PLS were similar, but sparse PLS highlighted some genome regions more clearly. Both PLS and sparse PLS were more accurate than pedigree-based BLUP and generally provided lower correlations between observed and predicted phenotypes than did genomic BLUP. Furthermore, PLS and sparse PLS required similar computing time to genomic BLUP for the study of 6 traits.
Key words: partial least squares regression, sparse partial least squares, genomic selection, French dairy cattle

INTRODUCTION

Genomic selection relies on computing genomic estimated breeding values (**GEBV**) using high-density SNP marker data. Meuwissen et al. (2001) suggested a 2-step approach to calculate GEBV. First, the effects of SNP are estimated to obtain a prediction equation using a reference population in which the animals are genotyped and phenotyped. Then, GEBV are predicted for the genotyped animals (without phenotypes) from this equation.

In the past few years, the accuracy of GEBV provided by genomic selection has been assessed using different methods in dairy cattle populations in the United States, New Zealand, Australia, the Netherlands, and France, among others. A simple BLUP, as described in Meuwissen et al. (2001) and known as genomic BLUP (**GBLUP**) in subsequent literature, was used as the reference method. The simple BLUP assumes that all SNP have an effect sampled from the same normal distribution. Hayes et al. (2009) treated Australian Holstein-Friesian bull data using a method derived from BayesA, which exploits the prior knowledge that many SNP have small individual effects on the trait and only a few have moderate to large effects. The Bayesian method was shown to be slightly more reliable (+0.02 to +0.07 compared with the reliability of BLUP) for most traits. Using New Zealand dairy cattle, Harris et al. (2009) also compared the BLUP approach with Bayesian methods (BayesA and BayesB), in which some SNP may have zero effect (Meuwissen et al., 2001). Bayesian methods slightly improved reliability (2%), whereas the use of regression methods such as least angle regression (Efron et al., 2004) did not lead to any improvement. VanRaden et al. (2009) compared the reliability of GEBV in US and Canadian young bulls, using a method similar to GBLUP that fits the allelic effects of each SNP as random effects with a normal distribution with known variance (VanRaden, 2008), and a similar method to BayesA with a heavier tail distribution for the SNP effects. As in the Australian and New Zealand results, the Bayesian approach slightly increased reliability (1% compared with the reliability of GBLUP).

Moser et al. (2009) compared 5 methods on dairy bull data including regression methods (least squares regression), shrinkage methods [Bayes regression (**Bayes-R**) similar to BayesA, and random regression BLUP (**RR-BLUP**), comparable to GBLUP], sup-

Received June 22, 2011.
Accepted December 9, 2011.
[1]Corresponding author: carine.colombani@toulouse.inra.fr

port vector machine learning methods (nonparametric support vector regression), and dimension reduction methods such as partial least squares (**PLS**) regression. The accuracy of Bayes-R, RR-BLUP, PLS, and support vector regression was very similar for the 2 traits studied by these authors. However, PLS and RR-BLUP required substantially less computation time than the Bayesian method.

Using simulated data, Coster et al. (2010) demonstrated the superiority of PLS over Bayesian methods with regard to the stability of results according to the number of QTL or the distribution of QTL variance. They also showed that the computation time for the PLS method required to fit, cross validate, and evaluate the models was less than that for the Bayesian method. However, the Bayesian method was more accurate. Solberg et al. (2009) also used simulated data to compare PLS and principal component regression with BayesB. They obtained the same results: BayesB was more accurate than other methods but PLS and principal component regression were computationally faster and simpler.

The PLS regression (Wold et al., 2001) appears to be an efficient method to deal with genomic selection data, both in its capacity to handle large data sets and its prediction ability. This approach is particularly suitable when the matrix of predictors has more variables than observations, and when multicollinearity exists among variables. The sparse PLS regression (**sPLS**, Lê Cao et al., 2008) is a recent approach that combines variable selection and modeling in a one-step procedure. Dimension reduction methods and variable selection approaches may be an attractive way to deal with the increasing number of markers used in genomic selection in dairy cattle by limiting computing time (Coster et al., 2010). Furthermore, even though PLS has already been studied in a genomic evaluation context, the authors used simulated data and did not compare PLS accuracy with current genomic selection methods such as BLUP and GBLUP (Solberg et al., 2009; Coster et al., 2010). Long et al. (2011) introduced sparsity in PLS and tested the predictive ability of sPLS versus principal component regression and PLS methods but did not apply other current genomic selection methods on their real data. They showed that combining dimension reduction and variable selection for accurate prediction of genomic breeding values was promising.

The aim of the present study was to compare PLS and sPLS on a real data set with other methods currently used in the evaluation of dairy cattle such as pedigree-based BLUP and GBLUP. Both PLS and sPLS regressions were compared based on their predictive abilities and then with pedigree-based BLUP and GBLUP results to evaluate their accuracy.

MATERIALS AND METHODS

Data

A data set of genotyped French Holstein bulls was split into a training data set and a validation data set using a cut-off birth date defined so that the validation set included the youngest 25% genotyped bulls. First, the prediction equation was estimated with the training data set, which comprised 2,976 genotyped and phenotyped Holstein bulls born before June 2002. Then, phenotypes were predicted for the bulls in the validation data set, which comprised 964 bulls (born between June 2002 and 2004).

Genotypes for 39,738 polymorphic SNP were used as independent variables. The selected SNP, provided by the Illumina Bovine SNP50 Beadchip (Illumina, San Diego, CA), had minor allele frequencies >3%. Mendelian segregation was checked. Missing genotypes were inferred from large family information with a low error rate using DualPHASE software (Druet and Georges, 2009).

Six traits with different heritability were used as dependent variables: milk yield, fat yield, and protein yield ($h^2 = 0.3$), fat content and protein content ($h^2 = 0.5$), and conception rate ($h^2 = 0.02$; Boichard and Manfredi, 1994). The bulls' phenotypes used in this study were daughter yield deviations (**DYD**, VanRaden and Wiggans, 1991; Mrode and Swanson, 2004) from a French national evaluation October 2009; that is, the average performance of the daughters of a sire, adjusted for fixed and nongenetic random effects and for the additive genetic value of their dam. For each DYD, a weighting was added in the form of the effective daughter contribution (**EDC**; VanRaden and Wiggans, 1991; Fikse and Banos, 2001). To be included in the analysis, each observation required an EDC >20.

BLUP and GBLUP

The general statistical model in BLUP and GBLUP is

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

where $\mathbf{y}$ is a vector of phenotypes (DYD), μ is the mean, $\mathbf{Z}$ is a design matrix allocating observations to breeding values, $\mathbf{g}$ is a random vector of additive genetic values, and $\mathbf{e}$ is a vector of random normal errors. In BLUP, $\mathrm{Var}(\mathbf{g}) = \mathbf{A}\sigma_g^2$, where $\mathbf{A}$ is the pedigree-based relationship matrix, and σ_g^2 is the additive genetic variance. In GBLUP, $\mathrm{Var}(\mathbf{g}) = \mathbf{G}\sigma_g^2$, and $\mathbf{G}$ is the genomic relationship matrix as defined by VanRaden (2008):

$$\mathbf{G} = \frac{\mathbf{WW}'}{2\sum_{j=1}^{p} q_j\left(1-q_j\right)},$$

where p is the number of loci considered, q_j is the frequency of an allele of the marker j, and $\mathbf{W}$ is a centered incidence matrix of SNP genotypes. The SNP marker effects are assumed to have a prior normal distribution and mixed model equations are used with the genomic relationship matrix (Cole et al., 2009; VanRaden et al., 2009).

PLS and sPLS Regression

PLS. The PLS regression introduced by Wold (1966) is a data analysis method that generalizes and combines principal component analysis and multiple regression. The method was mainly developed for industrial applications (petroleum and food processing industries) and the social sciences. The PLS method was designed to deal with the "$p \gg n$ problem"; that is, when the number of independent variables (p) is much larger than the number of observations (n). Partial least squares regression is very useful to predict dependent variables from a very large number of predictors that might be highly correlated.

In its general form, the PLS regression replaces the initial independent variable space ($\mathbf{X}$) and the initial response variable space ($\mathbf{Y}$) by smaller spaces that rely on a reduced number of variables named latent variables, which are included one by one in an iterative process. These factors will be the new variables of a usual linear regression. The main idea is to perform successive regressions by projections onto latent structures to reveal hidden or latent underlying biological effects (Wold et al., 2004; Lê Cao et al., 2008).

Using the same notation as in Lê Cao et al. (2008), the PLS regression looks for a decomposition of centered data matrices $\mathbf{X}$ and $\mathbf{Y}$ in terms of component scores, called latent variables: ($\xi_1, \ldots, \xi_h, \ldots, \xi_H$) and ($\omega_1, \ldots, \omega_h, \ldots, \omega_H$), which are linear combinations of the columns of $\mathbf{X}$ and $\mathbf{Y}$ respectively, and associated loading vectors: ($\mathbf{u}_1, \ldots, \mathbf{u}_h, \ldots, \mathbf{u}_H$) and ($\mathbf{v}_1, \ldots, \mathbf{v}_h, \ldots, \mathbf{v}_H$), where H is the number of latent variables retained in the final model. However, the regression coefficients that define these components are not linear, as they are solved via successive local regressions on the latent variables. The loading vectors are estimated to solve the following optimization problem:

$$\max_{\mathbf{u_h}=1,\mathbf{v_h}=1} \operatorname{cov}\left(\mathbf{X}_{h-1}\mathbf{u_h}, \mathbf{Y}\mathbf{v_h}\right),$$

where $\mathbf{X}_{h-1}$ is the residual $\mathbf{X}$ matrix in the regression of $\mathbf{Y}$ on ($\xi_1, \ldots, \xi_{h-1}$) for each dimension $h = 1, \ldots, H$, and the associated latent variables are denoted $\xi_h = \mathbf{X}_{h-1}\,\mathbf{u}_h$ and $\boldsymbol{\omega}_h = \mathbf{Y}\mathbf{v}_h$.

As in principal component analysis, the loading vectors and the latent variables are directly interpretable. The loading vectors $\mathbf{u}_h$ and $\mathbf{v}_h$ indicate how the x_j and y_i variables explain the relationship between $\mathbf{X}$ and $\mathbf{Y}$. The latent variables contain information regarding similarities or dissimilarities between individuals (Wold et al., 2004).

sPLS. The sPLS regression (Lê Cao et al., 2008) aims at combining variable selection and modeling in a one-step procedure. It was first proposed to handle transcriptomic data and was adapted for genomic data in this study. To understand the sPLS approach, it is helpful to describe first the principle of the PLS-singular value decomposition (Lorber et al., 1987) that solves PLS problems efficiently by decomposing the $\mathbf{X}'\mathbf{Y}$ matrix into singular values and vectors.

For a real matrix $\mathbf{M}$ ($p \times q$) of rank r, the singular value decomposition of $\mathbf{M}$ can be obtained as follows:

$$\mathbf{M} = \boldsymbol{\Gamma}\boldsymbol{\Delta}\boldsymbol{\Theta}'$$

where $\boldsymbol{\Gamma}$ ($p \times r$) and $\boldsymbol{\Theta}$ ($q \times r$) are orthonormal and $\boldsymbol{\Delta}$ ($r \times r$) is a diagonal matrix with singular values δ_k ($k = 1 \ldots r$).

The loading vectors $\mathbf{u_1}$ and $\mathbf{v_1}$ of $\mathbf{X}$ and $\mathbf{Y}$, respectively, correspond to the first singular vectors γ_1 and θ_1 if $\mathbf{M} = \mathbf{X}'\mathbf{Y}$. Then, for $h = 2, \ldots, H$, $\mathbf{M}_h$ is directly deflated by its rank-one approximation, as explained in Lê Cao et al. (2008): $\mathbf{M_h} = \mathbf{M_{h-1}} - \delta_h \mathbf{u_h} \mathbf{v_h}'$.

Sparsity of the loading vectors is introduced iteratively by penalizing both $\mathbf{u_h}$ and $\mathbf{v_h}$ with a soft-thresholding penalization, as for sparse principal component analysis (Shen and Huang, 2008). The optimization problem becomes

$$\min_{\mathbf{u},\mathbf{v}} \mathbf{M} - \mathbf{u}\mathbf{v}'^2_F + g_{\lambda_1}(\mathbf{u}) + g_{\lambda_2}(\mathbf{v}),$$

where $g_{\lambda_1}(x) = sign(x)\left(|\, x \,| - \lambda_1\right)_+$ is the soft-thresholding penalty function.

When no sparsity is required, the same results are obtained as in classical PLS. The details of this algorithm are presented in Lê Cao et al. (2008). Although PLS and sPLS can perform multi-trait analyses, each trait in this study was considered independently with $\mathbf{X}$ the matrix of SNP and $\mathbf{y}$ the vector of phenotypes of one trait. In this case, the sPLS used here is similar to the sPLS introduced by Chun and Keles (2010) and used by Long et al. (2011), in a genomic selection con-

text. When one trait is studied, PLS and sPLS are easy to implement and not time consuming because they are based on successive regressions and do not require matrix inversion.

Parameter Tuning

In both PLS and sPLS, the optimal number H of dimensions has to be determined. The parameter H can be tuned by cross-validation as in the original PLS and as proposed by Chun and Keles (2010). Coster et al. (2010) also proposed to use cross-validation to find the number of dimensions that minimized the prediction error. In this study, the root mean squared error of prediction (**RMSEP**) was minimized with 10-fold cross-validation in the training data set and for each given dimension h (Mevik and Cederkvist, 2004):

$$RMSEP = \sqrt{\frac{1}{10}\sum_{k=1}^{10}\left(\hat{\mathbf{y}}_k - \mathbf{y}_k\right)},$$

where $\hat{\mathbf{y}}_k$ is the vector of predicted values for the sample k. Solberg et al. (2009) suggested keeping the number of dimensions leading to the highest correlation between predicted values and observed values in the validation data set. This approach was also tested in this study and the results of the 2 ways used to fix H are discussed in the Results and Discussion section.

In sPLS regression, in addition to H (that was selected as above), the number of variables selected in each dimension of the model has to be fixed. Long et al. (2011) tested different values of H and different values of the number of variables selected in each dimension, based on results of their PCR study, to maximize the cross-validation correlation. In our study, we chose to minimize the RMSEP with 10-fold cross-validation in the training data set, for the previously fixed number of dimensions H. In practice, for each trait, several sPLS were performed depending on the number of selected SNP in each latent variable or dimension (assumed to be constant), as a percentage of the number of SNP in the whole data set. By construction, the same SNP could be selected in several dimensions. Ten sPLS regressions (keeping 0.2 to 10% of all SNP for each dimension considered) and the PLS regression were tested using dimensions 1 to 100.

Importance of SNP Effects

To enable better interpretation of the models, coefficients that represent the power of x_j to explain $\mathbf{y}$ has to be defined. The "variable importance in projection" (**VIP**) coefficients measure the contribution of x_j to the construction of $\mathbf{y}$ through latent variables $\boldsymbol{\xi}_\mathbf{h}$, ($h = 1,\ldots, H$) and is defined by

$$VIP_{Hj} = \sqrt{\frac{p}{\sum_{h=1}^{H} cor^2\left(\mathbf{y}, \boldsymbol{\xi}_\mathbf{h}\right)}\sum_{h=1}^{H} cor^2\left(\mathbf{y}, \boldsymbol{\xi}_\mathbf{h}\right)\omega_{hj}^2},$$

with

$$\sum_{j=1}^{p} VIP_{Hj}^2 = p.$$

The sum of squares of the VIP coefficients of all the SNP in one dimension of the PLS models is equal to the number of independent variables. Thus, the VIP coefficient of a SNP is related to the number of SNP that have a nonzero effect in the model.

The contribution of x_j to the construction of $\xi_\mathbf{h}$ is measured by its weight ω_{hj}, provided by PLS or sPLS. Although the weight ω_{hj} of x_j is interpretable, it does not account for the contribution of the latent variable $\xi_\mathbf{h}$. The VIP coefficients are able to classify the variables x_j according to their weight in each latent variable and the weight of each latent variable in the construction of $\mathbf{y}$. So they could be considered as an evaluation of the effects of SNP on the prediction of $\mathbf{y}$. Both PLS and sPLS were performed using the R package named "mixOmics" (Lê Cao et al., 2009).

Comparison of Methods and EDC

Prediction Equation. In this study, we compared 2 currently used methods for the evaluation of dairy cattle, BLUP and GBLUP, with PLS and sPLS regressions. The application of the different methods followed the same pattern, regardless of the method. The prediction equation was estimated using the training data set. The $\mathbf{y}$ phenotypes were DYD. One EDC was associated with each DYD, reflecting its uncertainty. This generates heterogeneity of variances, so that the ith DYD has a (pseudo-residual) variance σ_e^2/EDC_i (VanRaden and Wiggans, 1991). An equivalent model was constructed in all cases (BLUP, GBLUP, PLS, and sPLS), multiplying y_i and the ith row of the incidence matrix by the square root of the EDC to obtain homogeneous variances.

Accuracy of the Methods. Two criteria were used to test the accuracy of the different methods and to compare their predictive ability: the correlation (ρ) between observed and predicted values and the regression slope (b) of observed to predicted values (a value of 1 is expected; Henderson, 1963). The bulls in the validation data set were predicted by the prediction

equations provided by the different methods. Bulls in the validation set were progeny tested, so that observed DYD and associated EDC were also available for this population. These weights were taken into account in the calculation of the correlation, using

$$\rho_{xy} = \frac{\sum_i w_i \left(x_i - \overline{x}_w\right)\left(y_i - \overline{y}_w\right)}{\sqrt{\sum_i w_i \left(x_i - \overline{x}_w\right)^2 \sum_i w_i \left(y_i - \overline{y}_w\right)^2}},$$

where

$$\overline{x}_w = \frac{\sum_i w_i x_i}{\sum_i w_i} \text{ and } \overline{y}_w = \frac{\sum_i w_i y_i}{\sum_i w_i},$$

and w_i are the weights (EDC) for each observation. In the regression of observed DYD onto predicted DYD, EDC were introduced in the same way as for the production of prediction equations.

The Hotelling-Williams procedure was used to test for differences between the correlations obtained from the different methods. It tests the null hypothesis of equality between 2 dependent correlations that share a variable (Steiger, 1980; VanSickle, 2003). Under the null hypothesis, the statistical test is distributed as t with $n - 3$ degrees of freedom. All the correlations discussed in this study were compared with one another using the Hotelling-Williams test with a 5% threshold.

RESULTS AND DISCUSSION

Parameter Tuning

Figure 1 presents RMSEP obtained by cross-validation in the training data set by PLS, according to the number of dimensions. For milk yield, fat yield, protein yield, and protein content, the pattern of the different curves became very stable after only 10 dimensions. For fat content and conception rate, RMSEP stabilized after about 30 dimensions. The minimum RMSEP was reached after around 20 dimensions for milk yield, protein yield, fat content, and conception rate and after around 30 dimensions for fat yield and protein content. However, the differences in RMSEP between 2 values for the number of dimensions were very small (the minimum of the curves was not accentuated), so cross-validation did not appear to be the best criterion to choose the number of dimensions. The same conclusions were reached using the sparse PLS approach, so the number of dimensions that led to the highest correlation between phenotypes and predicted values in the validation data set was kept. This was also done for practical reasons. In fact, creating pseudo-training and pseudo-validation data sets within the training population to calibrate H was difficult, especially if a time structure (old vs. young) had to be used. Partial least squares regression was tested up to dimension 100 but the correlations obtained after more than 50 dimensions no longer increased for most traits. Figure 2 shows the correlations between observed phenotypes and predicted values in the validation data set obtained by PLS as a function of the number of latent variables built for each trait. The pattern of the different curves was the same whatever the trait: the correlation continued to increase with the number of dimensions until a plateau was reached at around dimension 30 for conception rate, 40 for milk yield, fat yield, and protein yield, and 80 for fat content and protein content. The number of dimensions in the final model was fixed at these minimum values to avoid overfitting the data (Abdi, 2010). As can be seen in Figure 2, the number of dimensions is not critical, and therefore the choice of H using the validation data set did not overestimate the predictive ability of PLS.

Sparse PLS required 2 parameters: the number of latent variables (H) and the proportion of SNP selected in each dimension. The same criterion as in PLS was used to fix the number of dimensions kept in the sparse PLS models. The pattern was the same as in PLS (results not shown): a plateau was reached at around dimension 25 for conception rate, 40 for milk, fat, and protein yields, and 50 for fat and protein contents.

The proportion of SNP selected by sPLS was tuned by cross-validation within the training data set. Table 1 shows the RMSEP provided by PLS and the different sPLS (according to the proportion of SNP selected in each dimension) for each trait and for the previously fixed number of dimensions. The selected proportion of SNP for sPLS was the one that minimized RMSEP (Table 1, in bold, based on 3 decimal places). The minimum RMSEP obtained from sPLS was close to that obtained from PLS. For example, for fat yield, the RMSEP obtained with sPLS with 2% of the total number of SNP for each dimension was the smallest (0.18) and was the same as the error of prediction obtained with PLS. The heritability of the trait played a role in the magnitude of the RMSEP: traits with the same heritability led to similar prediction errors. Milk, fat, and protein yield ($h^2 = 0.3$) obtained an RMSEP of around 0.15. Fat and protein content ($h^2 = 0.5$) gave an RMSEP of 0.56. Conception rate had the highest error (0.79) but this result was the same as the error with the PLS model (0.78). As DYD were pseudo-phenotypes calculated from the performance of the bull's daughters, low heritability traits were difficult to predict.

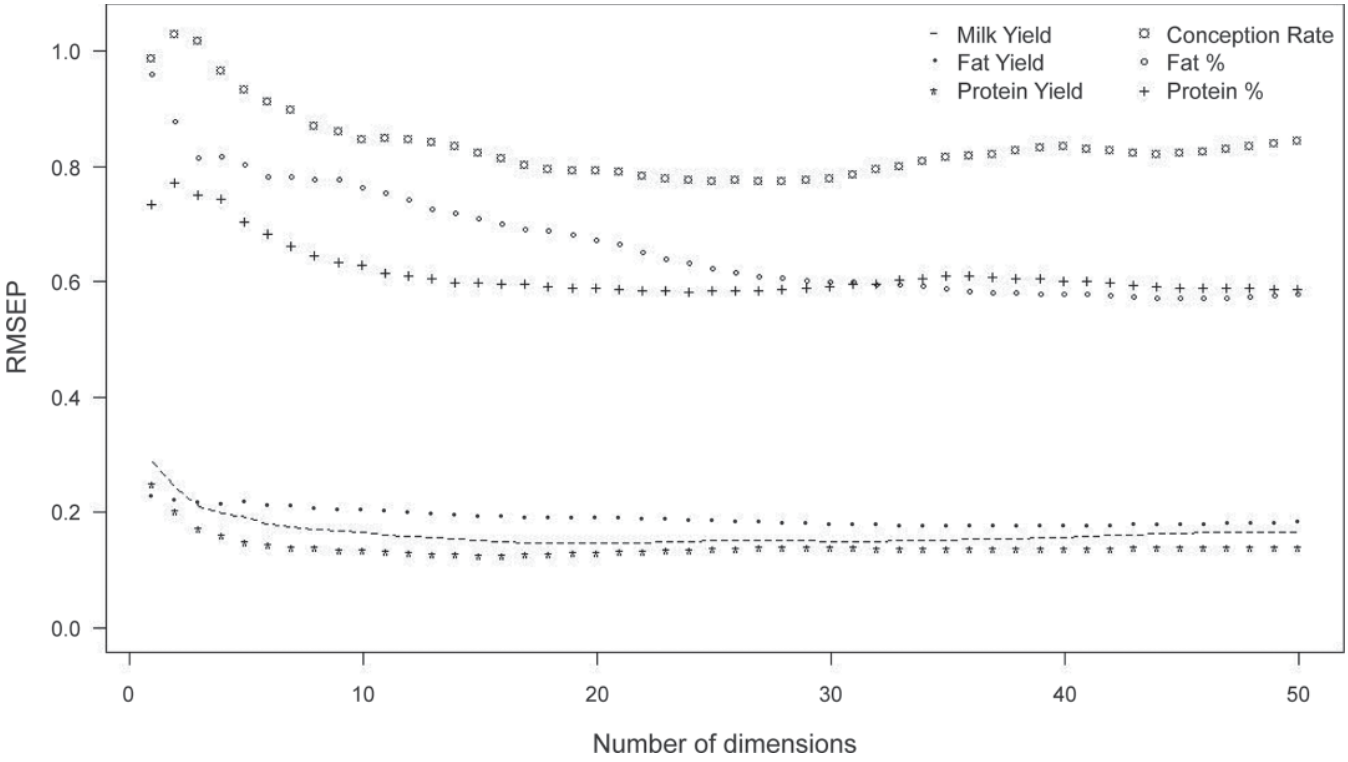


Figure 1. Root mean square error of prediction (RMSEP) for the 6 traits studied plotted against the number of dimensions for partial least squares regression.

Genomic Predictions with PLS and sPLS

Table 2 shows the accuracy of PLS and sPLS regressions for the different traits, with the number of SNP in sPLS that led to the minimum RMSEP, and the number of dimensions that led to the maximum correlation. The PLS regression gave significantly higher correlations whatever the trait with an average increase of 0.04 compared with sparse PLS. The best correlations were obtained for fat and protein content (0.70 and 0.71 in PLS, 0.66 and 0.65 in sPLS, respectively). Using the Hotelling-Williams test with a 5% threshold, production traits with a higher heritability, such as milk, fat, and protein yield ($h^2 = 0.3$), and fat and protein content ($h^2 = 0.5$) were predicted more accurately (from 0.53 in PLS and 0.48 in sPLS for milk yield to 0.71 in PLS for protein content and 0.66 in sPLS for fat content) than traits with lower heritability, such as conception rate ($h^2 = 0.02$) with an accuracy of 0.33 in PLS and 0.29 in sPLS. Thus, accuracy of prediction and heritability of the trait are closely related. Moser et al. (2010) processed data from 2,144 Holstein-Friesian bulls and compared accuracy between traits of different heritability. Production traits such as protein content, fat content, and milk yield, which have high heritability (0.56, 0.52, and 0.28, respectively), achieved higher accuracy than survival ($h^2 = 0.03$), which showed similar heritability to conception rate in the present study. For conception rate, a larger reference population is required to achieve the same level of accuracy as for production traits (Hayes et al., 2009). However, unlike the data sets used by Hayes et al. (2009), which contained only 332 Australian Holstein bulls for fertility and 798 for the other traits, the data sets of this study for conception rate used a similar number of bulls with as many daughters to evaluate DYD as the number of bulls and daughters used for production traits. Therefore, for conception rate, the power of the analysis was not reduced by a smaller data set but by low heritability. Regarding the regression slope b, both PLS and sPLS gave values below 1, with PLS values closer to 1 (from 0.60 for conception rate to 0.83 for protein content) than those for sPLS (from 0.53 for milk yield to 0.76 for protein content). Furthermore, the relationship between the heritability of the trait and slope was less clear than between heritability and the correlation. For example, the sparse PLS slopes for milk yield ($b = 0.53$) and for conception rate ($b = 0.54$) were similar but milk yield is a trait with a moderate heritability ($h^2 = 0.3$), whereas conception rate is a low heritability trait (h^2

Table 1. Root mean square error of prediction (RMSEP) in partial least squares (PLS) and each sparse PLS tested as a function of the percentage of SNP selected in each latent variable [minimum RMSEP (3 decimal places) in bold]

	Sparse PLS (% of the SNP data set selected)										PLS
Variable	0.2	0.4	0.6	0.8	1	2	3	4	5	10	100
Milk yield	0.17	0.16	0.17	0.16	0.16	0.16	0.15	**0.15**	0.15	0.15	0.15
Fat yield	0.18	0.20	0.19	0.18	0.21	**0.18**	0.18	0.19	0.18	0.18	0.18
Protein yield	0.14	0.15	0.14	0.14	0.14	0.14	0.13	0.13	0.13	**0.13**	0.13
Fat content	0.60	0.67	0.58	**0.56**	0.57	0.58	0.57	0.59	0.63	0.60	0.57
Protein content	0.64	0.62	0.59	0.60	0.58	0.57	0.56	0.59	**0.56**	0.56	0.58
Conception rate	0.94	0.83	0.81	0.83	0.86	0.81	0.87	**0.79**	0.83	0.82	0.78

= 0.02). Two traits with 2 different heritabilities would lead to 2 different correlations but not necessarily to 2 different regression slopes.

Sparse PLS gave significantly less accurate predictions than PLS. However, sPLS performed a variable selection by allowing the number of variables in the final model to be reduced by 50% (Table 2). The number of selected SNP was reduced to 9,832 for fat content. One explanation could be the presence of *DGAT1* (Grisart et al., 2004), a gene on bovine chromosome 14 that leads to a mutation that has a major effect on fat content in milk in Holstein dairy cattle. Therefore, a small number of SNP, of which many were located around this QTL, was sufficient to obtain accurate predictions. However, the number of selected SNP could have been smaller, but the large number of latent components (50) used in the model led to a high number of selected SNP, irrespective of the presence of a large QTL.

Indeed, the number of SNP in the final model was directly related to the percentage of SNP kept for each latent variable and to the number of latent variables. The large number of SNP selected for protein yield was the consequence of the large proportion of SNP (10%) selected for each of the 38 latent variables. Fifty latent variables were used for fat content but the number of SNP was reduced because the sPLS that led to the

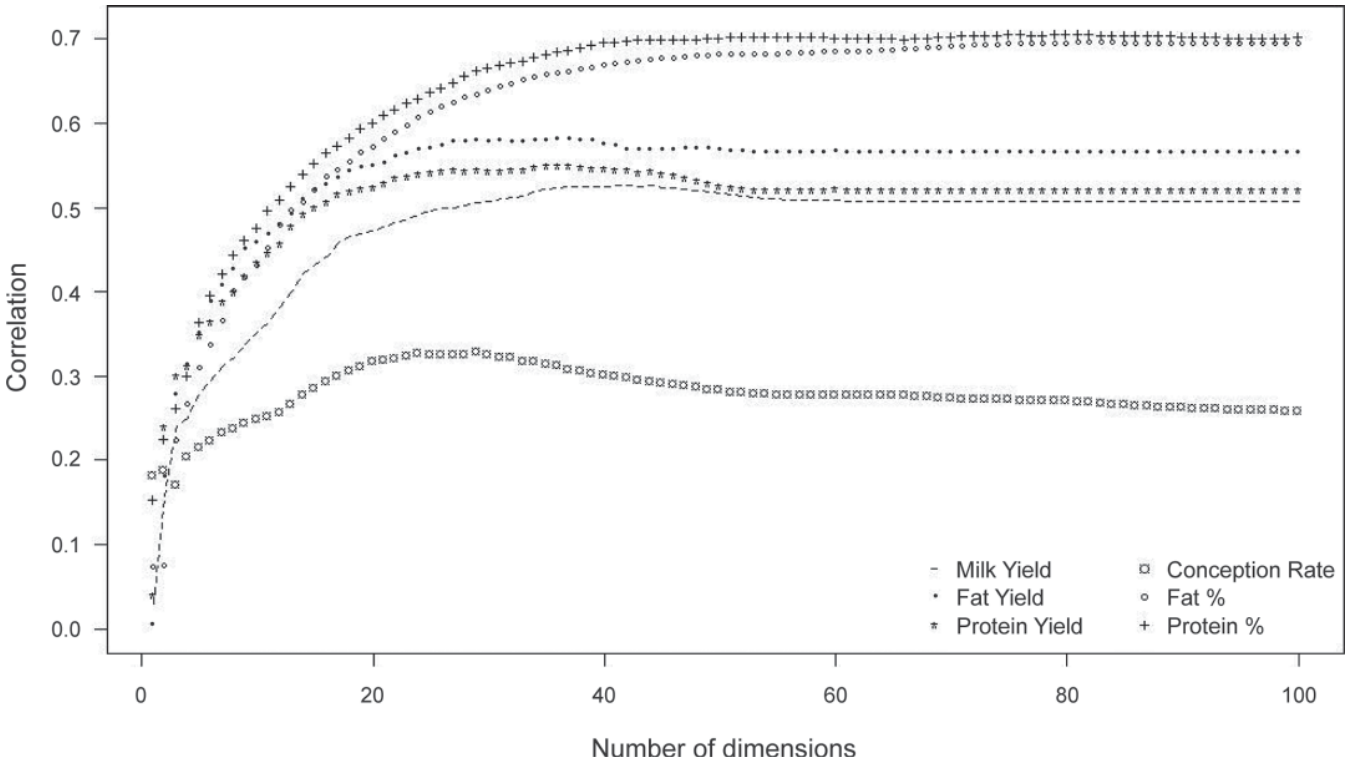


Figure 2. Correlation between observed and predicted daughter yield deviations for the 6 traits studied plotted against the number of dimensions for partial least squares regression.

Table 2. Effective daughter contribution-weighted correlations (ρ) and regression slopes (*b*) provided by partial least squares regression (PLS) and sparse PLS (sPLS)

Item	Milk yield	Fat yield	Protein yield	Fat content	Protein content	Conception rate
PLS						
ρ	0.53	0.58	0.55	0.70	0.71	0.33
b	0.65	0.83	0.67	0.80	0.83	0.60
Dim[1]	42	37	36	83	75	29
sPLS						
ρ	0.48	0.54	0.51	0.66	0.65	0.29
b	0.53	0.70	0.60	0.69	0.76	0.54
Dim	44	43	38	50	51	27
No. of SNP[2]	22,948	16,296	32,578	9,832	26,034	20,150

[1]Number of latent variables or dimensions included in the final model.

[2]Number of SNP selected by sPLS.

minimum RMSEP kept only 0.8% of SNP at each dimension. Finally, for both milk yield and conception rate, 4% of SNP on each dimension were required and similar numbers of SNP were kept (22,948 and 20,150, respectively) but with a much larger number of latent variables for milk yield (44 vs. 27 for conception rate) because the same SNP can be selected for several latent variables.

The number of latent variables required by PLS and sPLS was very high (between 27 and 83 dimensions depending on the trait). Long et al. (2011) evaluated PLS, sPLS developed by Chun and Keles (2010), and principal component regression in Holstein bulls. They showed that to predict milk yield in Holsteins by PLS, only 15 latent components were sufficient to obtain the largest predictive correlation, suggesting a strong predictive power of the latent variables. The lack of predictive power of the additional latent components in the present study seems to be due to the presence of highly EDC-weighted bulls in the training data set, which had a major effect on the distribution of the phenotypes.

Effect of EDC on PLS and sPLS

The distribution of the phenotypes in the training data set was normal, but applying EDC disturbed the normal distribution of the phenotypes (results not shown). Furthermore, the correlation between observed DYD and observed weighted DYD was surprisingly small: about 0.20 for milk, fat, and protein yield, about 0.45 for fat and protein content, and about 0.35 for conception rate. Consequently, we investigated the effect of the use of EDC on the training population.

Figure 3 shows the distribution of EDC in the training data set for the 6 traits studied. The computation of EDC relies on the number of daughters per bull and trait parameters (heritability and repeatability). The EDC were the same for milk yield, fat yield, and protein yield and fat content and protein content. In order for the information content in conception rate to be consistent with that of production traits, a large number of daughters per bull was assumed both for production traits and conception rate. The graphs show that some bulls differed from others in their very high EDC. These bulls were generally older than the other bulls in the training population and did not obtain stronger DYD. The significant difference in weights between bulls resulted in a bias and had a major effect on the number of latent components introduced in the PLS and sPLS models. To test this hypothesis, the same study was performed without considering EDC either in the weighting of DYD or in the calculation of accuracy and of the regression slope.

Table 3 shows the accuracy of PLS and sPLS regressions with respect to the different traits without EDC. For production traits, with PLS, the correlations were very similar to the results of PLS with EDC (Table 2). With sPLS, accuracy was significantly better in the study without EDC (fat yield $\rho = 0.59$ and protein content $\rho = 0.72$, for example) than in the study including EDC (fat yield $\rho = 0.54$ and protein content $\rho = 0.65$). For conception rate, using EDC reduced significantly the accuracy of both PLS and sPLS. Sparse PLS and PLS gave no significantly different correlations without EDC, whereas sPLS was shown to be significantly less accurate than PLS with EDC. The regression slopes were differently affected by the use or nonuse of EDC, with regression slopes lower than or close to those in the previous study with PLS, and regression slopes greater than or equal to those in the previous study with sPLS. Irrespective of the trait, the number of dimensions was considerably reduced with both methods, which led to a stronger variable selection in sPLS and a restricted number of SNP in the prediction equation. Only 10 or fewer latent variables were needed to obtain the best correlations for fat yield, protein yield, and conception rate with both PLS and sPLS. These results are in

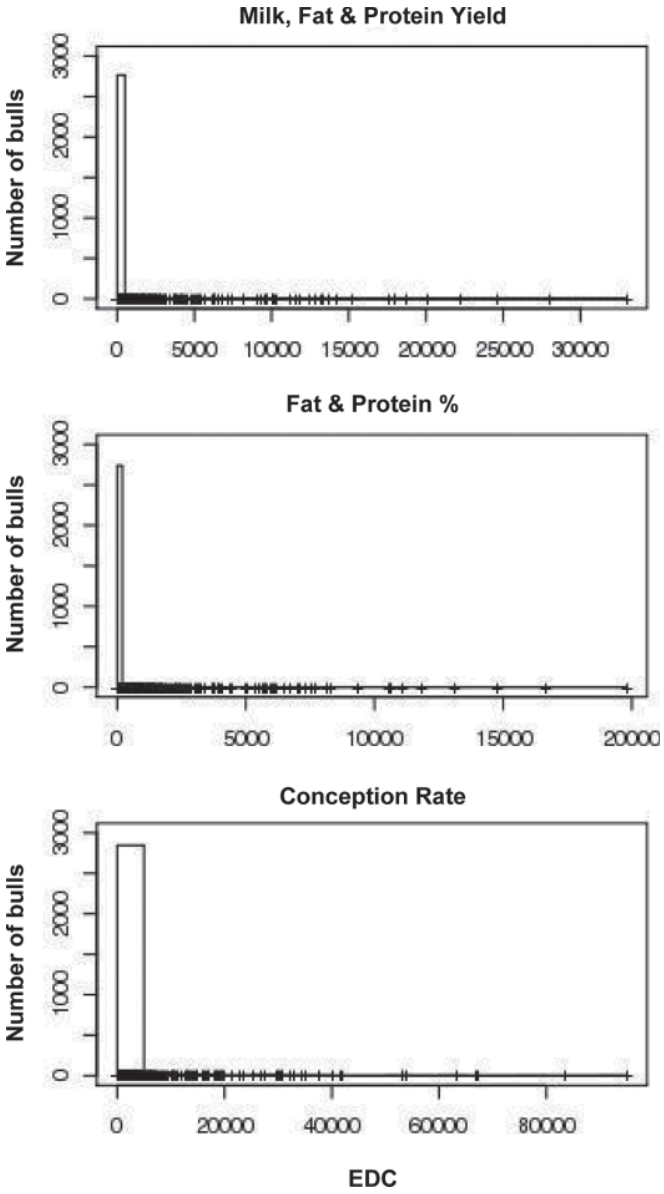


Figure 3. Distribution of effective daughter contribution (EDC) in the training data set.

agreement with the number of dimensions obtained by Long et al. (2011). For protein yield and protein content, the number of SNP remained high because the sPLS, with the minimum RMSEP by cross-validation, kept 10% and 5% of the SNP in each dimension. Therefore, introducing EDC did not have a major effect on the predictive ability of PLS but did affect the number of latent variables of the model. With PLS methods, it is probably wiser to have a more homogeneous distribution of the weights to favor the dimension-reducing ability of the PLS variants and hence to reduce computation time. In the remainder of the study, EDC were not included.

Both PLS and sPLS seemed to fit well in the context of genomic selection, but sPLS led to slightly smaller correlations than PLS but with no significant differences when EDC were not considered. However, sPLS favored a variable selection that can highlight the most important SNP, if required. A secondary aim of the present study was to underline the explanatory power of PLS regarding biological processes. However, to interpret the model in a biological context, coefficients that represent the explanatory power of the variables in the construction of the response variable are needed.

VIP Coefficients of PLS and sPLS

Figure 4 shows the VIP coefficient computed for each variable according to the position of the SNP on the genome. All the SNP variables are shown in the graph. A large number of VIP coefficients were set to zero in sPLS, whereas all the coefficients differed from zero in PLS. Therefore, sPLS was able to select variables based on VIP coefficients. The scale of the y-axis was the same for all the traits except fat content, which had the highest VIP coefficients.

Excluding the use of EDC, the variable selection performed by sPLS highlighted areas of interest. For fat yield, some SNP on chromosomes 2 and 5 were already highlighted by PLS but were clearly weighted up in sPLS (VIP coefficients of around 2.2 and 2.5 in PLS and of 12 and 16 in sparse PLS, respectively). For fat content, the SNP located at the beginning of chromosome 14 were highlighted by both methods. This location corresponds to a region of the genome that hosts the QTL *DGAT1* (Grisart et al., 2004). Furthermore, chromosomes 5 and 20 were more clearly revealed by sPLS than by PLS. The same comments can be made for most traits. The differences between PLS and sPLS were obvious for all traits with higher VIP coefficients in sPLS due to the reduced number of SNP considered. Conception rate showed many regions of interest with both methods. The aim of this study was to highlight differences and similarities between PLS and sPLS. We are not yet able to affirm that the different genome areas localized by PLS or sPLS correspond to QTL locations. This is currently under study (Colombani et al., 2011).

Comparing PLS and sPLS with Current Methods

The aim of this study was to compare PLS variants with currently used methods in the evaluation of dairy cattle. Table 4 shows the correlation between observed and estimated DYD for all the traits, with 3 methods

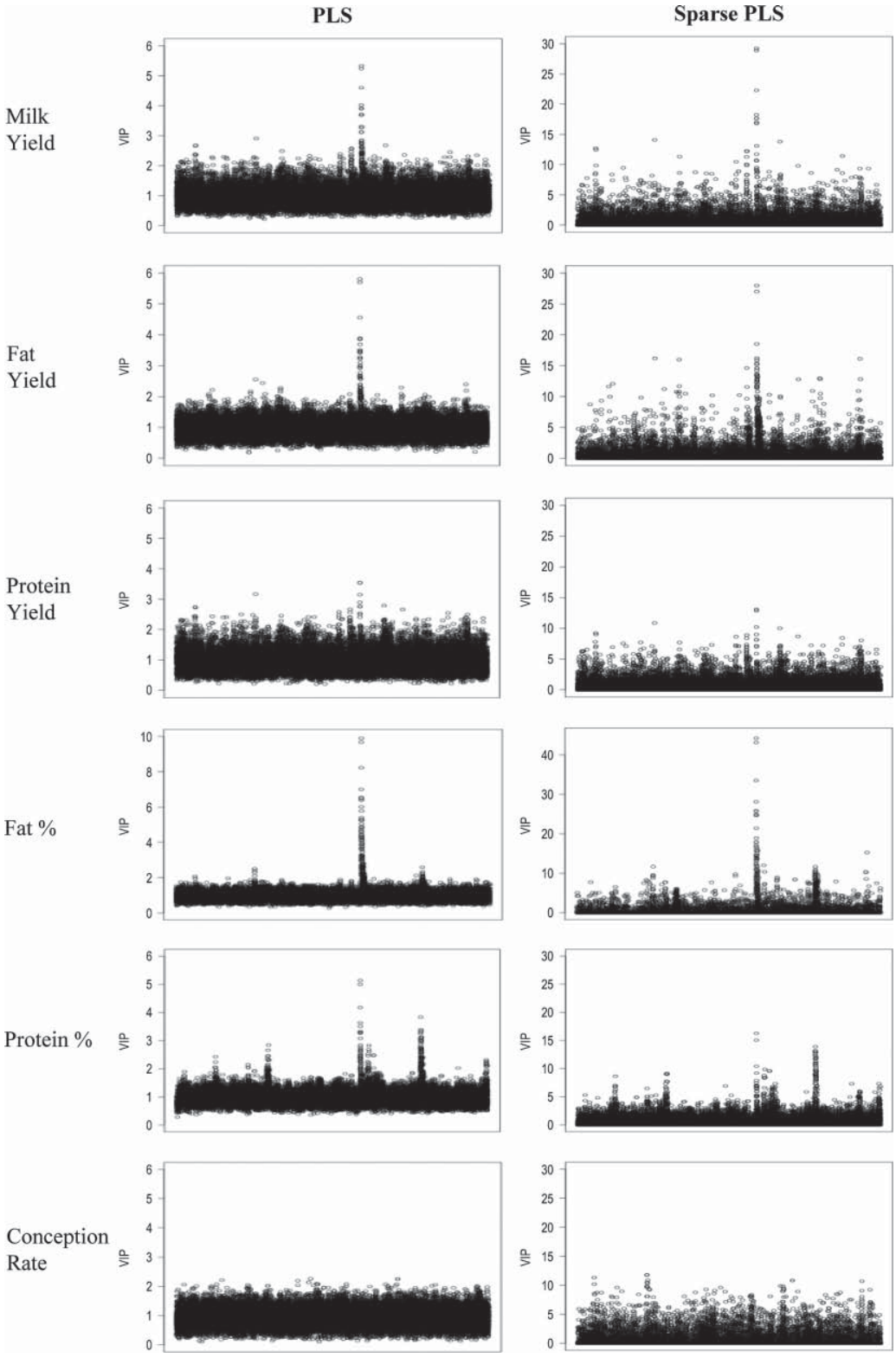


Figure 4. Variable importance in projection (VIP) coefficients from partial least squares regression (PLS) and sparse PLS for the 6 traits without effective daughter contribution (EDC).

Table 3. Correlations (ρ) and regression slopes (b) provided by partial least squares regression (PLS) and sparse PLS (sPLS) without effective daughter contribution

Item	Milk yield	Fat yield	Protein yield	Fat content	Protein content	Conception rate
PLS						
ρ	0.52	0.58	0.55	0.71	0.71	0.25
b	0.60	0.77	0.64	0.80	0.82	0.62
Dim[1]	20	9	10	23	26	3
sPLS						
ρ	0.50	0.59	0.53	0.72	0.72	0.21
b	0.56	0.76	0.59	0.81	0.77	0.49
Dim	14	10	10	11	23	3
No. of SNP[2]	17,408	6,779	26,458	2,870	27,191	4,592

[1]Number of latent variables or dimensions included in the final model.

[2]Number of SNP selected by sPLS.

of genomic selection: PLS, sPLS, and GBLUP, compared with pedigree-based BLUP. On average, the correlations obtained by genomic selection methods were significantly higher than with pedigree-based BLUP for the 5 production traits concerned (0.426 for pedigree-based BLUP and 0.614, 0.612, and 0.630 for PLS, sPLS, and GBLUP, respectively). The differences between pedigree-based BLUP and genomic selection methods were not as clear for the conception rate trait, with a correlation of 0.28 for BLUP and 0.21 and 0.35 for sPLS and GBLUP. For conception rate, the BLUP correlation and the PLS correlation were not significantly different. Genomic BLUP (from 0.35 to 0.73), PLS (from 0.25 to 0.71), and sPLS (from 0.21 to 0.72) gave similar results for all traits concerned except for milk yield and conception rate, for which GBLUP gave significantly better results.

As expected, the genomic selection methods tested in this study were more efficient than pedigree-based BLUP. Genomic BLUP was accurate for use with French Holstein data with significantly higher accuracy for some traits. However, PLS and sPLS methods were comparable to GBLUP if we considered one trait at a time. Regarding computing time, GBLUP requires one inversion of the genomic relationship matrix for all traits, which took about 1 h. Then, once the genomic relationship matrix was inverted, computation was a matter of seconds. For each trait, PLS took about 10 min and sPLS took less than 10 min, depending on the number of SNP selected. The disadvantage of PLS and sPLS with respect to GBLUP for some traits could be overcome by building a multi-trait model.

CONCLUSIONS

Sparse PLS regression was compared with PLS and with 2 currently used methods in the evaluation of dairy cattle: pedigree-based BLUP and GBLUP. The results demonstrated that PLS variants were more efficient than pedigree-based BLUP but less accurate than GBLUP for 2 out of 5 traits. Furthermore, GBLUP provided a clear biologic interpretation by the use of a genomic relationship matrix that PLS and sPLS may lack, and PLS and sPLS do not provide reliabilities of GEBV, in contrast to GBLUP. Sparse PLS enabled easier identification of relevant variables than PLS, which are possibly associated with QTL regions. Currently, more and more markers are being genotyped, forcing the handling of increasing quantities of data and consequently a critical need exists for methods that perform well in this context. Sparse PLS could be used as a preliminary step in genomic selection to reduce the number of SNP used in the prediction equations provided by other genomic selection methods, such as Bayesian methods. Most importantly, the sPLS algorithm is fast to compute even with a large reference

Table 4. Correlations between observed daughter yield deviations (DYD) and predicted DYD provided by partial least squares regression (PLS), sparse PLS (sPLS), pedigree-based BLUP (BLUP), and genomic BLUP (GBLUP)

Item	Milk yield	Fat yield	Protein yield	Fat content	Protein content	Conception rate
BLUP	0.38	0.40	0.44	0.44	0.47	0.28
PLS	0.52	0.58	0.55	0.71	0.71	0.25
sPLS	0.50	0.59	0.53	0.72	0.72	0.21
GBLUP	0.56	0.59	0.55	0.72	0.73	0.35

population and a large number of explanatory variables in a one-trait evaluation.

ACKNOWLEDGMENTS

This work was supported by the French project AMASGEN, financed by the French National Research Agency (ANR, Paris France) and ApisGene (Paris, France). Labogena (Jouy-en-Josas, France) is gratefully acknowledged for providing the genotypes. We thank the reviewers for their constructive suggestions that enabled major improvement of the manuscript

REFERENCES

Abdi, H. 2010. Partial least squares regression and projection on latent structure regression (PLS regression). Comput. Stat. 2:97–106.

Boichard, D., and E. Manfredi. 1994. Genetic analysis of conception rate in French Holstein cattle. Acta Agric. Scand. A Anim. Sci. 44:138–145.

Chun, H., and S. Keles. 2010. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. J. R. Stat. Soc. Series B Stat. Methodol. 72:3–25.

Cole, J. B., P. M. VanRaden, J. R. O'Connell, C. P. Van Tassell, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and G. R. Wiggans. 2009. Distribution and location of genetic effects for dairy traits. J. Dairy Sci. 92:2931–2946.

Colombani, C., P. Croiseau, C. Hozé, S. Fritz, F. Guillaume, D. Boichard, A. Legarra, V. Ducrocq, and C. Robert-Granié. 2011. Could genomic selection be efficient to detect QTL? 15th QTLMAS Workshop, Rennes, France. BMC Proceedings, London, UK.

Coster, A., J. W. M. Bastiaansen, M. P. L. Calus, J. A. M. van Arendonk, and H. Bovenhuis. 2010. Sensitivity of methods for estimating breeding values using genetic markers to the number of QTL and distribution of QTL variance. Genet. Sel. Evol. 42:9.

Druet, T., and M. Georges. 2010. A hidden Markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. Genetics 184:789–798.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani. 2004. Least angle regression. Ann. Stat. 32:407–451.

Fikse, W. F., and G. Banos. 2001. Weighting factors of sire daughter information in international genetic evaluations. J. Dairy Sci. 84:1759–1767.

Grisart, B., F. Farnir, L. Karim, N. Cambisano, J. J. Kim, A. Kvasz, M. Mni, P. Simon, J. M. Frere, W. Coppieters, and M. Georges. 2004. Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. Proc. Natl. Acad. Sci. USA 101:2398–2403.

Harris, B. L., D. L. Johnson, and R. J. Spelman. 2009. Genomic selection in New Zealand and the implications for national genetic evaluation. Pages 325–330 in Proc. 36th ICAR Biennial Session: Identification, Breeding, Production, Health and Recording of Farm Animals. International Committee for Animal Recording (ICAR), Rome, Italy.

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

Henderson, C. R. 1963. Selection index and expected genetic advance. Statistical genetics and plant breeding. Nat. Acad. Sci. Nat. Res. Counc. Pub. 983:141–163.

Lê Cao, K. A., I. Gonzalez, and S. Dejean. 2009. integrOmics: An R package to unravel relationships between two omics datasets. Bioinformatics 25:2855–2856.

Lê Cao, K. A., D. Rossouw, C. Robert-Granie, and P. Besse. 2008. A sparse PLS for variable selection when integrating omics data. Stat. Appl. Genet. Mol. Biol. 7:35.

Long, N., D. Gianola, G. J. M. Rosa, and K. A. Weigel. 2011. Dimension reduction and variable selection for genomic selection: Application to predicting milk yield in Holsteins. J. Anim. Breed. Genet. 128:247–257.

Lorber, A., L. E. Wangen, and B. R. Kowalski. 1987. A theoretical foundation for the PLS algorithm. J. Chemometr. 1:19–31.

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.

Mevik, B. H., and H. R. Cederkvist. 2004. Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR). J. Chemometr. 18:422–429.

Moser, G., M. Khatkar, B. Hayes, and H. Raadsma. 2010. Accuracy of direct genomic values in Holstein bulls and cows using subsets of SNP markers. Genet. Sel. Evol. 42:37.

Moser, G., B. Tier, R. Crump, M. Khatkar, and H. Raadsma. 2009. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genet. Sel. Evol. 41:56.

Mrode, R. A., and G. J. Swanson. 2004. Calculating cow and daughter yield deviations and partitioning of genetic evaluations under a random regression model. Livest. Prod. Sci. 86:253–260.

Shen, H. P., and J. H. Z. Huang. 2008. Sparse principal component analysis via regularized low rank matrix approximation. J. Multivariate Anal. 99:1015–1034.

Solberg, T. R., A. K. Sonesson, J. A. Woolliams, and T. H. E. Meuwissen. 2009. Reducing dimensionality for prediction of genome-wide breeding values. Genet. Sel. Evol. 41:29.

Steiger, J. H. 1980. Tests for comparing elements of a correlation matrix. Psychol. Bull. 87:245–251.

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

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92:16–24.

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

VanSickle, J. 2003. Analysing correlations between stream and watershed attributes. J. Am. Water Resour. Assoc. 39:717–726.

Wold, H. 1966. Estimation of principal components and related models by iterative least squares. Pages 391–420 in Multivariate Analysis. P. R. Krishnaiah, ed. Academic Press, New York, NY.

Wold, S., L. Eriksson, J. Trygg, and N. Kettaneh. 2004. The PLS method—Partial least squares projections to latent structures and its application in industrial RDP (research, development, and production). Technical report. Umea University, Sweden.

Wold, S., M. Sjostrom, and L. Eriksson. 2001. PLS-regression: A basic tool of chemometrics. Chemom. Intell. Lab. Syst. 58:109–130.

4.2.2 Étude de l'influence des EDC sur les méthodes PLS et sparse PLS

L'application de la régression PLS et de la sparse PLS pour l'évaluation génomique de taureaux Holstein, a soulevé de nombreuses questions. La première partie du travail a été d'adapter les méthodes à ce contexte particulier. Lors de cette étape, le principal problème a été l'intégration des poids (les EDC) des taureaux de l'ensemble d'apprentissage dans la construction des équations de prédiction, que ce soit dans la méthode de régression PLS ou dans la sparse PLS. Il a été montré dans l'article 1 que la prise en compte ou non des EDC dans les modèles ne modifie pas la valeur de la corrélation c'est-à-dire qu'elle n'impacte pas la capacité prédictive des méthodes. En revanche, les modèles sont plus complexes quand on introduit les EDC, c'est-à-dire qu'ils reposent sur un nombre plus important de variables latentes, d'où un nombre de variables sélectionnées par sparse PLS plus grand (voir les tableaux 2 et 3 de l'article 1). Les EDC ont aussi un impact très fort sur l'estimation des effets des SNP. Dans la régression PLS ou la sparse PLS, les coefficients VIP sont utilisés pour représenter les effets individuels des SNP sur le caractère étudié. Ils prennent en compte non seulement le poids individuel de chaque SNP sur chaque variable latente construite mais aussi la contribution de chaque variable latente dans le modèle. Le coefficient VIP associé à chaque marqueur représente ainsi l'importance de chaque marqueur à la construction du modèle final. Plus la valeur du VIP est importante, plus la variable contribue au phénotype étudié.

Prenons l'exemple de deux caractères : la quantité de matière grasse (MG) et le taux protéique (TP). La figure 4.1 présente le graphe des coefficients VIP pour la régression PLS et la sparse PLS en prenant en compte les EDC (ligne « *Avec EDC* ») ou sans prendre en compte les EDC (ligne « *Sans EDC* », graphes présentés dans l'article 1). On remarque que la prise en compte des EDC dans les modèles perturbe le calcul des VIP et cache les zones du génome mises en avant dans l'étude « sans EDC ». Les graphes des « effets » SNP sont donc moins facilement « interprétables » : ils ne mettent pas en avant les variables SNP les plus significatives ou pertinentes.

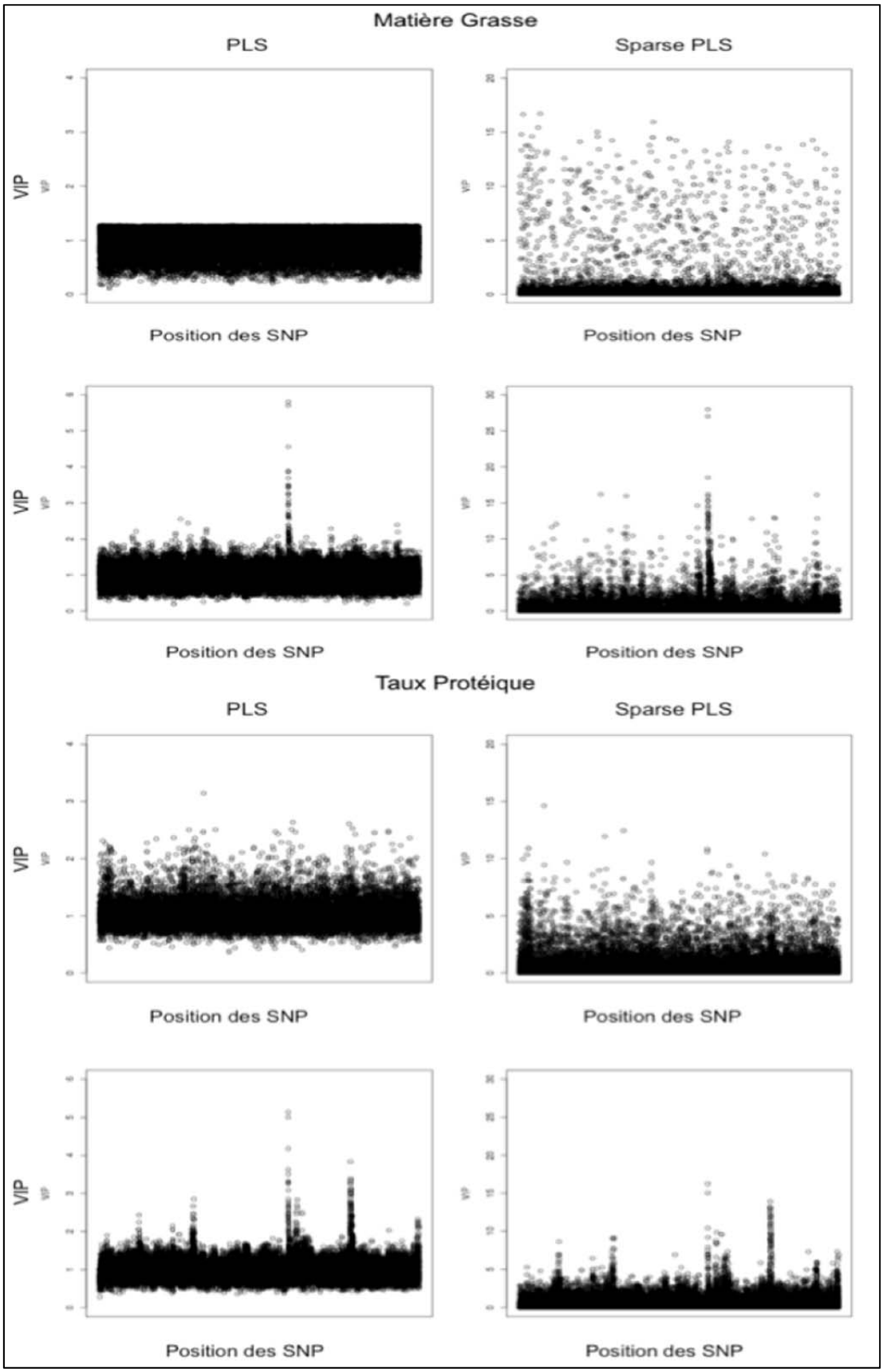


Figure 4.1 : Coefficients VIP associés aux marqueurs SNP en fonction de leur position sur le génome, en prenant en compte ou non les EDC dans la modélisation PLS et sparse PLS pour les caractères matière grasse et taux protéique

Le profil des EDC associés aux taureaux d'apprentissage (ensemble A) pour chaque caractère étudié est présenté dans la figure 3 de l'article 1). Cette distribution met en avant que certains taureaux se détachent des autres par leur forte valeur EDC. Ce sous-ensemble de taureaux est composé, en moyenne, de taureaux plus âgés que le reste de la population d'apprentissage. Leur fort EDC est dû au nombre très important de filles considérés dans le calcul du DYD. Ces taureaux ont donc été fortement diffusés car ils sont parmi les plus performants de leurs contemporains mais ils ne sont pas les plus efficients de notre population d'apprentissage. Pour étudier l'effet de ces taureaux dans la modélisation PLS et sparse PLS, deux sous-ensembles de la population d'apprentissage A ont été créés : les taureaux ayant un EDC inférieur à 50 sur le caractère quantité de lait (soit un seuil de 50 sur le MG et le MP, 30 sur le TB et le TP et 150 sur la fertilité) composent l'ensemble A- (environ 2600 taureaux) et les taureaux dont l'EDC est supérieur ou égal à ce même seuil composent l'ensemble A+ (environ 350 taureaux). Ce seuil a été choisi arbitrairement de façon à obtenir une distribution des EDC de l'ensemble A- homogène d'un taureau à l'autre. D'autres valeurs de seuil ont été testées (de 30 à 80 sur le lait) sans modifier les résultats.

Dans un premier temps, l'ensemble A est confronté à l'ensemble A-. La figure 4.2 compare la distribution des DYD avec la distribution des DYD pondérés ($\sqrt{EDC} * DYD$) dans l'ensemble d'apprentissage complet A et dans l'ensemble réduit A- (ensemble des taureaux avec des poids homogènes les uns par rapport aux autres). Dans l'ensemble A-, les distributions sont similaires entre DYD et DYD pondérés et sont proches d'une loi gaussienne. Dans l'ensemble A, les DYD pondérés présentent une distribution très asymétrique avec une queue de distribution importante : pour le caractère MG, par exemple, les valeurs s'étendent jusqu'à 20 alors que dans l'ensemble A-, elles restent inférieures à 4. Une EDC représente la précision associée au phénotype considéré (DYD) pour chaque individu. Plus un taureau a une mesure phénotypique précise, plus son EDC sera élevé et plus ce taureau aura d'influence sur la modélisation par rapport à d'autres taureaux avec des EDC moins importants. Ainsi, le fait d'avoir des valeurs d'EDC extrêmes par rapport à la moyenne des taureaux, perturbe la distribution statistique des DYD pondérés en allongeant la queue de distribution.

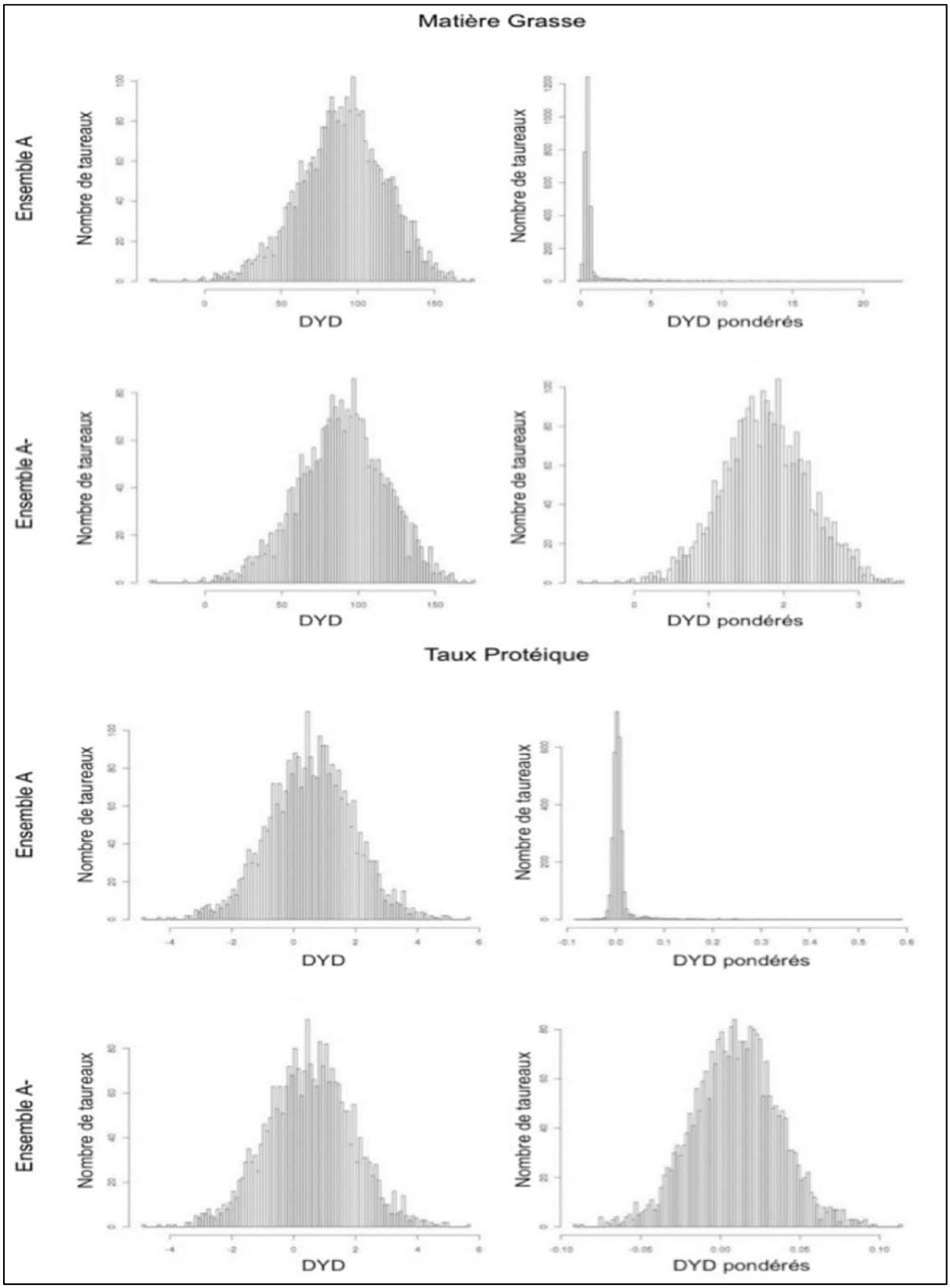


Figure 4.2 : Distribution des DYD et des DYD pondérés associés aux taureaux des ensembles A (ensemble complet des taureaux de la population d'apprentissage) et A- (ensemble réduit de taureaux de la population d'apprentissage), pour les caractères matière grasse et taux protéique

Le tableau 4.1 présente les corrélations entre DYD et $\sqrt{EDC} * DYD$, pour chacun des ensembles des taureaux considérés (ensemble complet A et ensemble réduit A-) pour les six caractères étudiés. Ces corrélations traduisent l'intensité de la liaison entre DYD et DYD pondérés dans le cas d'une distribution hétérogène (ensemble A) ou homogène (ensemble A-) des EDC. Elles sont faibles pour tous les caractères sur l'ensemble A et particulièrement pour le lait (ρ_{DYD}= 0,16), la matière grasse (ρ_{DYD}= 0,19), et la matière protéique (ρ_{DYD}= 0,18). Sur l'ensemble A-, les DYD et les DYD pondérés sont fortement corrélés avec des corrélations supérieures à 0,9 sauf pour la fertilité (ρ_{DYD}= 0,70). Cela montre de nouveau que les taureaux qui ont de très forts EDC sont très influents. Ils changent la distribution phénotypique (DYD) globale originale (apparition de données extrêmes associées aux taureaux à forts EDC). Il est donc important d'évaluer l'impact des pondérations hétérogènes sur les méthodes PLS et sparse PLS.

Tableau 4.1 : Corrélations entre DYD et DYD pondérés des ensembles A et A- pour les 6 caractères étudiés

	Lait	**MG**	**MP**	**TB**	**TP**	**Fer**
Ensemble A	0,16	0,19	0,18	0,46	0,45	0,35
Ensemble A-	0,93	0,93	0,92	0,99	0,99	0,70

Pour cela, nous avons appliqué la régression PLS et la sparse PLS pondérées sur l'ensemble d'apprentissage A-, c'est-à-dire en écartant les taureaux avec un fort EDC. L'ensemble de validation reste inchangé (ensemble V) par rapport à l'étude présenté dans l'article 1. Le tableau 4.2 présente les résultats de la régression PLS et de la sparse PLS pour l'ensemble A- et les compare aux résultats précédemment obtenus sur l'ensemble A. Malgré une différence maximale de corrélations entre ensemble A et ensemble A- de 0,06 pour la régression PLS (TP et fertilité) et pour la sparse PLS (fertilité), le test de Hotelling-Williams (seuil de 5%) indique que la précision des estimations génomiques n'est pas significativement affectée par la restriction de l'ensemble d'apprentissage. Les pentes de régression sont moins proches de 1 pour la régression PLS sur l'ensemble A-. Le biais de prédiction est donc plus important sur l'ensemble A- que sur l'ensemble A. Pour la sparse PLS,

cela n'est pas observé sur tous les caractères : seules les estimations des valeurs génomiques du caractère MP et du caractère TP sont plus biaisées sur l'ensemble A- mais avec des écarts très faibles (0,02 pour le MP et 0,04 pour le TP).

Tableau 4.2 : Corrélations pondérées (ρ) et pentes de régression (b) entre DYD observés dans l'ensemble de validation et DYD estimés à partir d'un modèle PLS ou sparse PLS à H dimensions construit sur A ou A- en race Holstein

PLS	**Ensemble A**			**Ensemble A-**		
	ρ	***b***	***H***	***ρ***	***b***	***H***
Lait	0,53	0,65	42	0,50	0,61	10
MG	0,58	0,83	37	0,56	0,77	10
MP	0,55	0,67	36	0,51	0,60	10
TB	0,70	0,80	83	0,65	0,77	21
TP	0,71	0,83	75	0,65	0,76	37
Fer	0,33	0,60	29	0,27	0,55	5

sPLS	**Ensemble A**			**Ensemble A-**		
	ρ	***b***	***H***	***ρ***	***b***	***H***
Lait	0,48	0,53	44	0,49	0,61	7
MG	0,54	0,70	43	0,53	0,71	10
MP	0,51	0,60	38	0,47	0,58	7
TB	0,66	0,69	50	0,70	0,78	14
TP	0,65	0,76	51	0,68	0,72	24
Fer	0,29	0,54	27	0,23	0,56	4

En revanche, le nombre de dimensions des modèles PLS et sparse PLS est drastiquement réduit sur l'ensemble A-. Les différents résultats de l'étude de l'ensemble A-, c'est-à-dire avec les taureaux dont les EDC sont homogènes, sont semblables aux résultats de l'étude sans EDC présentée dans l'article. Les graphes VIP sur l'ensemble A- sont très proches des graphes VIP de l'étude sans EDC (résultats non montrés). On peut donc conclure de la même façon pour l'étude de l'ensemble A- que pour l'étude sans EDC. Rappelons que les taureaux qui ont les plus forts EDC sont les plus diffusés et parmi les plus anciens. Ils sont donc fortement liés aux plus jeunes taureaux de la population d'apprentissage et de la

population de validation. Si on enlève ces taureaux de la population d'apprentissage, les relations de parentés entre les animaux d'apprentissage et ceux de validation sont maintenues par la présence des fils de ces taureaux dans la population d'apprentissage.

Ainsi, l'efficacité d'une évaluation génomique (jugée avant tout sur la précision des DYD prédits) n'est pas affectée par la présence ou non de ces taureaux particuliers. Cependant cela a un impact sur la complexité des modèles (nombre de variables latentes introduites dans les modèles finaux) et donc sur le temps de calcul. De ce fait, la sélection de variables par sparse PLS sera plus efficace dans le cadre d'une étude sans EDC ou avec un ensemble d'apprentissage sans données extrêmes. L'ensemble de validation V est composé de taureaux avec des EDC homogènes. Pour valider les modèles de prédiction, la corrélation pondérée entre DYD prédits et DYD observés est calculée. Cependant, la corrélation simple entre DYD observés et DYD prédits (sans prendre en compte les EDC de l'ensemble de validation) ou la corrélation pondérée par les EDC donne des valeurs très proches car il n'y a pas de valeurs extrêmes dans les taureaux de validation.

L'ensemble A+ est composé des taureaux avec les EDC les plus forts (environ 350 animaux). Les conclusions faites jusqu'à présent pour la régression PLS et la sparse PLS étant identiques, une seule méthode (la régression PLS) a été utilisée pour l'étude de l'ensemble A+ car elle est plus rapidement paramétrable. L'objectif de cette analyse est d'étudier de façon plus approfondie l'impact des animaux à forts EDC dans l'établissement du modèle de prédiction : les équations de prédiction et les critères de validation (corrélations et pentes de régression) sont donc pondérés. Dans un premier temps, les taureaux de l'ensemble de validation ne sont pas considérés. L'ensemble A+ est utilisé en apprentissage (et donc A- en validation) puis en validation (et A- en apprentissage). Le tableau 4.3 présente les résultats de cette étude.

Comme précédemment évoqué, quand on intègre des taureaux de l'ensemble A+ en apprentissage, le nombre de dimensions devient très élevé. Utiliser les taureaux les plus anciens comme ensemble de validation revient à évaluer certains taureaux de l'ensemble de validation à partir des génotypes de quelques-uns de leurs fils. Il n'est donc pas surprenant d'obtenir de bonnes corrélations dans ce cas-là, voire très bonnes pour les caractères lait, MP et TB (>0,8). Même pour la fertilité,

la corrélation est pratiquement doublée en utilisant A- en apprentissage plutôt que A ou A+. La fertilité est un caractère peu héritable ce qui signifie qu'il est difficile à sélectionner. De plus, on utilise pour représenter le phénotype fertilité, des DYD : ce sont des performances corrigées donc si le modèle de correction de ces performances n'est pas adapté aux performances brutes, cela peut avoir un impact sur le modèle de prédiction génomique. Il est donc étonnant d'obtenir des estimations génomiques aussi précises (0,65) pour ce caractère.

Tableau 4.3 : Corrélations (ρ) et pentes de régression (b) entre DYD observés et DYD estimés à partir d'un modèle PLS ou sparse PLS à H dimensions construit sur A+ (et validé sur A-) ou construit sur A- (et validé sur A+)

	A+ en apprentissage			**A+ en validation**		
	ρ	***b***	***H***	***ρ***	***b***	***H***
Lait	0,74	1,13	50	0,87	0,74	7
MG	0,57	1,10	30	0,69	0,59	29
MP	0,79	1,26	38	0,88	0,72	7
TB	0,55	0,88	50	0,81	0,90	26
TP	0,48	0,86	44	0,71	0,73	22
Fer	0,35	0,78	21	0,65	0,74	6

En revanche, quand on utilise A+ en apprentissage, les pentes sont améliorées par rapport à l'étude sur A (plus proches de 1). Pour les caractères MG et fertilité, les corrélations passent de 0,58 et 0,33 sur A à 0,57 et 0,35 sur A+ et ne présentent donc pas de différence significative bien que la taille de la population passe de 2 900 animaux environ à 350. Les corrélations pour le TB et le TP (0,70 et 0,71 sur A, 0,55 et 0,48 sur A+) sont très affectées par la forte diminution du nombre de taureaux constituant l'ensemble d'apprentissage : en effet, A+ compte moins de 400 taureaux contre environ 2 900 dans A. La quantité de lait et de matière protéique sont nettement mieux prédites si on considère une population d'apprentissage constituée des taureaux de l'ensemble A+ que des taureaux de l'ensemble A (+0,21 et +0,24 respectivement). Le nombre très limité des taureaux qui composent l'ensemble A+ n'est pas une contrainte pour l'évaluation de ces deux caractères. Cela est probablement dû au fait que la sélection des reproducteurs en race laitière a

d'abord reposé sur ces deux caractères avant de s'étendre à un ensemble plus large de caractères. L'étude de l'ensemble A+ a mis en évidence que le lien de parenté entre les populations d'apprentissage et de validation a un impact direct sur la qualité des prédictions des DYD. L'ensemble A+ est composé des pères des taureaux des générations suivantes : ils ont de forts EDC c'est-à-dire qu'ils sont fortement diffusés. Utiliser A+ en ensemble de validation et donc, les fils des taureaux de l'ensemble A+ en apprentissage, revient à estimer la valeur génétique des animaux sur l'information apportée par leurs fils. Il est donc logique d'obtenir de très bonnes corrélations.

Enfin, dans le but d'évaluer la régression PLS d'un point de vue strictement statistique, les taureaux des ensembles A- et V (environ 3 500 taureaux) ont été réunis : les taureaux les plus diffusés (ensemble A+) sont éliminés dans le cadre de cette étude. L'objectif est de tester les capacités prédictives de la régression PLS dans un contexte général, sans considérer de structure familiale ni de poids (EDC) associés aux individus et pour travailler sur un ensemble de données le plus homogène possible. Dix ensembles d'apprentissage ont été tirés au sort (75% des données soit environ 2 625 taureaux) pour construire un modèle PLS non pondéré. Puis les dix modèles construits ont été appliqués sur l'ensemble restant des taureaux (25% des données soit environ 875 taureaux) pour tester la capacité prédictive de la méthode utilisée.

Tableau 4.4 : Moyenne et intervalle des valeurs des corrélations non pondérées (ρ) et pentes de régression (b) entre DYD observés et DYD prédits à partir d'un modèle PLS non pondéré sur 10 tirages au sort des ensembles d'apprentissage et de validation

	ρ	***b***	***H***
Lait	0,79 [0,77;0,82]	0,95 [0,89;0,99]	8
MG	0,69 [0,67;0,72]	0,86 [0,81;0,94]	11
MP	0,82 [0,81;0,84]	0,96 [0,92;1,00]	7
TB	0,75 [0,72;0,77]	0,92 [0,84;0,97]	23
TP	0,68 [0,66;0,70]	0,83 [0,80;0,90]	18
Fer	0,44 [0,39;0,56]	0,67 [0,58;0,75]	6

Le tableau 4.4 présente la moyenne sur les dix tirages de la corrélation non pondérée ρ, la pente de régression b et le nombre de variables latentes H ainsi que l'intervalle des valeurs obtenues en corrélations et pentes. Le nombre de variables latentes construites étant très stable d'un tirage à un autre (même identiques pour la plupart des tirages), l'intervalle des valeurs de H n'apporte aucune information complémentaire et ne figure donc pas dans ce tableau. L'amplitude des valeurs obtenues est de 0,03 pour le caractère MP jusqu'à 0,17 pour la fertilité, pour les corrélations et de 0,08 pour le caractère MP jusqu'à 0,17 pour la fertilité, pour les pentes. On obtient en moyenne, de très bonnes pentes sur les cinq caractères de production et des corrélations proches ou supérieures à 0,7. L'amplitude des valeurs obtenues sur les dix tirages est grande en fertilité : la corrélation et la pente moyennes sont donc plus faibles que pour les caractères de production. Ainsi, la régression PLS est une méthode de prédiction fiable, peu sensible au changement de population d'apprentissage pour un ensemble de données homogènes et sans valeurs extrêmes.

La régression PLS permet d'obtenir des estimations précises des valeurs génomiques des animaux des différents ensembles de validation sur la plupart des caractères. En revanche, elle reste très sensible à la présence de valeurs extrêmes, aux particularités des caractères étudiés et à la structure des populations considérées. Dans le cadre d'une évaluation génomique grandeur réelle, la population de référence sera composée des animaux les plus anciens, l'objectif étant de prédire la valeur génomique des animaux les plus jeunes à partir de leur génotype aux marqueurs SNP et de l'information généalogique disponible. Il est donc important d'en tenir compte dans la stratégie de validation des méthodes en utilisant les animaux les plus jeunes comme ensemble de validation.

4.3 Application des méthodes de régression PLS et sparse PLS aux données bovines laitières françaises : la race Montbéliarde

Un autre jeu de données mais avec des effectifs plus réduits a ensuite été analysé. Comme mentionné dans la littérature et dans le chapitre 1, les performances des méthodes d'évaluation génomique reposent, en grande partie, sur la taille de la population de référence. Les méthodes PLS et sparse PLS ont donc été

évaluées sur des données bovines laitières de la race Montbéliarde afin de confirmer ou d'infirmer les conclusions obtenues en race Holstein. En race Montbéliarde, la population d'apprentissage A regroupe 950 taureaux et la population de validation rassemble les 222 plus jeunes taureaux génotypés et phénotypés.

Comme pour la Holstein, la première étape des études méthodologiques sur les données Montbéliarde a été de paramétrer les méthodes. Pour cela, les EDC ont été pris en compte dans l'établissement des équations de prédiction et dans la phase de validation comme expliqué dans le chapitre 2. Le nombre de variables latentes construites en PLS et sparse PLS a été choisi de manière à maximiser la corrélation pondérée par les EDC entre DYD prédits et DYD observés dans l'ensemble de validation. Pour choisir le nombre de variables latentes, le critère Q^2 (voir chapitre 2) a également été considéré. Sur les données de janvier 2009, pour les caractères Lait et TB, pour la régression PLS et la sparse PLS, ce critère nous a conduit à conserver un modèle à une seule dimension donc avec un pouvoir prédictif très faible. Le critère RMSEP (erreur de prédiction moyenne) par validation croisée (tirage au sort de dix échantillons de l'ensemble d'apprentissage) a aussi été testé mais les résultats obtenus basés sur ce critère et ceux obtenus en calculant la corrélation entre les DYD prédits et observés de l'ensemble d'apprentissage aboutissent à un même nombre de variables latentes à introduire dans le modèle final. Ainsi, la maximisation de la corrélation pondérée entre DYD observés et prédits dans l'ensemble de validation est apparue comme étant la meilleure stratégie pour le choix du nombre de variables latentes *H* des méthodes PLS et sparse PLS car plus rapide que la validation croisée.

Une fois le nombre de dimensions *H* fixé pour chaque caractère, le nombre de SNP sélectionnés par dimension en sparse PLS a été fixé par validation croisée. Le tableau 4.5 présente les valeurs de la moyenne de l'erreur de prédiction (RMSEP) obtenues pour les différentes sparse PLS testées et pour la régression PLS à *H* fixé.

Les valeurs RMSEP sont similaires d'une sparse PLS à l'autre ; les différences apparaissent aux troisièmes chiffres après la virgule. La sparse PLS pour laquelle la valeur du RMSEP est la meilleure (c'est-à-dire la plus petite), est présentée en gras dans le tableau. Pour les caractères lait, MG et MP, le modèle optimal est celui qui conserve 10% des SNP sur chacune des variables latentes ; pour les caractères de taux TB et TP, c'est celui avec 3% des SNP et pour la fertilité,

le meilleur modèle est celui retenant 5% des SNP. On remarque que sur les données Montbéliarde, pour des caractères de même héritabilité, le même pourcentage de SNP par dimension est choisi en sparse PLS. La régression PLS conduit à des erreurs de prédiction légèrement inférieures pour trois caractères : l'erreur est de 0,69 en PLS et 0,76 et 0,71 en moyenne en sparse PLS pour le TB et le TP respectivement et pour la fertilité elle est de 0,70 pour la régression PLS mais de 0,79 en moyenne en sparse PLS. Ainsi, le critère RMSEP permettant de définir le pourcentage de SNP à conserver par variable latente n'est pas optimal pour l'analyse de données génomiques. En effet, que ce soit sur les données Holstein (voir article 1) ou Montbéliarde, le critère RMSEP varie très peu entre les différentes sparse PLS considérées.

Tableau 4.5 : Valeurs RMSEP pour la régression PLS et pour chaque sparse PLS testée selon le pourcentage de SNP conservés sur chaque dimension

	Sparse PLS										**PLS**
% de SNP sélectionnés par dimension	**0,2**	**0,4**	**0,6**	**0,8**	**1**	**2**	**3**	**4**	**5**	**10**	**100**
Lait	0,23	0,22	0,22	0,21	0,21	0,21	0,23	0,21	0,22	**0,21**	0,21
MG	0,22	0,22	0,21	0,21	0,21	0,21	0,21	0,21	0,21	**0,21**	0,20
MP	0,20	0,21	0,20	0,21	0,22	0,21	0,20	0,20	0,20	**0,19**	0,18
TB	0,80	0,79	0,76	0,73	0,79	0,72	**0,72**	0,79	0,76	0,74	0,69
TP	0,72	0,76	0,77	0,77	0,76	0,74	**0,71**	0,75	0,75	0,72	0,69
Fer	0,84	0,84	0,82	0,80	0,80	0,75	0,78	0,75	**0,74**	0,77	0,70

Il semble très difficile de fixer les deux paramètres (% de SNP sélectionnées par variable latente et nombre de variables latentes à introduire dans le modèle) de la sparse PLS, indépendamment l'un de l'autre. En effet, si on augmente le nombre de dimensions en fixant le nombre de SNP par dimension ou si on augmente le nombre de SNP sélectionné sur un nombre fixe de dimensions, on obtiendra une fiabilité des modèles semblable. Finalement, la corrélation entre DYD prédits et DYD observés dans l'ensemble d'apprentissage pour le nombre de dimensions (pour la

régression PLS et sparse PLS), et la RMSEP pour le nombre de SNP (pour la sparse PLS) ont été choisies afin d'automatiser la construction des modèles.

Les modèles optimaux de régression PLS et sparse PLS ont ensuite été appliqués sur l'ensemble de validation afin d'estimer la capacité prédictive des méthodes. Les résultats présentés dans le tableau 4.6 montrent qu'il n'y a pas de différence significative entre les deux méthodes en termes de corrélation (confirmé par le test de Hotelling-Williams). La pente de régression est plus proche de 1 pour la régression PLS que pour la sparse PLS sur tous les caractères. La sparse PLS réalise une sélection de variables assez faible sur les caractères de production (plus de 70% de SNP conservés sur le lait, le MG et le MP et plus de 35% pour le TB et le TP) dû au nombre élevé de dimensions conservées dans les modèles (de 20 à 35 variables latentes selon le caractère).

Tableau 4.6 : Corrélations (ρ) et pentes de régression (b) entre DYD observés et DYD estimés à partir d'un modèle PLS ou sparse PLS à H dimensions

		Lait	**MG**	**MP**	**TB**	**TP**	**Fer**
PLS	***ρ***	0,44	0,50	0,46	0,54	0,43	0,43
	b	0,64	0,79	0,70	0,98	0,65	1,79
	H	39	37	35	24	49	7
sPLS	***ρ***	0,38	0,47	0,41	0,56	0,35	0,43
	b	0,63	0,75	0,59	0,81	0,49	2,27
	H	24	35	33	20	28	2
	Nb SNP	28 837	30 389	30 212	14 447	16 822	3 808

Pour avoir une idée plus précise de la capacité prédictive des modèles de prédiction, il aurait été intéressant de disposer de trois ensembles de données différents. Un premier ensemble scindé en deux sous-populations : une population d'apprentissage et une population de validation pour étalonner les modèles afin d'établir les équations de prédiction. Et un second ensemble de test, indépendant, permettant d'évaluer la qualité prédictive des modèles établis. Cependant, dans le cadre des évaluations génomiques, on cherche à évaluer de jeunes taureaux à partir de leurs données génomiques et des informations de parenté de leurs ascendants et

collatéraux. Cette pratique statistique n'est donc pas applicable ni judicieuse dans ce contexte.

Le nombre de dimensions est très élevé quel que soit la méthode utilisée, pour les cinq caractères de production ce qui pourrait traduire, si on s'en réfère aux résultats en race Holstein, un trop fort impact des EDC sur la modélisation. La même démarche que celle décrite pour l'étude des données Holstein a alors été suivie. La distribution des EDC (résultats non montrés) est très semblable à celle obtenue en race Holstein : un groupe de taureaux se détache des autres par leurs forts EDC.

Dans un premier temps, une étude sans EDC a été réalisée et a apporté les mêmes conclusions que pour la race Holstein : entre A avec EDC et A sans EDC, les résultats mesurant la capacité prédictive des modèles sont proches mais la complexité des modèles est moindre dans l'étude sans EDC ce qui implique des graphes des coefficients VIP plus clairs et une meilleure sélection des SNP par sparse PLS (résultats non montrés).

Puis deux nouveaux groupes d'apprentissage ont été créés selon la valeur des EDC des taureaux. Pour la race Montbéliarde, le seuil choisi pour créer les groupes A+ (groupe de taureaux ayant de forts EDC) et A- (groupe de taureaux ayant des EDC plus homogènes) est de 40 pour l'analyse du caractère lait (40 pour les matières MG et MP, 25 pour les taux TB et TP et 100 pour la fertilité). L'ensemble A+ compte environ 140 taureaux et l'ensemble A- en compte environ 810.

Tableau 4.7 : Corrélations pondérée (ρ) et pentes de régression (b) entre DYD observés dans l'ensemble de validation et DYD prédits à partir d'un modèle PLS à H dimensions construit à partir de l'ensemble des données A ou A- en race Montbéliarde

PLS	**A**			**A-**		
	ρ	***b***	***H***	***ρ***	***b***	***H***
Lait	0,44	0,64	39	0,31	0,42	16
MG	0,50	0,79	37	0,32	0,47	12
MP	0,46	0,70	35	0,39	0,53	13
TB	0,58	0,98	24	0,40	0,87	8
TP	0,43	0,65	49	0,37	0,62	13
Fer	0,43	1,79	7	0,42	4,22	1

En prenant en compte les EDC pour l'établissement des équations de prédiction et pour la validation des modèles sur l'ensemble V, si on utilise l'ensemble A- comme ensemble d'apprentissage au lieu de l'ensemble A, on réduit très fortement (de plus de la moitié pour lait et MP, et de plus d'un tiers pour MG, TB, TP et fertilité) le nombre de dimensions du modèle (tableau 4.7). Les graphes des coefficients VIP pour la régression PLS et la sparse PLS (non montrés) sur A et sur A- ne mettent en avant certaines zones du génome que sur A-, les graphes obtenus sur A étant très brouillés et semblables à la Holstein sur A. Cependant, et contrairement aux résultats obtenus en race Holstein, l'utilisation d'un groupe réduit de taureaux selon leur EDC impacte la capacité prédictive des modèles : les corrélations et les pentes sont significativement moins bonnes sur l'ensemble de données A- que sur l'ensemble complet A. Le fait de réduire autant la taille de la population de référence diminue fortement la précision des DYD prédits. Ainsi, d'autres valeurs de seuil pourront être testées afin de sélectionner au mieux les taureaux trop influents avant de pouvoir conclure quant à l'impact des EDC en race Montbéliarde.

4.4 Comparaison des régressions PLS et sparse PLS au BLUP et au GBLUP sur les données Montbéliarde

Les résultats de la régression PLS et de la sparse PLS en race Montbéliarde ont été comparés à deux méthodes répandues et classiquement utilisées chez les bovins laitiers : le BLUP et le GBLUP. Le tableau 4.8 présente les corrélations pondérées entre les DYD observés dans l'ensemble de validation V et les DYD prédits par les quatre méthodes. Le modèle de prédiction utilisé pour chaque méthode a été construit sur l'ensemble complet d'apprentissage A en prenant en compte les EDC. Les paramètres de chaque modèle sont fixés de manière à maximiser la corrélation entre DYD observés et DYD prédits dans l'ensemble de validation V.

Comme dans l'étude des données de la race Holstein, les méthodes génomiques (PLS, sPLS et GBLUP) donnent de meilleures corrélations (test de Hotelling-Williams significatif au seuil 5%) donc des estimations des valeurs génomiques plus fiables qu'avec la méthode BLUP sauf pour le caractère fertilité, où on observe des corrélations identiques pour toutes les méthodes, y compris le BLUP.

Dans le prochain chapitre, nous verrons que la supériorité des évaluations génomiques sur les évaluations BLUP basé sur pedigree, était attendue, car cela est souvent observé dans la littérature. Parmi les méthodes utilisant l'information génomique, il semble que la sparse PLS soit un peu moins performante. Cependant, ces différences se révèlent non significatives selon le test de Hotelling-Williams (α = 5%).

Tableau 4.8 : Corrélations entre DYD observés dans l'ensemble de validation et DYD prédits par BLUP, PLS, sPLS et GBLUP en Montbéliarde

	Lait	MG	MP	TB	TP	Fer
BLUP	0,28	0,40	0,27	0,40	0,25	0,43
PLS	0,44	0,50	0,46	0,54	0,43	0,43
sPLS	0,38	0,47	0,41	0,56	0,35	0,43
GBLUP	0,44	0,50	0,46	0,51	0,44	0,43

Les deux méthodes PLS et sparse PLS sont très rapides (moins de 8mn en Montbéliarde et moins de 10mn en Holstein) avec des performances très proches du GBLUP. Le GBLUP nécessite une quinzaine de minutes en Montbéliarde et une heure de calcul pour la plus grande race (Holstein) : l'augmentation des tailles des populations de référence à l'avenir risque donc de représenter un très gros obstacle calculatoire pour cette méthode à moins de mettre au point des algorithmes efficaces dans la résolution de gros système d'équations.

4.5 Présélection des marqueurs SNP

Dans le cadre de ce travail de thèse, le nombre de marqueurs disponibles après la phase de contrôle de qualité des données reste très important et supérieur au nombre d'observations. Malgré notre intérêt pour des méthodes adaptées au « *p>>n* » problème, nous avons cherché à savoir si l'utilisation d'ensembles réduits de SNP permettait d'obtenir de meilleurs prédictions des valeurs génomiques. Croiseau *et al.* (2011) présentent l'étude de l'impact d'une présélection des SNP en comparant la méthode Elastic Net au GBLUP sur le jeu de données du projet AMASGEN d'octobre 2009. La présélection des SNP se fait sur la base des résultats d'une étude de détection de QTL reposant sur des analyses LDLA (Linkage

Disequilibrium Linkage Analysis, Druet *et al.,* 2008 ; Meuwissen et Goddard, 2001). L'analyse d'association et le déséquilibre de liaison entre QTL et marqueurs sont combinés pour tester chaque ensemble (ou haplotypes) de 6 SNP et ainsi, révéler la présence éventuelle d'un QTL. Dans cette approche, on calcule pour chaque haplotype, une valeur **LRT** (Likelihood Ratio Test) correspondant à la statistique du test de rapport de vraisemblance entre les deux modèles avec ou sans QTL à l'haplotype considéré. Pour définir un « pic LRT » qui permet de pointer la présence d'éventuels QTL le long du génome, deux valeurs de seuil LRT ont été utilisées dans cette étude (seuils : 3 et 5) avec des fenêtres de 25 ou 50 SNP. Une fois les pics LRT révélés, on construit des ensembles de ± 25 SNP autour de chaque pic. Pour chaque race (Holstein, Montbéliarde et Normande) et pour chaque caractère (Lait, MG, MP, TB, TP et Fertilité), on obtient 4 sous-ensembles différents de marqueurs (2 seuils LRT et 2 tailles de fenêtres pour définir les pics) qui contiennent quelques centaines voire quelques milliers de marqueurs. Le tableau 4.9 contient le nombre de pics détectés pour chaque combinaison de ces deux paramètres. Un seuil de LRT de 5 pour une taille de fenêtre de 50 est la combinaison de caractère la plus contraignante. Il est donc normal que le nombre de pics LRT détectés dans ce cas-là soit réduit. On remarque également que le nombre de pics est très lié à l'effectif de la race considéré : plus la taille de la population de référence est grande et plus il y aura de pics LRT détecté car l'analyse est plus précise.

Tableau 4.9 : Nombre de pics LRT détectés pour le caractère Lait, selon la taille de la fenêtre de définition des SNP et le seuil LRT pour les trois races étudiées (Croiseau et al., 2011)

	Taille de la fenêtre	**Seuil LRT**	
		3	**5**
Montbéliarde	25	432	265
	50	273	180
Normande	25	363	197
	50	219	142
Holstein	25	481	350
	50	268	204

Les approches Elastic Net et GBLUP ont été appliquées sur chaque sous-ensemble dans le cadre de l'article de Croiseau *et al.* (2011). J'ai, à mon tour, appliqué la méthode PLS sur les mêmes ensembles réduits de SNP, sur les mêmes caractères et sur les trois races présentées par Croiseau *et al.* (2011). La population de référence Normande est composée de 1 218 taureaux soit une taille équivalente à la population de référence Montbéliarde (1 172 taureaux).

Le tableau 4.10 résume les corrélations pondérées par les EDC entre les DYD observés et les DYD prédits par Elastic Net, GBLUP et PLS, avec ou sans présélection de SNP. Sur les différentes corrélations obtenues sur les 4 fichiers de SNP sélectionnés, seule la plus forte corrélation est présentée (c'est-à-dire sur un seul fichier de SNP présélectionnés). L'article de Croiseau *et al.* (2011) donne les conclusions suivantes : l'Elastic Net peut être utilisé comme une méthode de sélection de variables pour réduire le nombre de SNP à introduire dans une évaluation génomique reposant, par exemple, sur une sélection assistée par marqueurs. Nous verrons, par la suite, que cette méthode (appelée SAM génomique) tend à s'imposer dans les évaluations françaises des bovins laitiers. En effet, les corrélations obtenues par l'Elastic Net sont souvent les meilleures (en gras dans le tableau 4.10) c'est-à-dire pour la plupart des caractères et des races étudiés. Cependant, les corrélations obtenues par la régression PLS sont très proches des résultats de l'Elastic Net : on pourrait donc utiliser de manière équivalent l'Elastic Net et la sélection de variables sur les coefficients VIP de la PLS (PLS-VIP, voir chapitre 2) pour réduire le nombre de SNP à considérer dans les évaluations génomiques.

Les corrélations ne sont pas significativement différentes quel que soit la race, la méthode, ou le caractère considéré, que l'on utilise ou pas une présélection des SNP. L'utilisation d'ensembles réduits de SNP n'améliore ni ne détériore la fiabilité des prédictions génomiques. La différence maximale observée sur les 6 caractères entre la corrélation sur l'ensemble complet et la corrélation sur les ensembles réduits de SNP est de ±0,04 en Montbéliarde, ±0,05 en Normande et de ± 0,08 en Holstein. De plus, les trois méthodes utilisées ici sont très sensibles à l'ensemble de SNP utilisé. Il est donc préférable de ne pas présélectionner les marqueurs SNP afin de simplifier les analyses, et d'éviter une étape supplémentaire passant par une étude LDLA (étape relativement complexe et couteuse).

Tableau 4.10 : Corrélations pondérées entre DYD observés et prédits par GBLUP, Elastic Net (EN), et PLS pour les trois races sur l'ensemble complet de SNP (54k) ou après une présélection de SNP (PS)

	GBLUP		EN		PLS	
	54k	**PS**	**54k**	**PS**	**54k**	**PS**
Montbéliarde						
Lait	0,44	0,43	**0,45**	0,42	0,44	0,44
MG	0,50	0,50	0,50	**0,51**	0,50	0,46
MP	0,46	0,47	0,46	0,47	0,46	**0,48**
TB	0,51	0,56	**0,59**	**0,59**	0,54	0,58
TP	**0,44**	0,42	**0,44**	0,42	0,43	0,41
Fer	0,43	0,42	0,47	**0,48**	0,43	0,46
Normande						
Lait	0,34	0,38	0,41	**0,42**	0,35	0,37
MG	0,39	0,38	**0,41**	**0,41**	**0,41**	0,40
MP	0,31	0,33	**0,37**	0,40	0,35	0,36
TB	0,61	0,63	0,71	**0,75**	0,61	0,64
TP	0,50	**0,55**	0,54	0,53	0,47	0,50
Fer	0,27	0,30	**0,31**	**0,31**	0,29	0,30
Holstein						
Lait	0,56	0,56	0,57	0,57	0,53	**0,61**
MG	0,59	0,59	**0,63**	**0,63**	0,58	0,59
MP	0,55	0,54	**0,57**	**0,57**	0,55	0,53
TB	0,72	0,74	**0,80**	0,79	0,70	0,72
TP	0,73	0,73	**0,75**	0,73	0,71	0,70
Fer	**0,35**	0,33	0,33	0,33	0,33	0,33

4.6 Conclusion

L'ensemble des travaux réalisés dans ce chapitre met en évidence que les méthodes utilisant l'information génomique sont supérieures à une méthode basée uniquement sur l'information pedigree (appelé BLUP dans notre étude) sur la plupart

des caractères. Nous verrons par la suite que cette remarque reste vraie sur les méthodes bayésiennes BayesCπ et Lasso bayésien : ce point sera donc discuté dans le prochain chapitre. L'utilisation de données pré-corrigées telles que les DYD ne semble pas optimale pour des caractères peu héritables, pour lesquels la variation phénotypique est principalement liée aux conditions d'élevage et non à la génétique, comme la fertilité. De plus, ce caractère est difficile à mesurer donc la performance brute mesurée reflète mal le caractère « fertilité ». Enfin, l'étape de précorrection des performances, c'est-à-dire le passage d'une performance brute y_i à un DYD_i finit de déformer la nature du caractère fertilité.

La régression PLS et la sparse PLS ont donné des résultats similaires en termes de corrélations et de pente de régression. Cependant, la sparse PLS présente l'avantage de sélectionner un nombre limité de variables, donc de simplifier les modèles et d'accélérer les temps de calcul.

La prise en compte des EDC dans les évaluations génomiques est rarement expliquée dans la littérature. L'utilisation de pseudo-performances sous la forme de DYD implique d'introduire dans les modèles d'évaluation des poids associés à chaque taureau. Cela n'a pas d'effet négatif ou positif sur la fiabilité des GEBV de la régression PLS et de la sparse PLS mais influence grandement la complexité des modèles, l'estimation des effets des SNP et la sélection de variables par sparse PLS. Par exemple, il semblerait que la régression PLS et la sparse PLS introduisent des dimensions supplémentaires pour représenter exclusivement les taureaux à forte influence. L'analyse des profils VIP de modèles PLS à 1, 2 ou 3 dimensions en comparant les résultats sur les ensembles A- (taureaux avec EDC homogènes), A+ (taureaux avec de forts EDC) et A (ensemble d'apprentissage complet) a montré que les premières dimensions construites à partir des ensembles A+ et A sont très similaires. Ainsi, les premières dimensions captureraient uniquement l'information des taureaux ayant de forts EDC. Pour vérifier cette hypothèse, il serait utile de comparer les coefficients associés à chaque SNP des ensembles A et A+, dimension par dimension afin d'identifier la présence ou non de variables latentes strictement identiques sur les deux ensembles. L'intégration des EDC dans les méthodes mérite d'être approfondie si l'on veut pouvoir utiliser les méthodes PLS à d'autres fins que celle de la prédiction des valeurs génomiques des reproducteurs. Par exemple, la sparse PLS pourrait être utilisée pour sélectionner un ensemble réduit de SNP afin

de créer des puces dédiées à coût réduit à utiliser dans les évaluations génomiques. Nous verrons par la suite que l'estimation des effets SNP par régression PLS et sparse PLS s'avère être complémentaire aux méthodes usuelles de détection de QTL.

En race Holstein, la régression PLS a été appliquée sur un ensemble réduit de taureaux (ensemble A+) correspondant en majorité aux pères des autres taureaux de la population de référence. En race Montbéliarde, cet ensemble était trop réduit (140 taureaux) pour permettre les mêmes investigations qu'en race Holstein. Cette étude a permis de mettre en avant l'importance de la structure génétique de la population de référence (liens de parenté entre les animaux de la population d'apprentissage et de la population de validation) sur la fiabilité des estimations génomiques comme d'autres études antérieures à ma thèse. Habier *et al.* (2010) montrent que les méthodes GBLUP et BayesB appliquées à des données de 3 863 taureaux Holstein allemands sont moins efficaces, en termes de corrélations, si les taureaux de l'ensemble de validation sont trop éloignés des taureaux de l'ensemble d'apprentissage. De plus, Bastiaansen *et al.* (2012) montrent qu'une population de référence composée d'animaux fortement liés entre eux permet aussi une bonne réponse à la sélection génomique au long terme. Pour démontrer cela, ils ont simulé 10 générations d'individus sélectionnés, entre autres, par GBLUP ou régression PLS. Il est donc indispensable de renouveler régulièrement les populations de référence afin de garder un lien fort entre les taureaux de la population de référence et les taureaux candidats aux évaluations génomiques. Enfin, Meuwissen (2009) montre que pour pallier le manque d'individus apparentés entre les populations, il est nécessaire d'augmenter la taille de la population de référence.

Ce premier travail sur la sparse PLS a également montré l'intérêt des méthodes de sélection de variables pour aider à l'interprétation des modèles, d'un point de vue biologique. L'avantage de cette méthode est qu'elle sélectionne les variables et en estime les effets en une seule étape. Certains auteurs ont utilisé une étape de présélection de SNP antérieure à l'estimation des valeurs génomiques sur une population simulée de 5 865 individus (Macciotta *et al.*, 2009) ou sur une population de 2 114 taureaux Holstein (Moser *et al.*, 2010). Dans le cas de l'article de Macciotta *et al.* (2009), la présélection des SNP est réalisée par simple régression linéaire puis les GEBV sont estimés par un modèle linéaire mixte de type BLUP. Les

auteurs mettent en avant la pauvreté de leurs résultats par rapport à des méthodes plus complexes reposant sur l'inférence bayésienne (travaux présentés lors du XII congrès QTLMAS). Moser *et al.* (2010) basent leur présélection de SNP sur les estimations des effets des marqueurs par régression PLS et régression Ridge. Cependant, le gain de corrélation entre DYD observés et prédits n'est que de 5% à 6% par rapport à une sélection de 3 000 SNP répartis uniformément sur le génome. Ces résultats sont donc assez décevants. Notre étude d'une présélection de SNP par détection de pics LDLA, présentée par Croiseau *et al.* (2011) et dans le paragraphe 4.5, conclut que cette méthode n'apporte pas d'augmentation de la corrélation entre valeurs vraies et prédites sur les données bovines laitières françaises. La régression sparse PLS et l'Elastic Net donnent des résultats tout aussi précis en une seule étape qu'une régression PLS, un GBLUP ou un Elastic Net sur SNP présélectionnés. Dans le cadre d'une étude utilisant des ensembles réduits de SNP, on préfèrera donc la sparse PLS aux autres méthodes de sélection de variables étudiées jusqu'à présent ou aux méthodes de présélection de variables.

Des méthodes bayésiennes comme BayesB (Meuwissen *et al.*, 2001) et BayesC (Kizilkaya *et al.*, 2010) permettent de sélectionner un sous-ensemble de variables qui ciblent les zones d'intérêt du génome. Elles sont très largement répandues dans la littérature (Hayes *et al.*, 2009b ; Harris *et al.*, 2009 ; Gredler *et al.*, 2009). Il a été montré qu'elles peuvent être plus efficaces que les méthodes GBLUP ou PLS sur certains caractères car très sensibles au nombre de QTL ou à la part de variabilité liée aux QTL qui influencent un caractère. La deuxième étape de ma thèse a donc été de tester sur les données bovines laitières françaises, deux méthodes bayésiennes encore peu répandues : l'approche BayesCπ (Habier *et al*., 2011) et le LASSO bayésien (Park et Casella, 2008).

Chapitre 5 Évaluation génomique et détection de QTL par les méthodes bayésiennes BayesCπ et LASSO bayésien

5.1 Introduction

Les méthodes de sélection de variables telles que la sparse PLS et l'approche BayesB fournissent des modèles de prédiction aussi performants que les approches faisant intervenir un nombre nettement plus élevé de variables explicatives. Elles ont l'avantage de baser leurs estimations sur des modèles plus simples qui peuvent aider à leur interprétation. De plus, les étapes de sélection de variables et d'estimation des effets se font simultanément. Habier *et al.* (2007) montrent qu'elles fournissent également des estimations des effets des marqueurs stables sur le long terme. Meuwissen *et al.* (2001) mettent en avant l'approche BayesB et suggèrent qu'elle peut prédire les valeurs génomiques plus précisément que l'approche BayesA et le GBLUP. D'autres auteurs comparent l'approche BayesB avec le GBLUP et l'approche BayesA et arrivent aux mêmes conclusions : l'approche BayesB présente des capacités prédictives meilleures ou similaires au GBLUP et à l'approche BayesA (Hayes *et al.,* 2009a ; Harris *et al.,* 2009), et supérieures à la régression PLS (Gredler *et al.,* 2009) tout en permettant une sélection de SNP. Il nous a donc semblé naturel dans la recherche d'une méthode de prédiction optimale sur les données bovines laitières françaises, de nous intéresser à une méthode comparable à l'approche BayesB.

Gianola *et al.* (2009) mettent en avant quelques limitations statistiques des approches BayesA et BayesB. Elles résident principalement dans le choix de la distribution *a priori* des variances des effets des marqueurs. La distribution χ^{-2} utilisée comme loi *a priori* dans les méthodes BayesA et BayesB a un nombre très limité de degrés de liberté et repose fortement sur son paramètre d'échelle. Ces auteurs montrent que la distribution conditionnelle *a posteriori* sous ces hypothèses génétiques n'aura qu'un degré de liberté supplémentaire par rapport à la loi *a priori*, indépendamment du nombre d'observations ou de variables explicatives : cela s'oppose au concept d'apprentissage bayésien. Pour surmonter ce problème, et afin de réduire l'influence du paramètre d'échelle, Kizilkaya *et al*. (2010) ont développé

une méthode similaire à la méthode BayesB mais reposant sur l'utilisation de variances des effets SNP homogènes, appelée BayesC. En utilisant un paramètre d'échelle commun à tous les marqueurs, la méthode BayesC permet de prendre en compte la contribution de tous les marqueurs pour l'estimation de sa loi *a posteriori*. Habier *et al.* (2011) révèlent le fait que le paramètre *π,* qui peut être vu comme l'intensité de sélection des variables, est fixé dans ces deux méthodes (à 0 pour le BayesA et arbitrairement, à une valeur strictement positive pour le BayesB) alors qu'il devrait être supposé inconnu. La méthode BayesC*π* corrige les défauts des méthodes bayésiennes BayesA et BayesB, en traitant le paramètre *π* comme inconnu. Ce paramètre est estimé à partir d'une loi *a priori* uniforme entre 0 et 1. Habier *et al.* (2011) utilisent des données simulées pour tester l'efficacité de leur méthode et la comparent aux approches BayesA et BayesB. Leurs résultats montrent que la corrélation entre les valeurs génétiques simulées et les valeurs prédites par l'approche BayesCπ est similaire aux approches BayesA et BayesB mais l'utilisation de l'algorithme de Gibbs dans l'estimation des composantes de la variance génétique du modèle BayesCπ est plus rapide que l'algorithme de Metropolis-Hastings utilisé dans l'approche BayesB. Par comparaison au GBLUP et au BLUP sur pedigree, sur l'ensemble des données simulées du congrès QTLMAS de 2010, la méthode BayesCπ montrent des capacités prédictives supérieures (Sun *et al.,* 2011).

Dans ce chapitre, la méthode BayesCπ a été comparée à une autre méthode d'inférence bayésienne : le LASSO bayésien (Park et Casella, 2008). Il a été montré que le LASSO bayésien présente un avantage sur le LASSO classique en procurant une interprétation plus naturelle des effets des SNP : il impose une loi *a priori* sur leurs variances. Ses capacités prédictives sont comparables au BayesA (Gonzales-Recio et Forni, 2011), et au GBLUP (Legarra *et al.,* 2011). Ces deux méthodes ont été appliquées aux données des deux races bovines laitières françaises présentées précédemment. Nous étudierons non seulement, leurs capacités prédictives par rapport aux autres méthodes de sélection génomique, mais aussi leur intérêt pour la détection d'éventuels QTL.

5.2 Application des méthodes BayesCπ et LASSO bayésien aux données bovines laitières françaises

5.2.1 Résumé de l'article

L'objectif de cette étude (Article 2) est de comparer deux méthodes bayésiennes (BayesCπ et LASSO bayésien) aux méthodes de régression PLS et sparse PLS, aux méthodes BLUP sur pedigree et GBLUP en les appliquant aux données des races Montbéliarde et Holstein et sur un ensemble d'environ 40 000 marqueurs SNP polymorphes. Les deux populations de référence, composée de taureaux génotypés et évalués sur descendance, comportent 3 940 et 1 172 taureaux de races Holstein et Montbéliarde respectivement.

Mon premier travail a été d'étudier la modélisation de l'approche BayesCπ. Deux modèles de BayesCπ ont été testés sur trois caractères (quantité de lait, taux butyreux et fertilité) : le modèle « BayesCπ » dans sa plus simple formulation ne prend en compte que l'information génomique fournie par les marqueurs SNP et le modèle « BayesCπPED » intègre à la fois les effets des marqueurs et les effets polygéniques estimés à partir de l'information du pedigree. Les résultats de mesure de précision du modèle BayesCπPED en termes de corrélations entre les DYD observés dans la population de validation et les DYD prédits, montrent que l'intégration des effets polygéniques (calculés à partir de l'information pedigree) n'améliore pas les estimations génomiques. Le modèle BayesCπ, sans information de parenté, a donc été retenu dans la suite de l'étude. Les composantes de la variance dans le modèle BayesCπ sont estimées à partir de chaînes de Markov MCMC de 200 000 itérations. La convergence des valeurs des variances génétiques et résiduelles est acceptable, quelque soit le caractère considéré et dans les deux races, ce qui indique une bonne modélisation des données. On observe cependant une forte variabilité dans l'estimation du paramètre π, pour le caractère fertilité en races Holstein et Montbéliarde et pour le caractère Lait en race Montbéliarde.

Le modèle BayesCπ ainsi construit a été comparé à l'approche LASSO bayésien et aux méthodes d'évaluation génomique étudiées dans l'article 1. La comparaison des capacités prédictives des différentes méthodes mettent en avant la supériorité des méthodes bayésiennes sur les régressions PLS et sparse PLS et les méthodes basées sur le modèle BLUP, en race Holstein seulement. Toutes les

méthodes donnent des précisions des estimations génomiques semblables en race Montbéliarde. Les deux méthodes bayésiennes ont des performances similaires sur la plupart des caractères et dans les deux races. Les meilleures performances des méthodes bayésiennes pour le caractère taux butyreux sont probablement dues à l'effet du gène bovin DGAT1, identifié pour ce caractère (Grisart *et al.,* 2004).

Pour avoir une meilleure visualisation des régions du génome impliquées dans les différents modèles de prédiction, les effets des marqueurs ont été étudiés. Les positions sur le génome des plus gros effets estimés des marqueurs SNP sont très semblables pour la sparse PLS, le LASSO bayésien et l'approche BayesCπ. Ce résultat pourrait signifier que les approches de sélection génomique sont capables de détecter les zones du génome qui ont une forte influence sur les caractères d'intérêt.

Cet article a été accepté pour publication dans *Journal of Dairy Science* en septembre 2012.

Application of Bayesian Lasso and BayesCπ methods for genomic selection in French Holstein and Montbéliarde breeds. *By Colombani et al.* In recent years, the number of SNP markers covering the entire genome used for genomic selection has been continuously increasing but with a number of observations which remains lower than the number of SNP variables. The challenge in genomic evaluation is to find the best prediction method for the estimation of breeding values in young animals. Bayesian methods are known to be accurate and efficient in a genomic selection context in dairy cattle. The objective of this study was to compare Bayesian methods with other genomic selection approaches in order to assess the advantages of Bayesian methods in French dairy cattle breeds.

BAYESIAN METHODS FOR GENOMIC SELECTION

Application of Bayesian Lasso and BayesCπ methods for genomic selection in French Holstein and Montbéliarde breeds

C. Colombani,[*] A. Legarra,[*] S. Fritz†, F. Guillaume,‡ P. Croiseau,[§] V. Ducrocq,[§] and C. Robert-Granié[*]

[*]INRA, UR631-SAGA, BP 52627, 31326 Castanet-Tolosan Cedex, France

† UNCEIA, 149 rue de Bercy, 75595 Paris, France

‡ Institut de l'Élevage, 149 rue de Bercy, 75595 Paris, France.

[§] INRA, UMR1313-GABI, 78352 Jouy en Josas, France

Corresponding author: Christele.Robert-Granie@toulouse.inra.fr

ABSTRACT

Recently, the amount of available SNP marker data has considerably increased in dairy cattle breeds, both for research purposes and for application in commercial breeding and selection programs. Bayesian methods are currently used in the genomic evaluation of dairy cattle to handle very large sets of explanatory variables with a limited number of observations. In this study, we applied two Bayesian methods, BayesCπ and Bayesian LASSO, to two genotyped and phenotyped reference populations, consisting of 3,940 Holstein bulls and 1,172 Montbéliarde bulls with approximately 40,000 polymorphic SNPs. We compared the accuracy of the Bayesian methods for the prediction of three traits (milk yield, fat content and conception rate) with pedigree-based BLUP, Genomic BLUP, PLS regression and sparse PLS regression, a variable selection PLS variant. The results showed that the correlations between observed and predicted phenotypes were similar in BayesCπ (including or not pedigree information) and Bayesian LASSO for most of the traits and whatever the breed. In the Holstein breed, Bayesian methods led to higher correlations than other approaches for fat content, were similar to Genomic BLUP for Milk Yield, and to Genomic BLUP and PLS for the conception rate. In the Montbéliarde breed, no method dominated the others, except BayesCπ for fat content. The better performances of the Bayesian methods for fat content in Holstein and Montbéliarde breeds are probably due to the effect of DGAT1. The SNPs identified by the BayesCπ, Bayesian LASSO and sparse PLS methods based on their effect on the different traits of interest were located at almost the same position on the genome. As the Bayesian methods resulted in regressions of direct genomic values on daughter trait deviations closer to one than for the other methods tested in this study, Bayesian methods are suggested for genomic evaluations of French dairy cattle.

Key words: genomic selection, Bayesian methods, variable selection, Holstein and Montbéliarde breeds

INTRODUCTION

In recent years, massive amounts of Single Nucleotide Polymorphism (**SNP**) marker data have been made available in dairy cattle for application in selection schemes. In the future, the increase of the density of SNP data will be ensured by the rapid decrease in genotyping costs. However, the number of genotyped and phenotyped animals which constitute reference populations remains limited. Reference populations provide the prediction equations which give genomic estimated breeding values (**GEBV**). GEBV are obtained through the estimation of SNP effects in a context where the number of independent variables (SNP markers) is much larger than the number of individuals of the reference population.

In the literature, several methods were proposed to estimate SNP effects assuming or not a prior distribution of SNP effects (bayesian *vs* frequentist methods). The bayesian methods differentiate in the assumed prior distributions of SNP effects.

For the estimation of SNP effects, Meuwissen et al. (2001) proposed two Bayesian methods, named BayesA and BayesB. The BayesA method assumes that the prior distribution of the SNP effects is a normal distribution with a 0 mean and a different variance for each SNP. The prior distribution of these variances in BayesA is proportional to a scaled inverted chi-square distribution noted $\chi^{-2}(\nu,S)$ with ν degrees of freedom and a scale parameter S. In the BayesB method, a stochastic search variable selection is used, which assumes that only part of the SNP involved provide information about the phenotype. A combination of normal distribution with a 0 mean and a large variance (with probability π) and a distribution with point mass only at zero (with probability $1-\pi$) is assigned to each SNP effect. Both BayesA and BayesB assume a t distribution at the level of SNP effects (Sorensen and Gianola, 2002). Since then, BayesA and BayesB methods have been widely used in animal breeding research.

Several studies have also used a simple BLUP approach (also often referred to as Genomic BLUP –**GBLUP-** or **SNP-BLUP**), as described in Meuwissen et al. (2001) as a reference method, to compare the gain in accuracy with Bayesian methods. BayesA and BayesB were shown to be similar or slightly more reliable than GBLUP in Australian Holstein-Friesian bulls (Hayes et al., 2009) and in the New Zealand reference population (Harris et al., 2009). Using a Fleckvieh reference population (Gredler et al., 2009), BayesB was found to be more accurate for three traits out of four than the BayesA method modified to include a polygenic effect (Hayes, 2009).

The BayesC model (Kizilkaya et al., 2010) differs from BayesB by using a common variance for SNPs with a non-zero effect, instead of a locus-specific variance. This variance is estimated, in contrast to GBLUP where it is supposed as known. Habier et al. (2011) extended the panel of Bayesian methods with BayesCπ, treating the probability π that a SNP marker has an effect as an unknown parameter, which can be estimated. BayesCπ was compared with BayesA and B using simulated and real data from North American Holstein bulls. The results showed that the accuracies of GEBV were similar for the different methods.

The Bayesian **LASSO** (Least Absolute Shrinkage and Selection Operator) method was also used in a genomic evaluation context (de los Campos et al., 2009, Weigel et al., 2009), but these studies did not compare Bayesian LASSO with other genomic selection methods. De los Campos et al. (2009) compared the predictive ability of different Bayesian LASSO models with respect to the choice of prior for the regularization parameters on simulated data; using pedigree information only, marker information only or considering pedigree and marker information jointly, in wheat line data sets and populations of mice. The results showed that a double-exponential prior may be a better choice than a *t* distribution prior (like BayesA) if most markers do not have any effect. They outlined that a t-distribution may place more density at zero than the Gaussian prior of standard Bayesian methods, the

density at zero is larger in the double-exponential. The Bayesian LASSO appears to be an interesting alternative to the BayesA method for performing regressions on markers. They also have shown, on real data sets that the model with both a polygenic and SNP effect was the most efficient. Legarra et al. (2011) and Ostersen et al. (2011) showed that Bayesian LASSO and GBLUP gave comparable results for most traits, on real data sets of Montbéliarde and Holstein bulls, and on Danish Duroc pigs, respectively.

Gredler et al. (2009) used Partial Least Squares regression (**PLS**) on a Fleckvieh reference population and they compared it with BayesA, BayesB and GBLUP. The PLS method reduces the dimension of the regression model by building orthogonal linear combinations of markers or components which have a maximal correlation with the trait. PLS and GBLUP gave similar results but with lower accuracies than those obtained with BayesB and higher accuracies than with BayesA. GBLUP was also shown to be similar to PLS for two traits with dairy bull data (Moser et al., 2009) and for three traits with French Lacaune dairy sheep data (Robert-Granié et al., 2011) or very slightly better than PLS, in the Fleckvieh breed (Gredler et al., 2009). Colombani et al. (2010) have shown that PLS and sparse PLS (method performing variable selection in addition to reducing dimensionality) provided the same correlations than GLUP for four traits with French Holstein bulls.

The main goal of this study was to compare BayesCπ and Bayesian LASSO to methods currently used in dairy cattle evaluation, such as pedigree-based BLUP and GBLUP, and methods recently used in genomic selection by Long et al. (2011) and Colombani et al. (2012) and known to perform well with large data sets (Le Cao et al., 2008; Chun and Keles, 2009) such as PLS and sparse PLS. After studying the statistical modeling and the convergence properties of BayesCπ, we compared the different methods, based on their predictive abilities, using two real data sets from Montbéliarde and Holstein breeds. Then, the positions of SNPs selected by BayesCπ, Bayesian LASSO and sparse PLS were compared.

MATERIAL AND METHODS

Data

Data sets consisted of 1,172 Montbéliarde bulls and 3,940 French Holstein bulls, progeny tested and genotyped with the Illumina Bovine SNP50K Beadchip. Training and validation data sets were defined for each breed, according to a cut-off birth date defined so that the validation set included the youngest 25% genotyped bulls. Consequently, two training data sets consisting of 950 Montbéliarde bulls and 2,976 Holstein bulls were available to provide prediction equations for both breeds and the three traits (milk yield, fat content and conception rate). Next, the phenotypes of the 222 Montbéliarde bulls and 964 Holstein bulls from the validation data sets, born between June 2002 and 2004, were predicted. Pedigree files for the two breeds included 4,717 and 12,142 bulls in the Montbéliarde and Holstein breeds, respectively.

The DualPHASE software (Druet and Georges, 2009) was used to check mendelian segregation and infer missing genotypes from large-family information. Minimum minor allele frequencies of 3% were required and resulted in 38,462 SNPs for the Montbéliarde breed and 39,738 SNPs for the Holstein breed, used as independent variables.

The response variables (phenotypes) were Daughter Yield Deviations (**DYD**, VanRaden and Wiggans, 1991; Mrode and Swanson, 2004) from the October 2009 national evaluation. The precision of DYD was accounted for through the weighting of DYD by their error variance, which is a function of the sire's Effective Daughter Contribution (**EDC**). Three traits were considered with different heritabilities h^2: milk yield ($h^2 = 0.3$), fat content ($h^2 = 0.5$) and conception rate ($h^2 = 0.02$) (Boichard and Manfredi, 1994).

Genomic BLUP

Two methods were used as reference methods to assess the predictive ability of PLS, sparse PLS, Bayesian LASSO and BayesCπ: pedigree-based BLUP and GBLUP. The general statistical model was: $\mathbf{y} = \mu\mathbf{1} + \mathbf{Za} + \mathbf{e}$ where **y** is a vector of phenotypes (DYD), μ is the overall mean, **Z** is a design matrix allocating observations to breeding values, **a** is a random vector of additive genetic values and **e** is a vector of random normal errors. In pedigree-based BLUP, $Var(\mathbf{a}) = \mathbf{A}\sigma_a^2$ where **A** is the pedigree-based relationship matrix and σ_a^2 is the additive genetic variance. In GBLUP, genomic information was included in the BLUP model assuming a prior normal distribution for SNP markers (VanRaden, 2008) and using mixed model equations with a genomic relationship matrix (Cole et al., 2009; VanRaden et al., 2009). Using genomic information implies that the relationship matrix **A** based on pedigree is substituted by **G** the genomic relationship matrix as defined by VanRaden (2008). We assumed that $Var(\mathbf{a}) = \mathbf{G}\sigma_a^2$ with

$$\mathbf{G} = \frac{\mathbf{XX}'}{2\sum_{j=1}^{p} q_j\left(1-q_j\right)}$$

where p is the number of loci considered, q_j is the frequency of an allele of the marker j and **X** is an centered incidence matrix of SNP effects, corrected for allele frequencies.

PLS and sparse PLS regressions

PLS. The PLS regression (Wold, 1966) is a dimension reduction method developed to deal with the "$p>>n$" problem (the number of predictors p is much larger than the number of observations n). It combines principal component analysis and multiple regressions to handle very large sets of independent variables, which can be highly correlated, such as our set of SNP predictors.

PLS regression relies on successive regressions of the response variable **y** onto substitutes of the initial independent variables (**X**) named latent variables which define a space of smaller dimension. The latent variables ($\boldsymbol{\xi_1}$,..., $\boldsymbol{\xi_H}$) are linear combinations of the independent variables (**X**) through loadings vectors ($\mathbf{u_1}$,…, $\mathbf{u_H}$), where *H* is the number of latent variables retained in the final PLS model. These parameters are estimated in order to solve the following optimization problem:

$$\max_{\|\mathbf{u_h}\|=1} \operatorname{cov}(\mathbf{X_{h-1}u_h}, \mathbf{y_{h-1}})$$

where $\xi_h = \boldsymbol{X_{h-1}}\mathbf{u}_h$ **;** $\mathbf{X}_h$ and $\mathbf{y_h}$ are the residual matrices of the regression of $\mathbf{X_{h-1}}$ and $\mathbf{y_{h-1}}$ onto $\boldsymbol{\xi_h}$ for each PLS dimension h=1,…,H where $\mathbf{X}_0 = \mathbf{X}$ and $\mathbf{y_0} = \mathbf{y}$.

The parameter *H* can be tuned by cross-validation, as proposed by Chun and Keles (2009) and Coster et al. (2010). Solberg et al. (2009) proposed an alternative to fix the parameter *H* in order to obtain the PLS prediction equation which leads to the highest correlation between observed and predicted phenotypes from the validation data set. Colombani et al. (2012) tested and discussed these two approaches. Following their approach, 39, 24, and 7 latent variables provided optimal models for milk yield, fat content and conception rate, respectively, in the Montbéliarde data set. In the Holstein data set, 42, 83, and 29 latent variables were included in the best PLS models for milk yield, fat content and conception rate, respectively.

Sparse PLS. Sparse PLS regression (**sPLS**), developed by Lê Cao et al., (2008) and later by Chun and Keles (2009), differs from PLS regression by adding a step of variable selection to each latent variable through the loading vectors ($\mathbf{u_1}$, ,…, $\mathbf{u_H}$). The sparsity of the loading vectors is introduced iteratively by penalizing $\mathbf{u_h}$ with a soft-thresholding penalization, as in sparse **PCA** (Principal Component Analysis; Shen and Huang, 2008). The optimization problem becomes:

$$\max_{\|\mathbf{u_h}\|=1} \operatorname{cov}(\mathbf{X_{h-1}u_h}, \mathbf{y_{h-1}}) + g_\lambda(\mathbf{u_h})$$

where $g_\lambda(x) = sign(x)(|x| - \lambda)_+$ is the soft-thresholding penalty function of the **x** vector, and λ represents the intensity of penalization.

The number of dimensions (*H*) is fixed as in PLS regression: 24, 20, and 2 latent variables in the Montbéliarde breed and 44, 50, and 27 latent variables in the Holstein breed, for milk yield, fat content and conception rate respectively.

The sparsity is set through the choice of the number of selected variables in each dimension. This choice can be made by examining the Root Mean Squared Error of Prediction (**RMSEP**) with K-fold cross-validation (K=10) within the training data set and for each given dimension *h* (Mevik and Cederkvist, 2004):

$$RMSEP = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(\hat{\mathbf{y}}_k - \mathbf{y}_k)}$$

Where $\mathbf{y_k}$ and $\widehat{\mathbf{y_k}}$ are the vectors of observed and predicted DYD, respectively. The adequate number of selected variables is the one which minimizes RMSEP. In the Montbéliarde breed and for milk yield, 10% of the initial number of SNPs were kept in each of the 24 dimensions, which amounts to 28,837 SNPs in the final model. For the other traits, the corresponding figures were 3% for a total number of 14,447 SNPs for fat content and 5% for 3,808 SNPs for conception rate. In the Holstein breed, 4%, 0.8%, and 4% of the initial number of SNPs, that is 22,948 SNPs, 9,832 SNPs, and 20,150 SNPs, were kept for milk yield, fat content, and conception rate respectively (Colombani et al., 2012).

PLS and sPLS were performed with the R package named “mixOmics” (previously named “intergrOmics”, Lê Cao et al., 2009).

Evaluation of SNP effects in PLS and Sparse PLS. After fitting the PLS and sPLS models, we obtained a vector of regression coefficients with respect to the original variables which

could be directly used for prediction. As a result of variable selection in the sPLS model, some of the estimated coefficients were exactly zero. To assess marker contributions, Variable Importance in Projection (**VIP**) coefficients were used to measure the contribution of each SNP x_j in the construction of **y** through latent variables $\boldsymbol{\xi_h}$. A VIP coefficient was defined for each SNP x_j and for a model with H dimensions by:

$$VIP_{Hj} = \sqrt{\frac{p}{\sum_{h=1}^{H} cor^2(\mathbf{y}, \boldsymbol{\xi_h})} \sum_{h=1}^{H} cor^2(\mathbf{y}, \boldsymbol{\xi_h}) w_{hj}^2} \quad \text{with} \quad \sum_{j=1}^{p} VIP_{Hj}^2 = p .$$

The contribution of x_j in the construction of $\boldsymbol{\xi_h}$ was measured by its weight ω_{hj}, provided by PLS or sPLS. Since the mean of squared VIP scores equals 1, the 'greater than one rule' is generally used as a criterion for variable selection.

Bayesian LASSO

In a genomic selection context, Legarra et al. (2011) proposed a general model for Bayesian LASSO equivalent to the original LASSO proposed by Tibshirani (1996) by splitting the sources of variation in a purely residual term (σ_e^2) and variation due to SNP (σ_g^2) . It can be seen as a hierarchical model, in which individual variances for each SNP effect are modeled upon a common exponential distribution. In this study, we applied Bayesian LASSO as defined by Legarra et al. (2011), where it was called BL2Var because it was the most accurate method for prediction and accommodated well major genes in their study, on similar data. The model considered is:

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{Xg} + \mathbf{e} \text{ with } \mathbf{g} \mid \lambda \sim \prod_j \frac{\lambda}{2} \exp\left(-\lambda \left| g_j \right|\right) \text{ and } \mathbf{e} \mid \sigma_e^2 \sim MVN\left(\mathbf{0}, \mathbf{I}\sigma_e^2\right)$$

where **y** is the vector of phenotypes of the n individuals of the training data set, μ is the overall mean, **X** is a (n x p) design matrix which consists of the genotypes of p SNP markers for each of the n individuals, **g={**g_j**}** is the random vector of SNP effect, **e** is a random vector

of residual effects and λ is the "sharpness" parameter. The parameterization of SNP genotypes (elements of **X)** is as in VanRaden (2008): -2q_j, 1-2q_j, and 2-2q_j for the genotypes 00, 01 and 11 respectively where q_j is the allelic frequency of "1". The prior distribution for the residual variance was an inverted chi-square distribution with four degrees of freedom and expectations equal to the value used in the regular genetic evaluation for σ_e^2. Prior for λ was considered vague, being uniform between 0 and 1,000,000.

Bayes Cπ

BayesCπ model: The last method tested in this study was BayesCπ. It derives from the BayesC method (Kizilkaya et al., 2010; Sun et al., 2011). The statistical model was again:

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{Xg} + \mathbf{e}$$

as in the Bayesian LASSO method.

BayesC modifies the BayesB method by replacing the locus-specific variance components by a common effect variance. BayesCπ is equivalent to BayesC model with an unknown fraction π (with uniform (0, 1) prior) of SNPs with a non-zero effect.

The probability π is defined so that the prior distribution for the additive SNP effect is zero, with a probability π and normal with a probability *(1- π)* so that:

$$g_j \left| \pi, \sigma_g^2 = 0 \right. \text{ with probability } \pi \text{ and } g_j \left| \pi, \sigma_g^2 \sim N\left(0, \sigma_g^2\right) \right. \text{ with probability } (1- \pi).$$

The variance σ_g^2 was assumed to have a scaled inverse chi-square prior with ν_g degrees of freedom and scale S_g^2. The marginal prior of $g_j \left| \nu_g, S_g^2 \right.$ was a univariate Student's t-distribution $t(0, \nu_g, S_g^2)$ with a probability *(1- π).* As suggested by Habier et al. (2011), we took $\nu_g = 4.2$; S_g^2 was equal to $E[\sigma_g^2](\nu_g - 2)/\nu_g$ where $E[\sigma_g^2] = \tilde{\sigma}_g^2$ and $\tilde{\sigma}_g^2$ was the variance

of the additive effect for a randomly sampled locus. As defined in the previous LASSO model, the prior distribution for the residual variance was an inverted chi-square distribution.

BayesCπPED model: The BayesCπPED model differs from the previous one (BayesCπ model) by the addition of a polygenic effect, as proposed by Habier et al. (2011). The statistical model became:

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{Zu} + \mathbf{Xg} + \mathbf{e}$$

with the same definitions as previously and where **u** is a random vector of the polygenic effects of all the individuals in the pedigree and **Z** is an incidence matrix of the polygenic effects. The prior of $\mathbf{u}\left|\mathbf{A},\sigma_u^2\right.$ was normal with mean 0 and variance-covariance matrix $\mathbf{A}\sigma_u^2$ with **A** the numerator relationship matrix and σ_u^2 the additive genetic variance not explained by SNPs. The prior distribution for σ_u^2 was an inverted chi-square distribution, as defined by Habier et al. (2011).

Estimation of variance components in Bayesian LASSO and BayesCπ

Markov Chain Monte Carlo (**MCMC**) was used to estimate the posterior distribution of variances and the model parameters: μ, **u** and σ_u^2 (in BayesCπPED model), g_j and σ_g^2, σ_e^2 and π (if unknown). A burn-in period of 20,000 cycles was run before saving results every 50 cycles out of 180,000. The starting value for π was 0.5.

The genetic variance in the population σ_u^2 (estimated using a pedigree based BLUP model) is proportional to the variance of SNP effects σ_g^2 (Gianola et al., 2009).

For the LASSO model, the relation is $\sigma_g^2 = \frac{\sigma_u^2}{2\sum_{j=1}^{p} q_j(1-q_j)}$ and for the two Bayes*C*π models: $\sigma_g^2 = \frac{\sigma_u^2}{(1-\pi)2\sum_{j=1}^{p} q_j(1-q_j)}$ where p is the number of loci considered and q_j is the frequency of an allele of the marker j.

Bayesian LASSO and the two BayesCπ methods were performed using the GS3 software developed by Legarra et al. (2011, http://snp.toulouse.inra.fr/~alegarra).

Comparison of methods

The predictive ability of the different methods was compared by considering the EDC-weighted correlation between the observed DYD and predicted DYD from the validation data sets, and the EDC-weighted regression slopes of observed DYD onto predicted DYD from the validation data sets. Ideally, values near 1 were expected (Meuwissen et al, 2001) indicating that the GEBV are unbiased. Marker contributions for each method were also measured via genetic standard deviation units for Bayesian LASSO and BayesCπ and VIP coefficients for PLS and sPLS.

The Hotelling-Williams procedure was used to test the difference between the correlations of the different methods. It tests the null hypothesis of equality between two dependent correlations that share a variable (Steiger, 1980; VanSickle, 2003). Under the null hypothesis, the statistical test is distributed as t with n-3 degrees of freedom. All the correlations discussed in this study were compared to one another using the Hotelling-Williams test with a 5% threshold.

RESULTS

Considering a polygenic effect in the BayesCπ model

Table 1 shows the correlations between observed DYD and predicted DYD in the validation data set, for both breeds and for the three studied traits. It compares the accuracy of a BayesCπ model with only marker effects (BayesCπ Model) and a BayesCπ model with marker and polygenic effects (BayesCπPED Model). The correlations were not significantly different between the two models considering the Hotelling-Williams test with a threshold of 5% for both breeds and for all of the traits.

Figure 1 presents the regression slopes (*b*) of observed DYD onto predicted DYD from the validation data set for the same BayesCπ models (BayesCπ Model in black and BayesCπPED Model in white), the two breeds (Montbéliarde on the left and Holstein on the right) and the three traits studied (represented on the x axis). A value of 1 was expected and is depicted with a horizontal line. The confidence intervals were represented by a vertical line at each point and were calculated by adding or subtracting two times the standard error. The standard errors were similar for the two models and equal to 0.03, 0.02 and 0.07 for milk yield, fat content and conception rate respectively in the Holstein breed; and they were stronger in the Montbéliarde breed (0.10, 0.07 and 0.20 for milk yield, fat content and conception rate respectively). The confidence intervals should contain 1. This was the case only in the Montbéliarde breed, for fat content BayesCπPED Model (b=0.89) and conception rate BayesCπ Model (b=1.31).

The differences between the slopes of BayesCπ Model and BayesCπPED Model were small for milk yield and fat content: 0.01 and 0.05 respectively in the Montbéliarde breed, and 0.02 and 0.01 respectively in the Holstein breed. The largest increase in the slope for BayesCπPED Model was observed for conception rate: +0.32 in the Montbéliarde breed, i.e. a regression slope closer to 1 for BayesCπ Model (b=1.31), and +0.10 in the Holstein breed

leading to a regression slope closer to 1 for BayesCπPED Model (b=0.82). The best values of slopes were obtained for fat content with values close to 0.9 for both breeds.

There was no evidence of the superiority of BayesCπPED Model over BayesCπ Model either in terms of accuracy or for the regression slope: the simpler model (BayesCπ Model) was retained hereafter.

Estimation of variance components with the MCMC algorithm

Figures 2 and 3 display the posterior density of genetic variance (at the top), residual variance (in the middle), and π (at the bottom) during their estimation with the MCMC algorithm in the Holstein (Figure 2) and Montbéliarde (Figure 3) breeds. The sampling is represented according to 200,000 MCMC iterations with one record every 50 cycles and without the burn-in period of 20,000 iterations.

Visual inspection of figure 2 and the trace plot of the statistical distribution of parameters (genetic and residual variances and π) during their estimation with the MCMC algorithm (results not shown) indicated that the convergence of both genetic and residual variances was almost reached for the three studied traits. The posterior distributions covered narrow intervals that were shorter than those defined by the prior distribution. The mean μ and standard deviation σ for the genetic variance were $\mu_{MY} \approx 396{,}000$, $\sigma_{MY} \approx 26{,}000$ for milk yield, $\mu_{FC} \approx 8$ and $\sigma_{FC} \approx 0.65$ for fat content, and $\mu_{CR} \approx 45$ and $\sigma_{CR} \approx 3.2$ for conception rate. The residual variance gave $\mu_{MY} \approx 3{,}300{,}000$, $\sigma_{MY} \approx 180{,}000$, $\mu_{FC} \approx 28$, $\sigma_{FC} \approx 1.6$, $\mu_{CR} \approx 2{,}600$ and $\sigma_{CR} \approx 116$.The π parameter was quite accurately estimated and was very low for milk yield and fat content (mean of 0.04 and 0.02 and standard deviation of 0.03 and 0.02, respectively). However for the conception rate, the convergence of π was not reached with a mean around 0.5 and a standard deviation of 0.3. Nevertheless, the estimation of both the genetic and residual variances was not affected by the poor estimation of π.

Figure 3 relates to the Montbéliarde breed. The statistical distributions of parameters (results not shown) indicated that the Markov chain was stabilized and appeared constant only for the genetic and residuals variances for the three traits. They were very chaotic for π parameter on milk yield and conception rate. For the genetic variance, $\mu_{MY} \approx 385{,}000$, $\sigma_{MY} \approx 36{,}000$ for milk yield, $\mu_{FC} \approx 5.5$ and $\sigma_{FC} \approx 0.7$ for fat content and $\mu_{CR} \approx 31$ and $\sigma_{CR} \approx 4$ for conception rate. The residual variance gave $\mu_{MY} \approx 1{,}600{,}000$, $\sigma_{MY} \approx 396{,}000$, $\mu_{FC} \approx 10$, $\sigma_{FC} \approx 2.6$, $\mu_{CR} \approx 1{,}900$ and $\sigma_{CR} \approx 165$. The value of π stabilized only for fat content with a value lower than 3%.

MCMC chains were also run in the Montbéliarde breed with 1,000,000 iterations and a burn-in of 50,000 iterations (results not shown). The evolution of the estimation of the different variances and π were similar to those observed with 200,000 chains (Figure 3). The correlations between observed and predicted DYD from the validation data set were almost the same with a maximum difference of ± 0.01 with the model with 200,000 iterations (results not shown). The regression slopes acquired from 1,000,000 iterations were also very close to those obtained with 200,000 iterations. However, the estimation of π fluctuated wildly between chains.

The parameter π was therefore arbitrarily set to 10% for the three traits in the Montbéliarde breed. Figure 4 shows the posterior density of MCMC chains for genetic variance and residual variance in the Montbéliarde breed for milk yield, fat content, and conception rate, with π fixed at 10%. The trace plot of the statistical distribution of genetic and residual variances during their estimation with the MCMC algorithm (results not shown) covered narrow intervals. The convergence of both genetic and residual variances was acceptable for the three traits. The correlations between observed and predicted DYD were exactly the same (ρ=0.44 for milk yield and ρ=0.42 for conception rate). On the contrary, for

fat content, the correlation decreased from 0.63 to 0.58 but the estimation of genetic variance was more stable.

Thereafter, only the BayesCπ model was used without restricting the estimation of π.

Comparison of BayesCπ with other methods

Predictive ability of the different methods. Tables 2 and 3 present the correlations (ρ) between observed and predicted DYD from the validation data, for the different methods, in Holsteins and in Montbéliardes, respectively. All correlations were compared one to another using the Hotelling-Williams test with a threshold of 5%. All the methods that rely on genomic information performed significantly better than pedigree-based BLUP, except as regards to the conception rate (a trait with low heritability) in the Montbéliarde breed for which no significant difference was observed between any of the correlations. It should be noted that such a high correlation for the pedigree-based BLUP method is not consistent with the lower accuracy of classical BLUP evaluations for conception rate, compared to correlations obtained on the other traits (with higher heritability) . This may reflect a high heterogeneity at the genetic level for conception rate among sire families. The correlation of sPLS was not significantly different from that of BLUP for the conception rate in Holsteins. In the Montbéliarde breed, the correlations given by all genomic selection methods were not significantly different, except as regards to the fat content for which GBLUP was significantly less accurate than BayesCπ. Bayesian methods gave the highest correlations, although the difference was non-significant, except for fat content in Holsteins. BayesCπ showed a non-significant advantage over Bayesian LASSO (+0.01 for milk yield and fat content in Holsteins and +0.09 for fat content in Montbéliardes). The better performances of the Bayesian methods for fat content in Holstein and Montbeliarde breeds is probably due to the effect of DGAT1.

Tables 4 and 5 display regression slopes for each of the three traits, in Holsteins and Montbéliardes, respectively. A value close to 1 is expected. In Holsteins (Table 4), standard errors were similar for all methods and equal to 0.03, 0.02 and 0.07 for milk yield, fat content and conception rate respectively. Genomic selection methods provided lower regression slopes than pedigree-based BLUP. Among the genomic selection methods for predicting milk yield and fat content, the Bayesian methods were the most efficient with a small advantage for Bayesian LASSO (+0.01 for milk yield and +0.03 for fat content). PLS methods were the least efficient. As regards to the conception rate, GBLUP gave a slope value close to that obtained with BLUP: 0.78 with GBLUP and 0.80 with BLUP. In Montbéliardes (Table 5), standard errors were similar for all methods and equal to 0.10, 0.07 and 0.20 for milk yield, fat content and conception rate respectively. PLS methods were also shown to be the least efficient methods, except for the prediction of the fat content for which the regression slope for PLS equaled the regression slopes for BLUP and GBLUP (± 0.02 away from 1). The slope obtained with GBLUP for the milk yield was 0.84, that is +0.10 compared to the regression slopes with Bayesian methods and BLUP. Finally, very bad results were obtained for the conception rate in Montbéliardes (1.35 with BayesCπ compared to 2.27 with sPLS) but standard errors were very large.

Estimation of SNP effects. Figures 5 and 6 show the estimation of SNP effects (in genetic standard deviation units) for Bayesian LASSO and BayesCπ, and VIP coefficients for PLS and sPLS. Emphasis was placed on the position of the SNPs with the largest effects so that VIP coefficients and genetic standard deviation units could be compared. All SNPs were represented on the graphs, even SNPs with zero effect. PLS and sPLS gave almost the same positions for important SNPs, so only the results obtained with sPLS are shown. Moreover, the variable selection performed by sPLS allows a simpler interpretation of VIP coefficients (Colombani et al., 2010).

When BayesCπ with Bayesian LASSO were compared, the positions of the most important SNPs were found to be similar for most cases. For milk yield in Holsteins, a genome region on chromosome 5 was particularly highlighted with BayesCπ but represented a very small peak in Bayesian LASSO. However, chromosome 14 stood out strongly with both Bayesian LASSO and BayesCπ and almost the same SNP were selected. For fat content, Bayesian LASSO weighted one SNP particularly on chromosome 14 that resulted in a weaker effect for the other SNP. This one SNP on chromosome 14 was also the most weighted with BayesCπ and sPLS but with a lesser difference between the effect of the first and the second most important SNP. For conception rate, the graphs were similar with BayesCπ and Bayesian LASSO, with a large number of peaks. sPLS provided similar results to BayesCπ: the most important peaks of Bayes Cπ being also strongly highlighted in sPLS. The SNPs with the largest effects were almost the same with sPLS and Bayes Cπ, except for one SNP on chromosome 21 that showed up strongly for the milk yield but only with BayesCπ.

In Montbéliardes, almost identical graphs were observed with BayesCπ and Bayesian LASSO as regards to the position of the peaks and the size of the effects for milk yield and conception rate. However for fat content, Bayesian LASSO identified only chromosome 14 but with a much smaller estimated effect than with BayesCπ, indicating that the shrinkage of SNP effect was greater with Bayesian LASSO than BayesCπ. All peaks detected with BayesCπ were found with sPLS at the same position but with a different ranking for all traits. Moreover, for the milk yield some regions (such as chromosome 7 and 15) were detected as having a strong effect with sPLS but not with BayesCπ. For fat content, the same peaks appeared with both BayesCπ and sPLS but sPLS included more SNPs in its peaks than BayesCπ. For the conception rate, the effect peaks were found at the same positions with all methods but were more accentuated with sPLS than in Bayesian methods.

DISCUSSION

The objective of this study was to analyze the predictive ability of Bayesian LASSO and BayesCπ methods in particular in comparison to other methods used in the genomic evaluation of dairy cattle. Our first step was to explore the BayesCπ method with different settings of the model, considering the inclusion of pedigree information and the handling of the π value. The results of this step of the study using real data (Holstein and Montbéliarde breeds) demonstrated that the addition of a polygenic component to the BayesCπ model or setting the value of π did not, in most cases, improve the correlation between observed phenotypes and GEBV. Therefore, we retained the simplest Bayes Cπ model. The predictive ability of BayesCπ was then compared to another Bayesian method (Bayesian LASSO), BLUP approaches (pedigree-based BLUP and GBLUP), and dimension reduction methods (PLS and its variable selection variant, sparse PLS). Pedigree-based BLUP was less accurate than genomic selection methods but provided better regression slopes. However the different ways of comparing genomic selection methods failed to demonstrate the systematic superiority of BayesCπ or any other approach.

Polygenic effects and genomic selection

The results of the present study show that genomic selection methods are more accurate than pedigree-based BLUP but with a limited gain of accuracy. If there are linkage disequilibrium (**LD**) between SNP and **QTL** (Quantitative Trait Locus) and sufficient records in the reference set to estimate SNP effects accurately, GEBV accuracy is higher than pedigree-based EBV (Meuwissen et al., 2001; Habier et al., 2007). This is due to the fact that the accuracies of GEBV estimated using a genomic model without pedigree information are affected by the genetic relationship among individuals of the reference population. Habier et

al. (2007) demonstrated that SNP markers are able to capture genetic relationships among genotyped animals. Habier et al. (2010) tested also the impact on the accuracy of GEBV of different values of maximum additive-genetic relationship (a_{max}) between bulls in training and validation populations. They showed that the accuracy of GEBV were the highest with a_{max}=0.6, i.e. a strong relationship between training and validation bulls, but the gain of genomic selection over pedigree-based BLUP was lesser than with smaller values of a_{max}. When the relationships between training and validation bulls were high, pedigree-based BLUP performed better and so the gain of genomic selection was smaller. The results of our study are in agreement with the conclusion of Habier et al. (2010). Our training data sets contained sires, full and half sibs of the bulls in the validation set, the relationships between the bulls of our reference population were high so pedigree-based BLUP was quite efficient and the gain of genomic selection methods over pedigree-based BLUP was limited (the mean gain of correlations of GBLUP over BLUP was equal to 0.18 in Holstein and 0.10 in Montbéliarde).

The correlations between observed DYD and predicted DYD in the validation data set were very close with BayesCπ model (including only SNP information) and BayesCπPED model (including both polygenic and SNP effects) whatever the trait or the breed (Table 1). The regression slopes were also very similar for both models (Figure 1). Mrode et al. (2011) presented a study that aimed at testing the inclusion of polygenic effects using 11,480 Holstein Friesian bulls in the United Kingdom in a linear model equivalent to a GBLUP model. They showed that the correlations for production traits decreased slightly but the regression coefficient increased by approximately 0.1 for all traits. Liu et al. (2011) showed that including a polygenic effect in a GBLUP model resulted in decreased correlations between direct genomic values and estimated breeding values. In our study, the inclusion of polygenic effects also led to slightly smaller correlations for fat content in both breeds but

with no significant differences. These results show that SNP marker information could contain a part of pedigree information. In Lacaune dairy sheep, similar conclusions were obtained with a reference population of approximately 2,500 proven rams and 44,000 SNPs (Robert-Granié et al., 2011). Inclusion of infinitesimal effects in the prediction model with BayesCπ had little impact on accuracies and led to slightly better slopes of regressions.

However a difference between BayesCπ model and BayesCπPED model appeared regarding the number of SNPs selected for all traits in Montbéliardes and for the conception rate in Holsteins. The difference between the number of SNPs included in BayesCπ model and BayesCπPED model was about -5,000 SNPs for milk yield, +9,000 SNPs for conception rate in Montbéliardes and +6,000 SNPs for conception rate in Holsteins. However the graphs of the posterior distribution of π during the MCMC algorithm (Figures 2 and 3) showed that it was difficult to obtain a good convergence of π in these cases. So, it was problematic to obtain an accurate estimation of the parameter π in these cases. The final value of π varied around the mean (i.e. 0.5) and the number of SNPs selected fluctuated between 15,000 and 25,000 in these cases. For milk yield and fat content in Holsteins, the number of SNPs selected was stable between the two models: approximately 1,600 SNPs for milk yield and 800 SNPs for fat content. In the Montbéliarde breed, the result was more surprising for fat content because π seemed to be well estimated during the MCMC process (around 0.01 in BayesCπ model but around 0.51 in BayesCπPED model, i.e. a difference of +19,000 SNPs between these two models). The inclusion of polygenic effects (BayesCπPED model) for fat content in Montbéliardes interfered with SNP selection in BayesCπ: the model could not identify the most important SNP for the prediction of this trait although there was no change to the correlation between observed DYD and predicted DYD in the validation set. Several situations were tested in which the weight allocated to the polygenic effects was modified but the correlation between observed and predicted DYD was the same.

Comparison of methods

The results obtained with MCMC chains on the BayesCπ Model were compared with Bayesian LASSO, GBLUP, PLS and sparse PLS regressions. Hayes et al. (2009) and VanRaden et al. (2009) presented two reviews of empirical results in dairy cattle which pointed out the similarity of GBLUP and BayesB, as far as predictive ability is concerned. Croiseau et al (2011) compared the Elastic net approach to GBLUP on the French data sets used in this study. They obtained the same correlations than BayesCπ with better regression slopes. Our results are in good agreement with these studies since the Bayesian methods reached the same accuracies as GBLUP for most traits. The regression slopes did not allow to differentiate between GBLUP and Bayesian methods either. PLS variants were the least efficient, both in regard to correlation and regression coefficients in most cases.

The regression slopes of observed DYD on estimated DYD were less than one in most cases. It is probably due to the fact that the reference populations were made up of strongly selected bulls. Vitezica et al. (2011) proposed that the strong selection of animals in dairy cattle scheme and so, on the animals of the reference populations, may result in the observed biases of the regression coefficients. Biases might also be introduced by the utilization of DYD, as suggested by Patry and Ducrocq (2011).

BayesCπ and Bayesian LASSO seem to follow the same pattern in the present study. The difference between GBLUP and BayesCπ was high for fat content both in Montbéliarde and Holstein breeds. To obtain a good idea of the genome regions involved in the prediction equation of each method we studied the graphs of SNP effects. The genome areas with the largest effects were almost identical with BayesCπ, Bayesian LASSO and sparse PLS whatever the trait or breed. The graphs for conception rate were similar for the Bayesian methods but showed some dissimilarity with sparse PLS: the peaks were found at the same positions but the ranking of these peaks was different. However this could be explained by the

fact that the conception rate seems to be a very polygenic trait and that its low heritability leads to limited prediction accuracy with all methods. One can note that the number of SNPs with non-zero effects is relatively small for fat content and this in both breeds. A specific peak stood out particularly with all methods for the fat content at the beginning of chromosome 14. In Holsteins, this genome region corresponds to the DGAT1 gene (Grisart et al., 2004) which, when mutated, has a major effect on the fat content in milk. Hence, the superior accuracy of BayesCπ and Bayesian LASSO against GBLUP ($\rho_{BAYESC\pi}$ = 0.80, $\rho_{BayesianLASSO}$ = 0.79 but ρ_{GBLUP} = 0.72) for the fat content trait in Holsteins could be explained by the small number of QTLs related to this trait. For the fat content trait in Montbéliardes, BayesCπ outperformed GBLUP ($\rho_{BAYESC\pi}$ = 0.62 and ρ_{GBLUP} = 0.52) but surprisingly Bayesian LASSO was closer to GBLUP than BayesCπ with $\rho_{BayesianLASSO}$ = 0.53. Concerning SNP effects, Bayesian LASSO highlighted only one region on chromosome 14 whereas BayesCπ and sparse PLS highlighted 4 or 5 regions. Therefore the equation of prediction in Bayesian LASSO was based on a very small number of SNPs, all positioned within the same genome region, and this seemed to affect the accuracy of prediction. For milk yield in Holsteins, a few SNPs were retained by BayesCπ (about 1,500 SNPs) but almost 23,000 SNPs were necessary with sparse PLS. This suggests that milk yield is affected by a large number of QTLs. The results of GBLUP and Bayesian methods for milk yield in Holsteins were almost identical.

The conclusions established in this work should be transferable to other studies if the characteristics of the traits studied are considered correctly. Bayesian methods seem to perform well whatever the trait or population but in some cases, GBLUP is as accurate as Bayesian methods. These results were confirmed on the study of three traits in Lacaune sheep breed (Duchemin et al., 2012), indicating the superiority of BayesCπ. Regarding computing time, Bayesian methods are the less efficient with about 12 hours per trait in our rather small

data sets. GBLUP requires the inversion of the genomic relationship matrix for all traits which took about one hour for our data set. Then, once the genomic relationship matrix was inverted, computation was a matter of seconds.

CONCLUSIONS

The first goal of this study was to explore the predictive ability of BayesCπ in a genomic evaluation context. According to the accuracy and regression slope, the inclusion of pedigree information in the BayesCπ model did not change the results, nor did fixing the value of π. BayesCπ did not show a large advantage over other methods except for traits with a small final number of selected SNPs such as fat content. No genomic selection method tested in this study outperformed the others but it is interesting to note that the position of the SNPs selected by the different models (LASSO, BayesCπ, sparse PLS) were close. The next step of our work will be to compare more precisely the SNPs selected by all the methods and to study whether the genome regions highlighted by some genomic selection methods correspond to the QTLs detected by specific QTL detection methods (i.e., Linkage Analysis, Linkage disequilibrium or linkage disequilibrium-linkage analysis).

ACKNOWLEDGEMENTS

This work was supported by the French project AMASGEN (2009-2011), financed by the French National Research Agency (ANR) and ApisGene. Labogena (http://www.labogena.fr/; Jouy-en-Josas, France) is gratefully acknowledged for providing the genotypes. The project was partly supported by the Toulouse Midi-Pyrénées bio-informatics platform genotoul (http://bioinfo.genotoul.fr/ ; Toulouse, France).

Table 1. Correlations between observed DYD and predicted DYD in the validation data set without (BayesCπ model) or with (BayesCπPED model) adding a polygenic component to the BayesCπ model

	Montbéliarde[1]		Holstein[2]	
	BayesCπ model[3]	BayesCπPED model[4]	BayesCπ model[3]	BayesCπPED model[4]
Milk Yield	0.44	0.44	0.57	0.57
Fat %	0.63	0.62	0.80	0.78
Conception Rate	0.42	0.44	0.34	0.34

[1]Montbéliarde: training set = 950 bulls, validation set = 222 bulls
[2]Holstein: training set = 2976 bulls, validation set = 964 bulls
[3]BayesCπ model: $\mathbf{y} = \mu\mathbf{1} + \mathbf{Xg} + \mathbf{e}$
[4]BayesCπPED model: $\mathbf{y} = \mu\mathbf{1} + \mathbf{Zu} + \mathbf{Xg} + \mathbf{e}$

Table 2. Correlations between observed DYD and predicted DYD in the validation data set provided by pedigree-based BLUP (BLUP), Genomic BLUP (GBLUP), PLS, sparse PLS (sPLS), Bayesian LASSO and BayesCπ (BayesCπ model) in the Holstein breed

	BLUP	GBLUP	PLS	sPLS	Bayesian LASSO	BayesCπ
Milk Yield	0.38	0.56	0.53	0.48	0.56	0.57
Fat %	0.44	0.72	0.70	0.66	0.79	0.80
Conception Rate	0.28	0.35	0.33	0.29	0.34	0.34

Table 3. Correlations between observed DYD and predicted DYD in the validation data set provided by pedigree-based BLUP (BLUP), Genomic BLUP (GBLUP), PLS, sparse PLS (sPLS), Bayesian LASSO and BayesCπ (BayesCπ model) in the Montbéliarde breed

	BLUP	GBLUP	PLS	sPLS	Bayesian LASSO	BayesCπ
Milk Yield	0.28	0.42	0.44	0.38	0.44	0.44
Fat %	0.40	0.52	0.58	0.56	0.53	0.62
Conception Rate	0.43	0.47	0.43	0.43	0.43	0.43

Table 4. Regression slopes of observed DYD on predicted DYD in the validation data set provided by pedigree-based BLUP (BLUP), Genomic BLUP (GBLUP), PLS, sparse PLS (sPLS), Bayesian LASSO and BayesCπ (BayesCπ model) in the Holstein breed

	BLUP	GBLUP	PLS	sPLS	Bayesian LASSO	BayesCπ
Milk Yield	0.79	0.68	0.65	0.53	0.74	0.73
Fat %	0.97	0.87	0.80	0.69	0.93	0.90
Conception Rate	0.80	0.78	0.60	0.54	0.72	0.72

Table 5. Regression slopes of observed DYD on predicted DYD in the validation data set provided by pedigree-based BLUP (BLUP), Genomic BLUP (GBLUP), PLS, sparse PLS (sPLS), Bayesian LASSO and BayesCπ (BayesCπ model) in the Montbéliarde breed

	BLUP	GBLUP	PLS	sPLS	Bayesian LASSO	BayesCπ
Milk Yield	0.74	0.84	0.64	0.63	0.74	0.74
Fat %	1.01	1.01	0.98	0.81	0.91	0.85
Conception Rate	1.78	1.76	1.79	2.27	1.36	1.35

Figure 1. Regression slopes of observed DYD on predicted DYD in the validation data set without (BayesCπ model) or with (BayesCπPED model) adding a polygenic component to the BayesCπ model

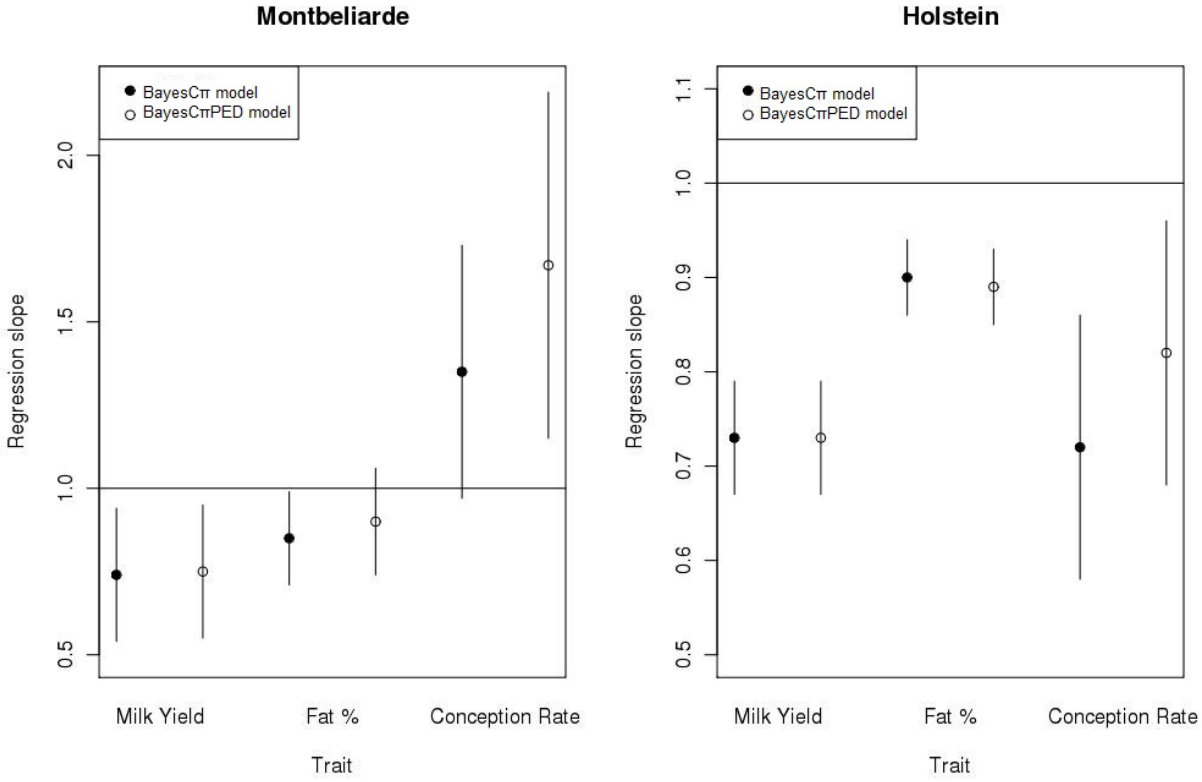
Figure 2. Density of genetic variance, residual variance and π during MCMC algorithm for milk yield, fat content and conception rate in the Holstein breed

Figure 3. Density of genetic variance, residual variance and π during MCMC algorithm for milk yield, fat content and conception rate in the Montbéliarde breed

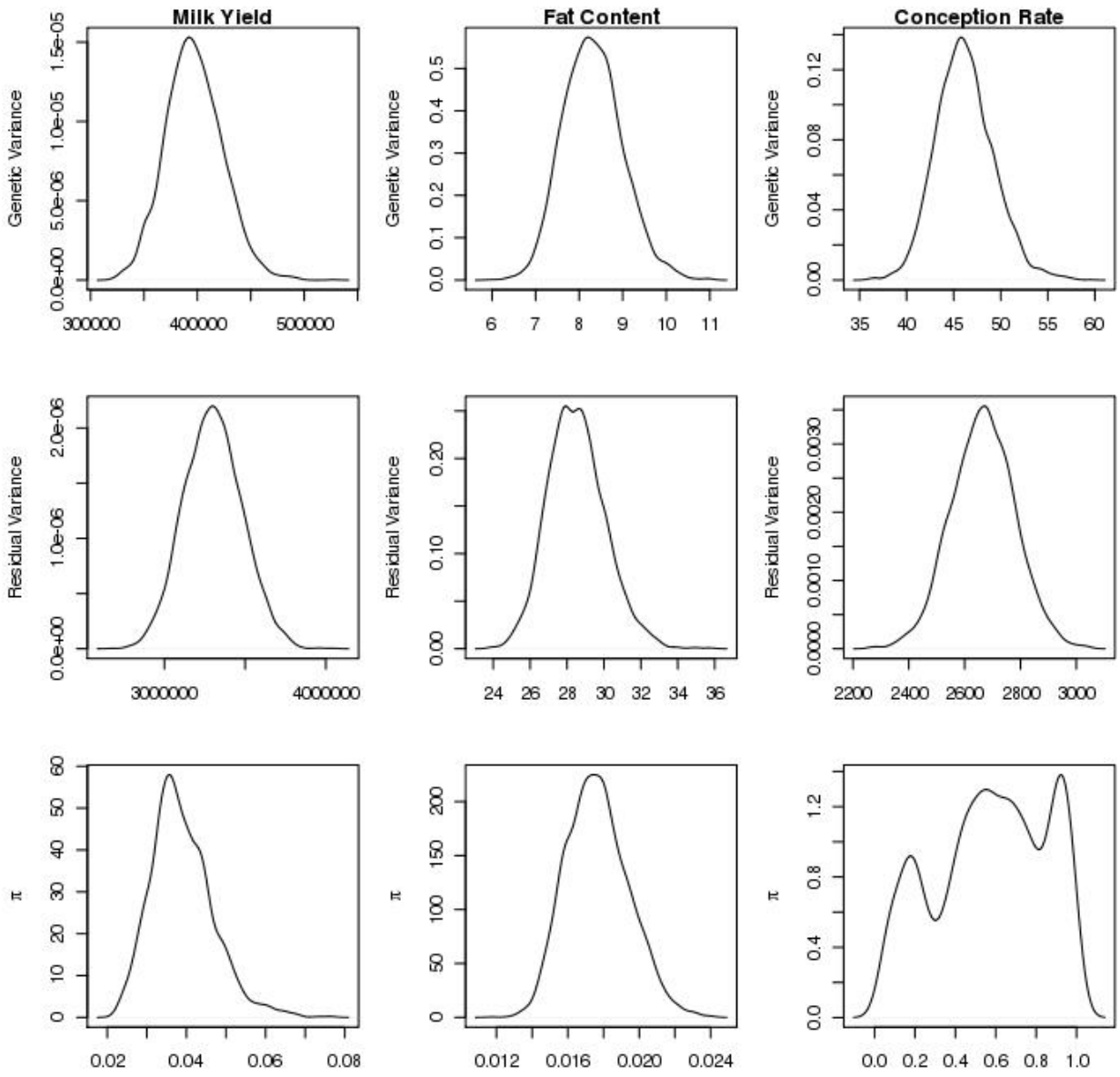
Figure 4. Density of genetic variance and residual variance during MCMC algorithm for milk yield, fat content and conception rate in the Montbéliarde breed with π=10%

Figure 5. Estimation of SNP effects by BayesCπ (BayesCπ model), Bayesian LASSO and VIP coefficients for sparse PLS for milk yield, fat content and conception rate in the Holstein breed according to the marker position along the genome

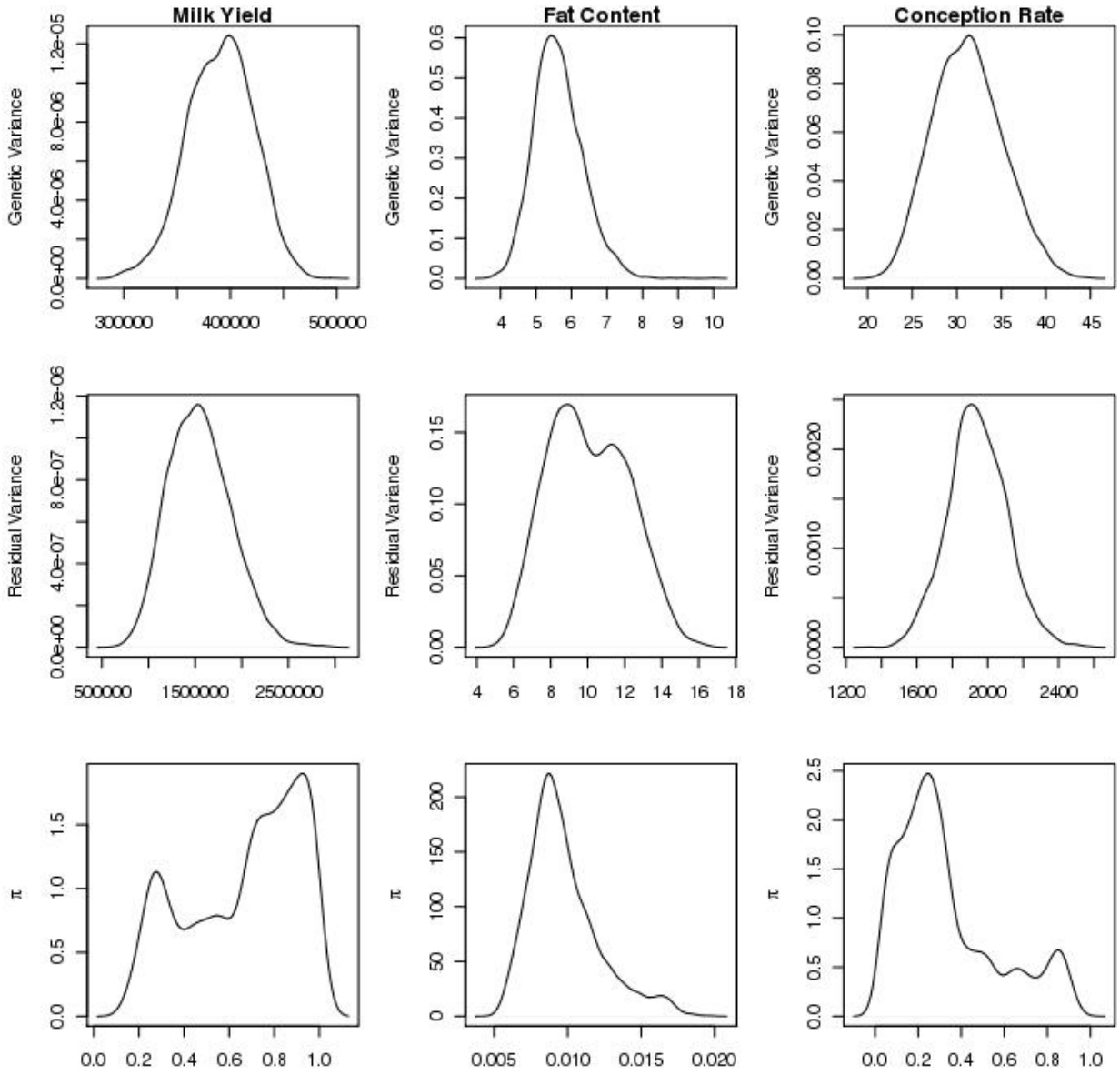
Figure 6. Estimation of SNP effects by BayesCπ (BayesCπ model), Bayesian LASSO and VIP coefficients in sparse PLS for milk yield, fat content and conception rate in the Montbéliarde breed according to the marker position along the genome



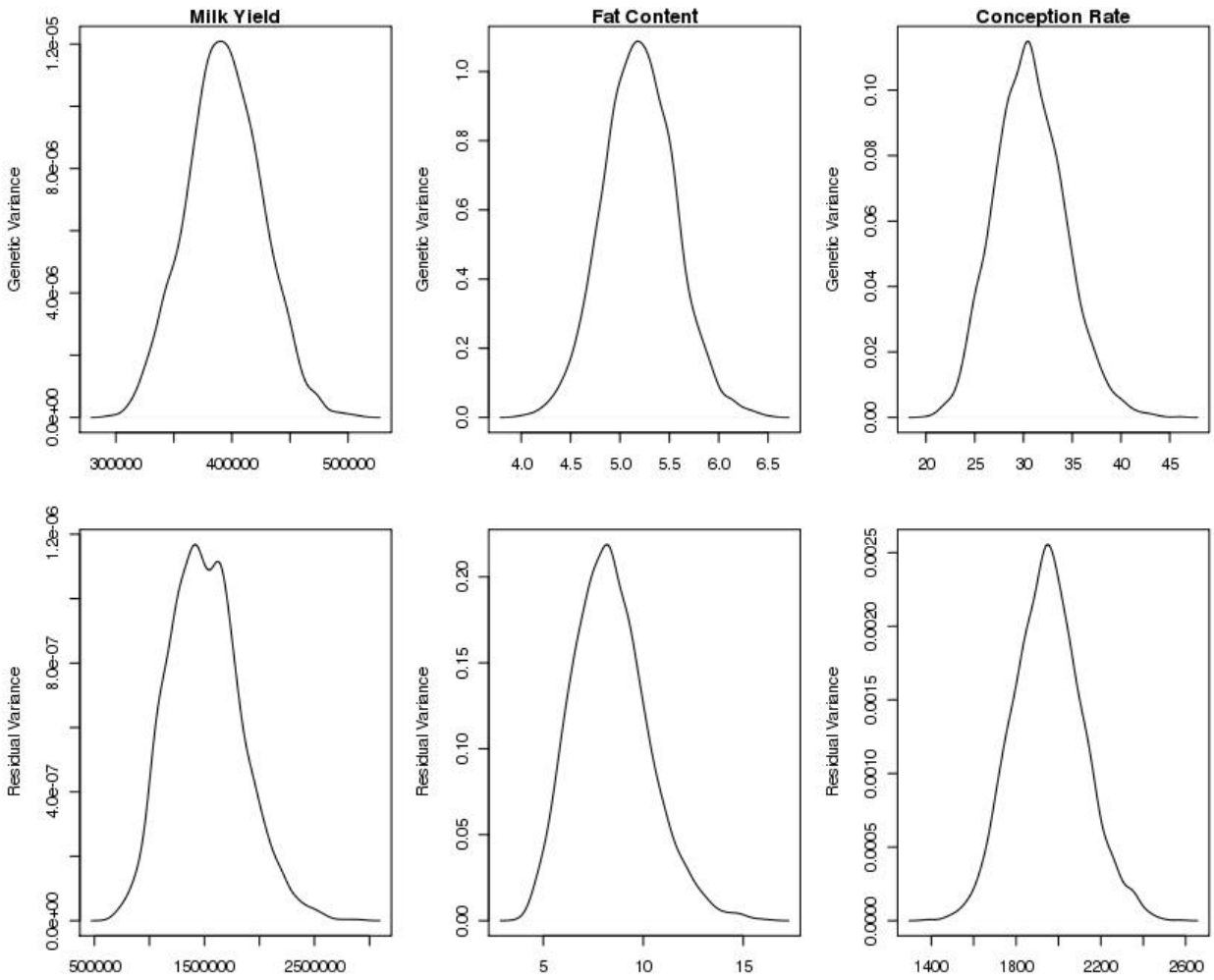
Colombani, Figure 1



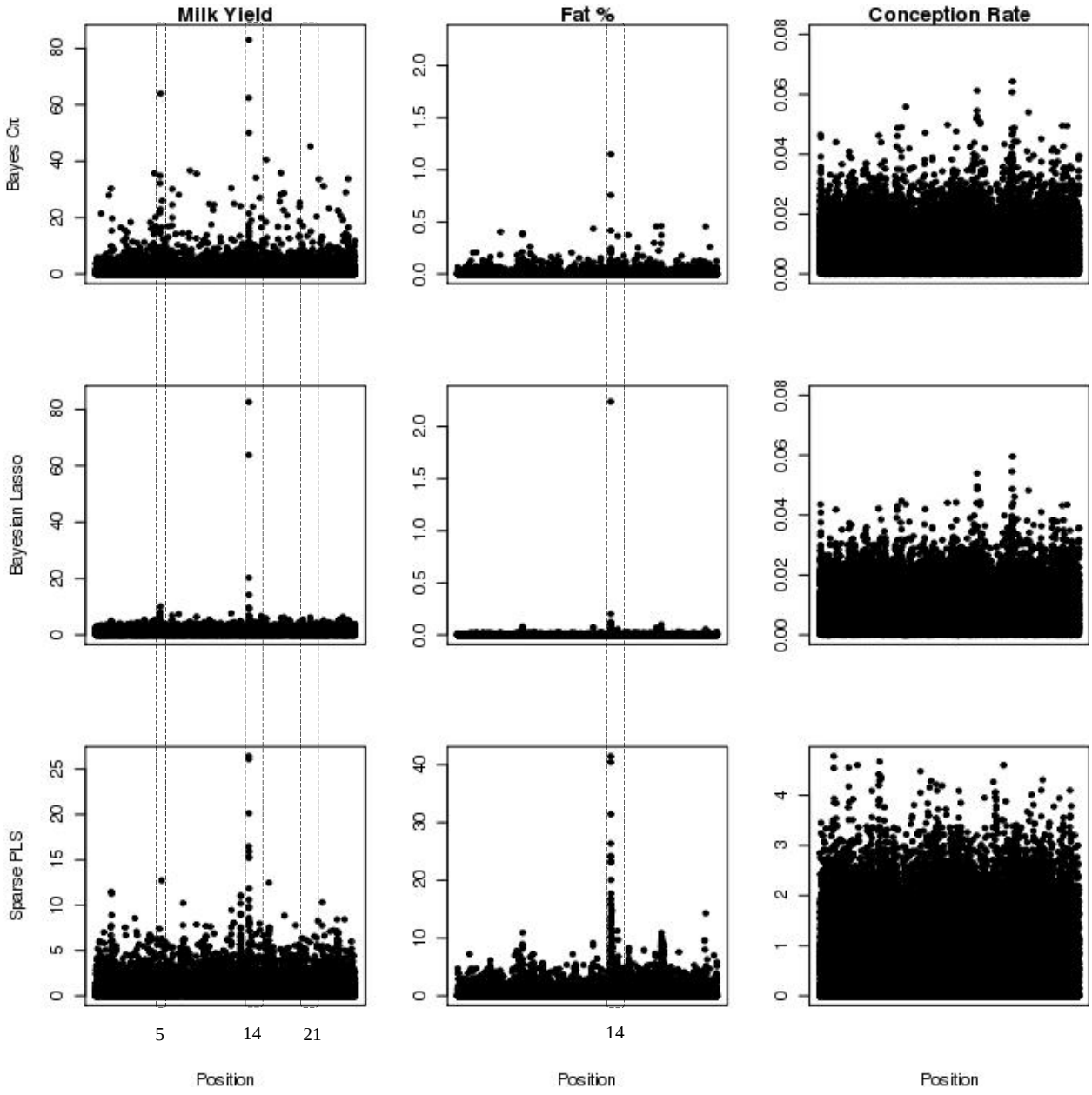
Colombani, Figure 2



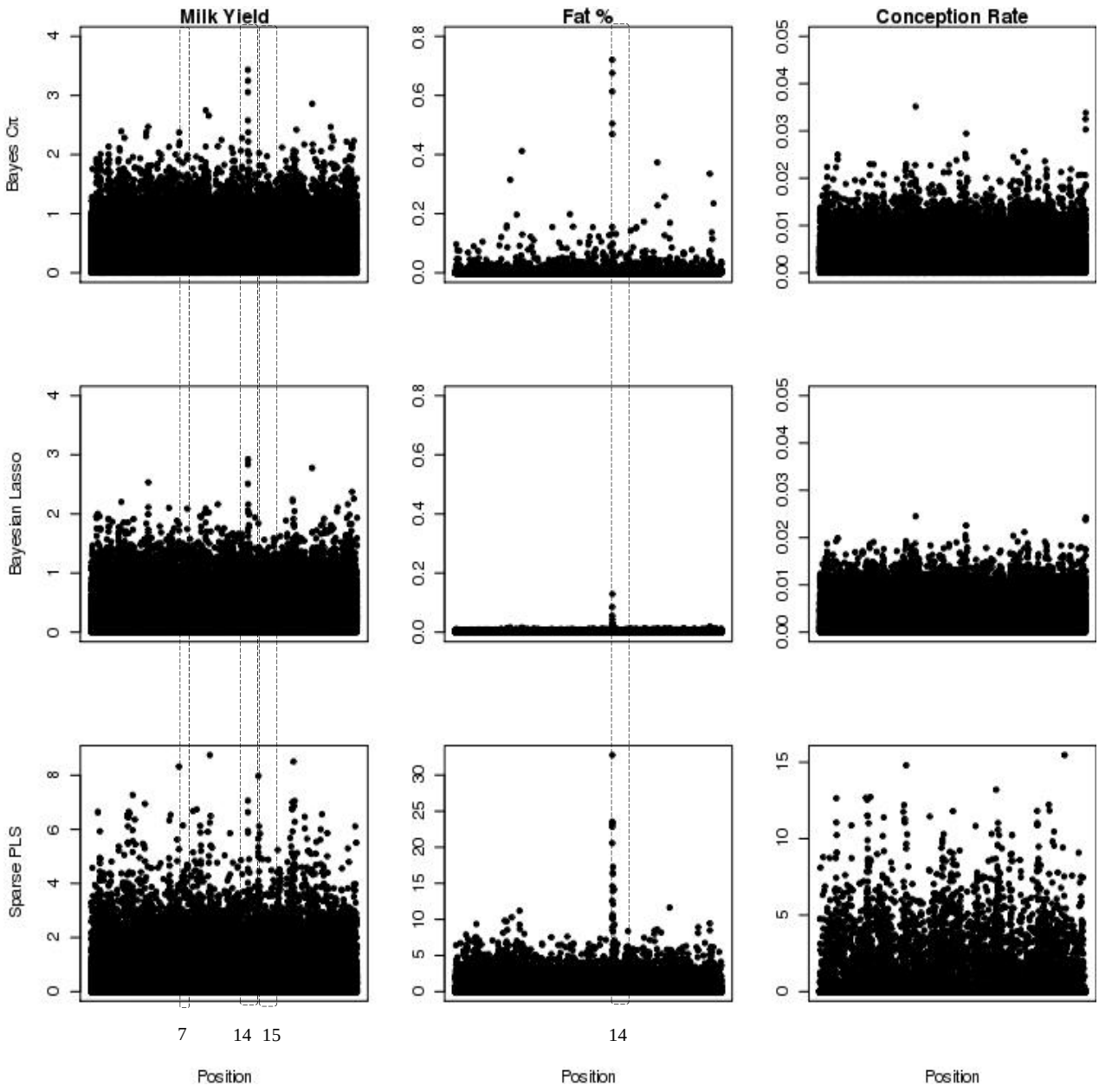
Colombani, Figure 3



Colombani, Figure 4



Colombani, Figure 5



Colombani, Figure 6

REFERENCES

Boichard, D. and E. Manfredi. 1994. Genetic analysis of conception rate in French Holstein cattle. Acta Agriculturae Scandinavica Section a-Animal Science 44(3):138-145.

Chun, H. and S. Keles. 2009. Expression Quantitative Trait Loci Mapping With Multivariate Sparse Partial Least Squares Regression. Genetics 182(1):79-90.

Cole, J. B., P. M. VanRaden, J. R. O'Connell, C. P. Van Tassell, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and G. R. Wiggans. 2009. Distribution and location of genetic effects for dairy traits. J. Dairy Sci. 92(6):2931-2946.

Colombani, C, A. Legarra, P. Croiseau, F. Guillaume, S. Fritz, V. Ducrocq, and C. Robert-Granié. 2010. Application of PLS and Sparse PLS regression in genomic selection. In 9th World congress on Genetics Applied to Livestock Production. Leipzig, Germany.

Colombani, C., P. Croiseau, S. Fritz, F. Guillaume, A. Legarra, V. Ducrocq, and C. Robert-Granié. 2012. A comparison of partial least squares (PLS) and sparse PLS regressions in genomic selection in French dairy cattle. J. Dairy Sci., 95 (4):2120-2131.

Croiseau, P., A. Legarra, F. Guillaume, S. Fritz, A. Baur, C. Colombani, C. Robert-Granié, D. Boichard and V. Ducrocq. 2011. Fine tuning genomic evaluations in dairy cattle through SNP pre-selection with the Elastic-Net algorithm. Genet. Res. 93:409-417.

Duchemin S.I., C. Colombani, A. Legarra, G. Baloche, H. Larroque, et al. 2012. Genomic selection in the French Lacaune dairy sheep breed. J. Dairy Sci. 95:2723–2733.

de los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel, and J. M. Cotes. 2009. Predicting Quantitative Traits With Regression Models for Dense Molecular Markers and Pedigree. Genetics 182(1):375-385.

Druet, T. and M. Georges. 2009. A Hidden Markov Model Combining Linkage and Linkage Disequilibrium Information for Haplotype Reconstruction and Quantitative Trait Locus Fine Mapping. Genetics 184(3):789-U237.

Gianola, D., G. de los Campos, W. G. Hill, E. Manfredi, and R. Fernando. 2009. Additive Genetic Variability and the Bayesian Alphabet. Genetics 183(1):347-363.

Gonzalez-Recio, O. and S. Forni. 2011. Genome-wide prediction of discrete traits using bayesian regressions and machine learning. Genet. Sel. Evol. 43:12.

Gredler, B., K. G. Nirea, T. R. Solberg, C. Egger-Danner, T. H. E. Meuwissen, and J. Solkner. 2009. Genomic Selection in Fleckvieh/Simmental -First results. In proceedings of the Interbull Meeting. Barcelona, Spain, 21-24 August, 2009.

Grisart, B., F. Farnir, L. Karim, N. Cambisano, J. J. Kim, A. Kvasz, M. Mni, P. Simon, J. M. Frere, W. Coppieters, and M. Georges. 2004. Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. Proc. Natl. Acad. Sci. U. S. A. 101(8):2398-2403.

Habier, D., R. L. Fernando, and J. C. M. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177(4):2389-2397.

Habier, D., J. Tetens, F.-R. Seefried, P. Lichtner, and G. Thaller. 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet. Sel. Evol. 42(1):5.

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick. 2011. Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics 12:12.

Harris, B. L., D. L. Johnson, and R. J. Spelman. 2009. Genomic selection in New Zealand and the implications for national genetic evaluation. Pages 325-330 in Identification, Breeding, Production, Health and Recording of Farm Animals. Proceedings of the 36th ICAR Biennial

Session, Niagara Falls, USA, 16-20 June, 2008. International Committee for Animal Recording (ICAR).

Hayes, B. J., A. J. Chamberlain, S. Maceachern, K. Savin, H. McPartlan, I. MacLeod, L. Sethuraman, and M. E. Goddard. 2009. A genome map of divergent artificial selection between Bos taurus dairy cattle and Bos taurus beef cattle. Anim. Genet. 40(2):176-184.

Hayes, B. J. 2009. Genomic selection in the era of the $1000 genome sequence. In Symposium Statistical Genetics of Livestock for the Post-Genomic Era. Wisconsin-Madison, USA.

Kizilkaya, K., R. L. Fernando, and D. J. Garrick. 2010. Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. J. Anim. Sci. 88(2):544-551.

Lê Cao, K. A., I. Gonzalez, and S. Dejean. 2009. integrOmics: an R package to unravel relationships between two omics datasets. Bioinformatics 25(21):2855-2856.

Lê Cao, K. A., D. Rossouw, C. Robert-Granie, and P. Besse. 2008. A Sparse PLS for Variable Selection when Integrating Omics Data. Stat. Appl. Genet. Mol. Biol. 7(1):32.

Legarra, A., C. Robert-Granie, P. Croiseau, F. Guillaume, and S. Fritz. 2011. Improved Lasso for genomic selection. Genet. Res. 93(1):77-87.

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

Long, N., D. Gianola, J. M. Rosa, and K. A. Weigel. 2011. Dimension reduction and variable selection for genomic selection: Application to predicting milk yield in Holsteins. J. Anim. Breed. Genet. 128(4):247-57

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(4):1819-1829.

Mevik, B. H. and H. R. Cederkvist. 2004. Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR). J. Chemometr. 18(9):422-429.

Moser, G., B. Tier, R. Crump, M. Khatkar, and H. Raadsma. 2009. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genet. Sel. Evol. 41(1):56.

Mrode, R. A. and G. J. Swanson. 2004. Calculating cow and daughter yield deviations and partitioning of genetic evaluations under a random regression model. Livest. Prod. Sci. 86(1-3):253-260.

Mrode, R. A., T. Krzyzelewski, K. Moore, M. Winters, M. Coffey. 2011. The implementation of genomic evaluations in the UK. In proceedings of Interbull meeting. Stavanger, Norway, 26-29 August, 2011.

Ostersen, T., O.F. Christensen, M. Henryon, B. Nielsen, G. Su, and P. Madsen. 2011. Deregressed EBV as the response variable yield more reliable genomic predictions than traditional EBV in pure-bred pigs. GSE. 43: 38.

Robert-Granié, C., Duchemin S., Larroque H., Baloche G., Barillet F., Moreno C., Legarra A. and E. Manfredi. 2011. A comparison of various methods for the computation of genomic breeding values in French Lacaune dairy sheep. EAAP Annual Meeting, Stavanger, Norway, 29 August-1 September, 2011.

Patry C., and V. Ducrocq. 2011. Evidence of biases in genetic evaluations due to genomic preselection in dairy cattle. J Dairy Sci. 94(2):1011-20.

Shen, H. P. and J. H. Z. Huang. 2008. Sparse principal component analysis via regularized low rank matrix approximation. J. Multivar. Anal. 99(6):1015-1034.

Solberg, T. R., A. K. Sonesson, J. A. Woolliams, and T. H. E. Meuwissen. 2009. Reducing dimensionality for prediction of genome-wide breeding values. Genet. Sel. Evol. 41:8.

Sorensen, D. and Gianola, D. 2002. Likelihood, Bayesian and MCMC methods in quantitative genetics. 740 pp. Springer, New York.

Steiger, J.H. 1980. Tests for comparing elements of a correlation matrix. Psychol. Bull. 87:245-251.

Sun, X., D. Habier, R. Fernando, D. Garrick, and J. Dekkers. 2011. Genomic breeding value prediction and QTL mapping of QTLMAS2010 data using Bayesian Methods. BMC Proceedings 5(Suppl 3):S13.

Tibshirani, R. 1996. Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. Ser. B-Methodol. 58(1):267-288.

VanRaden, P. M. and G. R. Wiggans. 1991. Derivation, calculation, and use of national animal-model information. J. Dairy Sci. 74(8):2737-2746.

VanRaden, P. M. 2008. Efficient Methods to Compute Genomic Predictions. J. Dairy Sci. 91(11):4414-4423.

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92(1):16-24.

VanSickle, J. 2003. Analysing correlations between stream and watershed attributes. J. AmWater Resour. Assoc. 39 (3):717-726.

Vitezica Z.G., I. Aguilar, I. Misztal et A. Legarra. 2011. Bias in genomic predictions for populations under selection. Genet. Res. 93:357-366.

Weigel, K. A., G. de los Campos, O. Gonzalez-Recio, H. Naya, X. L. Wu, N. Long, G. J. M. Rosa, and D. Gianola. 2009. Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. J. Dairy Sci. 92(10):5248-5257.

Wold, H. 1966. Estimation of principal components and related models by iterative least squares.

Pages 391-420 in Multivariates Analysis, Krishnaiah P.R. (ED.), Academic Press, New York.

5.2.2 Étude méthodologique de l'approche BayesCπ

Capacités prédictives du BayesCπ avec ou sans pedigree. Habier *et al.* (2011) proposent de construire le modèle suivant (noté BayesCπPED dans ce manuscrit comme dans l'article 2):

$$\mathbf{y} = \mu\mathbf{1_n} + \mathbf{Zu} + \mathbf{Xg} + \mathbf{e}$$

où **y,** μ, **X, g** et **e** sont tels que définis au chapitre 2. La matrice **Z** est une matrice d'incidence reliant les observations du vecteur **y** aux effets aléatoires individuels **u** ; la distribution conditionnelle de **u** est telle que $\mathbf{u}|\mathbf{A}, \sigma_u^2 \sim \mathrm{N}(0, \mathbf{A}\sigma_u^2)$ où **A** représente la matrice de parenté calculée à partir des informations généalogiques et σ_u^2 la part de variance génétique additive non expliquée par les marqueurs SNP. La distribution *a priori* de σ_u^2 est une loi χ^{-2}. La seule différence avec le modèle BayesCπ classique (présenté dans le chapitre 2) est l'ajout du vecteur **u** des effets aléatoires polygéniques de tous les individus du pedigree. Habier *et al.* (2011) appliquent le modèle BayesCπPED sur des données réelles de 7 094 taureaux Holstein américains. Ces taureaux sont génotypés sur les marqueurs de la puce 54k. Leurs phénotypes sont représentés par les valeurs génétiques dérégressées obtenues à partir d'évaluations génétiques officielles de 4 caractères (quantité de lait, quantité de matière grasse et de matière protéique et comptage des cellules somatiques). La précision des estimations génomiques de leur modèle est donnée par la corrélation entre les GEBV et les valeurs dérégressées divisée par la précision moyenne des estimations de la population de validation, composée des 113 taureaux les plus jeunes. Le modèle BayesCπPED obtient, en moyenne sur les 4 caractères étudiés, une précision similaire au GBLUP (0,33 pour le BayesCπPED contre 0,34 pour le GBLUP), supérieure à l'approche BayesB (0,30) mais étonnamment inférieure à l'approche BayesA (0,36) qui, comme nous l'avons vu dans l'introduction de ce chapitre, présente souvent une capacité prédictive inférieure aux autres méthodes.

Un premier travail a donc consisté à vérifier l'impact de l'ajout de l'information pedigree au modèle BayesCπ sur la qualité du modèle de prédiction en l'appliquant sur les données françaises des races Holstein et Montbéliarde. Les tableaux 5.1 et 5.2 présentent les corrélations entre DYD observés dans la population de validation et les DYD prédits par les modèles BayesCπ et BayesCπPED et la pente de la

régression des valeurs observées sur les valeurs prédites, pour la race Holstein et pour la race Montbéliarde respectivement.

Tableau 5.1 : Capacités prédictives (corrélation ρ et pente de régression b entre DYD observés et DYD prédits) des modèles BayesCπ et BayesCπPED en race Holstein

	BayesCπ		**BayesCπPED**	
	ρ	***b***	***ρ***	***b***
Lait	0,56	0,73	0,57	0,73
MG	0,63	0,91	0,62	0,99
MP	0,56	0,70	0,54	0,79
TB	0,80	0,90	0,78	0,89
TP	0,75	0,93	0,74	0,91
Fertilité	0,34	0,72	0,34	0,82

Selon le test de Hotelling-Williams au seuil de significativité de 5%, les corrélations ne sont pas significativement différentes entre ces deux modèles, quelque soit le caractère considéré et dans les deux races. Les coefficients de régression ne sont que très rarement proches de 1, ce qui implique des estimations génomiques biaisées. Les biais observés dans les études sur données réelles pourraient s'expliquer par la forte sélection imposée aux animaux des populations de référence bovines, comme exposé par Vitezica *et al*. (2011). Patry et Ducrocq (2011) ont également suggéré que l'utilisation de données phénotypiques sous la forme de DYD (qui implique l'hypothèse fausse d'une transmission mendélienne des caractères) peut introduire un biais dans les estimations génétiques. Les pentes de régression *b* sont améliorées, car plus proches de 1, pour la race Holstein pour les caractères MG, MP et fertilité (+0,08, +0,09 et +0,10 pour ces trois caractères, respectivement).

Tableau 5.2 : Capacités prédictives (corrélation ρ et pente de régression b entre DYD observés et DYD prédits) des modèles BayesCπ et BayesCπPED en race Montbéliarde

	BayesCπ		**BayesCπPED**	
	ρ	***b***	***ρ***	***b***
Lait	0,44	0,74	0,44	0,74
MG	0,50	0,87	0,42	1,10
MP	0,46	0,77	0,46	0,76
TB	0,62	0,85	0,62	0,90
TP	0,44	0,72	0,45	0,82
Fertilité	0,42	1,32	0,44	1,67

Pour la race Montbéliarde, la modélisation des caractères MG, TB et TP est affectée positivement par l'ajout de l'information pedigree, alors que le caractère de fertilité, qui présentait déjà une très mauvaise pente de régression (*b*=1,32) dans le modèle BayesCπ, subit un biais très important avec l'introduction de l'effet polygénique, faisant passer la pente à 1,67. Ce problème a été soulevé par Habier *et al.* (2007) dans le cadre des évaluations par la méthode GBLUP en démontrant que les effets estimés des marqueurs SNP pourraient ne pas modéliser correctement les relations de parenté. Dans ce cas, Solberg *et al.* (2009) proposent d'améliorer le modèle GBLUP en y incorporant un effet polygénique « résiduel ». Dans cette optique, Liu *et al.* (2011) estiment les valeurs génomiques d'un ensemble de plus de 17 000 taureaux Holstein, provenant des collaborations du consortium EuroGenomics (Lund *et al.,* 2010) à partir d'un modèle GBLUP avec ou sans effet polygénique. Ils montrent que l'inclusion d'un effet polygénique conduit à des estimations génomiques moins biaisées. Sun *et al.* (2011) étudient à leur tour, l'effet de l'ajout d'un terme polygénique résiduel dans l'approche BayesCπ, mais aussi BayesB et GBLUP, sur les données simulées du 14ème congrès QTLMAS. Dans leur étude, ils montrent que les modèles n'incluant aucun effet polygénique sont semblables en termes de corrélations (environ 0,02, 0,01 et 0,01 de gain de corrélations pour les méthodes GBLUP, BayesB et BayesCπ respectivement) mais

produisent des estimations génomiques moins biaisées (0,05, 0,04 et 0,03 de différence entre les pentes de régression d'un modèle sans et avec effet polygénique, sur les trois méthodes). L'approche BayesCπ semble être la moins affectée des trois méthodes, par l'inclusion d'un effet polygénique. Cette stabilité du modèle BayesCπ correspond aux résultats observés sur les données Holstein et Montbéliarde françaises. Ces mêmes approches ont été appliquées sur des données ovines Lacaune laitières françaises, dans le cadre du projet Roquefort'In réunissant 2 567 béliers génotypés et phénotypés (Duchemin *et al.,* 2012). L'inclusion d'un effet polygénique induit des corrélations supérieures de 1% sur le caractère quantité de lait et de 1,4% sur le caractère de comptage des cellules somatiques. Sur ces données, l'impact est donc légèrement plus important sur le caractère à faible héritabilité (score de comptage des cellules somatiques d'héritabilité égale à 0,14). Au niveau du biais de prédiction, le modèle BayesCπPED produit des coefficients de régression plus proches de 1 (0,99 pour le lait, 0,93 pour le TB et 0,94 pour les cellules avec un modèle avec effet polygénique contre 0,93, 0,90, 0,88 pour un modèle sans effet polygénique).

Convergence des estimations des composantes de la variance. En raison du peu de différence entre les capacités prédictives des modèles BayesCπ et BayesCπPED, le modèle le plus simple, c'est-à-dire sans considérer d'effets polygéniques, a été conservé dans le reste de l'étude. Les résultats présentés ne porteront plus que sur les trois caractères considérés dans l'article 2 afin de faciliter les comparaisons entre les conclusions des deux études : quantité de lait (Lait), taux butyreux (TB) et fertilité (Fer).

Dans l'article 2 (figures 2 et 3), sont présentées les courbes de densité des valeurs estimées au cours de l'algorithme MCMC des variances génétiques et résiduelles et du paramètre π. La figure 2 permet de témoigner de la bonne convergence des variances pour les trois caractères (avec une distribution proche d'une distribution gaussienne). L'estimation du paramètre π, qui représente le taux de SNP avec un effet non nul, converge très rapidement et ce vers une valeur très faible, pour le lait (π égal à 0,04) et le taux butyreux (π égal à 0,02). Pour la fertilité, la convergence n'est pas atteinte pour le paramètre π, les estimations variant entre 5% et 95% mais sans affecter les convergences des composantes de la variance.

L'estimation de π est finalement proche de 50%. La figure 3 montre que la convergence des variances n'est pas totalement atteinte pour les trois caractères avec des distributions non asymétriques ou bimodales. Même en augmentant le nombre d'itérations de l'algorithme, les résultats restent semblables. Pour le paramètre π, seul le caractère TB mène à une bonne convergence avec une estimation inférieure à 1%. La figure 3 montre également qu'il est difficile de fixer ce paramètre pour le lait et la fertilité en race Montbéliarde, mais que cela a peu d'impact sur la convergence et les estimations des variances génétiques et résiduelles pour ces caractères.

L'estimation des paramètres sur les données Montbéliardes est plus difficile que sur celles de la race Holstein, probablement en raison de la taille de la population de référence (1 172 taureaux en race Montbéliarde et 3 940 en race Holstein). L'estimation de la valeur de π entre les deux races pour un même caractère est différente (sauf pour le TB avec une valeur de π autour de 1% ou 2%), impliquant un nombre différent de SNP retenus dans chaque race. Cependant sur les caractères étudiés, nombreux sont ceux pour lesquels la valeur de π est mal estimée (mauvaise convergence), donc la comparaison reste difficile. Pour la race Holstein, le nombre de SNP sélectionnés est d'environ 1 600 pour la quantité de lait et 800 pour le taux butyreux. Pour la race Montbéliarde, le paramètre π est environ égal à 1% pour le taux butyreux, soit 400 SNP sélectionnés environ. Avec une petite population de référence, la capacité de détection des SNP les plus significatifs est réduite. Dans le cas où la convergence de π n'est pas atteinte (fertilité Holstein et Montbéliarde et quantité de lait en race Montbéliarde), la valeur finalement conservée pour ce paramètre se situe autour de la moyenne 0,5 ce qui correspond à un nombre de SNP sélectionnés compris entre 15 000 et 25 000 SNP. Malgré une convergence difficile pour le paramètre π et une mauvaise estimation de ce dernier, la corrélation entre les DYD observés et les DYD prédits dans l'ensemble de validation n'est pas modifiée, par rapport aux autres méthodes comparées dans l'article 2.

Dans l'étude de Habier *et al.* (2011), le nombre de SNP sélectionnés augmente avec la taille de la population pour les quantités de lait et de matière grasse mais diminue pour les deux autres caractères (quantité de matière protéique et comptage des cellules somatiques). Ces auteurs s'intéressent également à

l'impact de l'héritabilité et du nombre de QTL du caractère : le nombre de SNP sélectionnés diminue avec le nombre de QTL mais le nombre de SNP est très supérieur au nombre de QTL quelque soit le scénario étudié. Moins le caractère est héritable et plus cette surestimation est forte d'où un nombre de SNP sélectionnés très important par rapport au nombre de QTL. Habier *et al.* (2011) montrent que les deux premiers caractères sont régis par un nombre plus important de QTL à forts effets. Nous verrons par la suite que c'est ce qui semble être le cas dans notre étude, pour les caractères Lait et TB en Holstein, et le caractère TB en Montbéliarde : le paramètre π est bien estimé de par la présence éventuelle de forts QTL. Le nombre de SNP sélectionnés pour le TB en race Holstein (800 SNP) est plus important que pour la race Montbéliarde (400 SNP) car la population de référence est plus grande. Le nombre de SNP estimés dans le modèle BayesCπ pourrait être principalement dû aux QTL avec les plus forts effets : les QTL avec des effets plus faibles ne seraient alors pas détectables par cette méthode. L'étude d'une population de référence de taille plus importante semble être nécessaire pour estimer et détecter les plus faibles effets QTL.

5.2.3 Comparaison de la méthode BayesCπ avec le BLUP sur pedigree et les autres méthodes génomiques

Capacités prédictives des méthodes d'évaluation génomique et du BLUP sur pedigree. La capacité prédictive de la méthode BayesCπ a été comparée avec une méthode classique d'évaluation génétique (le BLUP sur pedigree) et les méthodes GBLUP, PLS, sparse PLS, et LASSO bayésien selon le même schéma que le chapitre 4 : c'est pourquoi nous ne nous attarderons que sur les méthodes bayésiennes. La première observation est que les méthodes bayésiennes, tout comme les méthodes GBLUP, PLS et sparse PLS, ont de meilleures performances prédictives que le BLUP basé exclusivement sur l'information pedigree. Au vu de ces résultats et des résultats présentés précédemment sur l'intégration d'un effet polygénique estimé à partir de l'information pedigree dans le modèle BayesCπ, il semble que les marqueurs SNP soient capables de rendre compte de l'information relative aux relations de parenté entre individus (Habier *et al.,* 2007). Ce gain de corrélation moyen des deux méthodes bayésiennes sur le BLUP, reste cependant

limité (pour la race Holstein +0,18, +0,35, et +0,06 et pour la race Montbéliarde +0,16, +0,17, et +0 pour les caractères lait, TB et fertilité respectivement), car les taureaux de nos ensembles d'apprentissage et de validation ont des liens de parenté très forts d'où une surestimation des résultats du BLUP (Habier *et al.,* 2010). Les méthodes bayésiennes donnent des corrélations significativement supérieures aux autres méthodes, sur le caractère taux butyreux seulement, probablement de par la présence du gène DGAT1 (Grisart *et al.,* 2004). En effet, nous verrons par la suite que les méthodes bayésiennes retiennent presque exclusivement des SNP autour de ce QTL qui explique une grande part de la variance génétique.

Estimation des effets des SNP par les méthodes BayesCπ, LASSO bayésien et sparse PLS. Les méthodes BayesCπ et LASSO bayésien sont deux méthodes de sélection de variables permettant l'estimation des effets d'un nombre réduit de variables. Les SNP sélectionnés sont ceux ayant le plus fort impact sur le caractère étudié. Il est donc intéressant de voir où ils se positionnent, car ils pourraient correspondre à des QTL, et de vérifier s'ils sont retrouvés de manière récurrente entre les différentes méthodes. La régression sparse PLS permet de sélectionner un nombre réduit de SNP par la construction de variables latentes. Nous avons donc souhaité comparer la position des SNP sélectionnés par ces trois méthodes. Les figures 5 et 6 de l'article 2 représentent l'estimation des effets des SNP obtenus par les méthodes bayésiennes et les coefficients VIP pour la sparse PLS pour les deux races et les trois caractères étudiés. L'intérêt porte principalement sur la position des SNP ayant les plus gros effets et non sur la valeur des effets.

Entre les deux méthodes bayésiennes et pour un caractère et une race donnés, les positions des SNP « très significatifs » sont très similaires. Pour la race Holstein et le caractère quantité de lait (1ère colonne sur la figure 5), les régions du génome des chromosomes 1, 5, 16 et 21 sont bien identifiées et importantes par l'approche BayesCπ et la régression sparse PLS mais ne représentent que de petits « pics » pour le LASSO bayésien. Cependant, le chromosome 14 est pointé très fortement par les trois méthodes, et les mêmes SNP sont sélectionnés. Au contraire, deux SNP sur le chromosome 7 ont de gros effets estimés par l'approche BayesCπ (>35) mais n'ont aucun d'effet détecté par le LASSO bayésien. Pour le taux butyreux (2ème colonne de la figure 5), 4 pics sont détectés par les méthodes BayesCπ et

sparse PLS sur les chromosomes 5, 11, 14 et 20 tandis que le LASSO bayésien n'attribue un poids très fort qu'à un seul SNP. Pour le caractère de fertilité (3ème colonne de la figure 5), les graphes des trois méthodes révèlent la nature très polygénique de ce caractère car peu de SNP se distinguent des autres et de très nombreux SNP ont des effets faibles. La régression sparse PLS donne des résultats semblables à l'approche BayesCπ sur les trois caractères : les plus gros pics sont détectés au même endroit en incluant le SNP du chromosome 7 pour la quantité de lait, qui n'est pas repéré par le LASSO bayésien.

Pour la race Montbéliarde, les graphes sont très semblables entre les trois méthodes : la position des SNP à fort effet est similaire, pour le lait et la fertilité. Cependant, pour le taux butyreux, les différences sont plus marquées : le LASSO bayésien ne met en valeur que quelques SNP sur le chromosome 14 alors que l'approche BayesCπ et la régression sparse PLS pointent aussi des régions importantes sur d'autres chromosomes (4, 5, 19, 21 et 27). Pour l'ensemble des caractères, tous les pics détectés par l'approche BayesCπ le sont aussi par la régression sparse PLS mais avec des rangs d'importance différents. Si on s'intéresse à la quantité de lait, le chromosome 14, n'est pas la région la plus importante pour la régression sparse PLS comme elle peut l'être pour l'approche BayesCπ. On peut noter que la régression sparse PLS met en avant certaines régions non détectées par l'approche BayesCπ : les SNP ayant les plus forts effets pour la sparse PLS pour la quantité de lait, sont sur les chromosomes 7, 10, 15 et 19 (VIP autour de 8). Pour le taux butyreux, les mêmes zones sont révélées par la sparse PLS et le BayesCπ mais la sparse PLS sélectionne plus de SNP autour de ces zones que l'approche BayesCπ. Par exemple, sur le chromosome 5, un seul SNP est sélectionné par l'approche BayesCπ alors que la régression sparse PLS en sélectionne 4 avec des coefficients VIP très proches. Pour le caractère de fertilité, les pics qui apparaissent sont aux mêmes positions pour toutes les méthodes mais ils semblent plus accentués avec la régression sparse PLS.

Les graphes obtenus pour les caractères TB et fertilité sont très semblables d'une race à l'autre. Le taux butyreux semble être gouverné par un nombre très réduit de gros QTL, avec notamment une zone d'importance majeure sur le chromosome 14. Cette zone est connue pour avoir un très fort impact sur le TB : c'est le gène DGAT1 (Grisart *et al.,* 2004). Au contraire, le caractère de fertilité

apparaît comme étant très polygénique dans les deux races, avec un nombre important de QTL à petits effets. Une différence majeure entre les deux races est relevée lors de l'étude du caractère lait : il semble être régi par peu de QTL à gros effets pour la race Holstein et par un plus grand nombre de QTL à effets moyens pour la race Montbéliarde ce qui laisse supposer de l'importance de la taille de la population de référence dans l'estimation fiable des effets des SNP.

5.3 Détection de QTL par des méthodes de sélection génomique

La découverte de ces SNP a non seulement bénéficié à une meilleure évaluation génomique des reproducteurs, mais également à la détection de QTL associés aux caractères d'intérêt pour la sélection animale. Les résultats présentés dans cette dernière partie sont issus d'une étude préliminaire concernant la capacité des méthodes génomiques à mettre en avant les zones d'intérêt du génome. Dans cette étude, les régions chromosomiques mises en évidence par la régression sparse PLS et l'approche Elastic Net ont été comparées entre elles et aux résultats d'une étude **LDLA** (Linkage Disequilibrium and Linkage Analysis).

5.3.1 Les études LDLA et l'approche BLUP-QTL

Les études LDLA. Plusieurs études ont montré que la distribution des effets aux loci ayant un impact sur les caractères d'intérêt, est telle qu'il existe peu de QTL ayant de gros effets et beaucoup ayant de petits effets (Shrimpton et Robertson, 1998, Hayes et Goddard, 2001). Un QTL est caractérisé par sa position sur le génome (ou par l'intervalle de confiance de sa position), la part de variance génétique du caractère qu'il explique et le nombre d'allèles et leurs effets. Les approches de détection de QTL permettent de localiser les régions du génome qui sont associées à des variations phénotypiques. Elles supposent que les vrais QTL qui affectent un caractère quantitatif ne sont pas connus. Les études de liaison (**LA** pour Linkage Analysis) utilisent les marqueurs moléculaires et cherchent des associations entre ces marqueurs et les variations phénotypiques. Si le nombre de marqueurs disponibles est limité, alors les associations entre marqueurs et QTL ne seront persistantes qu'intra-familles sur un nombre limité de générations. Ces approches ont été exploitées dans presque toutes les espèces d'élevage et pour un grand

nombre de caractères (Andersson et Georges, 2004). Mais le principal problème de ces études est que les QTL sont cartographiés sur des gros intervalles de confiance sur un chromosome. De nos jours, les outils de génotypage permettent d'analyser un grand nombre de marqueurs répartis sur l'ensemble du génome. Ces marqueurs sont suffisamment proches pour exploiter le déséquilibre de liaison qui existe entre ces marqueurs et d'éventuels QTL. Les analyses de liaison LA ne considèrent que le déséquilibre de liaison existant au sein d'une même famille alors que les études de déséquilibre de liaison **LD** (pour Linkage Disequilibrium) exploitent le déséquilibre de liaison au niveau d'une population entière. Ces associations sont dues à la présence de segments chromosomiques au sein de la population actuelle qui proviennent d'un même ancêtre commun. Ces segments chromosomiques vont porter des allèles aux marqueurs identiques. Ainsi, si un QTL est positionné entre deux marqueurs, ces deux marqueurs porteront également les mêmes allèles au QTL. Dans l'approche linkage disequilibrium and linkage analysis (LDLA), 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.

L'approche BLUP-QTL. Depuis Juin 2010, la sélection génomique chez les bovins laitiers français repose sur le modèle de Fernando-Grossman (1989, voir chapitre 1) dans lequel la liste des QTL inclus est fournie par deux sources différentes : une approche LDLA et une approche génomique de sélection de variables (figure 5.1). Ce modèle, désigné sous le nom de BLUP-QTL, a été testé en utilisant deux méthodes de sélection de variables, présentées dans ce document : l'approche Elastic Net et l'approche PLS.

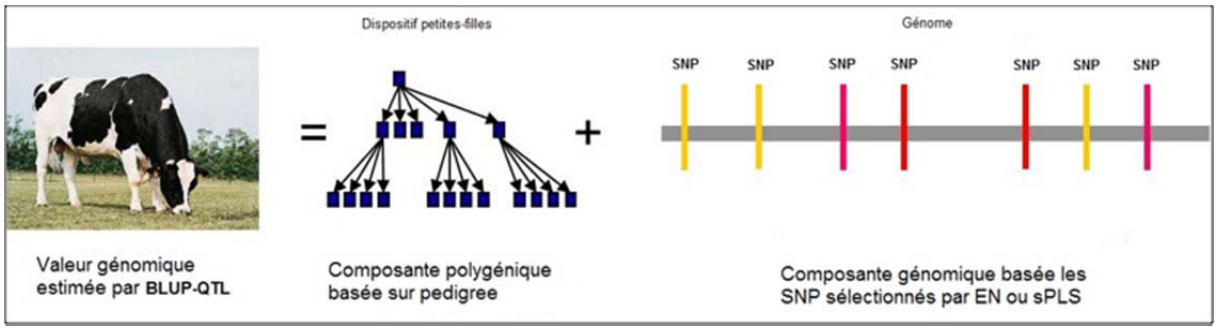


Figure 5.1 : Modélisation du BLUP-QTL français

Les différents essais réalisés sur les données bovines laitières françaises ont menés à des résultats très similaires entre un BLUP-QTL basé sur l'approche Elastic Net et un BLUP-QTL basé sur la régression sparse PLS (résultats non publiés). Fritz *et al.* (2010) présentent les résultats du BLUP-QTL basé sur l'approche Elastic Net, sur les populations de référence d'EuroGenomics (environ 16 000 taureaux Holstein et 1 250 taureaux de race Montbéliarde). Pour la race Holstein, les corrélations entre DYD observés dans leur population de validation et les DYD prédits par le BLUP-QTL sont de 0,60 pour le caractère lait, 0,59 pour le caractère matière protéique et 0,43 pour la fertilité. Pour la race Montbéliarde, les corrélations par le BLUP-QTL sont de 0,47 pour le caractère lait, 0,46 pour le caractère matière protéique et 0,41 pour la fertilité. Les résultats du GBLUP sur la même population Montbéliarde et sur les caractères lait, MP et fertilité sont similaires (0,44, 0,46 et 0,43 respectivement, Croiseau *et al.,* 2011). Les résultats sur cette grande population Holstein n'ont pas été publiés mais sont de 0,71 pour le lait, 0,63 pour le MP et 0,33 pour la fertilité donc proches des résultats du BLUP-QTL sur deux caractères sur trois. On peut donc supposer que les SNP sélectionnés par l'Elastic Net et introduits dans le modèle BLUP-QTL (quelques centaines de marqueurs) aident à expliquer les différents caractères considérés. Nous avons voulu savoir si ces SNP étaient équivalents aux QTL détectés par une étude LDLA, afin de voir si les méthodes de sélection de variables Elastic Net et sparse PLS pouvaient apporter une aide ou être complémentaires aux méthodes de détection des QTL.

5.3.2 Méthodologies pour la détection de QTL par les méthodes d'évaluation génomique

Stratégies de détection de QTL par sparse PLS et Elastic Net. Nous avons testé deux stratégies pour la détection des SNP importants basées sur deux méthodes de sélection génomique : l'Elastic Net et la sparse PLS. L'objectif est de créer des listes de SNP sélectionnés à partir de ces méthodes puis de les comparer entre elles et aux SNP mis en avant par une étude LDLA.

Comme décrit au chapitre 2, ces deux méthodes requièrent l'estimation de plusieurs paramètres qui influencent le nombre de SNP retenus dans le modèle final. Une première stratégie a consisté à paramétrer les méthodes pour que dans le modèle final établi sur la population d'apprentissage, ne soit retenu qu'un nombre

fixé de SNP (10, 100 et 500 SNP) sans se soucier de la capacité prédictive du modèle ainsi construit. Cependant, les précisions des estimations génomiques sur la population de validation se sont révélées trop basses pour pouvoir considérer que les SNP ainsi sélectionnés étaient bien choisis. Une seconde stratégie, consistant à évaluer le modèle optimal sur chacune des méthodes de sélection, a alors été adoptée, en choisissant les paramètres pour maximiser la corrélation entre DYD observés et DYD prédits de l'ensemble de validation. Nous appellerons les modèles optimaux SG-EN (pour sélection génomique Elastic Net) et SG-sPLS (pour la sparse PLS) pour les approches basées sur les méthodes de sélection génomique par opposition à l'approche LDLA. Les effets des SNP sont ainsi prédits selon les deux méthodes puis des listes réduites de SNP sont créées en sélectionnant les 10, 100 et 500 marqueurs avec les plus gros effets pour chacune des approches SG.

Le but de cette étude étant de pouvoir vérifier les capacités de détection de QTL des méthodes de sélection génomique, nous avons comparé les résultats des approches SG-EN et SG-sPLS avec les QTL détectés par une étude LDLA. Elles ont toutes été appliquées sur les données de la race Holstein, sur trois caractères : la matière protéique, le taux butyreux et la fertilité afin d'avoir un panel de caractères d'héritabilités différentes et de profil des effets SNP différents (voir chapitre 4).

Critères de comparaison entre les résultats de l'approche LDLA et des approches génomiques. Dans les études LDLA, chaque région chromosomique est testée afin d'obtenir une valeur **LRT** (Likelihood Ratio Test), obtenue par un test de rapport de vraisemblances. On considère qu'un QTL existe si la position de ce locus correspond à un haut pic LRT, c'est-à-dire si la valeur du coefficient LRT à ce SNP est supérieure à un certain seuil. Deux valeurs de seuil ont été testées (LRT > 5 et LRT > 9).

Un exemple est présenté par la figure 5.2 : elle montre les valeurs LRT des SNP du chromosome 5, pour le caractère de matière protéique. On remarque que si l'on considère un seuil LRT de 5 dans la définition des pics LDLA alors on dénombre 7 pics (en vert sur la figure 5.2) sur le chromosome 5 pour le caractère MP. Mais si on considère un seuil de 9, alors seuls trois pics sont conservés (en bleu sur la figure

5.2). Dans la suite de cette étude, nous nous appliquerons donc à conserver ces deux valeurs de seuil pour définir les pics LDLA.

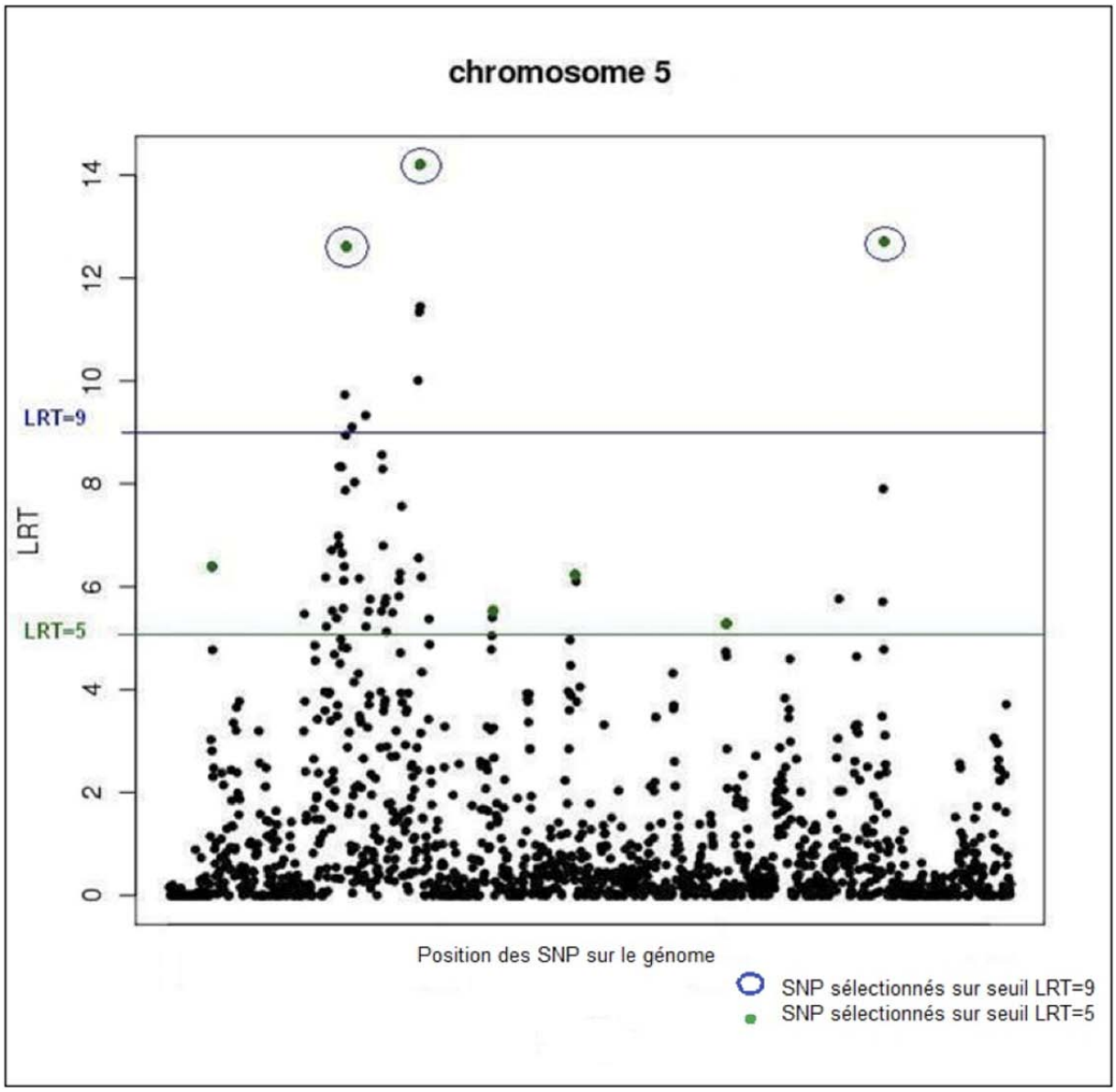


Figure 5.2 : Valeurs des coefficients LRT pour les SNP du chromosome 5, sur le caractère matière protéique (MP)

Dans un premier temps, les régions détectées par les approches SG-EN et SG-sPLS ont été comparées. Un pic LRT est considéré comme étant détecté par les méthodes SG s'il y a au moins un SNP des listes SG-EN ou SG-sPLS de 10, 100 ou 500 marqueurs dans une fenêtre de taille 1cM autour de ce pic (figure 5.3).

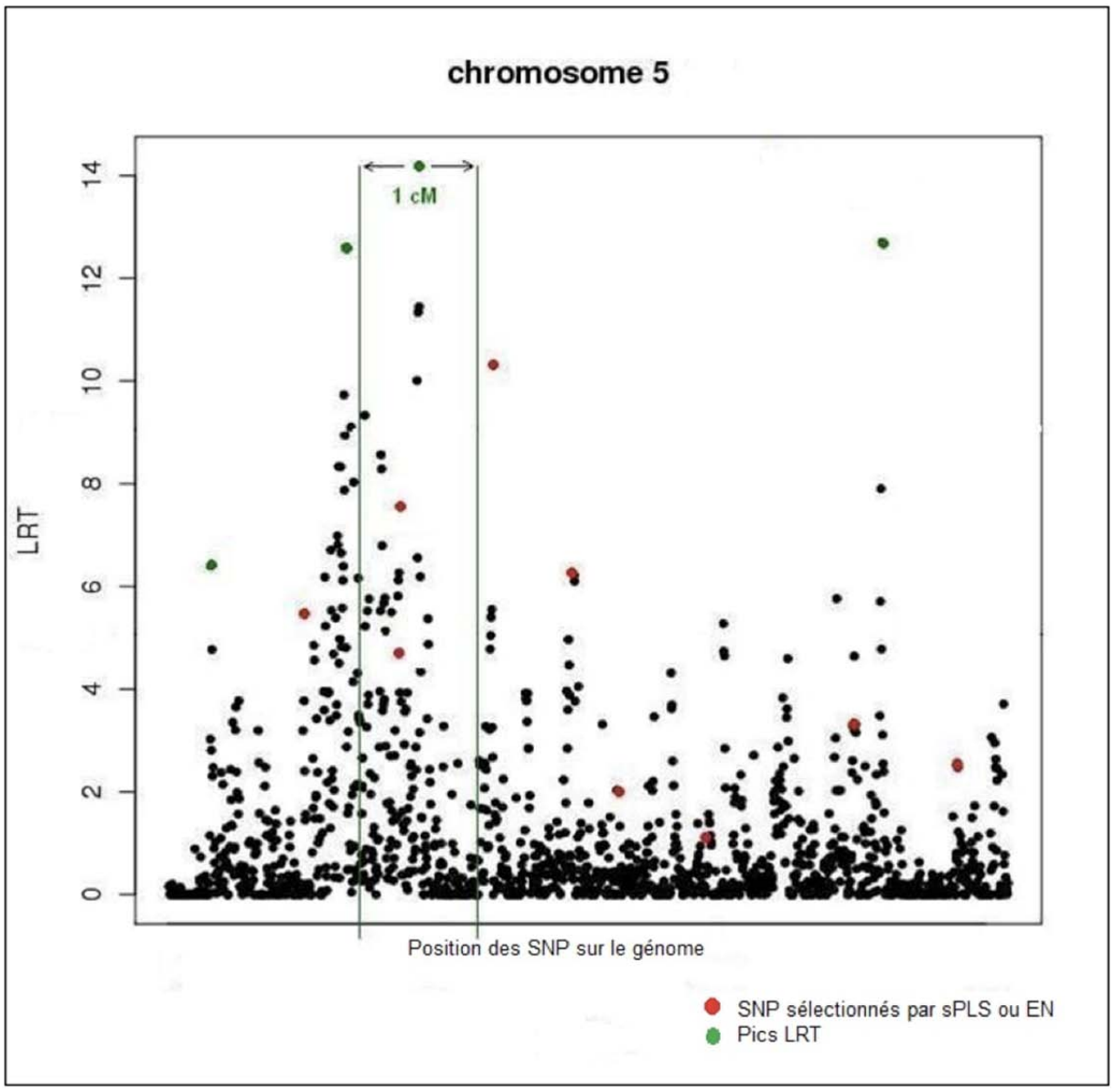


Figure 5.3 : Méthode de comparaison entre les pics LDLA et les SNP des approches SG

Un premier critère de comparaison entre les résultats LDLA et SG consiste donc à compter le nombre de SNP sélectionnés par les approches SG et qui sont compris dans cette fenêtre de proximité avec les pics LRT. Afin de vérifier la supériorité d'une approche par rapport à l'autre les listes SG-EN et SG-sPLS ont été comparées. Les pics LRT sont utilisés comme résultats témoins : la distance moyenne entre chaque pic LRT et le SNP des listes SG qui a le plus gros effet estimé, est calculée afin de savoir laquelle des deux méthodes SG se rapproche le plus des résultats LDLA.

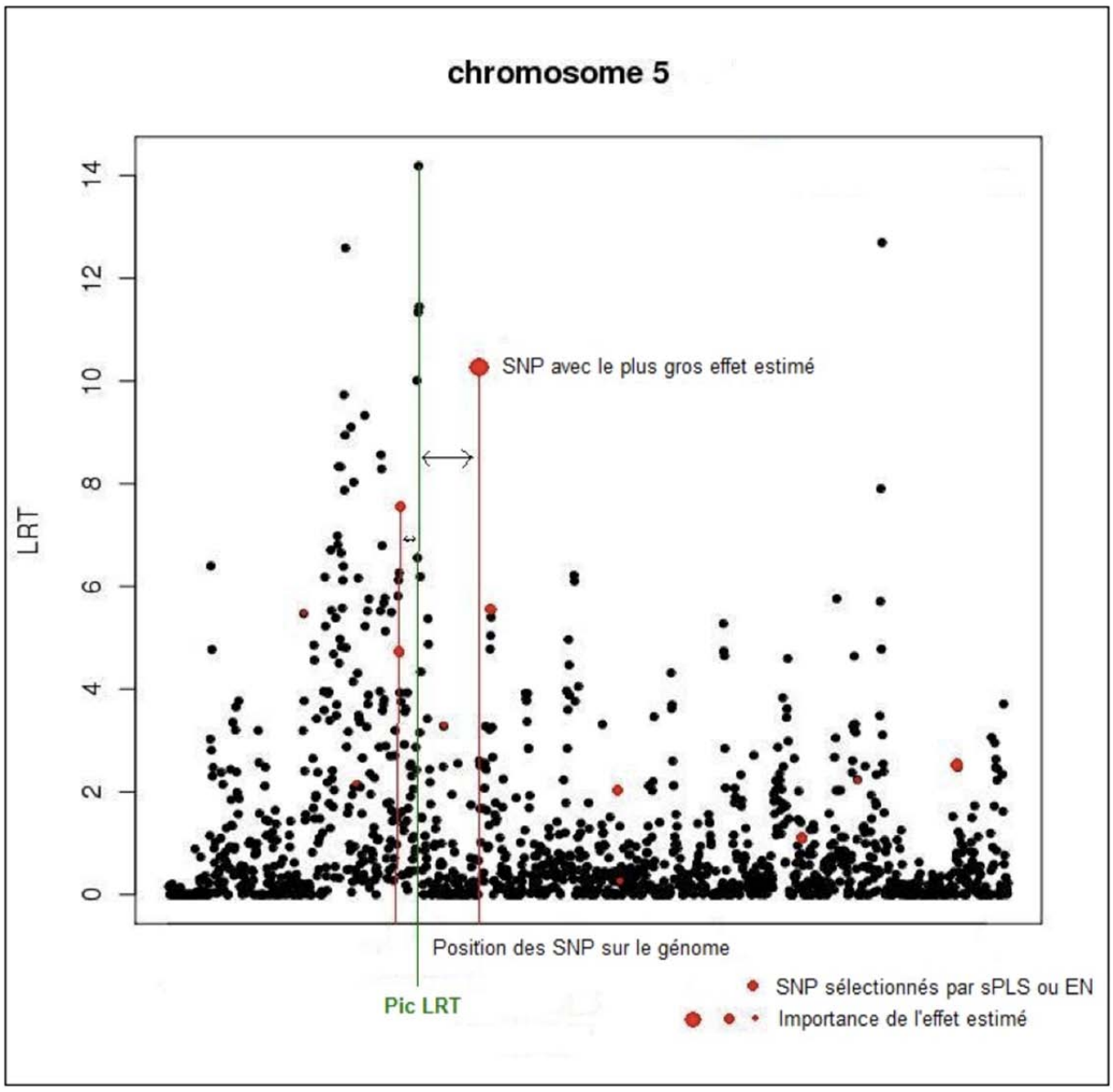


Figure 5.4 : Calcul de la distance entre les SNP de l'approche SG-EN et les SNP de l'approche SG-sPLS, par rapport aux pics LDLA

La figure 5.4 donne un exemple du calcul de la distance entre pics LRT et SNP sélectionnés par les approches SG. Elle représente les valeurs LRT obtenues par les SNP du chromosome 5. Autour du pic signalé en vert se trouvent plusieurs SNP (en rouge) sélectionnés par les méthodes SG. Le SNP le plus proche (à gauche du pic) n'est pas celui qui sera pris en compte dans le calcul de la distance car son effet estimé est plus faible que le SNP à droite du pic.

5.3.3 Étude comparative des approches d'évaluation génomique et de l'approche LDLA

Comparaison entre approches LDLA et approches SG. Le tableau 5.3 présente le nombre de pics LRT obtenu avec chacune des méthodes SG, en utilisant des listes de taille 10, 100 et 500 SNP, pour les deux seuils de LRT. Les deux dernières colonnes indiquent le nombre de pics LRT total pour chaque valeur de seuil, sur les trois caractères.

Tableau 5.3 : Nombre de pics LRT détectés par les approches SG-EN et SG-sPLS avec les 10, 100 et 500 SNP avec les plus gros effets estimés

	LRT>5						**LRT>9**						**Nombre total de pics LRT**	
	SG-EN			**SG-sPLS**			**SG-EN**			**SG-sPLS**				
	10	100	500	10	100	500	10	100	500	10	100	500	LRT>5	LRT>9
MP	1	5	30	3	6	17	1	5	27	3	6	15	186	82
TB	1	5	27	1	11	20	1	5	17	1	11	17	139	40
Fer	2	7	26	1	2	14	2	6	15	0	0	1	88	22

Sur les plus petites listes de SNP (10 et 100 SNP) et pour les caractères MP et TB, la sparse PLS donne de meilleurs résultats que l'Elastic Net : par exemple, pour le caractère TB et la liste 100 SNP, la sparse PLS repère 1/4 des pics LRT les plus importants (LRT>9) alors que l'Elastic Net n'en détecte que 1/8 (11 pics sur 40 pour la régression sparse PLS et 5 pics sur 40 pour l'Elastic Net). Ainsi, les SNP les plus proches des pics LRT pour la régression sparse PLS sont parmi les 100 SNP ayant les plus gros effets estimés. Sur la plus grande liste de SNP (500), l'Elastic Net présente les meilleures performances. Sur le caractère de fertilité, l'Elastic Net détecte 15 pics parmi les plus forts coefficients LRT sur 22 alors que la régression sparse PLS n'en détecte qu'un seul. Le profil très polygénique de ce caractère, déjà dévoilé par les graphes des effets estimés par les différentes méthodes présentées dans ce manuscrit, est confirmé par le nombre très réduit de valeurs LRT dépassant le seuil de 5 (88 pics) par rapport aux autres méthodes : cela signifie que beaucoup de QTL régissent ce caractère mais avec des effets trop petits pour ces fortes valeurs de seuil. Les résultats du tableau 5.3 sont similaires entre les deux valeurs

de seuil LRT et cohérents par rapport aux listes de SNP considérées ce qui laisse penser que les approches SG-EN et SG-sPLS détectent en priorité les plus gros QTL.

Comparaison entre approches de Sélection Génomique. Les méthodes SG sont donc utiles pour repérer certains QTL détectés par une approche LDLA. Il est maintenant intéressant de s'interroger sur les capacités individuelles des deux méthodes. La figure 5.5 présente les résultats du deuxième critère de comparaison : la distance entre un pic détecté par LDLA (avec un seuil LRT > 5) et le SNP sélectionné par les méthodes SG présentant le plus fort effet. La distance moyenne sur tous les pics détectés par SG-EN est plus petite que par SG-sPLS, pour les caractères matière protéique et fertilité. Ainsi, quelque soit la taille de la liste considérée, les SNP sélectionnés par l'Elastic Net sont donc plus proches des QTL détectés par l'étude LDLA. Pour le caractère taux butyreux, la distance est semblable entre les deux approches à part pour la liste de 10 SNP où elle est nulle pour la méthode SG-sPLS. Cependant, si on s'en réfère au tableau 5.3, on remarque que dans ce cas, la sparse PLS et l'Elastic Net ne détectent qu'un seul pic LRT.

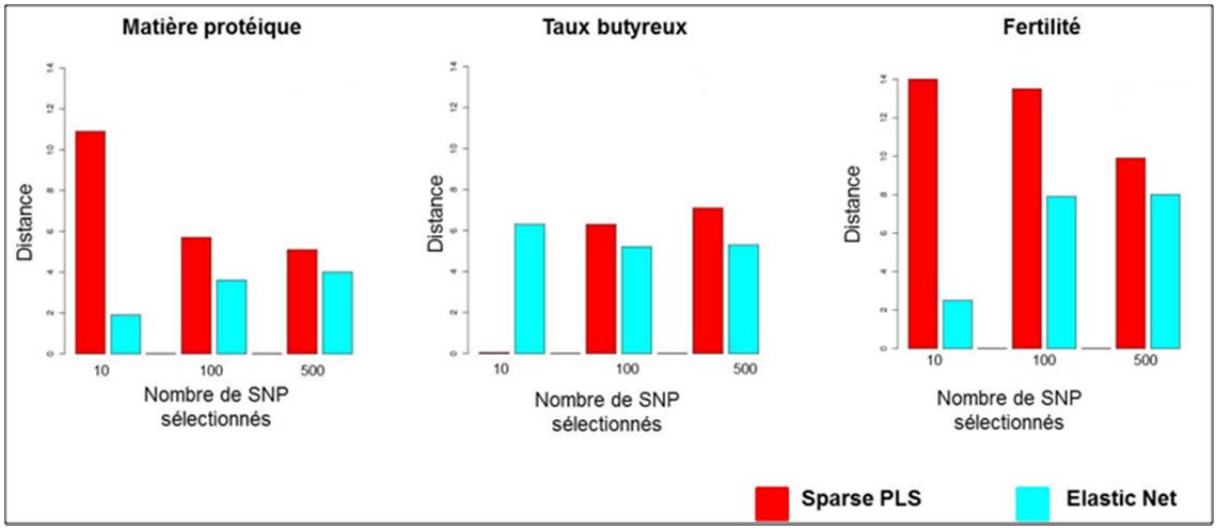


Figure 5.5 : Distance moyenne entre les SNP sélectionnés par SG-EN et SG-sPLS et le pic LRT le plus proche

La distance représentée dans la figure 5.5 est donc la distance entre ce pic LRT et le SNP qui a le plus fort effet estimé par chacune des méthodes SG autour de

cet unique pic. La sparse PLS sélectionne donc en priorité le SNP qui marque le pic LRT alors que l'Elastic Net en choisira un plus éloigné.

La plus forte valeur de LRT pour le caractère taux butyreux correspond au gène DGAT1, situé au début du chromosome 14. La figure 5.6 représente cette région du génome et les valeurs LRT des SNP qui s'y trouvent.

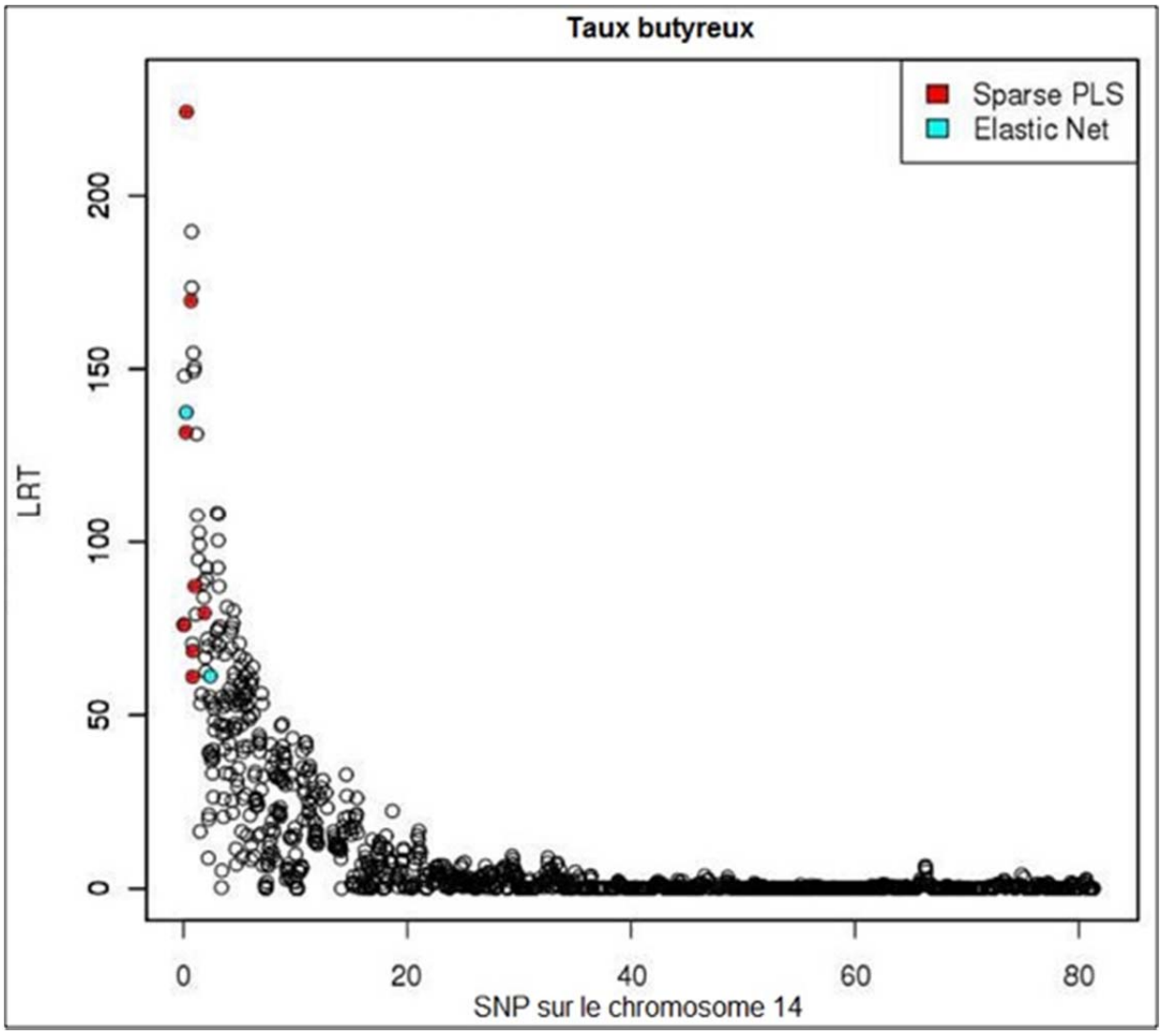


Figure 5.6 : Valeurs LRT des SNP du chromosome 14

Les six points rouges sont les six SNP qui obtiennent les plus forts effets estimés par la régression sparse PLS : ils sont très proches ou sont sur le gène DGAT1. Les deux points bleus représentent les SNP sélectionnés par l'Elastic Net. La sparse PLS semble donc plus efficace pour repérer les gros QTL mais utilise plus de SNP pour marquer les régions d'intérêt que l'Elastic Net ce qui pourrait empêcher

la détection d'autres régions d'importance si des listes de SNP trop réduites sont utilisées.

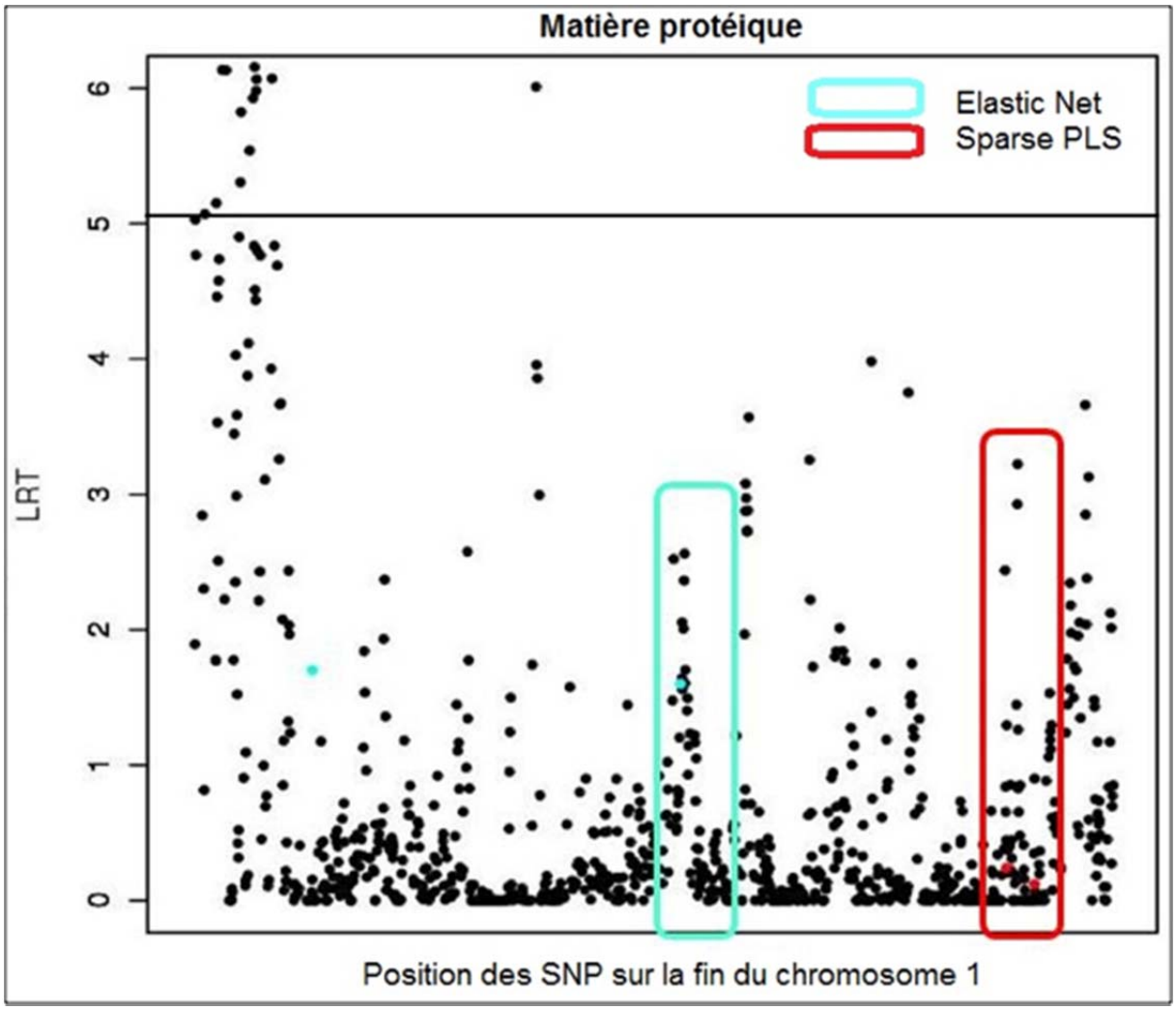


Figure 5.7 : Mise en évidence des régions à la fin du chromosome 1 détectées par les approches SG seulement

Intéressons-nous à une autre zone du génome. La figure 5.7 présente les valeurs LRT des SNP présents sur la fin du chromosome 1, pour le caractère de matière protéique. Les points bleus représentent les SNP sélectionnés par la méthode SG-EN limitée à 100 SNP et les points rouges les SNP sélectionnés par la méthode SG-sPLS limitée à 100 SNP. On observe deux régions sur ce fragment de chromosome qui sont pointées par la méthode SG-EN ou la méthode SG-sPLS mais qui obtiennent des valeurs LRT moyennes, bien inférieures au seuil LRT de 5. Ainsi, les approches SG sont capables de détecter d'éventuels QTL sur des régions non

considérées comme d'importance par l'approche LDLA. Des études supplémentaires sont nécessaires afin de vérifier la possibilité de la présence de QTL dans ces régions.

Ces résultats montrent que les méthodes de sélection génomique sont capables de localiser des SNP proches des plus gros QTL identifiés. La régression sparse PLS révèle plus de pics LRT que l'approche Elastic Net mais l'approche Elastic Net sélectionne des SNP qui sont plus proches de ces pics que la régression sparse PLS. Les deux approches permettent aussi de mettre en avant des zones du génome non détectées par les études LDLA. Contrairement à l'étude LDLA uni-QTL utilisée ici, les méthodes d'évaluation génomique estiment tous les effets des SNP simultanément ce qui pourrait améliorer la précision de ces estimations. Cependant, les SNP localisés par les différentes approches ne sont pas strictement identiques. De plus, il n'y a aucun moyen de savoir si les résultats LDLA sont plus justes que les résultats d'évaluation génomique, ou l'inverse. Des études par simulations devraient permettre d'identifier l'approche optimale pour la détection de QTL entre une méthodologie génomique et/ou LDLA.

5.4 Conclusion

L'étude des méthodes bayésiennes, et des méthodes de sélection de variables en général, ont permis d'aborder plusieurs aspects des méthodes d'évaluation génomique. Tout d'abord, même si l'approche BayesCπ permet de prendre en compte les liens de parenté entre les animaux comme source d'information supplémentaire, cela n'améliore pas la précision des estimations génomiques de ce modèle. Des études utilisant d'autres méthodes de prédiction intégrant un effet polygénique ont été réalisées mais mènent à des conclusions différentes. Avec l'approche LASSO bayésien, de los Campos *et al.* (2009) testent plusieurs scénarios reposants sur des distributions *a priori* des paramètres du modèle et utilisent l'information pedigree seulement, l'information des marqueurs seulement ou les deux, sur des données issues de lignées de blé et de populations de souris. Sur ces données réelles, les auteurs montrent que le modèle utilisant à la fois les informations provenant du pedigree et des marqueurs est le plus efficace.

L'étude de Calus et Veerkamp (2007) conclut que l'importance de l'effet polygénique varie selon l'héritabilité du caractère étudié. Les précisions de leurs estimations génomiques augmentent en incluant un effet polygénique seulement sur les caractères à faible héritabilité ce qui n'est pas le cas dans l'étude des données bovines laitières françaises. Liu *et al.* (2011) proposent aussi d'ajouter un effet polygénique résiduel dans une évaluation de type GBLUP, sur les taureaux Holstein de la population d'EuroGenomics, afin d'estimer par les informations de parenté classique ce qui n'a pas pu l'être par les marqueurs SNP.

Combiner les informations phénotypiques, génomiques et de pedigree est un véritable défi dans le cadre de la sélection génomique. 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, ou ABCG2 (Cohen-Zinder *et al.,* 2005) sur le chromosome 6, le nombre de mutations causales identifiées chez les bovins laitiers reste encore très réduit. Le nombre de QTL avérés sur les caractères issus des évaluations bovines reste donc très limité. En pratique, on dispose d'une localisation plus ou moins fine sur le génome, contenant un ou plusieurs gènes candidats, ainsi qu'une estimation de leur effet. Dans certains cas, les modèles d'évaluation génomique comme les modèles BLUP-QTL et SAM tirent profit de cette information et conduisent à des estimations plus fiables. Une méthode de sélection génomique telle que le BLUP-QTL repose pourtant sur des QTL pré-identifiés. Ainsi, l'apparition des données SNP et le développement des méthodes de prédiction génomique n'ont pas diminué l'intérêt porté aux méthodes de détection de QTL dans le domaine de l'amélioration génétique. Démontrer que les méthodes d'évaluation génomique permettent de détecter efficacement ces QTL sans passer par une phase d'analyse LDLA, longue et coûteuse en temps de calcul et en moyens informatiques, est un enjeu important pour établir un modèle de sélection optimal, combinant à la fois les informations des QTL et des marqueurs, ainsi que l'information phénotypique de l'animal et le pedigree.

Chapitre 6 Discussion générale et perspectives

La sélection génomique est basée sur l'estimation de la valeur génétique d'un ensemble d'individus, à partir de l'information génomique de milliers de marqueurs moléculaires. Le principal avantage à utiliser de l'information moléculaire est de pouvoir prédire la valeur génomique d'un animal dès sa naissance, sans attendre de connaître ses performances ni celles de ses apparentés. L'aléa de méiose ne peut pas être déduit de la valeur génétique des parents alors que la génomique permet d'en avoir une estimation précoce. Les reproducteurs potentiels peuvent être sélectionnés de façon beaucoup plus fiable par sélection génomique que sur la base d'une sélection traditionnelle sur ascendance et aussi efficacement que sur la base d'un testage sur descendance. De plus, le testage sur descendance, même s'il implique une estimation fiable des valeurs génétiques des mâles, tend à allonger l'intervalle de génération et s'accompagne de coûts de testage très élevés. Chez les ruminants laitiers particulièrement, l'utilisation des évaluations génomiques permet de réduire les coûts des schémas de sélection tout en augmentant le progrès génétique car elle diminue l'intervalle de génération.

L'impact de la sélection génomique est très variable selon les espèces considérées et les caractères d'intérêt. Il est d'autant plus important quand les caractères sont difficiles ou couteux à mesurer car le nombre de candidats à une sélection classique basée sur ces caractères reste limité. La sélection génomique peut s'appliquer sur des caractères qui ne sont pas directement mesurables sur le candidat : par exemple, si l'expression du phénotype est limitée à un sexe (production laitière), ou s'il nécessite la mort de l'animal (qualité de la viande). Elle est aussi particulièrement utile pour la sélection de caractères peu héritables, impliquant un dispositif de mesures des performances lourd car reposant sur de nombreux descendants, et des caractères parfois exprimés tardivement qui induisent un intervalle de génération plus long.

Actuellement, beaucoup de pays ont adopté le concept de sélection génomique dans leurs évaluations génétiques officielles, chez les bovins laitiers principalement. Depuis 2009, le Canada et les Etats-Unis collaborent afin de développer des évaluations génomiques officielles sur la base des marqueurs de la

puce 54k (Van Doormaal *et al.,* 2009 ; Wiggans *et al.,* 2011). Les Pays-Bas utilisent à la fois les informations de la puce 54k et d'une puce customisée (de Roos *et al.,* 2009). La Nouvelle-Zélande (LIC, 2008) encourage une utilisation large de jeunes « taureaux génomiques » évalués sur la puce 54k. Aujourd'hui, l'Australie, l'Allemagne (Reinhardt *et al.,* 2009), l'Italie et la Suisse (Loberg *et al.,* 2009) sélectionnent différentes races de bovins laitiers sur la base d'évaluations génomiques officielles.

La sélection génomique en France, repose sur le calcul d'index génomiques à partir d'une méthode combinant les avantages de la sélection assistée par marqueurs et des approches de sélection génomique : la « SAM Génomique » ou SAMG. Sa modélisation est proche du BLUP-QTL présenté dans le chapitre 5. Développée dans le cadre du projet AMASGEN, elle permet de prendre en compte l'information des QTL identifiés et cartographiés, depuis de nombreuses années, avec précision, en les introduisant dans le modèle de prédiction sous forme d'haplotypes de 4 à 6 marqueurs. Ces haplotypes sont définis pour suivre de façon précise les transmissions entre générations. Chaque QTL est affecté d'une part de variance estimée sur la population de référence et variant de 15% à 20% selon la race et le caractère considéré. La composante polygénique calculée sur l'information pedigree représente entre 35% et 50% de la variance génétique additive. Enfin, entre 200 et 600 marqueurs SNP, sélectionnés par la méthode Elastic Net et de variances homogènes, sont utilisés pour expliquer la part de variance génétique restante.

Les résultats de la SAM2 étaient déjà suffisamment probants pour justifier dès juin 2009 l'arrêt du dispositif de testage classique dans les races Montbéliarde, Normande et Holstein. Cette date marque le début de la diffusion, en France, de taureaux « génomiques », c'est-à-dire sélectionnés sur la base de leurs index génomiques. Le passage à la SAMG en juin 2010 se traduit par une amélioration importante de la précision des index génomiques (augmentation significative de leur CD). Fritz *et al.* (2010) comparent les précisions des prédictions en calculant les corrélations entre les DYD observés dans la population de validation et les DYD prédits par un modèle basé sur pedigree ou par SAMG. Le gain de précision moyen entre les évaluations sur pedigree et les évaluations SAMG sur l'ensemble des 7 caractères présentés dans cette étude, est de 0,19 en Holstein, 0,11 en Montbéliarde et 0,12 en Normande. Les meilleurs résultats sont obtenus en race Holstein car la taille de la population de référence (plus de 16 000 taureaux réunis grâce à des

collaborations européennes) est nettement plus importante que dans les autres races. Cependant, les récents efforts de génotypage en races Montbéliarde et Normande ont conduit à disposer aujourd'hui d'évaluations génomiques efficaces dans ces deux races également.

De nombreuses études ont été menées afin de développer des méthodes adaptées à l'évaluation génomique des reproducteurs et d'améliorer la qualité des estimations génomiques. Les travaux réalisés dans le cadre du projet AMASGEN et de mon travail de thèse, contribuent à l'enrichissement de cet ensemble de connaissances. Le chapitre 4 compare les méthodes basées sur le BLUP, couramment utilisées pour l'amélioration génétique des animaux, aux méthodes de régression PLS et sparse PLS. Les résultats démontrent la supériorité des méthodes génomiques par rapport à la méthode reposant exclusivement sur l'utilisation d'informations de pedigree. Les résultats du chapitre 4 permettent également de juger de l'impact des poids associés aux données de performances (EDC) et, indirectement, des relations de parenté entre animaux de l'ensemble d'apprentissage et animaux de l'ensemble de validation, sur la qualité des prédictions génomiques. Dans le chapitre 5, l'accent est mis sur l'avantage des méthodes de sélection de variables pour l'interprétation des effets associés aux marqueurs. Il expose également une étude préliminaire sur l'utilisation des méthodes «génomiques » pour la détection de QTL. Il existe un large panel de méthodes statistiques dans la littérature mais aucune approche en particulier ne parvient à s'imposer. Les méthodes bayésiennes sont les plus prometteuses en termes de précision des prédictions mais exigent de gros moyens de calcul ce qui freine leur application en routine. À l'avenir, les méthodes étudiées devront combiner les capacités prédictives des méthodes bayésiennes avec des temps de calcul et des besoins informatiques raisonnables. Cependant, les enjeux ne sont plus seulement d'ordre méthodologique mais reposent sur l'utilisation optimale des données génomiques et l'intégration intelligente de cette technologie dans les nouveaux schémas d'amélioration génétique.

6.1 Comment faire progresser la sélection génomique chez les bovins laitiers ?

Les résultats obtenus au cours de ma thèse, comme l'ensemble de la littérature sur le sujet, mettent en avant les avantages des méthodes d'évaluations génomiques sur les évaluations classiques. Cependant, de nombreux facteurs affectent la qualité des prédictions (Solbert *et al.,* 2010). Au cours de mon travail de thèse, j'ai pu montrer que la taille et la structure de la population de référence avaient un impact important sur la précision des prédictions génomiques. Différentes études sur des données simulées montrent que les deux principaux facteurs sont la taille de la population de référence et la densité de marqueurs (Meuwissen *et al.,* 2001 ; Calus *et al.,* 2008 ; Goddard et Hayes, 2009 ; Daetwyler *et al.,* 2008 ; de Roos, 2011).

6.1.1 En augmentant la taille de la population de référence…

Dans l'article conceptuel de la sélection génomique, Meuwissen *et al.* (2001) se sont interrogés sur l'influence de la taille de la population de référence sur la précision des estimations génomiques en termes de corrélation entre les valeurs génomiques observées et les valeurs génomiques prédites. Quand la taille de la population d'apprentissage est réduite de 2 200 individus à 500, la corrélation diminue de 60%, 21% et 17% pour la méthode des moindres carrés et les approches GBLUP et BayesB respectivement. La régression sur les moindres carrés est nettement plus affectée par la diminution de la taille de la population d'apprentissage que les autres méthodes car elle n'est pas adaptée aux études où le nombre d'observations est très inférieur au nombre de variables. La précision des estimations de la méthode des moindres carrés est en moyenne trois fois moins grande (ρ = 0,22) que la méthode GBLUP (ρ = 0,66) et BayesB (ρ = 0,78), sur les trois tailles de population testées (n = 500, 1 000 et 2 200). Le caractère étudié a une héritabilité simulée de 0,3. Une héritabilité plus élevée aurait nécessité moins d'observations pour obtenir une précision des prédictions génomiques égale (Hayes, 2011).

La taille de la population d'apprentissage joue donc un rôle prépondérant sur la qualité des prédictions. Cependant, au niveau des observations individuelles, on remarque que la précision n'est pas la même pour tous les individus de l'ensemble

de validation. Il a ainsi été montré (Legarra *et al.,* 2008 ; Habier *et al.,* 2010) que plus le candidat est apparenté aux individus de la population de référence et plus sa prédiction génomique est précise. En effet, les relations de parenté proches induisent de très forts déséquilibres de liaison entre QTL et marqueurs. Ces déséquilibres peuvent être rompus à chaque génération à cause des recombinaisons ce qui compromet l'efficacité d'une évaluation génomique sur un grand nombre de générations. D'un point de vue méthodologique, certaines approches (BayesB) semblent légèrement plus robustes que d'autres telles que le GBLUP (Habier *et al.,* 2007, 2010). Dans le chapitre 4, en utilisant la méthode PLS, j'ai montré que lorsque l'ensemble d'apprentissage est exclusivement constitué de taureaux très diffusés, les précisions des prédictions obtenues pour un ensemble de taureaux apparentés, sont dans certains cas, supérieures aux précisions de l'étude sur l'ensemble d'apprentissage complet. Tous ces résultats montrent la nécessité d'entretenir une population de référence relativement jeune et de ré-estimer régulièrement les effets des SNP.

Comme nous l'avons vu dans le chapitre 1, la population de référence ne peut pas être indéfiniment enrichie. Elle est, pour l'instant, limitée aux taureaux testés sur descendance au niveau national. Quand tous les taureaux testés seront génotypés, il deviendra nécessaire de s'intéresser à de nouvelles sources de données provenant par exemple, de collaborations internationales ou de la voie femelle.

Les collaborations internationales. L'enrichissement des ensembles de données pour les populations de référence pour les grandes races internationales passe en premier lieu par les collaborations internationales. En effet, plusieurs pays se sont associés pour partager les génotypages de leurs taureaux et ainsi augmenter la précision de leurs évaluations génomiques nationales en augmentant la taille de leurs populations de référence. La race Holstein a été la première concernée. On distingue, pour cette race, deux grandes collaborations : l'Amérique du Nord (États Unis et Canada), avec une population de référence d'environ 14 000 taureaux, et EuroGenomics (France, Belgique, Danemark, Finlande, Suède, Allemagne et Pays Bas), dont la population de référence a atteint les 18 300 taureaux (David *et al.,* 2010 ; Lund *et al.,* 2010). Chaque pays utilise comme performances les index

Interbull pour les taureaux étrangers. Il a donc fallu développer un système permettant l'échange des données et leurs correspondances entre les génotypages issus de la puce 54k et de la puce customisée des Pays-Bas (Druet *et al.,* 2010). La collaboration européenne réunit le plus grand nombre d'animaux Holstein mais des discussions sont en cours pour étendre les collaborations nord-américaines à d'autres pays. Afin d'établir des collaborations efficaces, il est indispensable d'impliquer des pays qui utilisent des mâles similaires afin de maintenir des connexions entre pays et une distance génétique faible entre les animaux de la population de référence et les animaux candidats.

L'exemple de la race Holstein est particulièrement intéressant car le nombre d'animaux disponibles pour construire les modèles de prédiction devient réellement important tout en répartissant entre les acteurs les coûts de constitution des populations de référence. Dans un tel système, tous les acteurs sont gagnants, et l'outil génomique devient un outil prédictif particulièrement efficace au service des éleveurs. Cet exemple de coopération internationale devrait faire des émules : à titre d'exemple, un accord international similaire, baptisé Intergenomics à récemment permis aux éleveurs de race Brune d'accéder à des prédictions génomiques en constituant une population de référence de plus de 5 000 taureaux génotypés.

L'évaluation génomique des femelles et l'analyse de nouveaux caractères. L'évaluation génomique des femelles est un enjeu important des travaux de sélection génomique. Dans les pays où tous les taureaux avec phénotypes ont déjà été génotypés, comme par exemple, la Nouvelle-Zélande (Spelman *et al.,* 2010), le génotypage des femelles est envisagé pour agrandir la population de référence. En effet, alors que le nombre de taureaux testés sur descendance reste limité, des performances sont contrôlées pour des millions de vaches au sein des troupeaux, et peuvent potentiellement alimenter les populations de référence. En France, depuis 2011, un outil commercial basé sur le génotypage 54k est proposé aux éleveurs afin d'évaluer précocement les caractéristiques d'intérêt des femelles. Si cet outil se développe et s'accompagne d'une diminution des coûts de génotypage, de nombreuses femelles pourront intégrer les populations de référence françaises. Le principal risque serait de ne génotyper que les animaux les plus performants et ainsi, d'introduire un biais dans les évaluations génomiques (Patry et Ducrocq, 2011).

La sélection génomique permet d'avoir des index aussi précis sur les jeunes femelles que sur les jeunes taureaux même pour des caractères faiblement héritables comme la fertilité. Ainsi, un enjeu majeur du génotypage des génisses est d'optimiser le renouvellement des troupeaux et d'augmenter l'intensité de sélection sur la voie femelle. Alors qu'elle est aujourd'hui délaissée au profit d'une sélection importante sur la voie mâle, notamment en raison de la diffusion massive des taureaux d'insémination, l'amélioration de la voie femelle va devenir, grâce aux prédictions génomiques, une source importante de progrès génétique dans les années à venir.

Enfin, l'utilisation de populations de référence femelles permettrait, en la couplant avec une organisation de la collecte de nouveaux phénotypes, de sélectionner les animaux sur de nouveaux caractères qui n'ont pas été enregistrés dans les schémas de testage sur descendance (par exemple, composition fine des laits, caractères de résistance aux mammites, *etc.*). Aujourd'hui, une quarantaine de caractères est recueilli, en routine, avec une grande fiabilité. L'introduction de nouveaux caractères se fera selon les demandes des éleveurs et des filières. Il sera bien sûr, indispensable de disposer de suffisamment d'observations pour une estimation précise des effets des SNP. La principale question est de savoir comment collecter les informations relatives à ces nouveaux caractères ou de récupérer des données existantes mais non utilisées. En France, le programme PhénoFinLait, démarré en 2008, s'intéresse aux composants fins du lait (acides gras et protéines) et sera une source importante de nouvelles données qui pourront être utilisées comme de futurs critères de sélection génomique. Dans le domaine de la santé, les données du carnet sanitaire ou de prophylaxie telle que la paratuberculose, peuvent être récupérées afin de sélectionner des caractères de résistance aux maladies, par exemple. Beaucoup d'autres caractères pourraient être évalués car facilement enregistrables comme par exemple, le comportement alimentaire ou la docilité. L'application des méthodologies de sélection génomique dans le domaine de la santé animale est un enjeu scientifique et industriel important : à terme, cela permettra de sélectionner des animaux plus résistants, et ainsi de réduire considérablement les frais d'élevages tout en maintenant une production efficace pour les éleveurs.

Les évaluations multiraciales. En pratique, les évaluations génomiques sont appliquées sur une population de candidats distincte de la population de référence. Cependant, ces animaux candidats restent issus de la même race mais sont plus jeunes. Théoriquement, ils peuvent aussi provenir d'une lignée différente ou même d'une autre race. Harris *et al.* (2008) montrent que les estimations des effets des SNP calculées à partir d'une population Holstein ne donnent pas des prédictions génomiques précises sur des taureaux de race Jersiaise (corrélations inférieures à 0,3). Pour prédire les valeurs génétiques d'animaux candidats d'une autre race que celle utilisée dans l'établissement des équations de prédiction, le déséquilibre de liaison entre marqueurs et QTL doit être similaire dans la population de référence et dans la population des candidats. En effet, certains allèles aux QTL ne se comportent pas de façon totalement similaire dans des races différentes (Spelman *et al,.* 2002 ; Kaupe *et al.,* 2004 ; Dunner *et al.,* 2003). Une solution est alors d'utiliser une population de référence multiraciale composée d'individus de toutes les races visées. Une mutualisation des données des animaux génotypés et phénotypés est donc à privilégier pour des populations de taille réduite ou moyenne, ou pour des caractères difficiles à phénotyper sur un grand nombre d'animaux d'une même race. L'objectif premier est de maximiser l'efficacité de la sélection génomique et de diminuer le coût global (Hayes *et al.,* 2009c ; Harris et Johnson, 2010 ; Kizilkaya *et al.,* 2010 ; de Roos *et al.,* 2009).

En France, une approche multiraciale est appliquée dans le cadre du projet ANR GEMBAL («GEnomique Multi-race des Bovins Allaitants et Laitiers»), rassemblant l'INRA et les filières : la valeur génétique d'un candidat d'une race donnée est prédite non seulement à partir de la population de référence de sa race mais également des populations de référence de toutes les autres races étudiées. Si les travaux en cours sur l'évaluation multiraciale sont concluants, un individu d'une race donnée pourrait alors être évalué à partir des populations de référence de toutes les races bovines pour lesquelles le caractère est disponible. Cette approche est préconisée, chez les races à petits effectifs chez les bovins allaitants car l'insémination artificielle y est très peu développée donc les populations de référence devront être de grande taille pour compenser la distance génétique trop forte entre les animaux de l'ensemble de référence et les candidats à la sélection. Le développement des approches multiraciales sera bénéfique pour les races à petits effectifs mais aussi pour les plus grandes populations.

En pratique, les résultats observés sont en général décevants lorsque les génotypes sont issus d'une puce 54k car la densité de marqueurs utilisée est suffisante intra-races mais pas entre races (Hayes *et al.,* 2009c). De Roos *et al.* (2008) analysent l'étendue du DL intra et entre races et concluent que pour des races très divergentes, 300 000 SNP sont nécessaires afin d'obtenir un nombre suffisant de marqueurs communs à deux races bovines différentes. Les études d'Ibánēz-Escriche *et al.* (2009) et Toosi *et al.* (2010) montrent que l'utilisation d'une population multiraciale n'est efficace que si la densité de marqueurs est suffisante pour compenser la divergence entre les populations. C'est pour cela qu'a été développée une puce à plus haute densité (« BovineHD® ») de plus de 777 000 marqueurs soit 259 SNP par Mb au lieu de 18 SNP par Mb pour la puce 54k (« BovineSNP50® »). Cette nouvelle puce doit permettre la constitution d'une population de référence multiraciale car la densité en marqueurs devrait être suffisante pour observer des déséquilibres de liaison inter-races et non plus seulement intra-races. Ainsi des évaluations génomiques pourront être développées dans chaque race de la population de référence multiraciale mais seulement sur des caractères mesurés de façon comparable dans les différentes races. Des premières évaluations génomiques, sous couvert d'efforts de génotypage dans les autres races laitières, devraient être disponibles fin 2012 grâce à ces travaux.

6.1.2 En utilisant des puces à SNP de densités différentes…

La puce HD d'Illumina (« BovineHD® », Illumina, 2010b) contient 777 962 SNP. Elle permet de mieux tracer les loci responsables des différences génétiques entre individus. Cependant, au sein d'une même race, les gains attendus, au niveau de la précision des évaluations génomiques, ne sont pas très importants (VanRaden et Tooker, 2010). Solberg *et al.* (2006) simulent une population de taille efficace égale à 100 pour l'étude de l'impact de la densité des marqueurs sur la précision des évaluations génomiques par la méthode BayesB. Ils montrent que, si les SNP sont espacés de 0,5 cM, la précision est 20 fois supérieure à un ensemble de marqueurs espacés de 4 cM. Cependant, les approches bayésiennes, déjà très couteuses en temps de calcul, peuvent devenir inapplicables à cause de la forte augmentation du nombre de paramètres à estimer. De ce fait, il est probable que les méthodes de sélection de variables soient les mieux adaptées au traitement de très gros ensembles de données.

Dans le contexte d'une étude multiraciale, une plus grande densité de marqueurs permet de détecter plus efficacement les segments chromosomiques conservés entre races afin de relier par des « ponts génomiques » les populations entre elles. Mais, pour des raisons de coût, il n'est pas possible de génotyper à haut-débit tous les individus de la population de référence multiraciale, ni tous les candidats à la sélection. Se poseront aussi les problèmes de volume de données à générer, stocker et traiter, ainsi que le problème d'hétérogénéité entre animaux. Toutefois, l'imputation, c'est-à-dire l'estimation statistique des données manquantes, à partir des données existantes de marqueurs et du déséquilibre de liaison dans la population, doit permettre de résoudre cette dernière difficulté. En pratique, une centaine d'individus choisis de façon à représenter au mieux les caractéristiques génétiques dans chaque race et qui permettent de décrire la structure du génome intra et inter-races (déséquilibre de liaison, fréquences alléliques) sont génotypés à haut débit et le reste avec des puces meilleurs marchés. En France, le projet GEMBAL va permettre de génotyper plus de 4 500 taureaux de 18 races avec la puce HD. En race Holstein, des premiers travaux montrent que la qualité d'imputation semble bonne (99% environ d'imputation exacte) à partir d'un premier échantillon de 500 taureaux génotypés à l'aide de la puce HD. Ces résultats sont à confirmer dans les différentes races. De plus, la façon optimale de prendre en compte l'hétérogénéité de la qualité des génotypes (imputés ou connus sans erreur) n'est pas encore clairement définie.

L'imputation permet aussi d'étendre l'utilisation de puces à basses densités de 3 000 ou 7 000 marqueurs SNP environ (Illumina, 2010a). L'arrivée de ces puces à basses densités a augmenté le nombre d'animaux génotypés : elles pourraient également remplacer les marqueurs microsatellites dans la vérification des filiations. Habier *et al.* (2010) démontrent que la prédiction des valeurs génétiques pourrait être aussi précise à partir d'une imputation de la puce 3k à la puce 54k que sur la population de validation génotypée sur une puce 54k directement. Weigel *et al.* (2010) montrent que la précision des imputations est limitée (<0,80) quand moins de 1 000 SNP sont utilisés mais grande (0,95) quand 4 000 SNP sont utilisés. Les résultats de Weigel *et al.* (2010) suggèrent que les estimations génomiques sont 5 fois plus précises sur les données imputées à partir d'une puce de 3 000 SNP que sur les 3 000 SNP seuls. Dassonneville *et al*. (2011) observent des taux d'erreurs

d'imputation de la 3k à la 54k inférieurs à 5% sur des données Holstein, ainsi qu'une précision des prédictions inférieures de seulement 0,02 en corrélation, en utilisant l'imputation sur une puce 3k au lieu des SNP de la puce 54k. Cette stratégie est particulièrement utile pour un génotypage à moindre coût, des femelles ou dans les races à petits effectifs qui disposent pour l'instant, de moyens limités.

L'évolution ultime du génotypage en termes de densité de marqueurs correspond au reséquençage total des génomes des animaux. Encore illusoire il y a quelques années, l'évolution incroyable des technologies de reséquençage et la très forte baisse des coûts associés à ces techniques permettent d'envisager aujourd'hui de reséquencer certains animaux afin de repérer les SNP correspondant aux mutations causales ou fortement liés à elles (Meuwissen et Goddard, 2010). L'identification de ces SNP conduirait à des évaluations génomiques plus précises tout en étant basées sur un nombre réduit de marqueurs. Plusieurs projets sont actuellement en cours pour reséquencer des animaux de différentes races dans ce but. Le projet 1 000 génome bovin (http://1000bullgenomes.com/), initié en 2011 a pour objectif de fournir à la communauté scientifique une large base de données constituée des génomes de taureaux et de vaches reséquencés. Outre la recherche de mutations causales associées aux caractères d'intérêt, ou l'identification de particularités génétiques rares, l'objectif de ce projet international est de pouvoir à terme imputer le génome complet de tous les animaux génotypés à l'aide de puces à ADN. Ce projet ambitieux fournira une base de données d'une richesse incomparable à la communauté scientifique.

6.2 Impact des évaluations génomiques sur les schémas de sélection

La sélection génomique va permettre un progrès génétique rapide et important mais elle doit être mise en place de manière raisonnée. Les travaux de Colleau *et al.* (2009) s'intéressent à l'utilisation optimale de la SAMG sur une population simulée soumise à une sélection sur index de synthèse de caractères laitiers et fonctionnels, pendant 20 ans. Le premier scénario étudié propose de continuer le testage sur descendance en utilisant les index SAM pour le choix des taureaux à mettre en testage. À ce scénario de référence, sont comparés des modèles de sélection basés sur des évaluations génomiques. L'objectif est de choisir

un scénario conduisant à une augmentation du progrès génétique annuel par rapport au scénario de référence tout en conservant un maximum de variabilité génétique. Il est ainsi conseillé d'utiliser de façon équilibrée un plus grand nombre de reproducteurs afin que le renforcement du progrès génétique soit sans conséquence sur la consanguinité. Le principal défi est donc la gestion correcte de la variabilité génétique, qui pourrait être intéressante dans le futur et, notamment, dans l'étude de nouveaux caractères. De plus, malgré une fiabilité des estimations génomiques élevée, les index représentent seulement des estimations des valeurs génétiques vraies des animaux et doivent donc être utilisés avec précaution.

En pratique, cela se traduit par la diffusion étendue d'une large gamme de pères à taureaux sélectionnés sur leurs index génomiques. Au sein d'un élevage, il est préférable d'utiliser 5 jeunes taureaux différents plutôt que 5 fois le même taureau même si son index génomique est bon. Dans ce but, certaines entreprises de sélection proposent, à présent, des lots de plusieurs taureaux de même profil. Cela limite l'utilisation exagérée d'un reproducteur et signe la fin du « star system » longtemps en vigueur chez les bovins laitiers.

6.3 Développement de la sélection génomique dans les autres espèces

La sélection génomique a d'abord été étudiée sur des données bovines laitières françaises, pour profiter des connaissances et des moyens importants de cette filière. Des puces à ADN de 50 à 60 000 SNP existent également chez le cheval, le porc, le mouton, la poule et le chien. Dans les filières ovines, porcines et chez les volailles, un impact majeur de la sélection génomique pourrait être d'augmenter le gain génétique pour des caractères difficiles à sélectionner tels que la résistance à certaines maladies et la qualité de la viande. Le principal avantage de la sélection génomique est le raccourcissement de l'intervalle de génération de par une évaluation précoce des reproducteurs et ce en maintenant ou en augmentant la précision des valeurs génomiques. Les retombées de la sélection génomique sont très variables selon les espèces. Les animaux ayant des cycles de reproduction longs, ont aussi une forte valeur commerciale : ainsi, le coût du génotypage est amorti et la constitution d'une population de référence de taille suffisamment importante est possible.

Pour les petits ruminants, les premiers travaux méthodologiques ont appliqué les méthodes étudiées dans le projet AMASGEN et décrits dans ce manuscrit (notamment la régression PLS, la sparse PLS et l'approche BayesCπ) sur des données issues de la race Lacaune laitière (Duchemin *et al.,* 2012). Les résultats mènent aux mêmes conclusions que chez les bovins laitiers français ce qui laisse supposer que l'application de la sélection génomique en race Lacaune Lait est tout à fait possible et pourrait apporter un gain non-négligeable à la filière. Ces travaux ont été réalisés dans le cadre du projet ANR SheepSNPQTL et du projet Roquefort'In, financé par le Fonds Unique Interministériel (FUI) dont l'objectif est de tester la sélection génomique en situation réelle pour la race Lacaune laitière. Parallèlement au projet Roquefort'In, le projet Genomia (France-Espagne) mutualise les connaissances et les moyens humains et financiers afin de bénéficier aux races franco-espagnoles à effectifs limités telles que les races Manech et Latxa, en testant des approches d'évaluation génomique intra-races et multiraciales.

Sur le plan économique, l'intérêt de cette approche demande à être confirmée. Le projet GENOVICAP (financement CASDAR) a été mis en place en 2010 par l'INRA et l'Institut de l'Elevage afin de modéliser des schémas de sélection ovins laitiers et allaitants mais aussi caprins. Le but sera de comparer des scénarios de sélection classique à des scénarios de sélection génomique afin de mesurer la rentabilité d'une sélection génomique pour ces espèces. Pour cela, les modalités d'application dans chaque filière sont étudiées en tenant compte des spécificités des schémas de sélection en termes d'effectifs, d'organisation, d'utilisation de l'IA et de rentabilité économique. L'objectif final est de déterminer la stratégie la plus appropriée pour chacun des cas : sélection génomique, SAM, sélection phénotypique et les différentes combinaisons de ces méthodologies.

Pour les espèces à cycle plus court, la faisabilité de l'application d'une sélection génomique doit être réfléchie car elle implique des coûts élevés. On peut cependant imaginer pour ces espèces d'utiliser les méthodologies développées en sélection génomique afin de détecter les zones d'intérêt du génome (voir chapitre 5). Ces « QTL » pourraient ensuite être intégrés dans un modèle de type SAM. Les méthodes de sélection de variables permettraient également, pour ces espèces, de sélectionner un ensemble réduit de marqueurs dans le but de créer des puces

customisées, c'est-à-dire contenant un nombre limité de SNP donc pour un coût moins important.

6.4 Conclusion

L'impact de l'utilisation de l'information moléculaire est d'ores et déjà considérable dans le domaine de la génétique animale. La génomique permet d'obtenir une estimation de la valeur génétique d'un potentiel reproducteur relativement précise à un âge très précoce et sans attendre des mesures phénotypiques sur l'animal ou ses apparentés. La France a été un des premiers pays à mettre en place une diffusion directe des jeunes taureaux «génomiques» dans les trois principales races bovines laitières françaises, en tirant profit du programme SAM démarré en 2001. Le projet AMASGEN, notamment au travers de mon travail de thèse, a rempli ses objectifs en améliorant les outils existants de façon à fournir des informations de plus en plus précises aux sélectionneurs. Les conséquences de ce projet sont nombreuses : diffusion différente des taureaux sélectionnés, travail de collecte des données amplifié, et potentiellement, progrès génétique doublé. Elles concernent aussi les autres maillons du dispositif génétique que sont le contrôle des performances et la certification des filiations. Il est très important de maintenir le contrôle des performances afin de garantir l'efficacité des évaluations génomiques sur le long terme et de renouveler les populations de référence pour maintenir une distance génétique faible entre les animaux de la population de référence et les animaux candidats. Les nouveaux projets de sélection génomique en bovins laitiers donnent maintenant la priorité aux nouveaux caractères, à la mise en place d'une évaluation multiraciale et à l'adaptation des pratiques de sélection. De plus, ces nouveaux outils bénéficieront aussi aux races bovines allaitantes et aux races bovines à effectifs modestes. On peut cependant s'attendre à un gain très variable selon les espèces et les caractères étudiés. Dans un avenir très proche, d'autres filières seront impactées par la sélection génomique notamment chez les ovins et les caprins laitiers. Toutes les espèces d'élevage pourront potentiellement profiter des avantages de la sélection génomique sous couvert de pouvoir constituer une population de référence de taille suffisante à des coûts abordables.

Liste des figures

Liste des tableaux

Liste des travaux

Articles Scientifiques :

Colombani C., Legarra A., Fritz S., Guillaume F., Croiseau P., Ducrocq V., Robert-Granié C.,

Application of Bayesian Lasso and BayesCPi for genomic selection in French Holstein and Montbéliarde breeds, soumis à Journal of Dairy Science (2nde lecture)

Duchemin S., **Colombani C**., Legarra A., Baloche G., Larroque H., Astruc J.M., Barillet F., Robert-Granié C., Manfredi E.,

Genomic Selection in French Lacaune dairy sheep breed. J. Dairy Sci. 95 :2723–2733

Colombani C., Croiseau P., Fritz S., Guillaume F., Legarra A., Ducrocq V., Robert-Granié C.,

A comparison of PLS and Sparse PLS regressions in genomic selection in dairy cattle. J. Dairy Sci. 95 :2120–2131

Croiseau P., Legarra A., Guillaume F., Fritz S., Baur A., **Colombani C**., Robert-Granié C., Boichard D., Ducrocq V.,

Fine tuning genomic evaluations in dairy cattle through SNP pre-selection with Elastic-Net algorithm. Genet. Res., Camb. (2011), 93, pp. 409–417.

Communications à des congrès :

Colombani C., Legarra A., Croiseau P., Fritz S., Guillaume F., Ducrocq V., Robert-Granié C. (2011)

Bayes Cπ vs GBLUP, PLS regression, Sparse PLS and Elastic Net: Genomic Selection in French dairy cattle, EAAP annual meeting 2011, Stavanger, Norvège, 29 Août- 2 Septembre 2011. (Poster, résumé 1 page)

Croiseau P., Hozé C., Fritz S., Guillaume F., **Colombani C.,** Legarra A, Baur A., Robert-Granié C., Boichard D., Ducrocq V. (2011)

Description of the French genomic evaluation approach, EAAP annual meeting 2011, Stavanger, Norvège, 29 Août- 2 Septembre 2011.

Colombani C., Croiseau P., Hozé C., Fritz S., Guillaume F., Boichard D., Legarra A., Ducrocq V., Robert-Granié C. (2011)

Could Genomic Selection methods be efficient to detect QTL? A study in French Dairy cattle, 15ème QTLMAS Workshop, Rennes, France, 19-20 Mai 2011. (Présentation orale, résumé 1 page)

Colombani C. (2011)

Genomic Selection : principles, current applications and perspectives, European College of Bovine Health Management, Toulouse, France, 11-15 Avril 2011. (Présentation orale)

Colombani C., Legarra A., Croiseau P., Guillaume F., Fritz S., Ducrocq V., Robert-Granié C. (2010)

Application of PLS and Sparse PLS regression in genomic selection. 9ème WCGALP, Leipzig, Allemagne, 1-6 Août 2010. (Présentation orale, article 4 pages)

Croiseau P., **Colombani C**., Legarra A., Guillaume F., Fritz S., Baur, A., Dassonneville R., Patry C., Robert-Granié C., Ducrocq V. (2010)

Improving genomic evaluation strategies in dairy cattle through SNP pre-selection. 9ème WCGALP, Leipzig, Allemagne, 1-6 Août 2010.

Guillaume F., Fritz S., Legarra A., Croiseau P., Robert-Granié C., **Colombani C.**, Patry C., Boichard D., Ducrocq V. (2009)

Modèles d'évaluations génomiques : Application aux populations bovines laitières françaises. 16ème Rencontres Recherches Ruminants, Paris, France, 2-3 Décembre 2009.

Bibliographie

Abdi H. 2007. PLS-Regression. Neil Salkind (Ed) : Encyclopedia of Measurement and statistics. Thousand Oaks : Sage.

Abdi H. 2010. Partial least squares regression and projection on latent structure regression (PLS Regression).Wiley Interdisciplinary Reviews: Computational Statistics. 2(1):97-106.

Andersson L. et M. Georges. 2004. Domestic-animal genomics: deciphering the genetics of complex traits. Nat Rev Genet. 5(3):202-212.

Barillet F. et B. Bonaïti. 1992. La production laitière des ruminants traits. INRA Prod. Anim. Hors-série Eléments de génétique quantitative et application aux populations animales. 117-121.

Bastiaansen J., A. Coster, M. Calus, J. van Arendong et H. Bovenhuis. 2012. Long-term response to genomic selection : effects of estimation method and reference population structure for different genetic architectures. Genet. Sel. Evol. 44:3.

Blott S., J.J. Kim, S. Moisio, A. Schmidt-Küntzel, A. Cornet *et al.* 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. et B. Bonaïti. 1987. Genetic parameters for first lactation dairy traits in friesian, Montbéliarde and Normande breeds. Genet. Sel. Evol 19(3):337-350.

Boichard D. et E. Manfredi. 1994. Genetic analysis of conception rate in French Holstein cattle. Acta Agriculturae Scandinavica Section a-Animal Science 44(3):138-145.

Boichard D., S. Fritz, M.N. Rossignol, M.Y. Boscher, A. Malafosse *et al.* 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., C. Grohs, F. Bourgeois, F. Cerqueira, R. Faugeras *et al.* 2003. Detection of genes influencing economic traits in three French dairy cattle breeds, Genet. Sel. Evol. 35:77–101.

Browning S.R. et B.L. Browning. 2007. Rapid and accurate haplotype phasing and missing data inference for whole genome association studies using localized haplotype clustering. Am. J. Hum. Genet., 81:1084–1097.

Browning S.R. 2008. Missing data imputation and haplotype phase inference for genome-wide association studies. Hum. Genet. 124:439–450.

Calus M.P.L. et R.F.Veerkamp. 2007. Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. J. Anim. Breed. Genet. 124 (2007) 362–368

Calus M.P.L., T.H.E. Meuwissen, A.P.W. de Roos et R.F. Veerkamp. 2008. Accuracy of genomic selection using different methods to define haplotypes. Genetics. 178 :553-561

Chun H. et S. Keles. 2010. Sparse Partial Least Squares Regression for Simultaneous Dimension Reduction and Variable Selection. J.R. Statist. Soc. B 72:3-25.

Cohen-Zinder M., E. Seroussi, D.M. Larkin, J.J. Loor, A.E. van der Wind et al. 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.

Cole J.B., P.M. VanRaden, J.R. O'Connell, C.P. Van Tassell, T.S. Sonstegard et al. 2009. Distribution and location of genetic effects for dairy traits. J. Dairy Sci. 92(6):2931-2946.

Colleau J.J, Y. Heyman et J.P. Renard. 1998. Les biotechnologies de la reproduction chez les bovins et leurs applications réelles ou potentielles en sélection. INRA Prod. Anim. 11: 41-56.

Colleau J.J, S. Fritz, F. Guillaume, A. Baur, D. Dupassieux et al. 2009. Simulating the potential of genomic selection in dairy cattle breeding. Renc. Rech. Ruminants. 16 :419.

Coster A., J.W.M. Bastiaansen, M.P.L. Calus, J.A.M. van Arendonk et H. Bovenhuis. 2010. Sensitivity of methods for estimating breeding values using genetic markers to the number of QTL and distribution of QTL variance. Genet. Sel. Evol. 42:9.

Croiseau P., A. Legarra, F. Guillaume, S. Fritz, A. Baur, C. Colombani, C. Robert-Granié, D. Boichard et V. Ducrocq. 2011. Fine tuning genomic evaluations in dairy cattle through SNP pre-selection with the Elastic-Net algorithm. Genet. Res. 93:409-417.

Daetwyler H.D., B. Villanueva et J.A. Woolliams. 2008. Accuracy of predicting the genetic risk of disease using a genome-wide approach. Plos One. 3(10):e3395.

Dassonneville R., R.F. Brondum, T. Druet, S. Fritz, F. Guillaume et al. 2011. Effect of imputing markers from a low-density chip on the reliability of genomic breeding values in Holstein populations. J. Dairy Sci. 94:3679-3686.

David X., A. de Vries, E. Feddersen et S. Borchersen. 2010. International genomic cooperation: EuroGenomics significantly improves reliability of genomic evaluations. Interbull Bull. 41

Dekkers J.C.M. 2004. Commercial application of marker- and gene-assisted selection in livestock : Strategies and lessons. J. Anim. Sci. 82: E313-E328.

De los Campos G., H. Naya, D. Gianola, J. Crossa, A. Legarra *et al.* 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics. 182:375-385.

Dempster A.P., N.M. Laird et D.B. Rubin. 1977. Maximum Likelihood from Incomplete Data via the EM algorithm. In: Journal of the Royal Statistical Society 39(1):pp. 1–38.

Denham M.C. 2000. Choosing the number of factors in PLS regression: estimating and minimizing the mean squared error of prediction. J. Chemometr. 14:351-361.

De Roos, A.P.W., B.J. Hayes, R. Spelman et M.E. Goddard. 2008. Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. Genetics. 179:1503-1512.

De Roos, A.P.W., C. Schrooten, E. Mullaart, S. van der Beek, G. de Jong et W. Voskamp. 2009. Genomic selection at CRV. Interbull Bull. 39:47–50.

De Roos S. 2011. Genomic selection in dairy cattle. PHD Thesis, Wageningen University.

Druet T., S. Fritz, M. Boussaha, S. Ben-Jemaa, F. Guillaume *et al.* 2008. Fine mapping of quantitative trait loci affecting female fertility in dairy cattle on BTA03 using a dense single-nucleotide polymorphism. Genetics 178:2227-2235.

Druet T. et M. Georges. 2009. A Hidden Markov Model Combining Linkage and Linkage Disequilibrium Information for Haplotype Reconstruction and Quantitative Trait Locus Fine Mapping. Genetics 184(3):789-U237.

Druet T. et M. Georges. 2010. A hidden markov model combining linkage and linkage disequilibrium information for haplotypes reconstruction and quantitative trait locus fine mapping. Genetics 184:789–798.

Druet, T., C. Schrooten, et A. P. W. de Roos. 2010. In silico genotyping of thousands of SNPs in dairy cattle for the EuroGenomics project. Commun. No. 0137 in Proc. 9th World Congr. Genet. Appl. Livest. Prod., Leipzig, Allemagne.

Duchemin S.I., C. Colombani, A. Legarra, G. Baloche, H. Larroque, et al. 2012. Genomic selection in the French Lacaune dairy sheep breed. J. Dairy Sci. 95:2723–2733.

Dunner S., M.E.Miranda, Y. Amigues, J.Canon, M. Georges *et al.* 2003. Haplotype diversity of the myostatin gene among beef cattle breeds. Genet. Sel. Evol. 35:103-118

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

Fikse W.F. et G. Banos. 2001. Weighting factors of sire daughter information in international genetic evaluations. J. Dairy Sci. 84:1759-1767.

Fischer R.A. 1918. The correlation between relatives on the supposition of Mendelian inheritance. Trans. R Soc. Edimbourg. 52:399-433.

Frank I et J. Friedman. 1993. A statistical view of some chemometrics regression tools (with discussion). Technometrics. 35(2):109-148.

Fritz S., T. Druet, F. Guillaume, M.Y. Boscher, A. Egger *et al.* 2007. Bilan du programme de Sélection Assistée par marqueurs dans les trois principales races bovines laitières françaises et perspectives d'évolution. Renc. Rech. Rum. pp 129-132.

Fritz S., F. Guillaume, P. Croiseau, A. Baur, C. Hozé, et al. 2010. Mise en place de la sélection génomique dans les trois principales races françaises de bovins laitiers. 17ème Rencontres Recherches Ruminants. Paris. Décembre 2010.

Fu W. 1998. Penalized regression: the bridge versus the lasso. J. Comput. Graph. Statist. 7 :397–416.

Geman S. et D. Geman.1984. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. IEEE Transactions on Pattern Analysis and Machine Intelligence 6 (6): 721–741

Gianola D., G. de los Campos, W.G. Hill, E. Manfredi et R. Fernando. 2009. Additive Genetic Variability and the Bayesian Alphabet. Genetics 183(1):347-363.

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

Goddard M.E. et B.J. Hayes. 2009. Mapping genes for complex traits in domestic animals and their use in breeding programs. Nature Reviews Genet. 10:381-391.

González-Recio O., D. Gianola, G.J. Rosa, K.A. Weigel et A. Kranis. 2009. Genome-assisted prediction of a quantitative trait measured in parents and progeny: application to food conversion rate in chickens. Genet. Sel. Evol. 41:3.

Gonzalez-Recio O. et S. Forni. 2011. Genome-wide prediction of discrete traits using bayesian regressions and machine learning. Genet. Sel. Evol. 43:12.

Gredler B., K.G. Nirea, T.R. Solberg, C. Egger-Danner, T.H.E. Meuwissen et J. Solkner. 2009. Genomic Selection in Fleckvieh/Simmental -First results. In proceedings of the Interbull Meeting. Barcelone, Espagne, 21-24 Août, 2009.

Gredler B., H. Schwarzenbacher, C. Egger-Danner, C. Fuerst, R Emmerling *et al.* 2010. Accuracy of genomic selection in dual purpose Fleckvieh cattle using three typer of methods and phenotypes. 9th World Congress on Genetics Applied to Livestock Production, Leipzig, Allemagne.

Grimard B., S. Freret, A. Chevallier, A. Pinto, C. Ponsart *et al.* 2006. Genetic and environmental factores influencing first service conception rate and late embryonic/fœtal mortality in low fertility herds. Anim. Reprod. Sci. 91 (1-2) : 31-44.

Grisart B., F. Farnir, L. Karim, N. Cambisano, J.J. Kim et al. 2004. Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. Proc. Natl. Acad. Sci. U. S. A. 101(8):2398-2403.

Guillaume F., S. Fritz, D. Boichard et 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.

Guo S.W. et E.A. Thompson. 1992. Performing the exact test oh Hardy-Weinberg proportion for multiple alleles. Biometrics. 48:361-372.

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

Habier D., R.L. Fernando, K. Kizilkaya et D.J. Garrick. 2010. Extension of the Bayesian alphabet for genomic selection. 9th World Congress on Genetics Applied to Livestock Production, Leipzig, Allemagne.

Habier D., R.L. Fernando, K. Kizilkaya et D.J. Garrick. 2011. Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics 12:12.

Hardy G. 1908. Mendelian proportions in a mixed population. Sciences. 28:49-50.

Harris B.L, D.L. Johnson et R.J. Spelman. 2008. Genomic selection in New Zealand and the implication for national genetic evaluation. Proc. Interbull Meeting. Niagara falls, Canada.

Harris B.L., D.L. Johnson et R.J. Spelman. 2009. Genomic selection in New Zealand and the implications for national genetic evaluation. Pp 325-330 in Identification, Breeding, Production, Health and Recording of Farm Animals. Proceedings of the 36th ICAR Biennial Session, Niagara Falls, USA, 16-20 Juin 2008. International Committee for Animal Recording (ICAR).

Harris B.L. et D.L. Johnson. 2010. Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. J. Dairy Sci. 93:1243–1252

Hastie T., R. Tibshirani et J. Friedman. 2001. The elements of statistical learning. Data mining, inference and prediction. Springer Series in Statistics, New York.

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

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

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

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

Hayes B.J. 2011. Course Notes: Genomic Selection. 12-16 Sept 2011, Toulouse, France.

Hazel L.N. 1943. The genetic basis for constructing selection indexes. Genetics. 28:476-490.

Henderson C.R., O. Kempthorne, S.R. Searle et C.M. von Krosigk. 1959. "The Estimation of Environmental and Genetic Trends from Records Subject to Culling". Biometrics (International Biometric Society) 15 (2): 192–218.

Henderson C.R. 1973. Sire evaluation and genetic trend. Proceedings of Animal Breeding and Genetics Symposium in Honor of Dr J.L. Lush, Blackburgh, Virginie, Août 1972, pp 10-41.

Huang X., W. Pan, S. Park, X. Han, L.W. Miller *et al.* 2004. Modeling the relationship between lvad support time and gene expression changes in the human heart by penalized partial least squares. Bioinformatics. 20: 888-894.

Ibánēz-Escriche N., R.L. Fernando, A. Toosi et J.C.M. Dekkers. 2009. Genomic selection of purebreds for crossbred performance. Genet. Sel. Evol. 41:12

Illumina. 2010a. GoldenGate Bovine3K Genotyping BeadChip. 14 Décembre 2010. http://www.illumina.com/Documents/products/datasheets/datasheets_bovine3K.pdf.

Illumina. 2010b. BovineHD Genotyping BeadChip. 14 Décembre 2010. http://www.illumina.com/Documents/products/datasheets/datasheets_bovineHD.pdf.

Kaupe B, A. Winter, R. Fries et G. Erhardt. 2004. DGAT1 polymorphism in Bos taurus and Bos indicus cattle breeds. J. Dairy Res. 71: 182-187

Kizilkaya K., R.L. Fernando et D.J. Garrick. 2010. Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. J. Anim. Sci. 88(2):544-551.

Lande R. et R. Thompson. 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics. 124:743-756

Lê Cao K.A., D. Rossouw, C. Robert-Granie et P. Besse. 2008. A Sparse PLS for Variable Selection when Integrating Omics Data. Stat. Appl. Genet. Mol. Biol. 7(1):32.

Lê Cao K.A., I. Gonzalez et S. Dejean. 2009. integrOmics: an R package to unravel relationships between two omics datasets. Bioinformatics 25(21):2855-2856.

Lê Cao K.A et C. Le Gall. 2011. Integration and variable selection of ‘omics’ data sets with PLS : a survey. J. de la SFdS. 152(2):77-96.

Ledoux D., P. Humblot, F. Constant, A. Pontet et B. Grimard. 2006. Echecs précoces de gestation chez la vache laitière. Point Vet. 37 (numéro spécial reproduction des ruminants):50-55.

Legarra A., D. Roussouw, C. Robert-Granié et P. Besse. 2008. Performance of genomic selection in mice. Genetics. 180:611-618.

Legarra A., I. Aguilar et I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. J. Dairy Sci. 92:4656-4663.

Legarra A., C. Robert-Granie, P. Croiseau, F. Guillaume et S. Fritz. 2011. Improved Lasso for genomic selection. Genet. Res. 93(1):77-87.

LIC. 2008. Inside LIC. C. Bayly, ed. Livestock Improvement Corporation, Hamilton, Nouvelle-Zélande.

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

Loberg, A. et J. W. Dürr. 2009. Interbull survey on the use of genomic information. Interbull Bull. 39:3–13.

Long N., D. Gianola, G.J.M. Rosa et K.A. Weigel. 2011. Dimension reduction and variable selection for genomic selection: Application to predicting milk yield in Holsteins. J. Anim. Breed.Genet. 128:247‑257.

Lorber A., L.E. Wangen et B.R. Kowalski. 1987. A theoretical foundation for the PLS algorithm. J. Chemometr. 1(1):19-31.

Luan T., J.A. Woollimas, S. Lien, M. Kent, M. Svendsen et T.H.E. Meuwissen. 2009. The accuracy of genomic selection in Norwegian red cattle assessed by cross-validation. Genetics. 183:1119-1126.

Lund M.S. A.P.W. de Roos, A.G. de Vries, T. Druet, V. Ducrocq et al. 2010. Improving genomic prediction by EuroGenomics collaboration. Proc 9ème WCGALP. Leipzig, Allemagne.

Macciotta N., G. Gaspa, R. Steri, C. Pieramati, P. Carnier *et al.* 2009. Pre-selection of most significant SNPs for the estimation of genomic breeding values. BMC proceedings. 3(Suppl l):S14.

Marchini J., D. Cutler, N. Patterson, M. Stephens, E. Eskin *et al.* 2006. A comparison of phasing algorithms for trios and unrelated individuals. Am. J. Hum. Genet., 78:437–450.

Metropolis N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller et E. Teller. 1953. Equation of state calculation by fast computing machines. J. Of Chemical Physics. 21(6):1087-1092.

Meuwissen T.H.E et M.E. Goddard. 1996. The use of marker haplotypes in animal breeding schemes. Genet. Sel. Evol. 28:161-176.

Meuwissen T.H.E. et M. Goddard. 2001. Prediction of identity by descent probabilities from marker-haplotypes. Genet. Sel. Evol. 33:605-634.

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

Meuwissen T.H.E., T. Solber, R. Sheperd et J. Woolliams. 2009. A fast algorithm for bayesb type of prediction of genome-wide estimates of genetic value. Genet. Sel. Evol. 41(1):2.

Meuwissen T.H.E., 2009. Accuracy of breeding values of unrelated individuals predicted by dense SNP genotyping. Genet. Sel. Evol. 41:35.

Meuwissen T.H.E. et M. Goddard. 2010. Accurate prediction of genetic values for complex traits by whole-genome resequencing. Genetics. 185(2):623-31.

Mevik B.H. and H.R. Cederkvist. 2004. Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR). J. Chemometr. 18(9):422-429.

Misztal I., A. Legarra, et I. Aguilar. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree and genomic information. J. Dairy Sci. 92:4648-4655.

Moser G., B. Tier, R.E. Crump, M.S. Khatkar et H.W. Raadsma. 2009. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genet. Sel. Evol. 41:56.

Moser G., M.S. Khatkar, B.J. Hayes et H.W. Raadsma. 2010. Accuracy of direct genomic values in Holstein bulls and cows using subsets of SNP markers. Genet. Sel. Evol. 42:37.

Park T. and G. Casella. 2008. The Bayesian Lasso. J. Am. Stat. Assoc. 103(482):681-686.

Patry C et V.Ducrocq. 2011. Evidence of biases in genetic evaluations due to genomic preselection in dairy cattle. J Dairy Sci. 94(2):1011-20.

Reinhardt, F., Z. Liu, F. Seefried et G. Thaller. 2009. Implementation of genomic evaluation in German Holsteins. Interbull Bull. 40:219–226.

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

Schaeffer L.R. 2006. Strategies for applying genome-wide selection in dairy-cattle. J. Anim. Breed. Genet. 123:218-223.

Scheet P. et Stephens, M. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotype phase. Am. J. Hum. Genet., 78:629–644.

Shen H.P. and J.H.Z. Huang. 2008. Sparse principal component analysis via regularized low rank matrix approximation. J. Multivar. Anal. 99(6):1015-1034.

Shrimpton A.E. et A. Robertson. 1998. The Isolation of Polygenic Factors Controlling Bristle Score in Drosophila melanogaster. II. Distribution of Third Chromosome Bristle Effects Within Chromosome Sections. Genetics 118: 445-459.

Solberg T. R., A. Sonesson, J. Woolliams et T.H.E Meuwissen. 2006. Genomic selection using different marker types and density. Proc. 8th WCGALP. Belo Horizonte, Brésil.

Solberg T.R., A.K. Sonesson, J.A. Woolliams et T.H.E. Meuwissen. 2009. Reducing dimensionality for prediction of genome-wide breeding values. Genet. Sel. Evol. 41:8.

Solberg T.R. A.K. Sonesson, J.A. Woolliams et T.H.E. Meuwissen. 2010. Genomic selection using different markers types and densities. J. Anim. Sci. 86:2447-2454.

Souverein O.W., A.H. Zwinderman et Tanck, M.W.T. 2006. Multiple imputation of missing genotype data for unrelated individuals. Ann. Hum. Genet. 70:372–381.

Spelman R.J., C.A. Ford, P. McElhinney, G.C. Gregory *et al.* 2002. Characterization of the DGAT1 gene in the New Zealand dairy population. J. Dairy Sci. 85: 3514-3517.

Spelman R.J., J. Arias, M.D. Keehan, V. Obolonkin, A.M. Winkelman, D.L. Johnson et B.L. Harris. 2010. Application of genomic selection in the New Zealand dairy cattle industry. Commun. No. 0311 in Proc. 9th World Congr. Genet. Appl. Livest. Prod., Leipzig, Allemagne.

Stella A., M.M. Lohuis, G. Pagnacco et G.B. Jansen. 2002. Strategies for continual application of marker-assisted selection in an open nucleus population. J Dairy Sci, 85(9):2358–2367.

Stephens M., Smith, N.J. et Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet., 68:978–989.

Sun X., D. Habier, R. Fernando, D. Garrick et J. Dekkers. 2011. Genomic breeding value prediction and QTL mapping of QTLMAS2010 data using Bayesian Methods. BMC Proceedings 5(Suppl 3):S13.

Tenenhaus M. 1998. La regression PLS. Théorie et pratique. Paris : Technip.

Thomsen H., N. Reinsch, N. Xu, C. Looft, S. Grupe *et al.* 2001. Comparison of estimated breeding values, daughter yield deviations and de-regressed proofs within a whole genome scan for qtl. J. of Animal Breeding and Genetics, 118(6):357–370.

Tibshirani R. 1996. Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. Ser. B-Methodol. 58(1):267-288.

Toosi A., R.L. Fernando et J.C.M. Dekkers. 2010. Genomic selection in admixed and crossbred populations. J. Anim. Sci. 88:32-46

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

Van Doormaal B.J., G.J. Kistemaker, P.G. Sullivan, M. Sargolzaei et F.S. Schenkel. 2009. Canadian implementation of genomic evaluations. Interbull Bull. 40:214–218.

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

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

VanRaden P.M., C.P. Van Tassel, G.R. Wiggans, T.S. Sonstegard, R.D. Schnabel, J.K. Taylor et F.S. Schenkel. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92:16-24.

VanRaden P.M. et M.E. Tooker. 2010. Gains in reliability from combining subsets of 500, 5 000, 50 000 or 500 000 genetic markers. J. Dairy Sci. 93(E-Suppl. 1):534

Van Sickle J. 2003. Analyzing correlations between stream and watershed attributes. Journal of the American Water Resources Association 39(3):717-726.

Visscher P.M. et C.S. Haley. 1998. Strategies for marker assisted selection in pig breeding programmes. in Proc. 6th World Cong. Genet. Appl. Livest. Prod., Armidale, Australie., 23:503–510.

Vitezica Z.G., I. Aguilar, I. Misztal et A. Legarra. 2011. Bias in genomic predictions for populations under selection. Genet. Res. 93:357-366.

Weigel K.A., G. de los Campos, O. González-Recio, H. Naya, X.L. Wu *et al.* 2009. Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of SNP markers. J. Dairy Sci, 92:5248-5257.

Weigel K. A., G. de los Campos, A. I. Vazquez, G. J. M. Rosa, D.Gianola *et al.* 2010. Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle. J. Dairy Sci. 93:5423–5435

Weinberg W. 1908. Uber den Nachweis der Vererbung beim Menschen. Jahreshefte Verein f. vaterl. Naturk, in Wurttemberg. 64:368–82.

Weller J.I., Y. Kashi et M. Soller. 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.

Whittaker J.C., R. Thompson et M.C. Denham. 2000. Marker-assisted selection using ridge regression. Genet. Res. 75:249-252.

Wiggans G.R., P.M. VanRaden et T.A. Cooper. 2011. The genomic evaluation system in the United States: Past, present, future. J. Dairy Sci. 94 :3202–3211.

Wigginton J.E., Cutler, D.J., et Abecasis, G.R. (2005). A note on exact tests of hardy-weinberg equilibrium. Am. J. Hum. Genet.. 76 :887–883.

Wold H. 1966. Estimation of principal components and related models by iterative least squares. Pp 391-420 in Multivariates Analysis, Krishnaiah P.R. (ED.), Academic Press, New York.

Xu S. 2003. Estimating polygenic effects using markers of the entire genome. Genetics. 163:789-801.

Zou H. et T.Hastie. 2003. Regression Shrinkage and Selection via the Elastic Net, with Application to Microarrays. Stanford University: Department of Statistics.