ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

AgroParisTech
INSTITUT DES SCIENCES ET INDUSTRIES DU VIVANT ET DE L'ENVIRONNEMENT
PARIS INSTITUTE OF TECHNOLOGY FOR LIFE, FOOD AND ENVIRONMENTAL SCIENCES

Ecole Doctorale ABIES
Agriculture Alimentation Biologie Environnement Santé
Agriculture Food Biology Environment Health

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

L'Institut des Sciences et Industries du Vivant et de l'Environnement

(AgroParisTech)

Spécialité : Génétique Animale

présentée et soutenue publiquement par

Chris HOZE

le 19 Juin 2014

Développement d'évaluations génomiques multiraciales chez les bovins laitiers

Directeur de thèse : **Vincent DUCROCQ**
Co-encadrement de thèse : **Pascal CROISEAU**

Jury

M. Etienne VERRIER, Professeur, AgroParisTech	Président
Mme. Christèle ROBERT-GRANIE, Directrice de recherche, INRA GenPhySE	Rapporteur
M. Leopoldo SANCHEZ, Chargé de recherche, Amélioration Génétique et Physiologie Forestières	Rapporteur
M. Tom DRUET, Research Associate, Université de Liège Unit of Animal Genomics, Belgique	Examinateur
M. Laurent SCHIBLER, Responsable développement et innovation, UNCEIA	Examinateur
M. Pascal CROISEAU, Chargé de recherche, INRA GABI	Examinateur

AgroParisTech
INRA GABI
Domaine de Vilvert, 78350 Jouy en Josas

Ce travail de thèse s'inscrit dans le cadre du projet **GEMBAL** (GEnomique Multiraciale Bovins Allaitants et Laitiers) financé l' **Agence Nationale de la Recherche** (ANR-10-GENM-0014), **APISGENE**, **Races de France** et l'appel à projets **INRA** "AIP Bioressources".







Il a été réalisé au sein de l'**UMT 3G** issue de la collaboration entre l'INRA **UMR 1313 GABI** (Génétique et Biologie Intégrative) et de l'équipe **G2B** (Génétique et Génomique Bovine), l'**UNCEIA** et **l'Institut de l'Elevage**.









Avec le soutien financier de l'**UNCEIA** dans le cadre d'une convention CIFRE (Agence Nationale de la Recherche et de la Technologie, convention n° 188/2011).





Remerciements

Ce document ne serait pas complet sans une pensée pour l'ensemble des personnes que j'ai pu rencontrer au cours de ces trois ans (ou plutôt quatre et demi).

Et oui, c'est en Février 2010 que je suis arrivée à Jouy en Josas pour mon stage de fin d'étude. Je n'aurais jamais pensé à ce moment là, rester à l'INRA et encore moins me lancer dans une thèse. Certains se rappelleront sûrement qu'au départ, j'étais plutôt réticente mais c'était trop tard : j'avais déjà attrapé le virus de la génétique animale… L'opportunité pour moi de valoriser ce travail comme une première expérience professionnelle, le défi représenté par le sujet et surtout ces applications potentielles m'ont fait me lancer dans cette aventure. Aujourd'hui même si je ne suis pas devenue le « Robin des bois de la génomique » (voler à la Holstein pour donner aux autres ;-)), je ne regrette pas mon choix. Cette thèse m'aura permis d'en apprendre beaucoup sur le monde de la génétique bovine mais également sur moi-même.

Je tiens donc à remercier tous ceux qui m'ont aidé dans ce travail :

Jean-Pierre Bidanel puis Claire Rogel-Gaillard pour m'avoir accueilli au sein de l'INRA Jouy en Josas et de l'unité GABI où j'ai découvert un environnement de travail très enrichissant.

Maurice Barbezant puis Xavier David qui m'ont permis de réaliser cette thèse avec l'UNCEIA. Ce lien avec le « terrain » et les professionnels m'a permis de donner un sens pratique à ce travail que je trouvais parfois un peu trop théorique.

Vincent Ducrocq, mon directeur de thèse et Pascal Croiseau mon encadrant, pour avoir su m'apporter leurs conseils et leur point de vue tout au long de ces trois ans tout en me laissant mon autonomie (ou devrais-je dire pour m'avoir laissée n'en faire qu'à ma tête ?). Merci également de m'avoir fait confiance et de m'avoir convaincu de me lancer dans cette aventure. Pascal, j'étais ta première thésarde, j'espère que l'expérience a été aussi bonne pour toi que pour moi. Merci pour ta bonne humeur, ta disponibilité et nos discussions qui m'ont permis de toujours relativiser.

L'ensemble des membres du projet GEMBAL pour m'avoir écouté, conseillé et aidé tout au long de ce projet. Je pense en particulier à Florence Phocas, la coordinatrice du projet, Eric Venot pour l'organisation de la base de données de génotypage, son aide avec le langage AWK et Shell et pour avoir subi sans broncher quelques unes de mes « blagues », Marie-Noëlle Fouilloux avec qui j'ai découvert les joies de la préparation des fichiers de génotypage et appris les codes de presque toutes les races bovines. Les « travailleurs » côté génomique allaitante qui m'ont souvent aidé à percer les mystères de GS3. Didier Boichard, responsable de l'équipe G2B, pour son accessibilité, ses réponses à mes questions et sa réactivité en particulier lors du début de thèse et de l'imputation.

Sébastien Fritz pour m'avoir fait découvrir la génétique et la génomique, pour les échanges autour de mon travail, ses implications pour les professionnels, mais également sur mon avenir professionnel. Merci également de m'avoir appris à « râler ».

François Guillaume pour son aide lors de mes débuts dans la génomique et l'informatique, pour avoir assurer le « service après-vente » même quand tu étais en Belgique et pour les discussions souvent improbables mais toujours divertissantes.

Je tiens également à remercier tous ceux qui, sans avoir été impliqué directement dans ce travail ont participé à son bon déroulement :

Les informaticiens (de GABI et du CTIG), les secrétaires et François Deniau pour leur gentillesse et leur efficacité.

La Dream Team du 211 (Amandine, Bérénice, Clotilde, Julie, Mathilde, Rachel) pour les discussions, les fous rires, les encouragements, la pause thé de 16h et bien sûr les WE filles.

Les occupants des bureaux de l'étage côté « café cochon » pour leurs soutiens, les réponses à mes questions et les craquages de fin de journée. Je pense en particulier à Alexis, Pauline (les thésards encore tout neufs) et Iola (future thésarde?), que j'ai souvent sollicité pour des points vocabulaire ou grammaire mais également pour des questions existentielles. J'aimerai également citer ici les thésards et stagiaires de l' (ex-) côté obscur qui font qu'aujourd'hui on a (presque) plus peur de franchir la porte coupe-feu.

Les membres de l'unité GABI pour l'ensemble des conversations sérieuses ou non qui contribuent à l'ambiance conviviale du bâtiment.

L'ensemble du personnel de l'UNCEIA pour leur accueil et leur gentillesse.

Pour conclure, j'aimerais remercier tous ceux qui, même s'ils ne comprennent pas tout à la génétique, m'ont soutenu dans ce projet.

Mes camarades PA à l'agro Rennes pour l'ambiance et l'accueil fait à la banlieusarde du 9-3 et surtout pour m'avoir conforté dans une idée qui me semblait folle : travailler dans les productions animales.

Ma famille et mes parents qui, je pense, n'ont jamais compris mon choix mais n'ont jamais essayé de m'en dissuader. Merci de vos encouragements.

Romain qui m'a supporté jusqu'au bout car je peux l'avouer maintenant : Garder le sourire et avoir l'air zen ne veux pas forcément dire qu'on ne stresse pas (mais ça aide !).

Je remercie également Etienne Verrier, Christèle Robert-Granié, Leopoldo Sanchez, Tom Druet et Laurent Schibler pour avoir accepté de faire partie de mon jury de thèse.

Enfin je remercie d'avance tous ceux qui liront ce manuscrit.

Résumé

L'efficacité de la sélection génomique étant principalement dépendante de la taille de la population de référence, seules les principales races laitières françaises bénéficient aujourd'hui d'évaluations génomiques. Pour contourner cette contrainte et développer des évaluations génomiques pour les races régionales, il a été proposé de créer une population de référence commune entre races.

Cependant, utiliser une population de référence multiraciale nécessite que le lien entre QTL et marqueurs soit conservé entre races, ce qui implique, sauf cas particulier, l'utilisation d'une puce haute densité. Les taux d'erreur d'imputation des génotypes haute densité à partir de génotypes moyenne densité (classiquement utilisés en bovins) ont été étudiés. La précision d'imputation étant supérieure à 99% dans la majorité des races laitières, l'imputation des génotypes des animaux des populations de référence des principales races laitières a été réalisée. Cette population de référence « haute densité » a ensuite été utilisée pour le développement d'évaluations génomiques multiraciales. Plusieurs stratégies d'évaluations génomiques (intra-race ou multi-race, à partir de génotypes moyenne ou haute densité) ont été comparées pour différentes tailles de populations de référence. Les évaluations multiraciales basées sur la puce haute densité permettent d'améliorer la précision des évaluations génomiques dans le cas d'une population de référence de 500 taureaux ou moins. L'efficacité d'une troisième stratégie utilisant des évaluations génomiques multiraciales à partir de la puce moyenne densité pour un groupe de races proches a donc été étudiée. L'augmentation de la précision des évaluations génomiques observée dans ce cas était trois fois supérieure à celle qui avait été observée dans le cas des principales races laitières. Par ailleurs, dans les deux cas étudiés ici, la précision des évaluations génomiques intra-races est relativement élevée, même dans le cas d'une population de référence réduite, lorsque les pères des candidats à la sélection sont inclus dans la population de référence.

Les résultats obtenus suggèrent donc que la mise en place d'évaluations génomiques dans les races régionales est envisageable. Il faudra toutefois poursuivre les travaux pour déterminer quelle stratégie optimale peut/doit être utilisée dans chacune des races.

Mots clés

Sélection génomique, Bovins laitiers, Evaluations multiraciales, Races régionales

Abstract

Within-breed genomic selection is now implemented in a number of large cattle breeds. However, building reference populations large enough remains a major challenge for smaller breeds. Combining reference populations and implementing a multi-breed approach appears to be an appealing alternative for small breeds.

Such an approach requires conserved linkage disequilibrium across breeds to maintain the association between QTL and markers. Therefore, the use of a high density chip is generally needed. Error rates for high density imputation from medium density genotypes, classically used in cattle, were estimated. The mean error rate was below 1% in most dairy cattle breed which implies that a large high density imputed reference populations can be available for genomic selection at low cost. Reference populations from the three major French dairy breeds were imputed to high density and used to develop a multi-breed genomic evaluation.

Several alternative genomic selection approaches (within-breed or multi-breed, based on medium or high density genotypes) were compared for breeds with different sizes of reference population. Improvement of genomic prediction accuracy due to the multi-breed evaluation was observed for breeds with 500 animals or less in their reference population. Accuracy of a third alternative using multi-breed evaluation based on medium density genotypes of closely related breeds was investigated. The benefit of multi-breed genomic evaluation was then three times higher in this situation than in the one where populations from major dairy cattle breeds were pooled. It can be noted that in both situations, using within-breed genomic information allowed a significant gain in accuracy compared to pedigree-based evaluations even for a small breed, when all sires of selection candidates belong to the reference population.

These results suggest that implementation of genomic selection is feasible in small dairy cattle breeds. However further work is required to determine which optimal strategy can/must be implemented in a given breed.

Key-words

Genomic selection, Dairy cattle, Multi-breed evaluation, Regional breed

Table des matières

Remerciements ____ *3*
Résumé ____ *5*

Abstract ____ *6*
Table des matières ____ *7*
Liste des abréviations ____ *10*

Introduction ____ *11*

I. Principes des évaluations génétiques et génomiques ____ *13*

A. Genèse des évaluations génétiques ____ **13**

1. Modélisation de la performance et le modèle polygénique ____ 13
2. Modèle polygénique et ses conséquences sur le modèle d'évaluation ____ 14
3. BLUP et modèle mixte ____ 15
4. Index et coefficients de détermination. ____ 16

B. Conséquence du modèle polygénique sur les schémas de sélection laitiers ____ **18**

1. Progrès génétique ____ 18
2. Mise en place du testage sur descendance ____ 19
3. Contraintes du testage sur descendance ____ 19

C. Génotypage et l'accès à l'information moléculaire ____ **21**

1. Marqueurs moléculaires ____ 22
2. Transmissions des allèles et déséquilibre de liaison ____ 23

D. De la sélection assistée par marqueurs à la sélection génomique ____ **26**

1. Sélection Assistée par Marqueurs ____ 26
2. Limites de la sélection assistée par marqueurs ____ 27
3. Sélection Génomique ____ 28

E. Perspectives offertes par les évaluations génomiques et conséquences sur les schémas de sélection ____ **29**

1. Accélération du progrès génétique ____ 30
2. Meilleure gestion de la variabilité génétique ____ 31
3. Augmentation du progrès génétique sur la voie femelle ____ 32

II. Développement des évaluations génomiques *33*

A. Constitution de la population de référence **33**

1. Population de référence 33
2. Phénotypes 34

B. Préparation des génotypes **35**

1. Contrôle de la qualité des génotypage 35
2. Construction des phases et imputation des génotypes manquants 36

C. Estimation des effets des marqueurs **39**

1. Méthodes des moindres carrés 39
2. Méthodes de régression pénalisée 40
3. GBLUP ou BLUP génomique 41
4. Méthodes bayésiennes 42

Méthodes de réduction de dimensions 44

5. Comparaison de l'efficacité des méthodes d'estimations des effets des marqueurs. 44

D. Facteurs influençant l'efficacité de la sélection génomique **47**

Effet de la taille efficace de la population 47

1. Effet de la taille de la population de référence 48
2. Effet de l'héritabilité du caractère 49
3. Effet du choix des animaux constituant la population de référence 49
4. Effet de la densité en marqueurs 53

E. Le développement des évaluations génomiques en France et à l'international **56**

1. Utilisation de la Sélection Assistée par Marqueurs : l'exception française 56
2. Utilisation de la sélection génomique 59
3. Extension des évaluations génomiques aux races régionales 61

III. Constitution de la population de référence « haute densité » *64*

A. Introduction **64**

1. Choix des animaux à génotyper en haute densité 64
2. Préparation des fichiers de génotypages 66

B. Article I : Etude de la qualité d'imputation de génotypes haute densité à partir du génotype moyenne densité dans 16 races bovines françaises. **66**

C. Erratum Article II **78**

D. Bilan **79**

IV. Comparaison de stratégies d'évaluations génomiques pour une race disposant d'une population de référence de taille limitée *80*

A. Introduction **80**

B. Article II : Efficacité d'une sélection génomique intra-race et multi-race pour différentes tailles de population de référence **81**

C. Bilan **94**

V. Etude du cas de deux races génétiquement proches : exemple des races Montbéliarde et Simmental ___ *95*

A. Introduction ___ **95**

B. Article III : Evaluations génomiques à partir d'une population de référence multiraciale composée d'animaux des races Montbéliarde et Simmental ___ **96**

C. Bilan ___ **100**

VI. Discussion et perspectives ___ *101*

A. Bilan et limites des études ___ **101**

1. Imputation des génotypes haute-densité ___ 101
2. Stratégie d'évaluations génomiques en fonction de la taille de population ___ 102
3. Evaluations génomiques multiraciales dans le cas de deux races proches ___ 103

B. Perspectives de développement d'évaluations génomiques multiraciales ___ **105**

1. Conservation du déséquilibre de liaison et présence de QTL communs entre races ___ 105
2. Développement de nouvelles approches d'évaluations génomiques multiraciales ___ 106
   a) Approches supposant que l'ensemble des QTL est commun entre races ___ 106
   b) Approches supposant qu'il existe une variabilité des QTL entre races ___ 107

C. Perspectives de développement d'évaluations génomiques intra-races ___ **108**

1. Une sélection assistée par marqueurs ou sur mutations ___ 108
2. Une sélection génomique utilisant uniquement les mâles génotypés ___ 109
3. Une sélection génomique utilisant à la fois les mâles et les femelles génotypés ___ 110
4. Une sélection génomique utilisant l'ensemble des individus, génotypés ou non ___ 112

D. Conséquences de la mise en place d'évaluations génomiques pour les races régionales. 115

1. Utilisation des évaluations génomiques et conséquences sur le progrès génétique annuel et l'évolution de la variabilité génétique ___ 116
2. Utilisation des évaluations génomiques, coût des schémas de sélection et précision des évaluations 120
3. Spécificités liés aux situations des races régionales françaises ___ 122
4. Utilisation de la sélection génomique et gestion du troupeau ___ 124

VII. Conclusion ___ *126*

Bibliographie ___ *130*

Liste des tableaux ___ *143*

Liste des figures ___ *144*

Liste des publications ___ *146*

A. Articles scientifiques ___ **146**

B. Communications à des congrès internationaux ___ **146**

Liste des abréviations

50K : 50 000 marqueurs

ACP : **A**nalyse en **C**omposantes **P**rincipales

ADN : **A**cide **D**ésoxyribo**N**ucléique

AOC : **A**ppellation d'**O**rigine **C**ontrôlée

AOP : **A**ppellation d'**O**rigine **P**rotégée

BLUP : « ***B**est **L**inear **U**nbiased **P**redictor* », meilleure prédiction linéaire non biaisée

CD : **C**oefficient de **D**étermination

DYD : « ***D**aughter **Y**ield **D**eviation*», déviation moyenne des filles

EDC : « ***E**ffective **D**aughter **C**ontribution*», équivalent nombres de filles

GBLUP : **BLUP g**énomique

H^2 : héritabilité

HD : haute densité (777 962 marqueurs)

INRA : **I**nstitut **N**ational de la **R**echerche **A**gronomique),

Kb : **K**ilo**B**ases

LDLA : *« **L**inkage **D**isequilibrium and **L**inkage **A**nalysis »*, analyse utilisant conjointement l'information de liaison intra famille et de déséquilibre de liaison

Ne : Taille **E**fficace

PLS : « ***P**artial **L**east **S**quare*»

QTL : « ***Q**uantitative **T**rait **L**oci*», régions ayant un effet sur un caractère quantitatif

SAM : **S**élection **A**ssistée par **M**arqueurs

SNP : « ***S**ingle **N**ucléotide **P**olymorphism*»

ssGBLUP : « ***S**ingle **S**tep* **GBLUP** »

UNCEIA : **U**nion **N**ationale des **C**oopératives d'**E**levage et d'**I**nsémination **A**nimale

UCEAR : **U**nion des **C**oopératives d'**E**levage **A**lpes **R**hône

Introduction

La production laitière française a connu une forte intensification durant la seconde moitié du vingtième siècle. En 1983, des quotas de production ont été instaurés afin de limiter la quantité de lait totale produite en Europe et ainsi stabiliser les prix. L'apparition de ces quotas n'a pas freiné l'intensification de la production mais a conduit, à production totale constante, à une augmentation de la productivité par animal accompagné d'une diminution du nombre de vaches par troupeau. Le nombre de vaches laitières en France est ainsi passé de 5,3 millions en 1991 à 3,7 millions en 2013, tandis que la production moyenne par animal a évolué de 5036 à 6838 kg de lait par animal et par an pendant la même période (GEB, 2013). L'intensification de la production s'est accompagnée d'une homogénéisation des pratiques d'élevage et une diminution de la variabilité raciale (Raboisson, 2004). En 2013, 94% des vaches laitières étaient issues des trois grandes races nationales, (Prim'Holstein (66%), Montbéliarde (17%) et Normande (10%) (GEB, 2013). Les six pourcents restants se répartissaient entre les races dites régionales (Abondance, Tarentaise, Vosgienne, Simmental, Brune, Pie Rouge des Plaines) et les races en conservation.

L'augmentation de la productivité a été rendue possible par l'amélioration de la conduite des troupeaux mais également par l'augmentation du niveau génétique des animaux. L'utilisation de l'insémination a grandement favorisé le potentiel de diffusion d'un taureau et l'augmentation du progrès génétique. Les protocoles de testage sur descendance ont permis de connaitre précisément la valeur génétique d'un taureau avant sa mise sur le marché. Depuis le début des années 2000, le développement des techniques de séquençage et de génotypage haut débit ont permis d'accéder à la connaissance de l'ADN des animaux pour un coût modéré. L'utilisation de cette information moléculaire en sélection est intéressante car il devient possible de prédire la valeur génétique d'un animal dès son plus jeune âge à partir d'un simple échantillon biologique. De nombreuses régions QTL (*Quantitative Trait Loci*) influençant un caractère d'intérêt ont pu être identifiées et peuvent être utilisées en sélection. La situation idéale serait d'identifier l'ensemble des gênes influençant un caractère d'intérêt afin d'estimer les valeurs génétiques des animaux. En pratique, cette identification est impossible. Il a alors été suggéré d'estimer l'effet de l'ensemble des segments chromosomiques sur un caractère et d'en déduire des équations de prédiction. Cette approche dite de sélection génomique nécessite de disposer d'une large population de référence (avec performances et génotypes) à partir de laquelle on peut prédire l'effet de segments chromosomiques.

En France, la constitution de telles populations de référence n'a été possible que dans les trois principales races laitières : Holstein, Montbéliarde et Normande. Initialement, les évaluations génomiques ont donc été mises en place uniquement dans les races laitières nationales.

L'utilisation de la sélection génomique permet une augmentation du niveau génétique des animaux deux fois plus rapide que la sélection génétique classique. Ce doublement de progrès génétique annuel possible dans les races nationales grâce à la sélection génomique, pourrait donc encore accentuer l'écart de production entre races nationales et races régionales. Aujourd'hui même s'il est inférieur à celui des races nationales, le rendement laitier des races régionales est compensé par une meilleure adaptation des animaux au milieu (élevage de montagne, système herbager) et une meilleure valorisation du lait liée en particulier à de nombreuses appellations protégées (AOC : Appellation d'Origine Contrôlée et AOP : Appellation d'Origine Protégée). La forte augmentation de la production laitière chez les races nationales permise par la sélection génomique remet en cause cet équilibre. La disparition des quotas laitiers prévue en 2015, pourrait pousser certains éleveurs à augmenter leur niveau de production et donc se tourner vers les races nationales plus productives. Dans ce contexte, il parait nécessaire, pour maintenir la diversité des races et des systèmes d'élevage, de mettre en place des évaluations génomiques dans les races régionales afin qu'elles bénéficient des mêmes outils que les races nationales.

L'objectif de ce travail de thèse est de développer des évaluations génomiques multiraciales qui devraient permettre de mutualiser les populations de référence entre races et ainsi développer des évaluations génomiques dans les races régionales. La première partie de ce document sera consacrée à la sélection animale telle qu'elle est en place actuellement chez les bovins laitiers. On s'attachera en particulier à la transition entre la sélection classique et sélection génomique. La seconde partie décrira les modèles d'évaluations et les facteurs influençant la précision de l'évaluation. Une troisième partie s'intéressera au programme ANR GEMBAL dans lequel ma thèse s'insère et la constitution de la population de référence multiraciale. La quatrième partie présentera une étude comparative des capacités de prédiction des évaluations intra- et multi-races en fonction de la taille de la population de référence. La cinquième partie sera consacrée au cas de deux races proches, la Simmental Française et la Montbéliarde. Enfin, la dernière partie dressera un bilan des travaux réalisés et discutera des perspectives de développement d'évaluations génomiques pour les races régionales.

I. Principes des évaluations génétiques et génomiques

A. Genèse des évaluations génétiques

Depuis la domestication des animaux et le développement de l'élevage, l'homme a cherché à sélectionner les animaux. La sélection s'est d'abord focalisée sur des critères plus ou moins subjectifs portant sur le phénotype extérieur de l'animal lui-même. Ce mécanisme a entre autre permis la sélection d'animaux homogènes conduisant à la création des races, chaque race présentant une apparence et des caractéristiques propres. Au XX^{ème} siècle, la redécouverte des lois de Mendel et la formalisation du modèle polygénique par Fisher ont permis de poser les bases des évaluations génétiques classiques. Par la suite, les importants progrès de l'informatique d'une part ainsi que la loi sur l'élevage de 1966 d'autre part ont permis d'organiser l'identification des animaux, le contrôle des filiations et des performances et ainsi mettre en place les évaluations génétiques.

1. Modélisation de la performance et le modèle polygénique

Les évaluations génétiques utilisent les phénotypes (ou performances) pour prédire les valeurs génétiques des animaux. Le modèle suppose que la performance d'un animal se décompose entre la somme d'un effet génétique et d'un effet de l'environnement. L'estimation de la valeur génétique d'un animal pour un caractère donné est obtenue à partir du phénotype d'un animal après correction pour les effets de milieu identifiés, par exemple, les effets influençant la production laitière sont le rang de lactation, l'âge au vêlage, la conduite de troupeau etc. Pour chaque caractère, la part de variance expliquée par la composante génétique est appelée héritabilité. Plus la valeur de l'héritabilité est forte, plus le caractère sera facile à sélectionner car une part importante de la performance d'un animal sera liée à sa valeur génétique.

En conséquence, la valeur génétique d'un animal est estimée à partir d'un modèle décrit par l'équation suivante :

$$\boldsymbol{y_i = x'_i\beta + g_i + e_i}$$

Où :

- y_i est la performance de l'individu i pour le caractère d'intérêt
- $\boldsymbol{x'_i}$ est le vecteur d'incidence permettant de relier la performance aux effets de milieux $\boldsymbol{\beta}$
- $\boldsymbol{\beta}$ est le vecteur correspondant aux effets de milieu identifiés
- $\boldsymbol{g_i}$ est la valeur génétique de l'individu i
- $\boldsymbol{e_i}$ est la résiduelle du modèle

La valeur génétique d'un animal peut elle aussi se décomposer entre une valeur génétique additive correspondant à la somme des effets des allèles influençant le caractère, une valeur de dominance correspondant aux effets d'interactions entre les allèles d'un même locus et la valeur d'épistasie correspondant aux effets d'interactions entre les allèles de loci différents. La valeur génétique transmise d'un animal à sa descendance étant principalement additive, on assimilera dans la suite de ce document, la valeur génétique d'un individu (g) à sa valeur génétique additive (u).

2. Modèle polygénique et ses conséquences sur le modèle d'évaluation

Le modèle polygénique suppose qu'un caractère est régi par un nombre infini (ou très élevé) de gènes ayant chacun un faible effet. La valeur génétique d'un individu i (u_i) se définit alors comme la somme des effets des allèles des gènes portés par ses chromosomes. Elle correspond à la somme de la moitié des valeurs génétiques parentales (u_p et u_m) et d'un aléa de méiose (ϕ_i). L'aléa de méiose est la différence entre la valeur génétique d'un animal et la moyenne de ses parents et explique les différences de niveau génétique existant entre deux pleins frères.

On définit la covariance entre les valeurs génétiques additives comme l'espérance de la proportion d'allèles communs entre deux individus, selon l'équation suivante :

$$\boldsymbol{cov}(\boldsymbol{u_i}, \boldsymbol{u_j}) = \boldsymbol{a_{i,j}\sigma_u^2}$$

Où :

- u_i et u_j sont les valeurs génétiques des individus i et j
- $a_{i,j}$ est la proportion espérée d'allèles communs entre deux individus
- σ_u^2 est la variance génétique additive de la population

La proportion espérée d'allèles communs entre deux individus peut être déduite des coefficients de parenté (Malécot, 1948 ; Wright, 1921). Sa valeur sera de 0,5 entre un animal et ses parents ou ses pleins-frères et 0,25 entre deux demi-frères ou un individu et ses grands parents. On peut ainsi affiner le modèle d'évaluation génétique en utilisant les performances d'un individu et de celles de l'ensemble des individus qui lui sont apparentés.

Le modèle est donc compliqué par l'inclusion et le calcul explicite de l'ensemble des covariances du modèle. Des solutions calculatoires ont été proposées par (Henderson, 1975, 1976) afin de permettre son utilisation à l'échelle des populations d'élevage.

3. BLUP et modèle mixte

Les évaluations génétiques polygéniques utilisent le pedigree et l'ensemble des performances disponibles pour un caractère pour estimer les valeurs génétiques. Elle nécessite la modélisation et l'estimation des effets de milieu ayant une influence un caractère.

On distinguera deux types d'effets : les effets fixes et les effets aléatoires. La distinction est liée à la connaissance *a priori* de la distribution des effets. Dans le cas d'un effet aléatoire, les effets à estimer seront le résultat implicite d'un tirage aléatoire dans une loi statistique. Les valeurs génétiques par exemple, se distribueront selon une loi normale. A l'opposé pour de nombreux autres effets, tels que l'effet de l'âge ou l'effet de la saison, la distribution *a priori* n'est pas connue. On les considérera donc comme des effets fixes.

Initialement, les évaluations génétiques étaient réalisées en deux étapes. Les données étaient d'abord pré-corrigées pour les effets de milieux et les effets aléatoires étaient estimés dans un second temps. L'inconvénient majeur de cette stratégie est qu'il devient impossible de comparer le niveau génétique d'animaux soumis à des effets de milieux différents comme par exemple des animaux nés à des années différentes.

Pour contourner cette difficulté, Henderson (1975) a développé une méthode d'estimation simultanée des effets fixes et aléatoires : le BLUP (*Best Linear Unbiased Predictor*), littéralement « meilleure prédiction linéaire non biaisée».

La modélisation des performances se réécrit alors selon l'équation suivante :

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

Où :

- $\boldsymbol{y}$ est le vecteur des observations y_i
- $\boldsymbol{\beta}$ est le vecteur des effets fixes considérés dans le modèle statistique
- **u** est le vecteur des valeurs génétiques
- $\boldsymbol{X}$ et $\boldsymbol{Z}$ sont les matrices d'incidence qui relient les observations aux effets fixes et aléatoires
- **e** le vecteur des résiduelles du modèle

Les estimations des effets fixes et des valeurs génétiques peuvent ainsi s'obtenir comme solution du système d'équation du modèle mixte :

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

Où :

- A^{-1} est l'inverse de la matrice des coefficients de parenté
- σ_e^2 et σ_u^2 sont les variances de la résiduelle du modèle et des valeurs génétiques

La prédiction des valeurs génétiques à l'aide du BLUP possède de nombreuses propriétés intéressantes. Elle permet l'estimation sans biais des effets des milieux et des effets génétiques mais également la prise en compte, par l'intermédiaire de la matrice de parenté **A**, de l'ensemble des performances mesurées sur un individu et ses apparentés.

Ce système d'équations fait intervenir l'inverse de la matrice de parenté et peut donc être rapidement difficile à mettre en œuvre dans le cas de généalogie complexe. Une méthode basée sur les corrélations partielles entre individus, permettant d'obtenir directement la matrice $\mathbf{A}^{-1}$ a été développée par Henderson (1976) et a permis au BLUP de devenir la référence internationale pour les évaluations génétiques.

4. Index et coefficients de détermination.

Les valeurs génétiques des animaux prédites au travers des évaluations génétiques sont appelés « index ». Elles s'expriment en écart par rapport à une base de référence. Ces valeurs ne sont ici que des estimations de la valeur génétique vraie, sous l'hypothèse que le modèle d'analyse soit correct. On leur associe donc une mesure de la précision nommée coefficient de détermination (CD).

Le CD traduit l'amplitude de l'intervalle de confiance associé à l'index. Il se définit comme le rapport entre la variance des valeurs génétiques estimées et la variance des valeurs génétiques vraies. Le CD prend donc des valeurs entre 0 et 1 : plus cette valeur sera proche de 1, plus la précision de l'index sera élevée et plus la variance des index se rapprochera de la variance génétique vraie.

La valeur du CD dépend essentiellement de deux paramètres : l'héritabilité du caractère ainsi que le nombre d'individus avec performances pris en compte dans l'évaluation. Son calcul prend en compte trois sources d'information : l'ascendance, les performances propres de l'animal et celles de sa descendance. Les trois composantes du CD et leur méthode de calcul sont décrites dans le Tableau 1. On définit la répétabilité d'un caractère comme la proportion de variance des phénotypes expliquée par l'environnement permanent (effet de milieu non contrôlé qui se répète d'une performance à l'autre). Elle peut également se définir comme la corrélation entre deux performances successives des mêmes animaux pour un même caractère et traduit la ressemblance attendue entre deux performances d'un même individu.

Pour un caractère moyennement héritable comme la production laitière, mesurer les performances d'une quarantaine de filles d'un taureau permet d'obtenir un CD de 0,7 pour l'index laitier. En revanche, obtenir cette même précision (CD=0,7) sur un caractère peu héritable comme la fertilité (h^2=0,02) nécessite de connaitre les résultats de réussite à l'insémination d'environ 500 femelles.

Tableau 1 : Les composantes du coefficient de détermination en évaluation génétique d'après (Minvielle, 1990)

CD sur ascendance	**CD sur performance**	**CD sur descendance**
$CD_a = \frac{CD_{père} + CD_{mère}}{4}$	$CD_{perf} = \frac{nh^2}{1 + (n-1)R}$	$CD_d = \frac{Nh^2}{4 + (N-1)h^2}$
Avec CD_a : le CD sur ascendance CD_{pere} : le CD du père $CD_{mère}$: le CD de la mère	Avec : CD_{perf} : le CD sur performances propres n : le nombre de mesures de cette performance h^2 : l'héritabilité du caractère R : la répétabilité du caractère	Avec : CD_d : le CD sur descendance N : le nombre de descendants avec performance h^2 : l'héritabilité du caractère

B. Conséquence du modèle polygénique sur les schémas de sélection laitiers

1. Progrès génétique

L'efficacité de la sélection génétique se mesure à travers le progrès génétique annuel espéré (ΔG, (Minvielle, 1990)) dont l'expression est la suivante :

$$\Delta G = \frac{iR\sigma_g}{T}$$

Où :

- ΔG est l'augmentation du niveau génétique réalisée en une année
- i est l'intensité de sélection égale à la différence exprimée en écart-type génétique entre la valeur génétique moyenne des individus sélectionnés et celle des candidats à la sélection.
- R est la précision de l'évaluation génétique et donc la corrélation entre valeur génétique vraie et valeur génétique estimée. Elle correspond à la racine carrée du CD.
- σ_g est l'écart type génétique et donc une mesure de la variabilité génétique du caractère d'intérêt
- T est l'intervalle de génération correspondant à l'âge moyen des parents à la naissance de leurs descendants

Chez les bovins laitiers, les techniques de l'insémination animale et la congélation de la semence permettent d'inséminer plusieurs milliers de vaches avec la semence d'un même taureau. Ainsi le nombre de taureaux réellement diffusés sur le territoire est faible mais chacun d'entre eux a un fort pouvoir de diffusion (avec parfois plus de 100 000 doses). Cette stratégie permet d'obtenir une forte augmentation du progrès génétique grâce à une forte intensité de sélection. Elle requiert cependant de connaitre avec une forte précision les valeurs génétiques des taureaux sélectionnés comme reproducteurs (diffusés). Pour garantir un progrès génétique élevé, en France, seuls les index pour lesquels le CD est supérieur à 0,5 sont publiables officiellement.

2. Mise en place du testage sur descendance

Connaitre précisément la valeur génétique d'un taureau laitier est difficile. En effet, la limite majeure du modèle polygénique concerne la prédiction de l'aléa de méiose. Il est impossible de le prédire pour un individu n'ayant pas de performances propres et/ou pas de descendants.

Cette situation est pourtant la plus fréquente chez les bovins laitiers où les caractères d'intérêt sont, pour leur majorité, uniquement mesurables chez les femelles adultes. Il a donc été nécessaire de mettre en place une étape de testage sur descendance dans les schémas de sélection. Le principe du testage sur descendance est simple : on fait produire à un taureau plusieurs dizaines de filles dont on mesurera ensuite les performances sur l'ensemble des caractères d'intérêt. Les taureaux présentant les meilleurs index après testage sont ensuite sélectionnés pour être utilisés intensément dans l'ensemble de la population.

La forte intensité de sélection et la précision des valeurs génétiques atteintes grâce au testage sur descendance ont permis une accélération du progrès génétique. Ainsi, en France pour la production laitière, le progrès génétique réalisé sur les trente dernières années correspond à une augmentation moyenne de 100kg de lait par vache et par an. Cependant, les contraintes de ce dispositif sont nombreuses.

3. Contraintes du testage sur descendance

La Figure 1 schématise le principe du testage sur descendance. On remarquera qu'il s'agit d'un processus long et coûteux puisque les performances des filles ne sont accessibles qu'en fin de première lactation. Les taureaux ne disposent donc d'index précis et diffusables (publiables) qu'à l'âge de 6-7 ans.

Le testage sur descendance a un fort impact négatif sur la variabilité génétique : cette étape coûteuse dans les programmes de sélection impose une forte pression de sélection dès le choix des mères et des pères à taureaux. Le nombre de taureaux testés par an est faible (Tableau 2), une dizaine de taureaux par an dans les races régionales, de l'ordre de cent cinquante pour les races Montbéliarde et Normande et six cent en race Holstein.

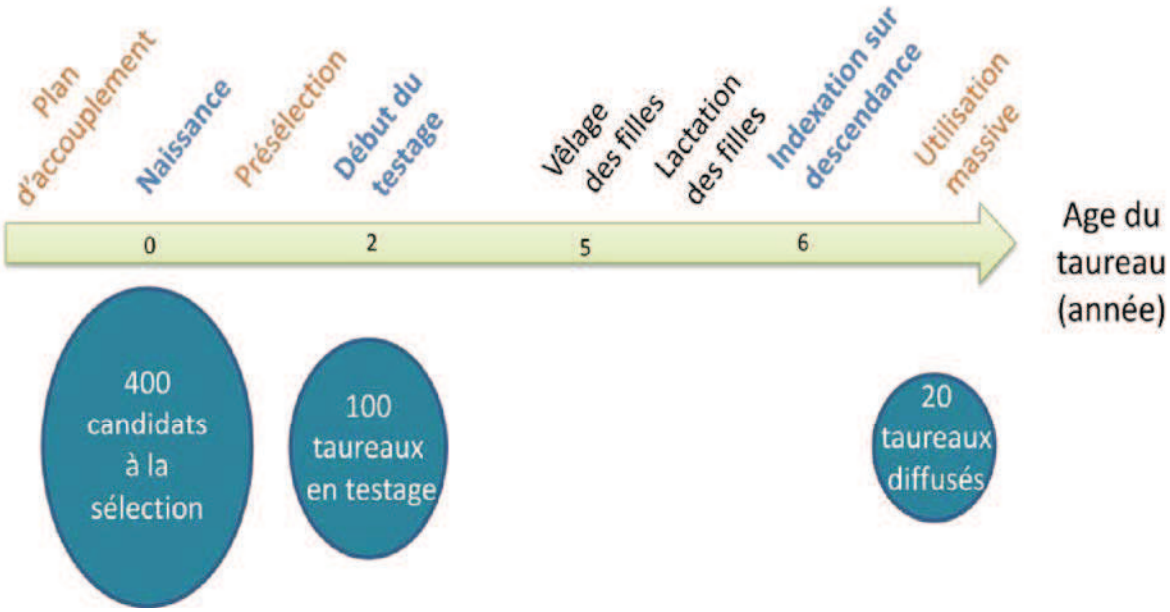


Figure 1 : Schéma de sélection avec testage sur descendance d'après S. Fritz, UNCEIA (communication personnelle)

Tableau 2 : Nombre de taureaux testés sur descendance par année d'après Boichard et al. (1996) et France Génétique Elevage

Race	**1980**	**1987**	**2008**	**2012**
Abondance	3	14	19	15
Pie Rouge des Plaines	8	7	6	-
Brune	7	8	9	-
Tarentaise	2	8	13	12
Simmental	8	6	10	6
Montbéliarde	67	110	154	6
Normande	146	139	143	107
Prim'Holstein	345	734	400	25

Le choix des animaux à mettre en testage est donc déterminant pour les entreprises de sélection qui ne disposent alors que d'un seul outil pour classer les animaux : l'index sur ascendance. Cet index, issu de l'information polygénique, reste peu précis en particulier pour des caractères faiblement héritables et est identique pour tous les animaux d'une même fratrie (pleins frères). Souvent, plusieurs veaux issus d'une même famille connue pour son niveau génétique sont donc mis en testage. Ainsi sur la période comprise entre les années 1986 et 1990, le nombre de pères ayant produit 80% des veaux mis en testage était respectivement de 11, 13 et 29 dans les races Montbéliarde, Normande et Prim'Holstein (Boichard et al., 1996). Cette concentration des origines des animaux ainsi que la forte utilisation des taureaux « stars » comme Jocko Besné (1.7 millions de doses produites et 122 000 filles en France) ont conduit à une importante chute de la variabilité génétique intra-race (Boichard et al., 1996 ; Danchin-Burge et al., 2012).

Cet ensemble de contraintes liées au testage sur descendance a poussé la filière bovine laitière à se tourner vers l'utilisation de l'information moléculaire pour prédire la valeur génétique individuelle des candidats à la sélection.

C. Génotypage et l'accès à l'information moléculaire

On conçoit aisément que connaitre le génome complet d'un animal peut, idéalement, permettre de prédire la valeur génétique d'un individu et en particulier son aléa de méiose. Si on prend l'exemple d'un caractère soumis à un gène unique, en connaissant les 2 allèles portés par un animal, il devient possible de connaitre sa valeur génétique. Dans un cas plus général, on pourrait estimer avec une grande précision la valeur génétique d'un animal en sommant l'effet estimé des allèles de l'ensemble des gènes influant le caractère d'intérêt.

Le premier séquençage du génome bovin a été réalisé en 2006 (The Bovine HapMap, 2009) mais en pratique, l'accès au génome complet d'un animal est, trop onéreux pour être utilisé en sélection. Il a en revanche, été possible dès les années 80 d'accéder au moins partiellement à l'ADN des animaux grâce à la découverte des marqueurs moléculaires.

1. Marqueurs moléculaires

Les marqueurs moléculaires se définissent comme des variations de l'ADN détectables à un coût abordable. Ils sont répartis sur le génome et servent de repère pour suivre la transmission, d'une génération à l'autre, d'un segment chromosomique. Les types de marqueurs les plus utilisés sont les microsatellites et les SNP *(Single Nucleotide Polymorphism)*.

Les microsatellites ont été identifiés à la fin des années 80. Ce sont des répétitions d'un motif d'un, deux ou trois nucléotides le long du génome. Le nombre de répétitions du motif varie selon l'allèle présent chez un animal donné. Près de 4000 marqueurs ont ainsi pu être identifiés et positionnés sur le génome bovin (Ihara et al., 2004). Leur typage n'est pas complètement automatisable ce qui entraîne un coût relativement élevé de l'ordre de 1 € par animal et par microsatellite.

Plus récemment des marqueurs plus nombreux et plus faciles à détecter ont pu être identifiés : les SNP. Un SNP correspond à un type de polymorphisme pour lequel seul un nucléotide diffère. Contrairement aux microsatellites, il n'existe que deux allèles pour chaque marqueur, mais cette faiblesse apparente est compensée par leur abondance sur le génome. En 2006, le séquençage complet du génome bovin a permis l'identification de plus d'un million de polymorphismes et ainsi autorisé la création de puces à SNP de densité élevée (Matukumalli et al., 2009). Des puces contenant environ 50 000 marqueurs SNP sont disponibles pour la plupart des espèces d'élevage (bovins, ovins, caprins, porcins etc.). Elles sont actuellement les puces les plus couramment utilisées chez les bovins car elles permettent pour un coût raisonnable (200€ par animal en 2009 et 100€ actuellement) d'accéder à un maillage relativement dense du génome : 1 SNP tous les 100Kb en moyenne (Matukumalli et al., 2009). En 2009 et 2010, des puces à plus faible et plus forte densité ont également été élaborées afin de s'adapter aux besoins des schémas de sélection et de la recherche, nous reviendrons sur leur utilisation ultérieurement.

L'utilisation de marqueurs moléculaires en génétique animale suppose qu'il existe une liaison génétique entre un marqueur et une mutation causale. La force de cette liaison génétique entre deux loci se mesure à travers le déséquilibre de liaison.

2. Transmissions des allèles et déséquilibre de liaison

Le déséquilibre de liaison se définit comme l'association préférentielle entre les allèles de deux loci. Il peut se calculer, dans le cas d'un locus bi-allélique de la façon suivante (Hill et Robertson, 1968) :

$$\mathbf{r^2 = \frac{((f_{A1B1} * f_{A2B2}) - (f_{A1B2} * f_{A2B1}))^2}{f_{A1} * f_{A2} * f_{B1} * f_{B2}}}$$

Où :

- f_{A1}, f_{A2}, f_{B1}, f_{B2} les fréquences dans la population des allèles 1 et 2 aux loci A et B
- f_{A1B1}, f_{A2B2}, f_{A1B2} f_{A2B1} les fréquences des haplotypes A_1B_1, A_2B_2, A_1B_2 et A_2B_1

Le déséquilibre de liaison prend des valeurs comprises entre 0 et 1. Si celui-ci vaut 1 on parlera de déséquilibre de liaison total, l'allèle 1 du locus A sera toujours associé à l'allèle 1 du locus B. A l'opposé s'il vaut 0, il n'existe pas d'association préférentielle entre les locus A et B, on parlera alors d'équilibre de liaison.

L'association entre deux allèles de deux loci peut être lié au hasard ou à un biais de sélection mais également à la transmission allélique. Un parent transmet un brin de chacun de ses chromosomes (après recombinaison) à ses descendants. Les allèles présents sur un chromosome d'un individu fondateur sont donc transmis en bloc à ses descendants. Au fil des générations, les recombinaisons aléatoires vont réduire la taille des segments chromosomiques conservés mais on retrouve trace de courts segments originaux, ou haplotypes fondateurs, conservés chez de nombreux individus de la population actuelle. Une association entre deux allèles présente chez un fondateur sera donc transmise à l'ensemble de ces descendants (Figure 2). Plus la distance entre les deux allèles sera faible, plus la probabilité qu'une recombinaison ait lieu est faible et plus l'association entre deux allèles présente chez un fondateur sera conservée dans la population actuelle. Le déséquilibre de liaison sera donc généralement plus élevé à courte distance qu'à grande distance (Figure 3). Si l'on considère une mutation causale entourée par deux marqueurs, il est fort probable que les individus ayant hérité des mêmes deux allèles aux marqueurs ait hérité du même allèle à la mutation. Cette hypothèse est d'autant plus vraisemblable que la densité en marqueurs est forte et que les marqueurs sont proches.

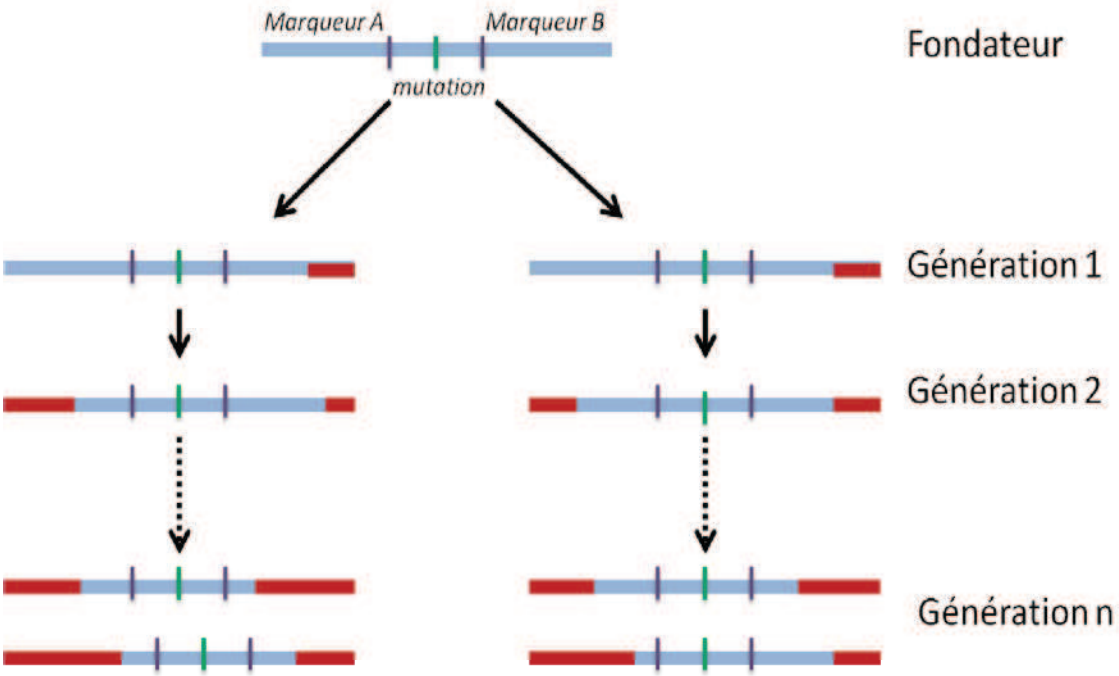


Figure 2 : Transmission d'une association entre une mutation et un marqueur d'un individu fondateur à ses descendants.
En bleu, segments chromosomiques du fondateur porteur de l'association entre marqueurs et mutations. En rouge, segments chromosomiques d'autres individus fondateurs apparus suite aux recombinaisons successives.

La Figure 3 représente l'évolution du déséquilibre de liaison moyen en fonction de la distance entre deux loci. Dans l'ensemble des races, le déséquilibre de liaison chute d'abord rapidement jusqu'à 100 kb et se stabilise ensuite. On pourra noter que le déséquilibre de liaison moyen est globalement plus élevé dans les races laitières (Holstein, Montbéliarde, Normande) que dans les races allaitantes (Blonde d'Aquitaine, Charolaise, Limousine). Cette différence est liée à une utilisation plus intensive de l'insémination animale dans les races laitières qui a conduit à une réduction de la variabilité génétique.

Le déséquilibre de liaison existant entre deux loci dépend du nombre de recombinaisons ayant eu lieu entre les deux loci. Le nombre de recombinaisons étant fonction de la distance entre deux loci et du nombre de générations écoulées, on distinguera le déséquilibre de liaison populationnel (ou association) correspondant à un lien entre deux loci observable sur l'ensemble de la population, du déséquilibre de liaison familial (ou liaison) correspondant à un lien entre deux loci observable uniquement au sein d'une même famille.

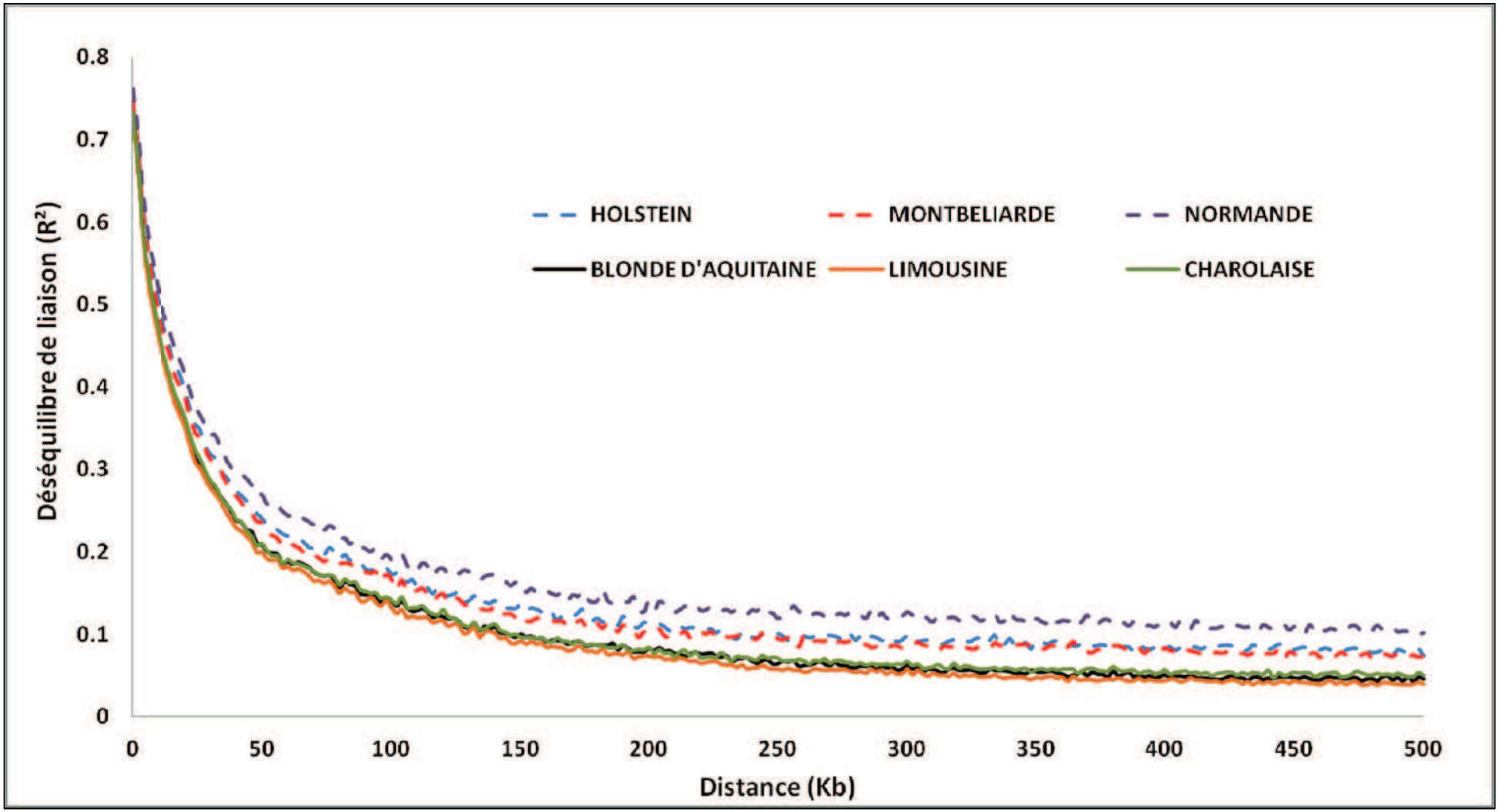


Figure 3 : Déséquilibre de liaison (r^2) moyen en fonction de la distance entre loci dans les principales races bovines françaises à partir des données de la puce haute-densité , d'après Hozé et al. (2012)

D. De la sélection assistée par marqueurs à la sélection génomique

Les marqueurs moléculaires peuvent donc être utilisés pour suivre la transmission de segments chromosomiques à l'échelle populationnelle ou familiale et donc pour prédire si un animal porte ou non un allèle favorable pour le caractère d'intérêt (Lande et Thompson, 1990 ; Haley et Visscher, 1998).

1. Sélection Assistée par Marqueurs

Dans une situation idéale, les mutations causales seraient toutes génotypées et utilisées directement dans une sélection assistée par marqueurs. En pratique, nous ne disposons généralement pas des mutations causales mais de marqueurs qui peuvent être en déséquilibre de liaison avec ces mutations causales. L'identification (à travers d'étude de liaison ou d'association) et la validation (à travers d'étude fonctionnelle) de mutations causales sont des étapes longues et délicates (Andersson, 2001). Seul un nombre limité de mutations affectant un gène majeur (ayant un effet fort sur le caractère) a pu être identifié et utilisé en sélection assistée par marqueurs à travers des tests directs recherchant la présence de l'allèle favorable ou défavorable à la mutation (Dekkers, 2004). En revanche, de nombreuses régions QTL, littéralement « régions ayant un effet sur un caractère quantitatif », ont été identifiées (Khatkar et al., 2004).

Les méthodes de Sélection Assistée par Marqueurs (SAM) permettent d'intégrer l'effet des QTL connus dans les modèles d'évaluations pour mieux estimer les valeurs génétiques des animaux (Lande et Thompson, 1990). On estime les effets des marqueurs (ou des groupes de marqueurs) à partir d'une population de référence constituée d'animaux pour lesquels un phénotype et un génotype sont disponibles. Ces effets sont ensuite extrapolés aux animaux pour lesquels seul le génotype est disponible. Toutes les régions influençant un caractère n'étant pas intégrées dans le modèle, les marqueurs n'expliquent qu'une partie de la variance génétique observée sur la population. L'estimation des valeurs génétiques des animaux se décompose donc en deux composantes : une composante polygénique correspondant à la variance expliquée par la généalogie et la composante génomique correspondant à la somme des effets des allèles aux marqueurs.

Un premier modèle de sélection assistée par marqueurs a été proposé par (Fernando et Grossman, 1989) :

$$\mathbf{y_i} = \mathbf{u_i} + \sum_{\mathbf{k=1}}^{\mathbf{Nqtl}} \left(\mathbf{v_{ik}^p} + \mathbf{v_{ik}^m}\right) + \mathbf{e_i}$$

Où :

- y_i est la performance de l'individu i
- u_i est l'effet polygénique de l'individu i
- v_{ik}^p et v_{ik}^m sont les effets des allèles d'origine paternelle et maternelle de l'individu i au QTL k
- e_i est la résiduelle du modèle

Si la densité en marqueurs est faible, ce qui est le cas pour les marqueurs microsatellites par exemple, la localisation d'une région QTL est peu précise et le lien entre QTL et marqueurs est relativement faible à l'échelle de la population. Il est alors nécessaire d'estimer l'effet des segments chromosomiques au sein de chaque famille d'animaux.

2. Limites de la sélection assistée par marqueurs

L'efficacité d'un programme de sélection comprenant une étape de sélection assistée par marqueur SAM dépend de la précision de la localisation, de l'ampleur des effets, et des fréquences des QTL intégrés dans le modèle (Spelman et van Arendonk, 1997). Le gain d'efficacité permis par un programme de SAM par rapport à un programme de sélection classique a été prouvé (Meuwissen et Goddard, 1997 ; Ruane et Colleau, 1995). En revanche, il a été démontré que si le nombre de QTL utilisé dans le modèle est faible, l'efficacité de la SAM peut rapidement se dégrader au fil des générations (Ruane et Colleau, 1995). La raison majeure est que, lorsque la SAM se focalise sur un faible nombre de QTL, elle augmente rapidement les fréquences des allèles favorables aux QTL. Une fois la fréquence du QTL élevée dans la population, il devient moins efficace voire inutile de l'utiliser en sélection.

Une première solution, pour contourner ce problème, est d'augmenter le nombre de QTL inclus dans le modèle afin de limiter la pression de sélection sur chacun d'entre eux et améliorer la précision du modèle (Stella et al., 2002). Cette solution bien qu'attrayante a longtemps été difficile à mettre en pratique. En effet, les dispositifs existants au début des années 2000 étaient basés sur des marqueurs microsatellites et ne permettaient de détecter uniquement les QTL à fort effet, alors que les caractères sont soumis à un nombre élevé de gènes à faible effet (Hayes et Goddard, 2001). De plus augmenter le nombre de QTL pris en compte dans un modèle de SAM, augmente aussi le risque de prendre en compte des faux QTL dans le modèle. Cette prise de risque permet d'augmenter la part de variance expliquée par les QTL, donc l'efficacité de la SAM, et est, en pratique, souvent préférable à une sélection stricte des QTL (Moreau et al., 1998 ; Guillaume, 2009).

L'efficacité de la sélection assistée par marqueurs est dépendante du choix, souvent assez arbitraire des QTL inclus dans le modèle d'évaluation. Afin d'éviter cette étape de détection des QTL, une stratégie alternative dite de sélection génomique a été proposée. Cette approche fait l'hypothèse que l'ensemble de la variabilité génomique peut être expliquée par les marqueurs sans qu'il soit nécessaire de connaitre précisément *a priori* la localisation des QTL et leur importance relative.

3. Sélection Génomique

Les modèles de sélection génomique supposent que l'ensemble des segments chromosomiques ont un effet sur le caractère (Haley et Visscher, 1998). Contrairement au modèle SAM, si la nature et le nombre des marqueurs permettent d'expliquer la totalité de la variabilité génétique, l'utilisation d'une composante polygénique devient inutile (Habier et al., 2007).

La principale difficulté est qu'il faut, ici, estimer l'effet d'un grand nombre de marqueurs (plusieurs dizaines de milliers) à partir d'un nombre limité d'individus (quelques centaines à plusieurs milliers). Plusieurs approches méthodologiques, qui seront détaillées dans le prochain chapitre, ont été proposées et comparées pour estimer ces effets de marqueurs et mettre en place une évaluation génomique. Elles se subdivisent en deux grandes familles. La première se rapproche du modèle polygénique et suppose que l'ensemble des segments chromosomiques ont un effet sur le caractère. La seconde se rapproche d'un modèle biologique en sélectionnant uniquement les segments chromosomiques ayant un effet sur le caractère.

Ces approches ont été formalisées et comparées sur données simulées par Meuwissen et al. (2001). Les premières études réalisées par simulation ont montré que la précision des valeurs génétiques estimées par un modèle de sélection génomique pouvaient atteindre des valeurs élevées (comprises entre 0,63 et 0,85) (Meuwissen et al., 2001; VanRaden, 2008). Ces valeurs sont proches des précisions obtenues pour des taureaux en sortie de testage et suggèrent qu'il est possible de sélectionner les animaux sur la base de leur information moléculaire uniquement. La sélection génomique nécessite cependant de disposer de marqueurs couvrant l'ensemble du génome et n'a donc été utilisée sur données réelles qu'à partir de 2008 avec la disponibilité de la puce Illumina BovineSNP50® (50K) (Matukumalli et al., 2009).

E. Perspectives offertes par les évaluations génomiques et conséquences sur les schémas de sélection

L'utilisation de l'information moléculaire et les évaluations génomiques permettent de connaitre avec une précision correcte les valeurs génétiques des animaux à partir d'un simple échantillon biologique. Cette possibilité a remis en cause les schémas de sélection laitiers qui reposaient jusqu'alors sur les index sur ascendance (pour la présélection des animaux) et le testage sur descendance.

1. Accélération du progrès génétique

L'information moléculaire a tout d'abord été intégré dans des évaluations de type SAM basées sur quelques QTL. Les valeurs génétiques estimées obtenues à partir de ces évaluations sont plus précises que les index sur ascendance mais restent beaucoup moins précises que les index d'un animal avec performances (propres ou celles des ses filles). Les évaluations SAM ont donc, dans un premier temps, été utilisées pour présélectionner les animaux avant leur entrée en testage (Kashi et al., 1990) et sélectionner des mères à taureaux (Schrooten et al., 2005). Plusieurs études ont cherché à mesurer par simulation le gain de progrès génétique lié à l'inclusion d'une étape de SAM dans les programmes de sélection laitiers. Les résultats obtenus étaient très variables en fonction des caractéristiques des QTL inclus dans le modèle d'analyse et du schéma de sélection simulé. Les estimations du progrès génétique additionnel variaient ainsi de -6% (Spelman et Garrick, 1997) à +105% (Spelman et al., 1999). A l'opposé, à progrès génétique constant, l'inclusion d'une étape SAM permet de réduire le nombre de taureaux testés par an (Schrooten et al., 2005). En France, Il a été estimé que les évaluations SAM permettaient de tester sans risque entre 5 et 40% de taureaux en moins par an (Boichard et al., 2002 ; Guillaume, 2009)

La disponibilité de puces moyenne densité a permis d'augmenter fortement la précision des valeurs génomiques obtenues en utilisant uniquement les marqueurs moléculaires et donc de les rapprocher des valeurs génétiques obtenues après testage sur descendance. Il est alors devenu envisageable de réorganiser complètement les schémas de sélection laitiers autour de l'utilisation de l'information moléculaire et de supprimer l'étape de testage organisée sur descendance. D'un point de vue théorique, l'utilisation directe d'un animal sur la base de son génotype aux marqueurs permet d'utiliser un taureau dès sa maturité sexuelle atteinte c'est-à-dire de réduire l'intervalle de génération de 5 à 2 ans et ainsi de doubler le progrès génétique annuel espéré (Schaeffer, 2006).

En pratique, les index obtenus à l'aide des marqueurs moléculaires restent légèrement moins précis que les valeurs génétiques estimées classiquement et sont donc moins fiables que ceux obtenus après testage ce qui nécessite une étude plus poussée de l'efficacité de cette stratégie.

Par rapport à un schéma classique, la présélection des animaux permettait, pour un caractère d'héritabilité moyenne, un gain de progrès génétique compris entre 9% (Lillehammer et al., 2011) et 16% (Pryce et al., 2010). Pour le même caractère, ce gain était compris entre 29% (Lillehammer et al., 2011) et 59% (Pryce et al., 2010) quand les jeunes animaux étaient utilisés directement. Sur les caractères faiblement héritables, le gain de progrès génétique permis par la sélection génomique était encore supérieur (Lillehammer et al., 2011).

En plus de la réduction de l'intervalle de génération, la suppression de l'étape de testage permet une forte réduction du coût d'un taureau avant sa mise en marché (Schaeffer, 2006). Il n'est alors plus nécessaire dans ce contexte, d'utiliser intensivement un animal pour rentabiliser l'investissement ce qui permet une meilleure gestion de la variabilité génétique.

2. Meilleure gestion de la variabilité génétique

Avec la diminution du coût associé à la mise en marché d'un animal, il devient possible de ne plus porter son choix sur un nombre limité d'individus issus de familles connues mais de sélectionner les animaux parmi un large panel de candidats représentatifs de l'ensemble des origines de la population (Bouquet et Juga, 2013). Il est donc possible au niveau des schémas de sélection de limiter l'accroissement de la consanguinité (Colleau et al., 2009)

Au niveau des élevages, il est également possible d'améliorer la gestion de la variabilité génétique. Avec l'augmentation du nombre de taureaux disponibles au niveau national, il devient plus facile d'inséminer les vaches d'un même troupeau avec un grand nombre de taureaux différents et non plus avec quelques taureaux stars. Cette stratégie a un second avantage, puisqu'elle diminue la prise de risque associée à l'utilisation d'un taureau non testé sur descendance. En effet, l'utilisation directe des taureaux sur la base de leur information moléculaire s'accompagne d'une diminution de la fiabilité des index et donc d'un risque accru que la valeur génétique vraie soit significativement plus faible que la valeur génétique estimée. L'utilisation d'un groupe de taureaux de valeur génétique homogène permet de contourner ce problème car les variations de la valeur génétique estimée moyenne d'un groupe sont d'autant plus faibles que le nombre d'individus constituant ce groupe est important (Pryce et Hayes, 2012). Il a ainsi été démontré que les variations associées à la valeur génétique moyenne d'un lot de cinq à 10 taureaux dont le CD est de 0,5 sont similaires à celles observées pour un taureau unique dont le CD serait de 0,8 (S. Fritz, UNCEIA, Jouy en Josas, communication personnelle).

3. Augmentation du progrès génétique sur la voie femelle

Les évaluations polygéniques, basées sur les performances et la généalogie des animaux, ne permettent pas d'obtenir des valeurs génétiques précises pour les femelles. Les coefficients de détermination associés aux index sont généralement de 0,5 après une première lactation pour les caractères laitiers et inférieurs à 0,3 pour les caractères de fertilité. Les évaluations génomiques permettent d'accéder à des valeurs génétiques estimées précises pour l'ensemble des individus, quel que soit leur sexe, et l'ensemble des caractères. Il devient donc possible de mettre en place une sélection efficace sur la voie femelle.

En raison des coûts de génotypages élevés, les évaluations génomiques ont d'abord été utilisées pour sélectionner les femelles « élites ». Cette sélection permet de déterminer quelles femelles seront utilisées comme mères à taureaux et donc, comme pour les mâles, d'augmenter la taille du « noyau de sélection ». La sélection génomique a donc permis d'augmenter l'intensité de sélection, la précision des évaluations des femelles qui sont deux composantes majeures du progrès génétique. L'association des évaluations génomiques avec les nouvelles techniques de reproduction (transplantation embryonnaire) permet encore d'accélérer le processus de sélection car il devient possible de faire produire plusieurs descendants à une femelle et de sélectionner celui ayant reçu la combinaison d'allèles la plus favorable.

Avec le développement de puces à plus faible densité (Boichard et al., 2012a; Wiggans et al., 2012) et la réduction des coûts de génotypage, les évaluations génomiques deviennent un outil accessible pour chaque éleveur et ne sont plus réservées qu'aux seules femelles élites. Le choix des femelles à utiliser pour le renouvellement du troupeau est alors facilité par la connaissance du niveau génétique de chacune des génisses. Ici encore, l'association avec les techniques de reproduction comme l'utilisation de la semence sexée, permet de s'assurer que les descendants des meilleures femelles seront des génisses (qui seront utilisées pour le renouvellement du troupeau) et à l'opposé de mieux valoriser les veaux issus des moins bonnes femelles au travers du croisement. Par ailleurs, la connaissance précise des valeurs génétiques d'un animal sur l'ensemble des caractères évalués permet de mieux gérer les accouplements et de sélectionner un taureau qui compensera les défauts de la femelle.

L'utilisation des marqueurs moléculaires en sélection des bovins laitiers semble donc particulièrement intéressante. Les évaluations génomiques offrent de multiples possibilités pour améliorer l'efficacité de la sélection tout en gérant le maintien de la variabilité génétique. Cependant, leur mise en place nécessite de lever de nombreuses contraintes techniques et en particulier, l'estimation de l'effet des marqueurs sur le caractère.

II. Développement des évaluations génomiques

La sélection génomique et la sélection assistée par marqueurs nécessitent d'estimer l'effet de segments chromosomiques à partir d'une population dite de référence et d'appliquer les solutions du modèle d'estimation pour prédire les valeurs génétiques d'animaux génotypés mais sans performance.

A. Constitution de la population de référence

1. Population de référence

Une population de référence est constituée d'animaux pour lesquels on dispose à la fois du génotype aux marqueurs et d'un phénotype pour les caractères que l'on cherche à sélectionner.

Une hypothèse implicite de la sélection génomique est que les effets aux marqueurs sont identiques entre les populations de référence et de validation. Il faut donc que le déséquilibre de liaison observé entre un marqueur et un QTL dans la population de référence soit conservé dans la population de validation. Cette hypothèse n'est valide que si un faible nombre de recombinaisons ont eu lieu entre la population de référence et les candidats à la sélection et donc que si les populations sont génétiquement proches et la densité en marqueurs élevée.

L'estimation des effets dépend également de la qualité de l'information disponible dans la population de référence. Pour estimer avec précision les effets des marqueurs, il faut que les phénotypes utilisés soient les plus représentatifs possible de la valeur génétique des animaux. Pour s'assurer de la qualité des phénotypes utilisés, chez les bovins laitiers la population de référence est généralement constituée de taureaux testés sur descendance pour lesquels la valeur génétique est estimée avec précision.

2. Phénotypes

Les taureaux testés sur descendance ne disposant pas de performance propre, on utilise des phénotypes « équivalents » calculés à partir des performances de leurs filles.

Le phénotype le plus couramment utilisé est la déviation moyenne des filles ou DYD (*Daughter Yield Deviation*) (Van Raden et Wiggans, 1991). Elle est égale à la performance moyenne des filles d'un taureau corrigée de l'ensemble des effets autres que l'effet génétique additif (effets liés à l'environnement et éventuellement effets d'environnement permanent) pris en compte dans l'indexation ainsi que du niveau génétique de sa mère. Ce phénotype est équivalent à une performance propre pour lequel l'héritabilité correspondrait à la précision (CD) de l'index.

Lorsque le DYD n'est pas disponible, ce qui est le cas pour les taureaux étrangers par exemple, il est nécessaire d'utiliser un autre phénotype. L'utilisation directe des index pour la détection de QTL et la sélection génomique est peu recommandée car les valeurs génétiques estimées sont régressées vers la moyenne de la population. Il est cependant possible d'utiliser un « index dérégressé ». Ce phénotype est lui aussi équivalent à une performance propre de l'animal. Si l'ensemble des index déregressés sont utilisés conjointement dans une évaluation BLUP, ils conduisent aux mêmes valeurs génétiques estimées que les performances de la population de départ. A la différence du DYD, l'index dérégressé n'est (implicitement) corrigé que pour les effets de milieu et pas pour le niveau génétique des mères. Ce phénotype a l'avantage de pouvoir être calculé directement à partir des index (Sigurdsson et Banos, 1995) et s'est révélé être équivalent ou très proche des DYD pour une utilisation en détection de QTL et en sélection génomique (Thomsen et al., 2001). Son utilisation est cependant moins préconisée car les index déregressés contiennent une partie de l'information génétique (en particulier l'ascendance) que l'on souhaite capturer au travers des SNP. De plus, contrairement aux DYD, les index déregressés ne sont pas indépendant les uns des autres.

Comme le coefficient de détermination, la précision associée au DYD (ou à l'index déregressé) d'un animal pour un caractère donné dépend de l'héritabilité du caractère ainsi que du nombre de filles de l'animal phénotypées pour ce caractère. Les « nombres équivalents de filles » ou EDC (*Effective Daughter Contribution*) permettent de prendre en compte cette précision (Fikse et Banos, 2001).

B. Préparation des génotypes

1. Contrôle de la qualité des génotypage

Les puces à ADN ont été développées afin d'accéder, pour un coût relativement faible à un génotype de qualité. Cependant, comme toutes données, les génotypes peuvent présenter des erreurs et des valeurs manquantes. De nombreuses procédures ont donc été mises en place pour préparer les fichiers de génotypages avant leur utilisation en sélection génomique. Plusieurs critères ont été développés pour s'assurer de la qualité des génotypes. On distingue deux grands types de filtres : un premier contrôle réalisé à l'échelle des SNP, un second à l'échelle des individus.

Les principaux critères de contrôle des SNP sont le pourcentage d'animaux génotypés pour un SNP (*call freq*) et le respect de l'équilibre de Hardy-Weinberg. Un trop grand nombre d'individus non génotypés à un marqueur est généralement révélateur d'une difficulté de lecture des génotypages (*clustering)* et donc une difficulté à discriminer les deux allèles d'un même SNP. Un premier filtre consiste donc à supprimer les marqueurs pour lesquels un pourcentage élevé d'individus (généralement 20%) n'est pas génotypé. Un second filtre concerne l'équilibre de Hardy-Weinberg. La théorie développée par Hardy (1908) et Weinberg suppose qu'au sein d'une population dite idéale, il existe un équilibre des fréquences alléliques et génotypiques d'un locus d'une génération à l'autre. Elle s'appuie sur plusieurs hypothèses qui supposent entre autres : un nombre infini d'individus, une reproduction panmictique, l'absence de mutation, de migration et de sélection. En pratique, ces conditions sont rarement atteintes dans les populations domestiques mais il a été démontré qu'une forte déviation aux conditions d'équilibre de Hardy-Weinberg correspond généralement à des difficultés de génotypage du marqueur (Hosking et al., 2004). On élimine donc les marqueurs présentant une déviation à cet équilibre ($p < 0,001$ ou $p < 0,0001$).

Enfin, un dernier filtre est souvent appliqué sur la fréquence de l'allèle minoritaire. Les marqueurs utilisés sur les puces à ADN ont été sélectionnés afin qu'en moyenne la fréquence de l'allèle rare soit supérieure à 20% sur l'ensemble des races *bos taurus*. Cependant, dans une race donnée, un nombre important de marqueurs sont peu polymorphes. La présence d'allèles à faible fréquence dans une population se traduit par un nombre limité de phénotypes disponibles pour prédire l'effet de cet allèle et donc l'introduction potentielle, d'erreurs dans les modèles de détection de QTL et de sélection génomique. Pour éviter ce problème, on élimine généralement les marqueurs dont la fréquence de l'allèle minoritaire est inférieure à 1, 3 ou 5% dans la population.

Au niveau de l'animal, deux critères de qualité sont majoritairement utilisés. L'un s'intéresse à la qualité du génotypage et en particulier au pourcentage de marqueurs manquants. Le second à la compatibilité entre le fichier de généalogie et les génotypes. Le pourcentage de SNP manquants pour un individu est mesuré à l'aide du *call rate.* Il se définit comme le rapport entre le nombre de marqueurs génotypés et le nombre de marqueurs sur la puce. Lorsque cet indicateur est « faible » (moins de 98 ou 95%), le génotypage de l'individu sera exclu. La vérification de la compatibilité entre le fichier de généalogie et les génotypes permet de détecter les éventuelles inversions de prélèvements, mais également les erreurs d'enregistrement de la généalogie. Cette étape est importante car une mauvaise association entre un individu, c'est à dire un phénotype, et un génotype pourrait induire des erreurs dans la détection des marqueurs ayant un effet sur le caractère étudié.

2. Construction des phases et imputation des génotypes manquants

Un traitement complémentaire des données peut également être réalisé afin de compléter, au moins en partie, les données manquantes restantes après le contrôle qualité mais également afin de déterminer pour chaque allèle aux marqueurs, son origine paternelle ou maternelle. Ces deux étapes nommées respectivement imputation et construction des phases se basent sur le mêmes principe : la reconstruction des segments chromosomiques (ou haplotypes) ancestraux. Deux grandes stratégies ont été développées pour réaliser cette opération : l'utilisation du déséquilibre de liaison à l'échelle de la population et l'utilisation de l'information familiale.

L'utilisation du déséquilibre de liaison permet de prédire à partir des allèles aux marqueurs adjacents et des observations réalisées sur l'ensemble de la population, l'allèle au marqueur manquant. Cette approche est particulièrement adaptée aux données de génétique humaine qui se caractérisent par une densité en marqueurs élevée (plusieurs millions de SNP) et une structure familiale limitée (moins de cinq individus par famille). Cette approche a été mise en œuvre pour le développement de plusieurs logiciels d'imputation tels que PHASE (Stephens et al., 2001), FastPHASE (Scheet et Stephens, 2006), ou BEAGLE (Browning et Browning, 2007) qui sont à présent utilisés dans de nombreuses espèces d'élevage. Leur principe se base sur une chaine de Markov à états cachés c'est-à-dire qu'ils se basent sur la succession des allèles aux marqueurs pour déterminer quel est l'allèle le plus probable au marqueur manquant. Ils considèrent que localement chaque région du génome correspond à un segment d'haplotype ancestral comme schématisé sur la Figure 4.

a)



b)

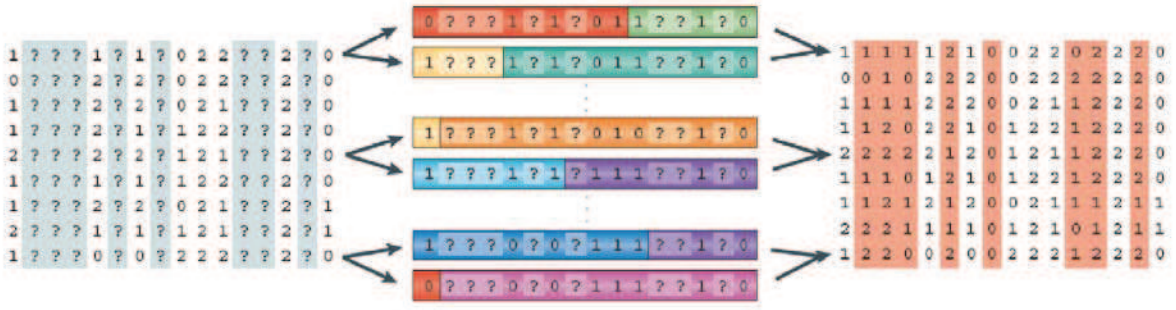


Figure 4 : Schématisation des méthodes d'imputation d'après Marchini et Howie (2010)

a) Haplotypes de référence correspondant aux haplotypes ancestraux
b) Les génotypes que l'on souhaite imputer sont d'abord phasés puis attribués à une mosaïque d'haplotype de référence, enfin les génotypes manquants sont imputés en utilisant les haplotypes de référence.

Le modèle de fonctionnement de Beagle est similaire à celui de FastPhase. La différence majeure entre les deux logiciels réside dans le nombre total d'haplotypes ancestraux possibles à un point donné : il est fixé dans FastPHASE alors qu'il varie en fonction du marqueur considéré dans Beagle.

Dans les espèces d'élevage et en particulier chez les bovins laitiers, les individus sont fortement apparentés et les pedigrees connus. Il semble donc sous-optimal de ne pas utiliser l'information familiale et le déséquilibre de liaison intra-famille pour l'imputation et la construction des phases. En effet lorsqu'un individu et l'un ou plusieurs de ses apparentés sont génotypés, il est possible en utilisant les règles mendéliennes de déterminer au moins partiellement quels sont les allèles reçus par l'individu. Par exemple, dans le cas simple où un individu est hétérozygote A/B et son père homozygote A/A, on en déduit facilement que l'animal a reçu l'allèle A de son père et l'allèle B de sa mère. Cette méthode utilisée par le logiciel LinkPHASE (Druet et Georges, 2010) ne permet pas de reconstituer intégralement les phases. Il a donc été proposé de combiner l'information familiale avec l'information populationnelle en utilisant dans une première étape le logiciel LinkPHASE puis dans un second temps un logiciel dérivé de BEAGLE : DAGPHASE (Druet et Georges, 2010).

L'imputation peut être utilisée pour compléter les marqueurs manquants suite à des problèmes techniques mais également, plus généralement, pour prédire les génotypes sur une puce donnée pour des animaux qui aurait été génotypés sur une autre. Elle a ainsi été largement utilisée pour prédire un génotype 50K à partir de génotypes de plus faible densité (Dassonneville et al., 2011; Mulder et al., 2012) mais également pour mutualiser des génotypes réalisés sur des puces différentes ce qui peut être le cas entre plusieurs pays, ou plusieurs protocoles expérimentaux, différents (Druet et al., 2010).

C. Estimation des effets des marqueurs

La population de référence, constituée d'individus disposant d'un phénotype et d'un génotype, est utilisée pour l'estimation des effets chromosomiques. Cette estimation pose de nombreux problèmes statistiques car il faut pouvoir estimer un grand nombre d'effets à partir d'un nombre d'individus relativement faible. De nombreuses méthodologies ont été proposées pour l'estimation des valeurs génétiques des animaux à partir de données génomiques. Plusieurs descriptions et comparaisons de l'efficacité de prédiction ont été réalisées (Daetwyler et al., 2013 ; de Los Campos et al., 2013 ; Robert-Granie et al., 2011). Nous décrirons ici les grands principes des méthodes d'évaluations génomiques les plus communes chez les bovins laitiers.

1. Méthodes des moindres carrés

La méthode des moindres carrées est une méthode statistique classique. Son principe de fonctionnement est simple : on cherche à minimiser le carré de l'écart entre valeurs prédites et valeurs observées. Cette méthode préconisée par Lande et Thompson (1990) pour la sélection assistée par marqueurs est difficilement applicable en sélection génomique. En effet, son inconvénient majeur est qu'elle ne peut s'appliquer uniquement dans les cas où le nombre de paramètres à estimer est plus faible que le nombre d'observations. Or, en évaluation génomique, on cherche à estimer l'effet de plusieurs dizaines de milliers de marqueurs à partir de quelques centaines ou milliers d'individus. Pour contourner ce problème, Meuwissen et al. (2001) a proposé une approche en trois étapes. Dans un premier temps, les effets individuels de chaque SNP sont estimés individuellement selon un modèle de régression simple. Dans un second temps, les marqueurs pour lesquels l'effet estimé est supérieur à un seuil préalablement fixé sont sélectionnés. Enfin, les effets des marqueurs sélectionnés sont ré-estimés simultanément en utilisant une régression multiple.

L'avantage de cette approche est qu'elle est très simple à mettre en œuvre d'un point de vue calculatoire. L'inconvénient est qu'elle nécessite une présélection des SNP à partir de leurs effets estimés individuellement. Le choix du seuil utilisé pour la présélection est ici déterminant. Un seuil trop élevé conduira à un faible nombre de SNP sélectionné et une précision élevée des estimations des effets des marqueurs mais la variance totale expliquée par les SNP sera faible. A l'inverse, un seuil trop faible conduira à la sélection de marqueurs inutiles et une faible précision des estimations des effets des marqueurs.

2. Méthodes de régression pénalisée

Afin d'éviter cette étape de présélection, plusieurs méthodes dites de régression pénalisée ont été développées. Il s'agit de méthodes de sélection de variables dont l'objectif est de retirer du modèle les variables les moins informatives pour mieux estimer celles qui ont un intérêt pour la prédiction. Appliquées en sélection génomique, les effets de l'ensemble des SNP sont estimés simultanément, et ce quel que soit le nombre d'observations, mais l'on ajoute une contrainte pour régresser vers zéro («pénaliser ») les estimations des effets. Autrement dit, on ajoute à la minimisation de la somme des carrés des résidus du modèle, une contrainte sur la somme des carrés des effets estimés. L'objectif est de faire tendre vers zéro les effets des SNP afin d'éliminer du modèle les SNP dont l'effet estimé est trop faible pour être réel. Nous décrirons ici, trois grandes méthodes de pénalisation des effets : la Regression Ridge, le Lasso et l'Elastic Net.

La Regression Ridge (Hoerl et Kennard, 1970) consiste à pénaliser les effets par la somme de leur carrés ce qui conduit à l'expression suivante :

$$\widehat{\boldsymbol{u}}_{\boldsymbol{Ridge}} = \boldsymbol{argmin} \left\{ \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \left(\boldsymbol{y_i} - \mu - \sum_{\boldsymbol{m=1}}^{\boldsymbol{p}} \boldsymbol{x_{im} g_m} \right)^{\boldsymbol{2}} + \boldsymbol{\lambda} \sum_{\boldsymbol{m=1}}^{\boldsymbol{p}} \boldsymbol{g_m^2} \right\}$$

Où :

- y_i la performance de l'individu i
- x_{im} le génotype de l'individu i au marqueur m
- g_m l'effet du marqueur m
- λ le paramètre contrôlant l'intensité de la pénalisation

Ce modèle considère que l'ensemble des marqueurs a un effet même si celui-ci peut être nul. En présence de variables fortement corrélées, ce qui est le cas de deux marqueurs en fort déséquilibre de liaison, la Regression Ridge retiendra plusieurs variables et répartira l'effet global sur chacune d'entre elles

Le LASSO (Least Absolute Shrinkage et Selection Operator) (Tibshirani, 1996) est une procédure très proche de la Regression Ridge. La différence entre les deux méthodes réside dans la pénalisation qui s'effectue non plus sur la somme des carrés des effets mais sur la somme de leurs valeurs absolues. Ici, le modèle ne retiendra parmi des variables corrélées, qu'une seule variable à laquelle il attribuera un effet fort. La contrainte majeure du LASSO, pour une utilisation en sélection génomique, est qu'il ne peut retenir un nombre d'effets à estimer supérieur au nombre d'observations.

Une méthode intermédiaire entre le LASSO et la Regression Ridge a été proposée par (Zou et Hastie, 2005). La pénalité utilisée ici est une combinaison linéaire de celle du LASSO et de la Regression Ridge. On introduit donc un paramètre α, compris entre 0 et 1, représentant la part de pénalité liée au LASSO et celle liée à la Regression Ridge. Si α vaut 1, le modèle est équivalent à un LASSO et si α vaut 0, le modèle est équivalent à une Regression Ridge. L'Elastic Net permet donc de combiner les avantages du LASSO et de l'Elastic Net. Il permet à la fois de pouvoir éliminer certains effets du modèle tout en conservant certaines variables corrélées. L'Elastic Net représente donc un bon compromis entre le LASSO et la Regression Ridge.

L'efficacité des méthodes de régressions pénalisées en sélection génomique a été démontrée. Ces méthodes sont cependant peu utilisées car elles nécessitent de déterminer par validation croisée, les valeurs des paramètres α et λ à utiliser.

3. GBLUP ou BLUP génomique

Une autre méthode de pénalisation des effets des SNP consiste à supposer que les effets des marqueurs sont distribués selon une loi normale centrée sur zéro. Dans ce cas, la méthode du BLUP, classiquement utilisé en évaluation génétique polygénique, peut être adaptée pour inclure l'estimation des effets des SNP. Ce modèle suppose que l'ensemble des marqueurs explique la variabilité du caractère et que chaque marqueur pris individuellement a une faible influence sur le caractère.

Plusieurs auteurs (Habier et al., 2007 ; VanRaden, 2008) ont montré que cette méthode est équivalente à un BLUP pour lequel on remplace la matrice de parenté espérée par la matrice de parenté réelle. Les éléments de cette matrice, dite de parenté génomique, mesurent la proportion d'allèles réellement partagée entre deux individus. Si le nombre de marqueurs utilisé pour son estimation est suffisamment important, la parenté génomique peut être plus précise que la parenté estimée sur pedigree car elle n'est pas dépendante de la connaissance exhaustive de la généalogie. D'autre part, elle permet de déterminer parmi une cohorte de pleins frères quels individus partagent le plus d'allèles en commun entre eux et avec d'autres individus (apparentés ou non).

Les deux méthodes étant équivalentes on préférera, pour des raisons calculatoires, utiliser une approche estimant les parentés génomiques lorsque le nombre d'individus sera faible et une approche basée sur l'estimation des effets des marqueurs lorsque le nombre d'individus sera supérieur au nombre de marqueurs.

4. Méthodes bayésiennes

Le BLUP génomique suppose que les effets des marqueurs sont tirés d'une loi normale de moyenne et variance unique. Ce modèle se rapproche d'un modèle polygénique et ne prend pas en compte le fait que certains marqueurs peuvent avoir un effet fort sur le caractère. Or, la variabilité génétique des caractères est le plus souvent la résultante d'un grand nombre de loci à effet faible et de quelques loci à effets forts (Hayes et Goddard, 2001), comme par exemple la mutation du gène DGAT1 sur le taux de matière grasse du lait (Grisart et al., 2002). L'utilisation des méthodes bayésiennes pour l'estimation des effets des marqueurs permet de lever cette contrainte en introduisant des distributions d'effets plus adaptées aux effets marqueurs.

On décrit alors la loi de distribution des effets des marqueurs avec une moyenne et une variance des effets qui peuvent varier d'un marqueur à l'autre. Une des caractéristiques principales des méthodes bayésiennes est qu'elles considèrent la distribution des effets comme un *a priori*. Cette information est combinée avec les observations afin d'obtenir des estimations *a posteriori* des effets qui combinent l'*a priori* avec ce qui aurait été obtenu en considérant uniquement les données. Plusieurs distributions *a priori* des effets ont été proposées. Nous décrirons ici les plus courantes d'entre elles.

La première méthode proposée, appelée « Bayes A », suppose que chacun des effets des marqueurs provient d'une distribution avec sa propre variance (Meuwissen et al., 2001), différente d'un marqueur à l'autre. Cette méthode revient à faire l'hypothèse que les effets des SNP suivent une loi t multivariée de faible degré de liberté (Gianola et al., 2009). Cette distribution ressemble à une loi normale avec des « queues » épaisses et autorise ainsi certains marqueurs à avoir un effet fort sur le caractère.

Cette méthode présente, elle aussi, un inconvénient car elle ne prend pas en compte le fait qu'un grand nombre de marqueurs ne sont liés à aucun QTL et n'ont donc aucun effet sur le caractère. Pour pallier à ce constat, Meuwissen et al. (2001) ont proposé une autre méthode, dite « Bayes B », qui suppose qu'une proportion (π) d'effets de marqueurs sont tirés d'une loi normale à variance nulle et n'ont donc pas d'effet sur le caractère. Les autres effets sont, comme pour le Bayes A, issus d'une distribution t. D'après Meuwissen et al. (2001), le Bayes B est la méthode permettant la meilleure qualité de prédiction génomique. Elle est donc souvent considérée comme la méthode de référence pour l'évaluation génétique mais son coût de calcul élevé a conduit au développement de nombreuses autres stratégies (Gianola, 2013).

Ces stratégies ont pour objectif de simplifier le modèle associé au BayesB pour réduire sa demande en ressources informatiques tout en maintenant une efficacité élevée. Le « Bayes C » suppose ainsi comme le BayesB, qu'une proportion π fixée des marqueurs a un effet nul sur le caractère mais considère que les effets des (1-π) marqueurs restants sont issus d'une loi de variance unique (Kizilkaya et al., 2010). Le « BayesCπ » repose sur les mêmes hypothèses mais, dans ce cas, la proportion π est estimée et non plus fixée (Habier et al., 2011). On peut également citer la méthode du « Bayes R » (Erbe et al., 2012) qui est une généralisation du BayesCπ introduisant un classement hiérarchique des effets des marqueurs et donc plusieurs classes de marqueurs selon la variance associée à leurs effets. Cette méthode permet de distinguer les marqueurs ayant un effet nul, un effet faible, moyen ou grand et se rapproche donc du BayesB. Le « BayesRS » développé par Brondum et al. (2012) est similaire au BayesR mais il prend en compte une information *a priori* sur l'importance de l'effet des marqueurs. Il a la particularité de pouvoir intégrer de l'information biologique et en particulier la localisation de gènes ayant un effet fort sur le caractère.

Méthodes de réduction de dimensions

Les méthodes de réduction de dimensions sont basées sur une autre stratégie. Elles supposent que même si le nombre de variables explicatives est élevé, un ensemble restreint de variables sous-jacentes permet d'expliquer une grande part de la variable réponse ce qui permet de réduire le nombre d'effets à estimer. Ces méthodes permettent ainsi d'établir des équations de prédiction à partir de variables latentes (sous-jacentes) issues de combinaisons linéaires des variables originelles. Elles permettent donc de réduire la complexité du modèle en réduisant le nombre de variables utilisées. Les méthodes de réduction de variables, les plus utilisées en génétique animale sont la PLS (Partial Least Square) et la Sparse PLS. Les deux méthodes diffèrent par leur manière de construire les variables latentes.

La régression PLS se rapproche d'une analyse en composante principale (ACP) à ceci près qu'elle prend en compte l'importance relative des variables latentes afin de prédire aux mieux la variable réponse. Ainsi, la régression PLS recherche les variables latentes qui ont une forte corrélation avec la variable réponse. La régression PLS est adaptée aux situations où le nombre d'observations est largement inférieur au nombre de variables à estimer et les variables sont fortement corrélées entre elles (Tenenhaus, 1998). Les effets des marqueurs présentant ces deux caractéristiques, l'utilisation de la régression PLS pour le traitement de données génomiques semble intéressante. Cependant dans le cas des données génomiques, on suppose généralement que seul un sous-ensemble de marqueurs a un effet sur le caractère. Cette particularité ne peut être prise en compte par la méthode régression PLS qui utilise l'ensemble des variables explicatives pour la construction des variables latentes. Une méthode dérivée de la régression PLS, la Sparse PLS, a donc été développée par Le Cao et al. (2008) pour le traitement des données génomiques et transcriptomiques. Cette méthode permet d'éliminer dans une première étape les variables initiales dont l'effet est négligeable sur la variable réponse.

5. Comparaison de l'efficacité des méthodes d'estimations des effets des marqueurs.

L'efficacité de la sélection génomique est généralement mesurée à travers la précision d'estimation des valeurs génétiques. De nombreux facteurs (dont la méthode utilisée pour la sélection génomique) que nous décrirons dans le prochain chapitre influencent l'efficacité de la sélection génomique. La comparaison des méthodes d'évaluations doit donc être réalisée sur des populations et caractères présentant les mêmes caractéristiques (idéalement une population et des caractères identiques).

La comparaison de méthodes sur données simulée apparait donc comme une stratégie intéressante mais elle nécessite de faire des hypothèses sur le déterminisme génétique des caractères qui peuvent influer sur le classement des méthodes d'estimations des effets marqueurs. Une première comparaison de la méthode des moindres carrés, du GBLUP, du BayesA et du BayesB a été réalisée pour un caractère soumis à de nombreux QTL à effet faible, un nombre limité de QTL à effet moyen et quelques QTL à effet fort (Meuwissen et al., 2001). Cette étude a démontré une forte supériorité (+60% de précision) des méthodes du GBLUP, BayesA, BayesB par rapport à la méthode des moindres carrés car elles ne nécessitent pas de présélectionner les effets à estimer. Il a également été observé une infériorité du GBLUP (-10%) par rapport aux méthodes bayésiennes et une légère supériorité du BayesB comparé au BayesA (+6%). Cependant, Daetwyler et al. (2010). a démontré que l'efficacité du GBLUP et du BayesB sont dépendantes du déterminisme génétique du caractère. On peut observer sur la Figure 5 que l'efficacité du GBLUP est très peu sensible à l'architecture génétique du caractère. A l'opposé, le BayesB est très efficace lorsqu'un faible nombre de QTL influence le caractère et devient moins précise que le GBLUP lorsque le nombre de QTL augmente.

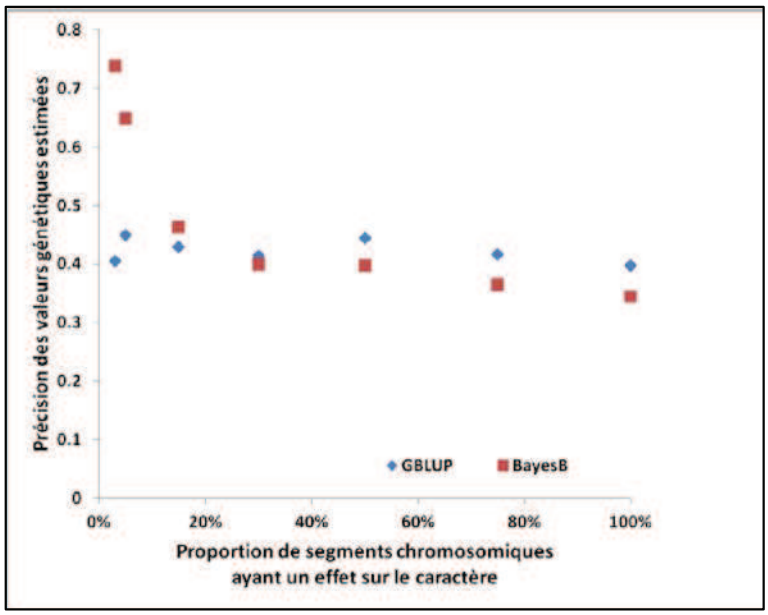


Figure 5 : Précision des évaluations génomiques en fonction de la méthode et de la proportion de segments chromosomiques ayant un effet sur le caractère, d'après Daetwyler et al. (2010)

Sur données réelles, de nombreuses études ont cherché à comparer l'efficacité des différentes méthodes d'estimations des effets des marqueurs. Une synthèse des résultats obtenus a été réalisée par De Los Campos et al. (2013). Tout d'abord, les différentes études ont montré que les méthodes de sélection génomique permettent une bien meilleure estimation des valeurs génétiques des animaux que les évaluations polygéniques classiques (utilisant la méthode BLUP). Elles ont confirmé les capacités prédictives supérieures des méthodes bayésiennes. Ce résultat peut s'expliquer par la faculté de ce type de méthode à s'adapter aux différents caractères et à leur architecture génétique (nombre de QTL, présence ou non de gènes à effet fort) (De Los Campos et al., 2013).

Contrairement à ce qui avait pu être observé sur données simulées, la méthode du GBLUP a également démontré de bonnes capacités de prédiction dans pratiquement tous les cas. Cette relativement bonne efficacité du GBLUP peut s'expliquer par l'architecture des caractères, plus complexe dans le cas de données réelles que de données simulées, et la moindre sensibilité du GBLUP au déterminisme génétique des caractères.

Il est important de noter que mêmes si les corrélations entre les effets SNP estimés par les différentes méthodes sont faibles (≈0,7), les corrélations entre valeurs génétiques estimées (c'est-à-dire la somme des effets SNP) sont plus élevées (0,8-0,9). On peut donc obtenir des valeurs génétiques similaires avec des estimations d'effets SNP variables (De Los Campos et al., 2013). On notera également que plus la taille de population de référence augmente, plus les estimations des valeurs génétiques issues des différentes méthodes de prédiction deviennent homogènes (De Los Campos et al., 2013).

D. Facteurs influençant l'efficacité de la sélection génomique

Plusieurs travaux ont cherché à estimer et à prédire l'efficacité des évaluations génomiques. Si l'on fait l'hypothèse que les marqueurs moléculaires expliquent l'intégralité de la variance génétique, la précision des évaluations génétiques peut être prédite à partir de la formule ci-dessous (Daetwyler et al., 2013) :

$$\rho = \sqrt{\frac{N_p h^2}{N_p h^2 + M_e}}$$

Où :

- ρ la corrélation entre valeurs génomiques estimées et valeurs génétiques vraies
- N_p est le nombre d'individus dans la population de référence
- h^2 est l'héritabilité du caractère
- Me est le nombre de segments chromosomiques indépendants (Daetwyler et al., 2008)

Le nombre de segments chromosomiques indépendants représente le nombre d'effets de marqueurs statistiquement indépendants à estimer avec la population dont on dispose. Il dépend de la structure de la population et notamment de la valeur du déséquilibre de liaison moyen entre deux marqueurs successifs. Le nombre de segments indépendants s'estime comme quatre fois la taille efficace (Ne) de la population multipliée par la longueur totale du génome (L) exprimé en Morgan (Goddard, 2009b ; Hayes et al., 2009b).

Effet de la taille efficace de la population

La taille efficace (Wright, 1931) de la population correspond au nombre équivalent d'individus dans une population idéale (c'est-à-dire non sélectionnée et pour laquelle la reproduction des individus est aléatoire) de même diversité génétique. Plus la taille efficace d'une population est importante, plus le déséquilibre de liaison décroit rapidement et donc, plus le nombre de segments chromosomiques indépendants (le nombre d'effets à estimer) augmente.

En utilisant la formule de Daetwyler et al. (2013), on peut estimer que pour une longueur du génome de 30Mb, une taille de population de référence de 1500 individus, la précision estimée est de 0,40 pour une taille efficace de 20, de 0,29 pour une taille efficace de 40 et de 0,19 pour une taille efficace de 100.

L'effet de la taille efficace sur la précision de la sélection génomique explique en partie pourquoi, en bovins allaitants, la sélection génomique a été mise en place plus tardivement qu'en bovins laitiers. En effet pour un caractère de même héritabilité et une même taille de population de référence, la précision estimée est deux fois inférieure en bovins allaitants où la taille efficace est comprise entre 100 et 200 individus, qu'en bovins laitiers où elle est comprise entre 20 et 30 (Danchin-Burge, 2009). Il faut donc, en bovins allaitants, augmenter la taille de la population de référence pour atteindre la même précision qu'en bovins laitiers.

1. Effet de la taille de la population de référence

L'augmentation de la taille de la population de référence permet d'augmenter le nombre d'observations utiles pour l'estimation des effets des segments chromosomiques et donc d'améliorer la qualité de prédiction.

Ainsi quand la précision théorique est de 0,39 pour une taille efficace de 20, une héritabilité de 0,3, un génome de 30 Morgan et une population de référence de 1500 individus, elle n'est que de 0,24 quand la population de référence compte 500 individus et de 0,49 lorsqu'elle en compte 2500 individus. L'effet de la population de référence n'est pas linéaire, l'ajout de 1000 individus à une population de référence de 500 augmente la précision estimée de 0,15 mais ajouter à nouveau 1000 individus ne l'augmente que de 0,10.

L'effet de la taille de la population de référence a également pu être observé sur données réelles. Liu et al. (2011) a ainsi démontré que les effets SNP estimés à partir d'une population de 5025 taureaux étaient plus stables d'une évaluation à l'autre et avaient une variance plus importante que ceux estimés à partir de 735 taureaux. On retrouve donc bien ici le fait qu'une population de référence plus importante permet une meilleure estimation des effets SNP.

La mutualisation des populations de référence Holstein entre pays européens en 2009 au sein du consortium EuroGenomics et l'augmentation de la taille de la population de référence de 4000 à 16 000 taureaux a, quant à elle, permise un gain de précision des évaluations génomiques de 10% (Lund et al., 2011).

2. Effet de l'héritabilité du caractère

Un autre facteur influant sur la qualité d'estimation est la qualité de l'information apportée par un individu. La corrélation théorique entre les phénotypes observés dans une population et les valeurs génétiques vraies est égale à la racine carrée de l'héritabilité. Plus l'héritabilité d'un caractère sera élevée, plus le phénotype d'un animal sera proche de sa valeur génétique et donc plus l'information sera utile pour l'estimation des effets des marqueurs.

En reprenant l'exemple précédent (Ne=20, N_p=1500, L=30), la précision estimée des évaluations est de 0,39 pour un caractère dont l'héritabilité est 0,3 (ex : production laitière), 0,11 pour un caractère dont l'héritabilité est de 0,02 (ex : taux de conception) et 0,50 pour un caractère dont l'héritabilité est de 0,50 (ex : hauteur sacrum). Cette relation engendre donc de forts écarts de précision entre les caractères et une faible capacité de prédiction des valeurs génétiques pour les caractères faiblement héritables. Pour contourner cette difficulté, on cherchera à utiliser non plus la performance brute de l'animal comme phénotype mais la performance moyenne de ses produits. Une population de référence constituée de taureaux testés sur descendance pour lesquels le phénotype est le DYD permet ainsi d'avoir une précision souvent équivalente à un caractère d'héritabilité 0,90.

La Figure 6 réalisée à partir de la formule ci-dessus illustre l'impact combiné de la taille de la population de référence, l'héritabilité du caractère et de la taille efficace de la population sur la précision des évaluations génomiques. On observe ainsi qu'une taille efficace élevée ou une héritabilité faible, peuvent être compensée par une population de référence de taille importante. Il faut donc prendre en compte ces facteurs lors de la constitution de la population de référence et le développement de la sélection génomique.

3. Effet du choix des animaux constituant la population de référence

Nous avons vu dans le paragraphe précédent que la taille de la population de référence était un facteur limitant majeur de l'efficacité des évaluations génomiques. La formule (Daetwyler et al., 2013; Goddard, 2009a) décrite au début du chapitre laisse à penser que le nombre d'animaux de la population de référence est le seul facteur sur lequel il est possible d'agir pour améliorer l'efficacité de la sélection génomique. Or, le choix des animaux constituant la population de référence est lui aussi déterminant.

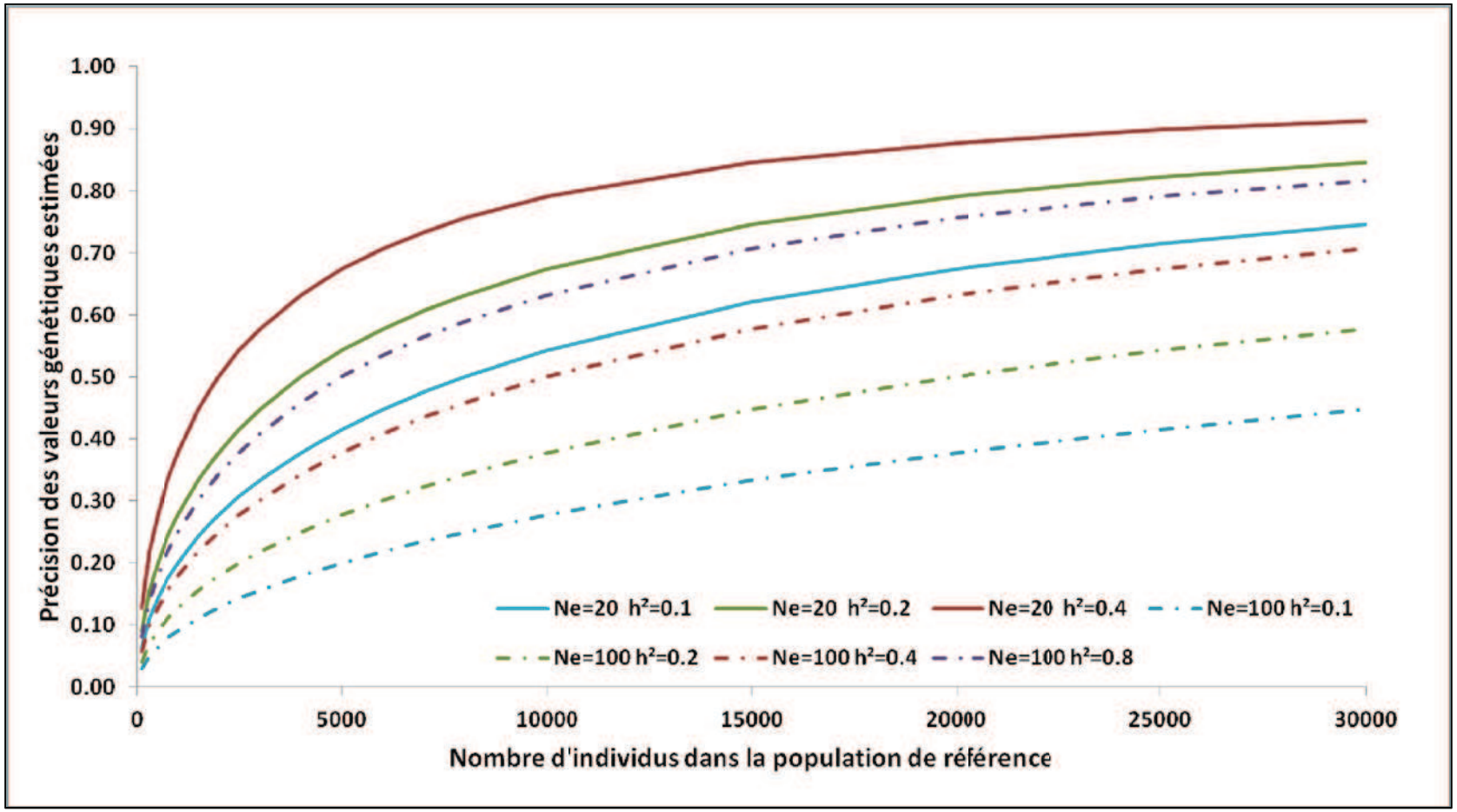


Figure 6 : Impact de la taille de la population de référence, de l'héritabilité (h^2) et de la taille efficace (Ne) de la population sur la précision des évaluations génomiques, d'après Goddard et Hayes (2009).

D'une part, l'utilisation de taureaux testés sur descendance permet d'augmenter artificiellement l'héritabilité du caractère et ainsi d'augmenter les capacités prédictives des évaluations génomiques. D'autre part, plusieurs études (Clark et al., 2012 ; Habier et al., 2007 ; Habier et al., 2010 ; Pszczola et al., 2012) ont montré que les parentés intra-population de référence, et entre les animaux de la population de référence et ceux de la population que l'on cherche à sélectionner, sont également à prendre en considération.

A taille de population fixée, il est préférable de sélectionner les animaux les moins apparentés possibles pour la constitution de la population de référence (Pszczola et al., 2012). Maximiser la diversité génétique de la population de référence permet de refléter au mieux la diversité des allèles présents dans la population et d'équilibrer leur représentation dans la population. Cette stratégie limite le risque qu'un allèle présent chez les candidats à la sélection soit absent (ou présent uniquement chez quelques individus) de la population de référence ce qui améliore l'estimation des effets des marqueurs et donc celle des valeurs génomiques.

A l'inverse, la précision des valeurs génétiques estimées est plus élevée lorsque l'apparentement entre la population de référence et la population de validation est fort (Clark et al., 2012 ; Pszczola et al., 2012). Cet apparentement permet d'éviter qu'un allèle soit présent chez la population de validation et absent de la population d'apprentissage mais également de profiter du déséquilibre de liaison à longue distance. En effet, Habier et al. (2013) a montré que pour un candidat donné, la précision de la valeur génomique estimée est liée au déséquilibre de liaison à courte distance observable sur l'ensemble de la population mais également au déséquilibre de liaison à plus longue distance observable uniquement intra-famille et correspondant à des segments chromosomiques transmis par un ascendant proche. La précision liée au déséquilibre populationnel correspond donc à la valeur minimale obtenue pour un individu non apparenté à la population de référence, le supplément de précision est, quant à lui, fonction de l'apparentement de l'individu à la population de référence.

Ces résultats ont été confirmés par les études de Rincent et al. (2012) chez le maïs qui a comparé les capacités prédictives de plusieurs populations de référence et Bapst et al. (2013) chez les bovins montrant que l'ajout de femelles dans la population de référence était plus utile lorsque celles-ci étaient génétiquement éloignées des taureaux de la population de référence. Les animaux de la population de référence doivent donc être les moins apparentés possibles entre eux pour représenter au mieux la variabilité de la population évaluée et contenir des individus proches des candidats à la sélection (idéalement leurs parents). Cette stratégie permet de maximiser à la fois le déséquilibre de liaison populationnel et le déséquilibre de liaison familial.

Au fil des générations de nombreuses recombinaisons tendent à faire disparaitre le lien entre un marqueur et un QTL. Le renouvellement régulier de la population de référence est donc nécessaire afin de maintenir un lien fort entre population de référence et candidats à la sélection. Plus généralement, le lien entre marqueur et QTL observé chez les animaux de la population de référence doit être conservé chez les candidats à la sélection. La précision des évaluations génomiques diminue donc lorsque la population de référence n'est pas régulièrement complétée (Bastiaansen et al., 2012 ; Solberg et al., 2009). Le déséquilibre de liaison familial est un déséquilibre de liaison à grande distance, il disparait donc rapidement d'une génération à l'autre. La précision des évaluations atteint alors son niveau lié au déséquilibre de liaison populationnel. Au cours des générations (et des recombinaisons), le déséquilibre de liaison populationnel diminue lui aussi, ce qui entraine une diminution de la précisions des valeurs génomiques.

La Figure 7, issue de l'étude de Bastiaansen et al. (2012), illustre cette diminution d'abord rapide puis plus lente de la précision des évaluations génomiques. Habier et al. (2007 ; 2010) a montré que selon les méthodes de sélection génomique, le rapport entre la précision liée au déséquilibre de liaison familial et au déséquilibre de liaison populationnel pouvait varier. Le GBLUP par exemple est une méthode basée sur le déséquilibre de liaison familial alors que les méthodes bayésiennes s'appuient sur le déséquilibre de liaison populationnel. En revanche, aucune différence sur la persistance de la précision des évaluations génomiques n'a été observée entre les méthodes bayésiennes et le GBLUP (Bastiaansen et al., 2012 ; Wolc et al., 2011).

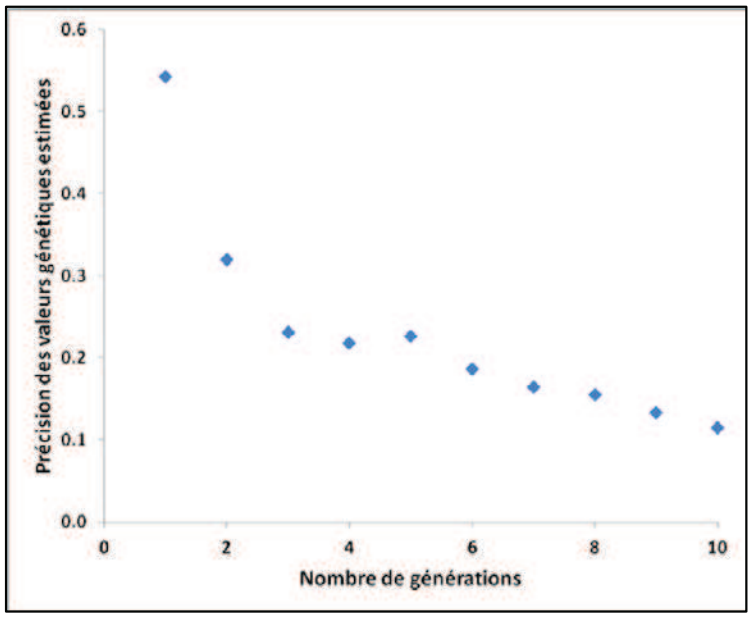


Figure 7 : Evolution de la précision des valeurs génomiques estimées en fonction du nombre de générations séparant la population de référence et les candidats à la sélection, d'après Bastiaansen et al. (2012)

4. Effet de la densité en marqueurs

La précision des évaluations génomiques étant liée au lien entre les marqueurs et le QTL. Plus la distance entre marqueurs diminue, plus le déséquilibre de liaison moyen observé entre deux marqueurs successifs est fort et donc la probabilité qu'un marqueur soit fortement lié à un QTL augmente.

Il existe donc une relation positive entre la densité en marqueurs et la précision des évaluations génomiques (Calus et al., 2008). VanRaden (2008) a observé un gain de précision en augmentant le nombre de marqueurs de 10 000 à 20 000 puis de 20 000 à 40 000. En revanche, un gain faible (1%) a été observé lors du passage d'une puce contenant 54 609 marqueurs (50K) à une puce haute-densité (HD) contenant 777 962 marqueurs (Erbe et al., 2012; Su et al., 2012 ; VanRaden et al., 2013). Sur données simulées, Calus et al. (2008) a montré que l'effet de la densité en marqueurs diminuait rapidement une fois une valeur critique atteinte. En sélection génomique intra-race laitière, ce plateau est probablement déjà atteint avec une densité d'un SNP tous les 70kb (50K). L'ajout de SNP additionnels apporte alors peu d'informations.

Lorsque la densité en marqueurs augmente, la distance entre un marqueur et un QTL diminue, ce qui réduit la probabilité qu'une recombinaison ait lieu entre les deux loci et augmente la persistance de la précision des évaluations génomiques au cours des générations. L'étude de Solberg et al. (2009) montre bien une diminution plus lente de la précision des évaluations génomiques lorsque la densité en marqueurs augmente (Figure 8).

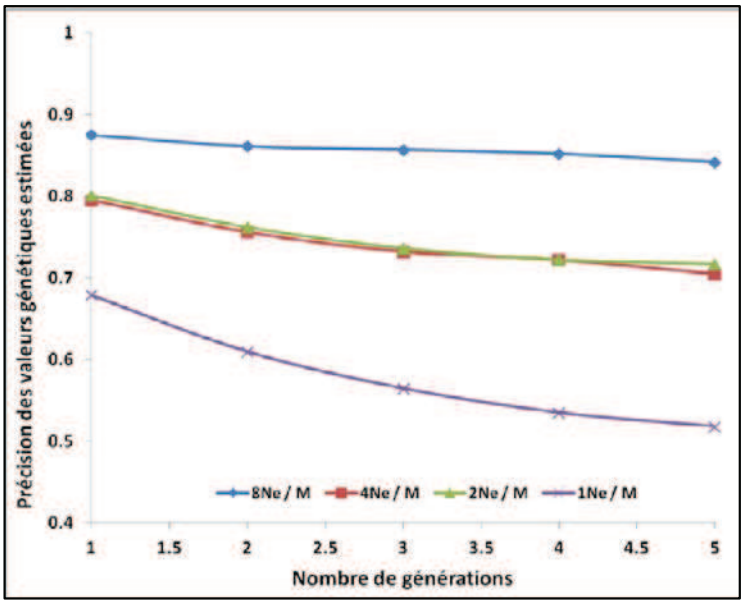


Figure 8 : Evolution de la précision des valeurs génomiques estimées en fonction du nombre de générations séparant la population de référence et les candidats à la sélection pour des densités en marqueurs égale à 8, 4, 2 ou 1 fois la taille efficace de la population (ici Ne=100) d'après Solberg et al., (2009)

Le lien entre marqueurs et QTL peut également être augmenté par l'utilisation d'haplotypes (groupe de marqueurs proches). En effet, la sélection génomique repose sur le suivi de segments chromosomiques et suppose que l'association entre un allèle à un QTL et un allèle à un marqueur, observée sur la population de référence, est conservée dans la population à évaluer. Du fait des recombinaisons aléatoires lors de la méiose, la probabilité que deux individus portent le même haplotype sans porter le même allèle au QTL est faible car cela impliquerait une double recombinaison autour du QTL. Considérer non plus un marqueur unique mais un groupe de marqueurs devrait donc permettre de mieux suivre l'haplotype ancestral porteur de la mutation causale et donc le QTL (Goddard et Hayes, 2007).

Les études de Calus et al. (2008) et Villumsen et al. (2009) confirment cette hypothèse. Le gain de précision permis par l'utilisation d'haplotypes est particulièrement important lorsque le déséquilibre de liaison entre les marqueurs adjacents et donc, la densité en marqueurs, sont faibles (Calus et al., 2008). Lorsque le déséquilibre de liaison entre le QTL et le marqueur est déjà élevé, l'utilisation d'haplotypes reste intéressante pour capturer un allèle rare au QTL à partir d'un haplotype constitué de marqueurs de fréquences plus élevées.

L'utilisation d'haplotypes peut donc permettre d'améliorer la précision des évaluations génomiques, mais également la persistance de cette précision au fil des générations (Figure 9) (Villumsen et al., 2009). A l'image de l'augmentation de la densité en marqueurs, l'utilisation d'haplotypes permet de mieux conserver le déséquilibre de liaison d'une génération à l'autre. La longueur optimale de l'haplotype semble se situer entre 5 et 10 marqueurs (Guillaume, 2009 ; Villumsen et al., 2009). Utiliser un nombre important de marqueurs par haplotype augmente le nombre total d'allèles observés et donc le nombre total d'effets à estimer ce qui réduit les précisions d'estimations de chacun des effets. Plusieurs stratégies que nous ne détaillerons pas ici ont été proposées pour regrouper les haplotypes entre eux (Calus et al., 2009) afin de réduire le nombre d'effets à estimer tout en maintenant une capacité prédictive plus élevée que celle obtenue par l'utilisation de marqueurs seuls.

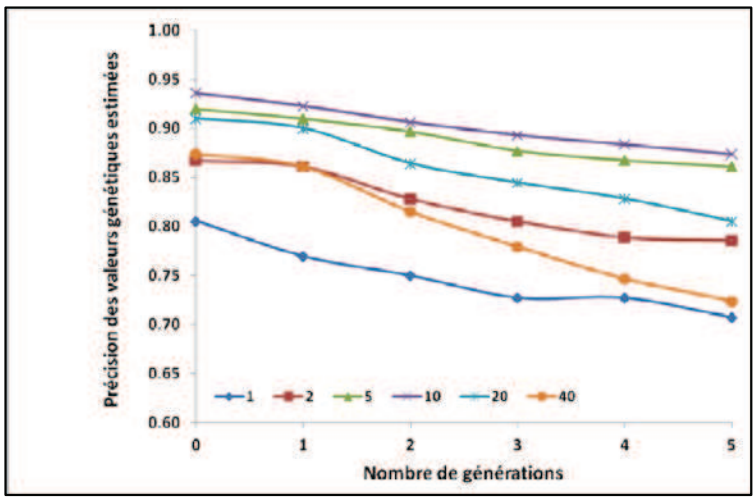


Figure 9 : Evolution de la précision des valeurs génomiques estimées en fonction du nombre de générations séparant la population de référence et les candidats à la sélection et de la longueur de l'haplotype (ici constitué de 1, 2, 5, 10, 20 ou 40 marqueurs) d'après Villumsen et al. (2009)

L'augmentation de la densité en marqueurs et/ou l'utilisation d'haplotypes permettent donc d'augmenter le déséquilibre de liaison entre un marqueur et un QTL. Plus le déséquilibre de liaison sera élevé, plus l'effet estimé pour le marqueur sera représentatif de l'effet du QTL. La formule de prédiction de la précision des évaluations génomiques (Daetwyler et al., 2013; Goddard, 2009a) suppose que l'ensemble de la variabilité génétique est expliquée par les marqueurs et néglige donc le déséquilibre de liaison incomplet entre les marqueurs et les QTL. Il faut donc y appliquer un terme correctif correspondant à la racine carrée de la part de la variance génétique totale expliquée par les marqueurs (Calus et al., 2013a). Une valeur généralement admise pour cette part de variance en bovins laitiers est de 80% (Calus et al., 2013a) mais elle est très dépendante du déterminisme génétique réel du caractère. Elle n'est par exemple, que de 55% pour les caractères de fertilité et de longévité (Haile-Mariam et al., 2013).

La variabilité génétique expliquée par les marqueurs est donc souvent élevée et permet d'atteindre une précision des évaluations génomiques supérieure à celles des évaluations polygéniques classiques. De nombreux pays ont donc choisi d'intégrer une étape de sélection génomique dans leur schéma de sélection.

E. Le développement des évaluations génomiques en France et à l'international

Les schémas de sélection mis en place chez les bovins laitiers permettent grâce au testage sur descendance de disposer de taureaux dont la valeur génétique est connue avec précision. Ce dispositif a pu être mis à profit pour constituer des populations de référence et détecter des QTL en lien avec les caractères de productions (Khatkar et al., 2004).

1. Utilisation de la Sélection Assistée par Marqueurs : l'exception française

Une première approche de sélection génomique consiste à intégrer explicitement les régions QTL identifiées dans les évaluations génomiques et les programmes de sélection. Historiquement, l'Allemagne (Bennewitz et al., 2003 ; Bennewitz et al., 2004), les Etats-Unis, la Nouvelle-Zélande et la France (Boichard et al., 2002) ont lancé de grands programmes de détection de QTL en vue de la mise en place d'une Sélection Assistée par Marqueurs (SAM). En pratique, les difficultés techniques liées à l'utilisation de marqueurs microsatellites les ont contraints à abandonner la SAM. Seule la France a développé et maintenu une évaluation SAM au niveau national dont nous décrirons ici les évolutions successives.

Un vaste programme de détection de QTL financé par les pouvoirs publics, l'INRA (Institut Nationale de la Recherche Agronomique), l'UNCEIA (Union Nationale des Coopératives d'Elevage et d'Insémination Animale) et les principales entreprises de sélection françaises a été conduit en France de 1996 à 1999 dans les principales races laitières (Holstein, Montbéliarde et Normande). Ce programme était basé sur les génotypes pour 157 marqueurs microsatellites de 1 548 taureaux issus de 14 familles de pères et a permis l'identification d'une centaine de QTL (Boichard et al., 2003).

Suite à cette étape de détection de QTL, un programme de sélection assistée par marqueurs a pu être mis en place dès 2001. La Sélection Assistée par Marqueurs de première génération (SAM1) se basait alors sur 43 marqueurs microsatellites pour suivre une quinzaine de QTL par caractère. La répartition hétérogène des marqueurs microsatellites sur le génome empêchaient une localisation précise des QTL. Par conséquent, la liaison entre le marqueur et le QTL n'était pas conservée au sein de l'ensemble de la population et l'estimation des effets des QTL devait être réalisée intra-famille. En 2007, à la fin de ce programme, ce modèle d'évaluation permettait d'obtenir pour un jeune animal et le caractère « production laitière », un coefficient de détermination d'une valeur de 0,44 contre 0,33 avec un modèle sur ascendance (Fritz et al., 2007). Il était donc possible de mesurer en partie l'aléa de méiose et ainsi mieux choisir quels animaux d'une fratrie devaient être testés sur descendance.

En 2008, avec la disponibilité d'un grand nombre de marqueurs SNP à travers la puce Illumina BovineSNP50® (50K) (Matukumalli et al., 2009), le programme de SAM1 a évolué vers une Sélection Assistée par Marqueurs de deuxième génération (SAM2). De grandes populations de taureaux testés sur descendance issus des trois grandes races laitières ont été génotypées sur la puce 50K et ont permis de localiser finement, par analyse utilisant conjointement l'information de liaison intra famille et de déséquilibre de liaison (LDLA, (Druet et al., 2008 ; Meuwissen et al., 2002), plusieurs centaines de QTL par race (Fritz et al., 2008). Une quarantaine de QTL par caractère ont alors été sélectionnés et intégrés dans le modèle d'évaluation génétique. Cette SAM2 a permis une forte augmentation de la précision des index pour les animaux sans performance. Si on prend l'exemple de la race Holstein, la corrélation moyenne entre valeur génétique estimée et phénotype observé sur les cinq caractères de production est passée de 0,38 avec une évaluation polygénique à 0,58 avec la SAM2 (Guillaume et al., 2008).

En 2010, le programme de SAM2 a évolué vers une « SAM génomique » : la SAMg. Ce modèle se rapproche de la sélection génomique puisque l'on considère plusieurs centaines de régions QTL par caractère (Boichard et al., 2012b). Les régions QTL sont sélectionnées à partir d'une étude de détection de QTL d'une part, et une méthode de sélection de variable, l'Elastic Net (Croiseau et al., 2012), d'autre part. La corrélation entre les phénotypes et les valeurs génomiques obtenues par cette méthode est comparable à celle observée avec les méthodes de sélection génomique classique : entre 0,6 et 0,8 pour les caractères de production et 0,3 pour la fertilité (Tableau 3). La précision de ces évaluations a permis aux entreprises de sélection françaises d'utiliser directement les taureaux sélectionnés sur la base de leurs index génomiques officiels dès Juin 2010 et ainsi d'abandonner le système de testage sur descendance organisé. La technologie a rapidement été adoptée par les entreprises de sélection, puisque dès 2011, les jeunes taureaux représentaient entre 30 et 40% des inséminations bovines laitières (Boichard et al., 2012b). Aujourd'hui, plus de 50% des inséminations bovines laitières sont réalisées avec de la semence de jeunes taureaux (Tableau 4). Cette adoption rapide a été permise par l'historique de sélection assistée par marqueurs et une logistique à laquelle les professionnels étaient familiarisée depuis 2001.

Tableau 3 : Corrélations entre valeurs génomiques estimées selon différentes approches et déviations moyennes des filles (DYD) en race Holstein pour les animaux de la population de validation, d'après Boichard et al. (2012b)

Méthode	**Production laitière**	**Matière protéique**	**Matière grasse**	**Taux protéique**	**Taux butyreux**	**Taux de conception (Fertilité)**
BLUP	0,38	0,44	0,40	0,47	0,44	0,29
Elastic-Net	0,57	0,57	0,63	0,75	0,80	0,34
GBLUP	0,56	0,55	0,59	0,73	0,72	0,35
SAMg	0,60	0,57	0,66	0,73	0,81	0,31

Tableau 4 : Utilisation des jeunes taureaux et des taureaux testés sur descendance en 2013 dans les trois grandes races laitières françaises, d'après S. Moureaux (Institut de l'Elevage, Jouy en Josas, communication personnelle)

Race	Catégorie	Nombre	Nombre moyen de doses par taureau	% d'inséminations de jeunes taureaux
Montbéliarde	Jeunes taureaux Taureaux confirmés	233 126	1975 3 500	56%
Normande	Jeunes taureaux Taureaux confirmés	191 81	1 550 5 560	55%
Holstein	Jeunes taureaux Taureaux confirmés	590 170	3 940 8 200	69%

2. Utilisation de la sélection génomique

Au niveau mondial, les approches de sélection génomique ont été privilégiées du fait de leur facilité de mise en œuvre (en particulier pour le GBLUP) et leur précision d'estimation des valeurs génétiques (Meuwissen et al., 2001 ; VanRaden, 2008).

De nombreux pays ont génotypé dès 2009, des animaux dont la valeur génétique était connue avec précision (majoritairement des taureaux testés sur descendance) afin de construire des populations de référence. Dans sa revue bibliographique Hayes et al. (2009a) a comparé les résultats obtenus dans les situations australienne, néo-zélandaise, hollandaise et américaine. En fonction du pays, du modèle utilisé et du caractère étudié, la précision des évaluations génomiques varient fortement de 0,20 à 0,67 (Hayes et al., 2009a). Ces valeurs obtenues grâce à l'information moléculaire étaient dans tous les cas significativement plus précises que celles obtenues par une évaluation classique utilisant uniquement l'information liée à l'ascendance.

Les perspectives offertes par la sélection génomique ont conduit au développement d'évaluations génomiques à travers le monde. Alors qu'en 2009, les évaluations génomiques ne concernaient qu'un nombre limité de pays, de races et de caractères (Loberg et Dürr, 2009), vingt pays et six races disposent aujourd'hui d'évaluations génomiques (Tableau 5). La plupart de ces évaluations répondent aux critères de validation des évaluations génomiques développés par Interbull (Loberg et al., 2011).

Tableau 5 : Liste des pays et des races disposant d'évaluations génomiques. Les évaluations génomiques validées par Interbull sont indiquées en souligné.
D'après Loberg et Dürr (2009) et Interbull, Uppsala (Suède)

Holstein	Brune	Jersiaise	Rouge Nordique	Simmental	Montbéliarde	Normande
Allemagne	Allemagne	Australie	Canada	Allemagne	France	France
Australie	Autriche	Canada	Danemark	Autriche		
Autriche	Canada	Danemark	Finlande	Italie		
Belgique	Etats-Unis	Etats-Unis	Norvège	Suisse		
Canada	Suisse	Finlande	Suède			
Danemark		Nouvelle Zélande				
Espagne	Allemagne*	Suède				
Etats-Unis	Autriche*					
Finlande	Etats-Unis*					
France	France*					
Grande-Bretagne	Italie*					
Irlande	Slovénie*					
Israël	Suisse*					
Italie						
Japon						
Nouvelle-Zélande						
Pays-Bas						
Pologne						
Suède						
Suisse						

*: Evaluations génomiques internationales: InterGenomics

De nombreux consortiums internationaux ont été mis en place afin de mutualiser les populations de référence nationales et augmenter la précision des évaluations. En race Holstein, on pourra citer le consortium Nord-Américain (Muir et al., 2010) regroupant initialement les Etats-Unis et le Canada, qui s'est élargi en 2011 au Royaume Uni et à l'Italie. Toujours en race Holstein, le consortium EuroGenomics (Allemagne, Danemark, Suède, Finlande, France et Pays-Bas initialement, puis Espagne en 2011 et Pologne en 2012) qui a permis à chacun des pays participant d'accéder à une population de référence de 16 000 taureaux en 2010 (Lund et al., 2011) et plus de 25 000 aujourd'hui (Liu et al., 2013). Dans les autres races, des coopérations internationales ont également vu le jour avec par exemple le développement du consortium Intergenomics regroupant sept pays (Allemagne, Autriche, Etats-Unis, France, Italie, Suisse, Slovénie) qui porte la population de référence de race Brune à plus de 7000 animaux génotypés sur la puce 50K (Dürr et Philipsson, 2012). On notera également l'existence d'un consortium regroupant les races rouges nordiques (Brondum et al., 2011).

3. Extension des évaluations génomiques aux races régionales

Les évaluations génomiques ont entrainé un fort remaniement des schémas de sélection. Cependant, leur efficacité étant très dépendante de la taille de la population de référence, elles ont majoritairement été mises en place dans les races (inter)nationales (Bouquet et Juga, 2013). Dans les races régionales, et en particulier les races régionales françaises, le nombre réduit de mâles testés sur descendance (souvent une dizaine, Tableau 2), empêche la constitution de la population de référence de taille suffisante et donc la mise en place d'évaluations génomiques efficaces.

Il a alors été proposé de regrouper les populations de référence de plusieurs races, ce qui permettait de constituer une population de référence de taille importante pour un coût plus modéré (mutualisation de l'investissement financier). Cette stratégie nécessite le développement d'évaluations génomiques multiraciales et soulève de nombreuses questions sur la précision des estimations des effets des marqueurs pour chaque race et donc la précision des valeurs génomiques.

En effet, utiliser une population de référence multiraciale suppose que les régions QTL (et les mutations causales) soient communes entre races et que le lien entre QTL et marqueur soit conservé entre races. Chez les bovins laitiers, il a été démontré par simulation que cette seconde hypothèse était vérifiée uniquement dans le cas où la distance entre marqueurs consécutifs est inférieure à 10kb (De Roos et al., 2008 ; Gautier et al., 2007a). Cette hypothèse n'est donc pas vérifiée avec la puce BovineSNP50® (50K) pour laquelle la distance moyenne entre deux marqueurs consécutifs est de 50Kb (Matukumalli et al., 2009).

En avril 2010, une puce haute densité, la BovineHD BeadChip® (HD), contenant 777 609 marqueurs et pour laquelle la distance moyenne entre deux marqueurs consécutifs est de 4Kb, a été développée par Illumina. Les travaux sur données simulées ont démontré que l'utilisation de cette puce haute densité peut permettre le développement d'évaluations génomiques multiraciales efficaces (De Roos et al., 2009; Toosi et al., 2010). Sur données réelles, Larmer et al., (2014) a démontré que le déséquilibre de liaison était mieux conservé entre race sur la puce HD que sur la puce 50K (Figure 10). Il devient donc envisageable de tirer avantage de l'information moléculaire dans les races régionales à l'aide des évaluations multiraciales.

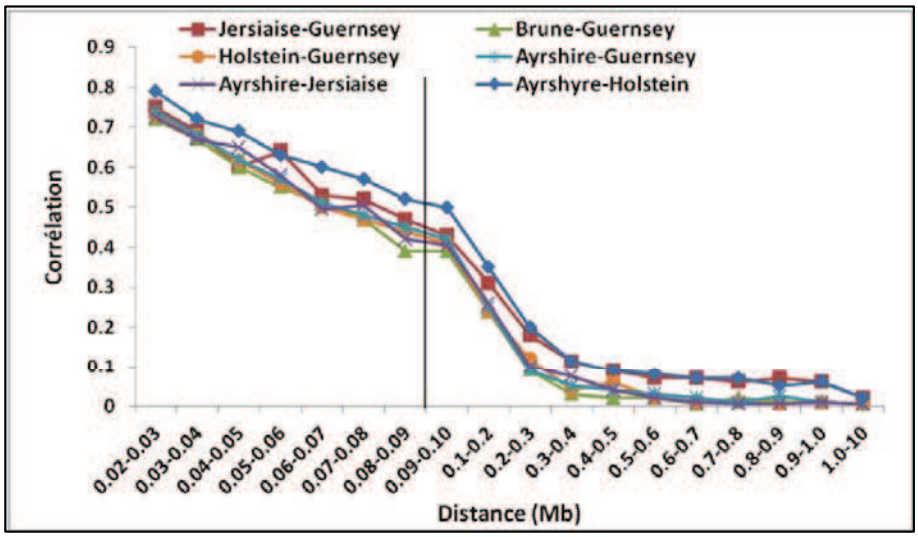
En France, le projet ANR GEMBAL (GEnomique Multiraciale Bovins Allaitants et Laitiers), fruit de la collaboration entre l'INRA, l'Institut de l'Elevage, l'UNCEIA et l'organisme Races de France a pour objectif premier de constituer une grande base de données multiraciales de génotypes haute densité. Cette base de données constituée de génotypes d'animaux issus de 20 races laitières et allaitantes doit pouvoir être utilisée pour des travaux de recherches en génomique bovine.

Ce projet se décompose en plusieurs tâches listées ci-dessous :

1. La définition de la population à génotyper en haute densité
2. Les études visant une meilleure compréhension du génome bovin
3. L'imputation des génotypes HD des animaux déjà génotypés sur la puce 50K
4. Le développement d'une méthodologie pour les évaluations multiraciales
5. Des applications en sélection génomique des bovins laitiers
6. Des applications en sélection génomique des bovins allaitants

Les travaux réalisés aux cours de cette thèse s'inscrivent dans le cadre des tâches 3, 4 et 5 de ce projet.

a)



b)

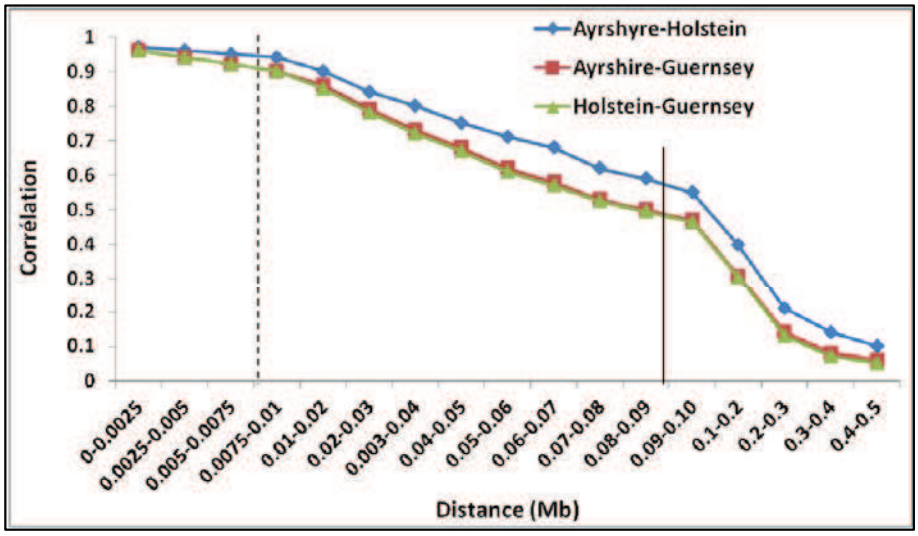


Figure 10 : Corrélations de Pearson entre déséquilibres de liaison (r) à différentes distances. Les barres verticales correspondent à la distance moyenne entre deux marqueurs sur la puce 50K (trait plein) ou HD (trait pointillé), d'après Larmer et al. (2014).
a) Résultats obtenus sur 5 races bovines à partir de données de génotypages 50K
b) Résultats obtenus sur 3 races bovines à partir de données de génotypages HD

III. Constitution de la population de référence « haute densité »

A. Introduction

Le projet GEMBAL s'appuie sur l'utilisation de génotypes obtenus à partir de la puce BovineHD BeadChip® (HD) contenant 777 962 marqueurs SNP. La première étape de ce projet a donc été la constitution d'une population de référence haute densité.

1. Choix des animaux à génotyper en haute densité

Le développement d'évaluations génomiques dans les principales races laitières et les travaux de recherches en génomique ont permis de constituer une base de données de plusieurs milliers de génotypes 50K. Regénotyper l'ensemble de ces individus en haute densité représente un investissement financier important. Sous réserve que le taux d'erreur d'imputation soit suffisamment faible, une stratégie alternative est de génotyper une partie des animaux sur la puce HD et d'imputer les génotypes HD des animaux qui ont déjà été génotypés en 50K. Dans ce cadre, 5072 taureaux de races laitières et allaitantes (Tableau 6) ont été génotypés sur la puce HD afin de construire une première population génotypée en haute densité qu'on appellera population d'imputation. Il faut ajouter à cette population, les génotypes HD de 548 taureaux Holstein génotypés dans le cadre du consortium EuroGenomics (Lund et al., 2011 ; Schrooten et al., 2014).

Tableau 6 : Répartition des génotypages réalisés dans le cadre du projet GEMBAL en fonction de la race et du type (laitier ou allaitant) de l'animal

Races laitières	Nombre de génotypes	Races allaitantes	Nombre de génotypes
Abondance	320	Aubrac	264
Pie Rouge des Plaines	41	Salers	248
Brune	100	Bazadaise	109
Bretonne	30	Limousine	463
Tarentaise	193	Charolaise	719
Simmental	128	Rouge des Prés	168
Montbéliarde	545	Parthenaise	309
Normande	557	Gasconne	169
Vosgienne	60	Blonde D'Aquitaine	352
Rouge Flamande	45		
Holstein	252 (+ 548*)		

* Génotypes issus du consortium EuroGenomics

Afin de minimiser le taux d'erreur d'imputation, la population a été constituée en sélectionnant les animaux ayant les plus grandes contributions à la race. En effet, l'imputation cherche à réattribuer à chaque segment chromosomique un haplotype ancestral (Figure 4). Si un individu a fortement contribué à la race, on devrait retrouver des segments chromosomiques issus de cet animal chez de nombreux individus de la population à imputer. Il a ainsi été démontré que le taux d'erreur augmente avec la distance entre l'animal à imputer et la population d'imputation (Druet et al., 2010 ; Schrooten et al., 2014).

La répartition des animaux entre les différentes races a été déterminée à partir de l'importance relative de chaque race dans le cheptel français, la disponibilité en échantillons biologiques mais également la taille efficace de la population. En effet, tout comme la précision des évaluations génomiques, la précision de l'imputation est fonction du nombre d'haplotypes différents observés dans la population et donc de sa taille efficace (Dassonneville et al., 2011). Les tailles efficaces étant plus élevées en bovins allaitants qu'en bovins laitiers, un nombre plus important d'animaux a été génotypé dans ces races.

Les données utilisées ici, étaient constituées de génotypes haute-densité d'animaux peu apparentés (dans les races autres que les trois principales races laitières). Elles ne permettaient donc pas de tirer pleinement profit de l'information familiale mais présentaient un déséquilibre de liaison populationnel élevé entre marqueurs adjacents.

Par ailleurs, une étude préliminaire réalisée dans le cadre d'EuroGenomics avait montré une supériorité du logiciel Beagle (Browning et Browning, 2007) sur le logiciel DAGPHASE (Druet et Georges, 2010) pour l'imputation de génotypes HD en race Holstein (Schrooten et al., 2014). Le choix du logiciel d'imputation s'est donc porté sur Beagle (Browning et Browning, 2007).

2. Préparation des fichiers de génotypages

De nombreuses étapes sont nécessaires pour la transformation des fichiers de génotypes obtenus en sortie de laboratoire en fichiers utilisables pour les évaluations génomiques. Elles sont composées d'étapes de contrôle qualité des marqueurs (*call freq*, équilibre de Hardy-Weinberg) et des génotypes (*call rate*, vérification de la compatibilité entre génotypes et pedigree) décrites précédemment. A cela s'ajoute des étapes spécifiques permettant de s'assurer de la compatibilité des typages aux marqueurs communs entre les puces HD et 50K mais également la vérification pour chaque marqueur de l'absence d'erreurs mendéliennes entre les génotypes d'un animal et ceux de ses apparentés.

La multiplicité des races étudiées dans ce projet et l'arrivée régulière de nouveaux génotypes nous ont conduits à mettre en place une chaine complète de traitement des génotypes HD en vue de leur utilisation en sélection génomique.

Une fois ces étapes de préparation de fichier réalisées, les taux d'erreur d'imputation dans chacune des races ont été estimés afin de déterminer si l'imputation haute densité était de bonne qualité ou non. La disponibilité d'un grand nombre de races présentant des caractéristiques différentes nous a permis d'identifier les facteurs influençant la qualité d'imputation et ainsi pouvoir proposer des solutions aux races pour lesquelles le taux d'erreur d'imputation serait trop élevé.

B. Article I : Etude de la qualité d'imputation de génotypes haute densité à partir du génotype moyenne densité dans 16 races bovines françaises.

Hozé C, Fouilloux MN, Venot E, Guillaume F, Dassonneville R, Fritz S, Ducrocq V, Phocas F, Boichard D, Croiseau P. 2013,

High-density marker imputation accuracy in sixteen French cattle breeds.

Genet. Sel. Evol. 2013, 45:33.

Hozé et al. Genetics Selection Evolution 2013, 45:33
http://www.gsejournal.org/content/45/1/33

GSE Genetics Selection Evolution

RESEARCH Open Access

High-density marker imputation accuracy in sixteen French cattle breeds

Chris Hozé[1,2,3*], Marie-Noëlle Fouilloux[4], Eric Venot[1,2], François Guillaume[1,2,4], Romain Dassonneville[1,2,4], Sébastien Fritz[3], Vincent Ducrocq[1,2], Florence Phocas[1,2], Didier Boichard[1,2] and Pascal Croiseau[1,2]

Abstract

Background: Genotyping with the medium-density Bovine SNP50 BeadChip® (50K) is now standard in cattle. The high-density BovineHD BeadChip®, which contains 777 609 single nucleotide polymorphisms (SNPs), was developed in 2010. Increasing marker density increases the level of linkage disequilibrium between quantitative trait loci (QTL) and SNPs and the accuracy of QTL localization and genomic selection. However, re-genotyping all animals with the high-density chip is not economically feasible. An alternative strategy is to genotype part of the animals with the high-density chip and to impute high-density genotypes for animals already genotyped with the 50K chip. Thus, it is necessary to investigate the error rate when imputing from the 50K to the high-density chip.

Methods: Five thousand one hundred and fifty three animals from 16 breeds (89 to 788 per breed) were genotyped with the high-density chip. Imputation error rates from the 50K to the high-density chip were computed for each breed with a validation set that included the 20% youngest animals. Marker genotypes were masked for animals in the validation population in order to mimic 50K genotypes. Imputation was carried out using the Beagle 3.3.0 software.

Results: Mean allele imputation error rates ranged from 0.31% to 2.41% depending on the breed. In total, 1980 SNPs had high imputation error rates in several breeds, which is probably due to genome assembly errors, and we recommend to discard these in future studies. Differences in imputation accuracy between breeds were related to the high-density-genotyped sample size and to the genetic relationship between reference and validation populations, whereas differences in effective population size and level of linkage disequilibrium showed limited effects. Accordingly, imputation accuracy was higher in breeds with large populations and in dairy breeds than in beef breeds. More than 99% of the alleles were correctly imputed if more than 300 animals were genotyped at high-density. No improvement was observed when multi-breed imputation was performed.

Conclusion: In all breeds, imputation accuracy was higher than 97%, which indicates that imputation to the high-density chip was accurate. Imputation accuracy depends mainly on the size of the reference population and the relationship between reference and target populations.

Background

The development of a high throughput chip of 54 001 single nucleotide polymorphisms (SNPs) for cattle, the Bovine SNP50 BeadChip® (50K), has drastically reduced genotyping costs and strongly contributed to the current implementation of genomic selection (GS). This approach, which was first proposed by Meuwissen et al. [1], uses a reference population (usually consisting of progeny-tested bulls) with both genotypes and phenotypes to estimate marker effects and then uses these estimates to predict breeding values for animals without phenotypes. The accuracy of prediction depends mainly on the size of the reference population, heritability of the phenotypes, and level of linkage disequilibrium (LD) between markers and quantitative trait loci (QTL) [2,3]. In some bovine breeds, a very large number of animals have been genotyped with the 50K chip and it is now possible to predict breeding values of animals at birth with high accuracy. In breeds

* Correspondence: chris.hoze@jouy.inra.fr
[1]INRA, UMR 1313 Génétique Animale et Biologie Intégrative, 78350 Jouy-en-Josas, France
[2]AgroParisTech, UMR1313 Génétique Animale et Biologie Intégrative, 75231 Paris 05, France
Full list of author information is available at the end of the article

BioMed Central

© 2013 Hozé et al.; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

with a limited number of progeny-tested bulls, assembling a large enough reference population is a real challenge. Under the assumption that LD is conserved across breeds, reference populations from different breeds can be combined to increase the size of the reference population. Conservation of LD, however, requires more than 300 000 informative SNPs and therefore, is not fulfilled with the classically used 50K chip [4]. The BovineHD BeadChip® (HD) that was developed in 2010 and contains 777K SNPs is expected to be sufficiently dense to detect conserved LD across breeds and allow multi-breed GS. Re-genotyping all animals on this HD chip is, however, not economically feasible but prediction (imputation) of HD genotypes from 50K genotypes is possible.

Several imputation methods have been implemented and are widely used to impute genotypes from one chip to another. They are based either on population LD (Fastphase [5], Beagle [6], MaCH [7], IMPUTE2 [8]), or on a combination of LD and family information (Fimpute [9], Dagphase [10], AlphaImpute [11], FindHap [12]). Many studies have compared these methods for imputation from low-density panels to the 50K chip. Beagle 3.3.0 has been shown to be an accurate software package [13-15] and is commonly used in bovine datasets.

In this study, Beagle 3.3.0 was used to study the accuracy of imputation from the Illumina 50K to the HD chip in 16 French cattle breeds and to investigate the main factors that affect this accuracy.

Methods

Genotypes

The dataset comprised 5153 animals genotyped with the HD chip. The cryopreserved semen, or blood samples of the animals included in our study, which were used for genotyping, were procured from various commercial AI organizations and breeder organizations through their routine practice in the framework of breeding programs. Therefore, no ethical approval was required for sampling of biological material. The HD chip contains 777 964 markers with an average probe spacing of 3.45 kilobases (kb) [16]. Animals belonged to 16 breeds (seven dairy breeds and nine beef breeds). The number of genotypes was not equally distributed across breeds (Table 1) but depended on population size. Animals genotyped with the HD chip were chosen based on their marginal contribution to the population, as defined by Boichard et al. [17], and computed using the PEDIG software [18]. If possible, two to three progeny of each founder were also genotyped in order to increase phasing accuracy, and the most influential progeny were chosen for HD-genotyping. In some situations, and in particular in breeds with small populations, the number of HD-genotyped animals was constrained by the availability of DNA from ancestors. The family structure of each breed is detailed in Table 1. The mean number of HD-genotyped male progeny per bull was higher in dairy breeds (2.5) than in beef breeds (2). In the Holstein breed, a large proportion of genotypes

Table 1 Number of high-density genotyped animals and population structure per breed

	Nb of genotyped animals	Nb of genotyped families (sire + progeny)	Mean nb of genotyped progeny per sire	Nb of effective ancestors
Dairy breeds				
Abondance (ABO)	209	54	3.69	15
Brown Swiss (BSW)	99	52	1.90	28
Holstein (HOL)	788	204	2.30	21
Montbéliarde (MON)	530	139	3.77	18
Normande (HOR)	551	138	3.82	23
Simmental (SIM)	125	55	2.24	39
Tarentaise (TAR)	185	65	2.77	15
Beef breeds				
Aubrac (AUB)	254	116	2.17	112
Bazadaise (BAZ)	89	60	1.45	46
Blonde d'Aquitaine (BLA)	327	187	1.74	78
Charolais (CHA)	672	310	2.14	249
Gasconne (GAS)	163	76	2.12	197
Limousine (LIM)	462	235	1.96	185
Parthenaise (PAR)	304	97	3.02	89
Rouge des Prés (RDP)	149	80	1.83	99
Salers (SAL)	246	186	1.31	99

originated from the Eurogenomics consortium [19,20]. Another dataset, containing 33 746 animals genotyped with the Bovine SNP50 BeadChip® [21], was also available. The genotypes originated either from the national genomic selection program or from complementary research programs, except for most of the Holstein genotypes, which were from the Eurogenomics consortium.

HD and 50k genotypes were used for parentage testing and to check the clustering quality of the HD genotypes by concordance analysis of the genotypes of the 1838 individuals genotyped on both chips.

Data editing

Quality control was performed within-breed on the HD and 50K genotypes, using the same criteria for both chips. Genotyped animals with a call rate lower than 0.95 were removed from the analysis. Only markers mapped on the UMD3.1 assembly covering the 29 bovine autosomes were used for analysis. SNPs showing departure from Hardy-Weinberg equilibrium (p-value < 0.001) or with more than 10% missing genotypes were removed. In addition, genotype consistency was checked using 1838 animals that were genotyped on both chips and 352 markers that were discordant for more than 1% of these animals were excluded. After editing, 708 771 and 44 580 SNPs were retained on the HD and 50K sets, respectively. The 37 634 SNPs present on both HD and 50K sets were used to mimic 50K genotypes. Parentage was tested following the French routine genomic selection procedure, using both 50K and HD datasets (S. Fritz, personal communication). This procedure uses 500 informative markers and a parentage error was concluded if more than 10 incompatibilities were detected. In case of inconsistency, the progeny was removed, except when at least two progeny from the same sire were found incompatible with their sire. In this case, the sire was removed. Genotypes were checked for Mendelian inconsistencies between compatible parents and offspring. The genotype of the male parent was deleted if more than 20% of its progeny showed contradiction, in other cases the genotype of the progeny was set to missing.

Assessing within-breed imputation accuracy

The accuracy of imputation from 50k to HD genotypes was computed within each breed. For this purpose, the HD-dataset was divided into two parts. The oldest animals were used as a training population to mimic a reference population genotyped with the HD chip. The 20% youngest animals formed the validation population. For animals in the validation population, markers that were only present on the HD chip were masked to mimic a target population genotyped with the 50K.

Beagle 3.3.0 was used to impute mimicked 50K genotypes up to high-density. This software uses a population-based method called "localized haplotype-cluster method", as described by Browning and Browning [6]. The method builds and clusters haplotypes along the whole chromosome and then uses an underlying variable length Markov chain based on haplotypes counts (and consequently on local LD patterns) to determine transition probabilities from one marker to the next. The scale and shift parameters were set to 2 and 0.1, respectively, and no pedigree information was taken into account.

Imputation accuracy was estimated based on the comparison between imputed and known HD genotypes and was defined as the allelic imputation error rate, computed as the ratio between the number of falsely imputed alleles and the total number of imputed alleles [22].

Identification of SNPs with high error rates is important because they may induce errors in QTL detection and genomic selection. To identify these SNPs, only the breeds with the largest populations (Blonde d'Aquitaine, Charolais, Holstein, Limousine, Montbéliarde and Normande) were considered in order to avoid erroneous identification of errors due to the use of a low number of genotypes when computing the error rate for a SNP.

Factors affecting imputation accuracy

Given the large number of breeds included here, it was possible to analyze the factors that may have an effect on imputation accuracy. Two types of variables that differ between breeds were studied: genetic diversity indicators such as the level of linkage disequilibrium and the number of effective ancestors, and variables specific to our datasets, such as the number of genotypes in the reference population and the relationship between reference and validation populations.

To avoid bias due to sampling, we used the indicators computed by Danchin-Burge [23] using the complete pedigree file of each breed. The number of effective ancestors was preferred to effective population size since the former is less sensitive to the depth of the pedigree [23].

Linkage disequilibrium was computed within-breed as a squared correlation coefficient based on phased genotypes for each marker pair [24], as defined in equation (1):

$$r^2 = \frac{(p_{A1B1} - p_{A1}p_{B1})^2}{p_{A1}p_{A2}p_{B1}p_{B2}} \quad (1)$$

where A1, A2, B1, B2 are the alleles of SNP A and B, $p_{A1,}$ p_{A2}, p_{B1}, and p_{B2} are the corresponding allelic frequencies, and p_{A1B1} is the frequency of the A1B1 haplotype.

For the Abondance breed, LD was computed for SNPs on *Bos taurus* (BTA) chromosomes 5, 10, 15, 20 and 25 and no differences between chromosomes were observed (data not shown). Therefore, for other breeds, LD computation was only performed on BTA5. Pairs of SNPs for which one SNP has a minor allele frequency (MAF) lower than 5% in the considered breed were discarded

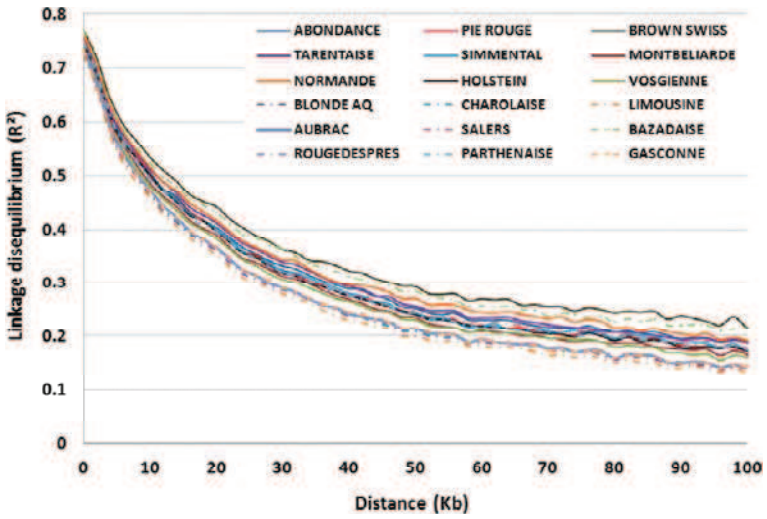


Figure 1 Linkage disequilibrium decay in 16 beef (dashed lines) and dairy (solid lines) cattle breeds.

because it has been shown that LD tends to be very low in such cases [25]. Furthermore, removing SNPs with low MAF facilitates comparisons between populations [25,26]. Then, LD for all marker pairs were averaged by the distance between SNPs. We focused particularly on the level of LD between SNPs 70 kb apart, which corresponds to the average distance between two informative SNPs on the 50K chip.

Average relationship levels between training and validation populations were measured based on pedigree information using the PEDIG [18] software. A relationship coefficient was computed for each pair of individuals in each population and then averaged for each breed.

The effect of the different factors on imputation accuracy was assessed by multiple regression across all evaluated datasets (including both dairy and beef breeds).

Table 2 Within-breed imputation error rate and others parameters affecting imputation error rate

	Training population size	Validation population size	Allelic imputation error rate (%)	LD level at 70 kb	Average $R_{T/V}$
Dairy breeds					
Abondance (ABO)	169	40	0.75	0.217	0.146
Brown Swiss (BSW)	79	20	1.92	0.255	0.074
Holstein (HOL)	634	154	0.73	0.255	0.078
Montbéliarde (MON)	424	106	0.51	0.196	0.116
Normande (HOR)	444	107	0.33	0.233	0.104
Simmental (SIM)	100	25	2.55	0.209	0.050
Beef breeds					
Aubrac (AUB)	204	50	2.03	0.177	0.028
Bazadaise (BAZ)	72	17	2.07	0.239	0.038
Blonde d'Aquitaine (BLA)	262	65	1.80	0.175	0.038
Charolais (CHA)	539	133	0.68	0.176	0.018
Gasconne (GAS)	131	32	2.26	0.174	0.026
Limousine (LIM)	370	92	1.09	0.164	0.014
Parthenaise (PAR)	245	59	1.88	0.161	0.024
Rouge des Prés (RDP)	119	30	2.39	0.206	0.028
Salers (SAL)	197	49	1.27	0.213	0.024

Size of the training and validation populations, within-breed imputation error rate, level of linkage disequilibrium (LD, r^2)) at 70 kb and average relationship between training and validation populations ($R_{T/V}$).

Tested factors were the reference population size, the number of effective ancestors, the linkage disequilibrium level at 70 kb and the relationship between training and validation populations. Factors were considered significant if the regression coefficient was different from zero (p-value < 0.05).

Computation of multi-breed imputation accuracy

Since our aim was to improve imputation accuracy in breeds with a small reference population size, multi-breed imputation was tested. Training and validation populations from different breeds were combined and then imputation accuracy was computed as described above. Considering the diversity of the breeds involved in this project, it did not seem relevant to combine all breeds. Instead, breeds were grouped based on Reynolds genetic distances between breeds, which were computed by Gautier et al. [27] using allele frequencies on the 50K chip and represented in a Neighbor-Joining tree. One branch, which was composed of the Abondance, Montbéliarde and Tarentaise breeds, was taken as an example of closely related breeds and these breeds were merged for the multi-breed study. Imputation accuracy was then computed per animal and averaged for each breed.

Results

Linkage disequilibrium decay

The results of LD decay are presented in Figure 1. Within the first 10 kb, LD strongly decreased in all breeds, thereafter it decreased at a lower rate and differences between breeds became noticeable. LD levels at 70 kb (Table 2) varied from 0.16 in the Parthenaise breed to 0.26 in the Holstein breed; on average LD levels at 70 kb were higher by 0.04 in dairy breeds than in beef breeds, in agreement with the difference in number of effective ancestors between these breeds.

Imputation accuracy

Error rates were low in most breeds, with an overall mean error rate of 1.36% (Table 2). However, large differences were observed between breeds, with a minimum error rate of 0.31% in the Normande breed and a maximum of 2.41% in the Simmental breed. The error rates were less than 0.7% for breeds with more than 500 animals in the reference population.

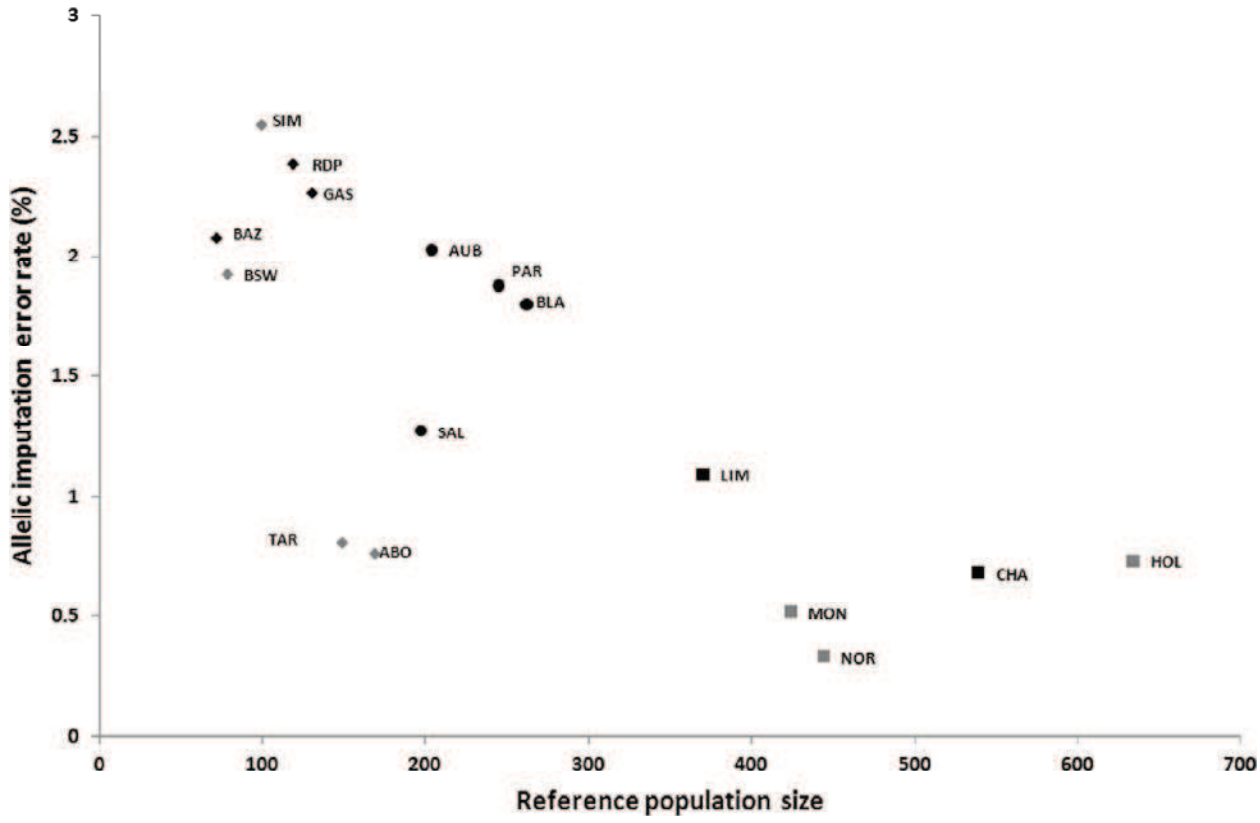


Figure 2 Relationship between allelic imputation error rate and reference population size in beef (black) and dairy (gray) cattle breeds. Breeds with more than 300 animals in the reference population are represented by a square, those with more than 200 animals by a circle, and those with 200 or less animals by a diamond.

Error rates were lower in dairy breeds (mean error rate = 1.02%) than in beef breeds (mean error rate = 1.62%). In beef breeds, the error rate ranged from 0.64% in the Charolais breed to 2.26% in the Rouge des Prés breed. In dairy breeds, error rates were higher in the Simmental and Brown Swiss breeds, which are regional breeds in France with many imported ancestors and limited numbers of genotyped animals.

Factors affecting imputation accuracy

Sizes of training and validation populations are presented in Table 2. Error rates decreased linearly when the number of animals in the reference population increased (Figure 2). Again, we observed a higher accuracy in dairy breeds: an error rate lower than 1% was achieved with only 100 to 200 genotyped animals in dairy breeds, whereas approximately 400 animals were required for beef breeds. With more than 400 genotyped animals, the impact of reference population size on accuracy leveled off, which suggests that the effect of size of the reference population is non-linear when a critical size is reached.

The relationship between the number of effective ancestors and imputation error rate is illustrated in Figure 3. The number of effective ancestors appeared to have an opposite effect on dairy versus beef breeds. In dairy breeds, the number of effective ancestors was low (< 50) and its relationship with imputation error rate was positive. However, the two breeds with the highest number of effective ancestors (Simmental and Brown Swiss) among the dairy breeds had the highest imputation error rate but also the smallest reference populations and many foreign ancestors were not genotyped. In beef breeds, a negative relationship between the number of effective ancestors and imputation error rate was observed, but the breeds with the highest number of effective ancestors (Charolais and Limousine) also had the largest reference populations. This suggests that the effect of the number of effective ancestors was masked by the effect of the reference population size. However, when considering populations with equal reference population sizes (e.g. Limousine and Montbeliarde), the error rates increased with number of effective ancestors.

Figure 4 presents the relationship of the imputation error rate with the degree of genetic relationship between training and validation populations ($R_{T/V}$). The first observation is that $R_{T/V}$ was clearly lower in beef than in dairy breeds,

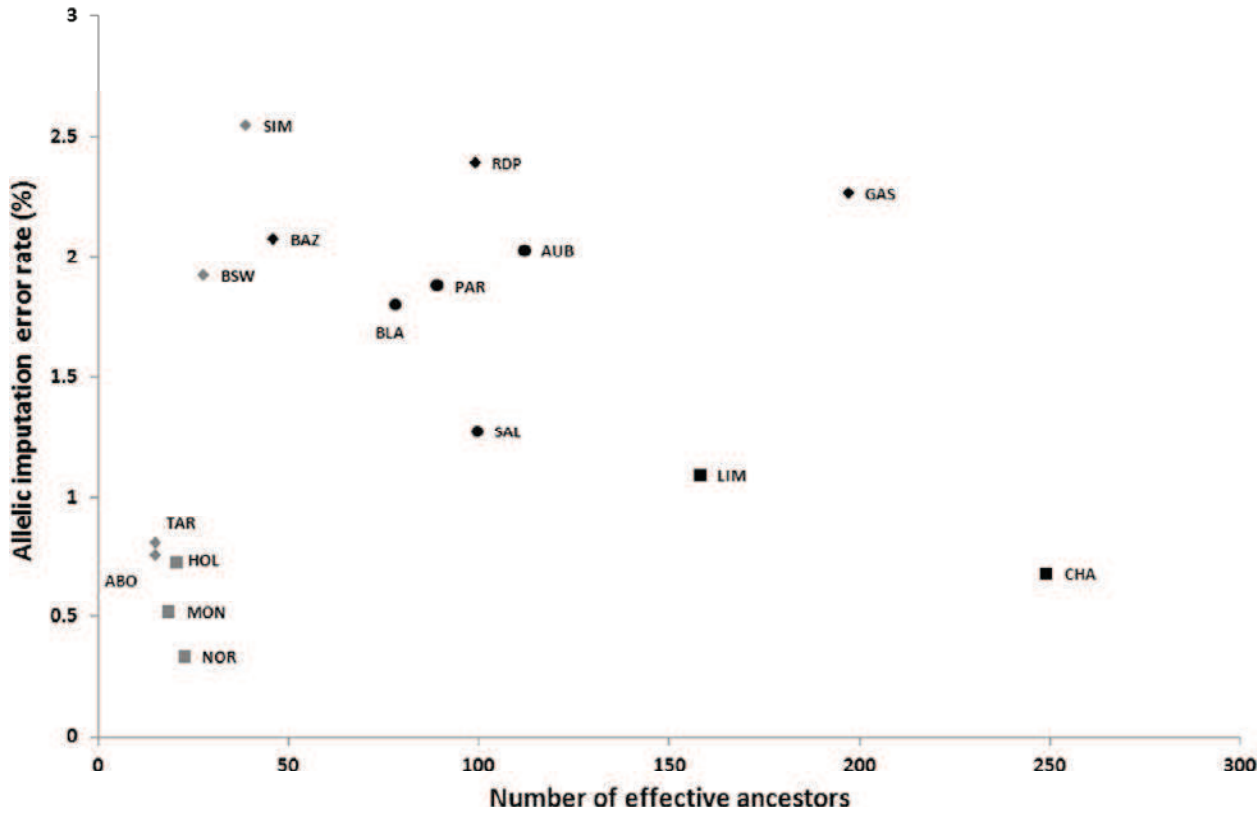


Figure 3 Relationship between allelic imputation error rate and number of effective ancestors in beef (black) and dairy (gray) cattle breeds. Breeds with more than 300 animals in the reference population are represented by a square, those with more than 200 animals by a circle, and those with 200 or less animals by a diamond.

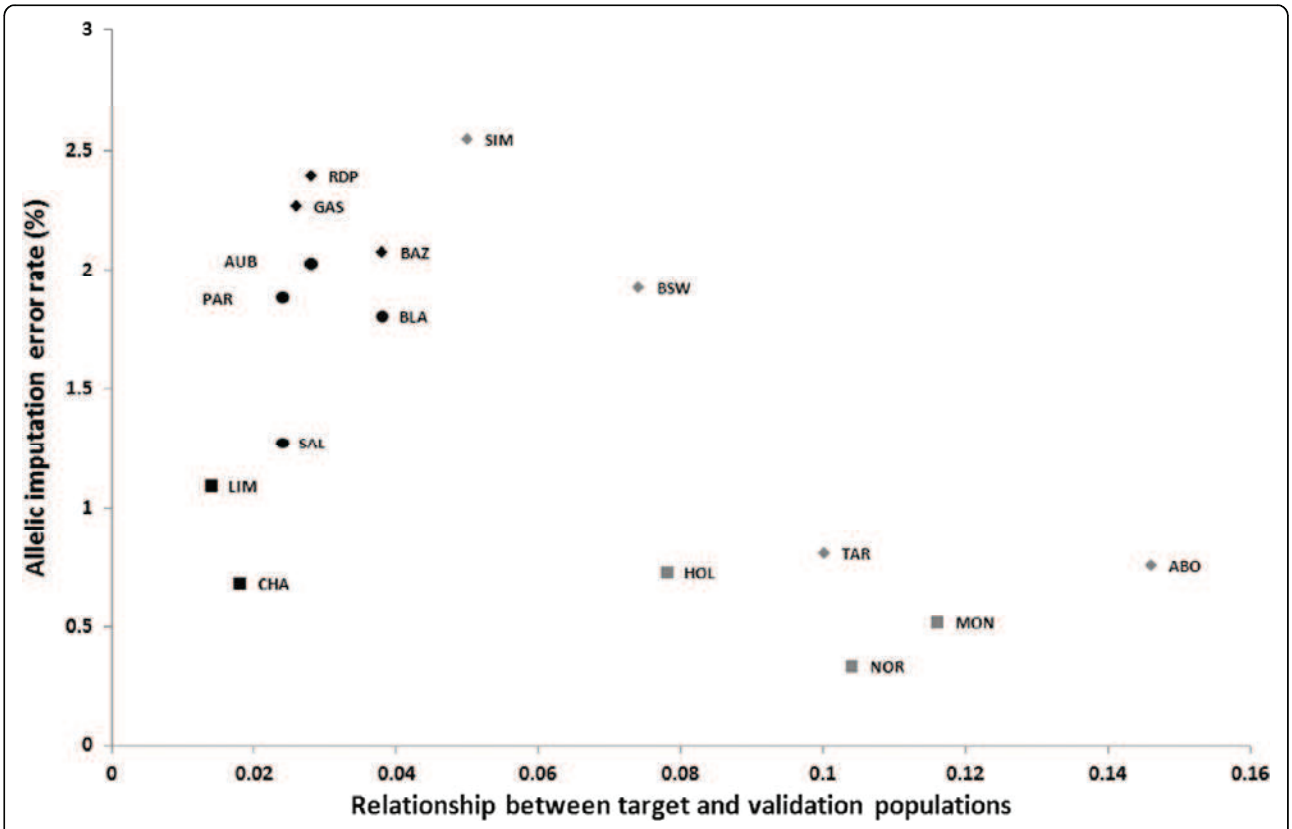


Figure 4 Relationship between allelic imputation error rate and the genetic relationship between target and validation populations in beef (black) and dairy (gray) cattle breeds. Breeds with more than 300 animals in the reference population are represented by a square, those with more than 200 animals by a circle, and those with 200 or less animals by a diamond.

which may be explained by a higher number of effective ancestors in beef breeds. The second observation was that when breeds with the largest populations were discarded, the relationship between $R_{T/V}$ and imputation error rate appeared to be linear. Among breeds with large populations, the relationship between training and validation populations had only a limited effect on the imputation error rate. The Charolais and Limousine breeds had low imputation error rates despite their lower $R_{T/V}$ but these breeds have large reference populations. However, the lower $R_{T/V}$ in the Holstein breed probably explains why the error rate was higher in this breed than in the Normande or Montbeliarde breeds.

Finally, multiple linear regression was performed to better quantify the impact of each factor on imputation accuracy. Results of the multiple linear regressions are in Table 3. Reference population size and $R_{T/V}$ had a significant effect ($p < 0.05$) on imputation error rate, in contrast to the number of effective ancestors and the level of LD. About half of the variation in imputation error rates was explained by reference population size and one quarter by $R_{T/V}$. Using the regression coefficients, we can predict that increasing $R_{T/V}$ by 1% reduces imputation error rate by 0.12% and that adding 100 animals to the reference population decreases the error rate by 0.26% (Table 3). This suggests that the size of the reference population is the major

Table 3 Results of the multiple linear regression model

	Part of variance explained	Regression coefficient	p-value
Reference population size	52%	−0.0026 ± 0.0006	0.002
Relationship between reference and validation populations	25%	−12.6042 ± 3.8937	0.008
Level of linkage disequilibrium at 70 kb	2.5%	−0.0028 ± 0.0026	0.305
Number of effective ancestors	0.03%	−0.0576 ± 4.4641	0.899

The model tests the effect of reference population size, relationship between reference and validation populations, level of linkage disequilibrium at 70 kb, and number of effective ancestors on imputation error rate.

factor affecting imputation error rate but the relationship between training and validation populations must also be taken into consideration. Despite the difference between breeds, our results suggest that the number of effective ancestors and the level of LD are not major factors affecting imputation accuracy.

SNP by SNP analysis of imputation accuracy

Imputation error rate was computed for each SNP in order to detect SNPs that were mismapped in the UMD3.1 assembly. This analysis was performed for the six major breeds (three beef breeds and three dairy breeds). Results in the Montbéliarde breed are presented in Figure 5. Despite an overall mean error rate of 0.51%, 13 104 and 6030 SNPs had error rates greater than 5 and 20%, respectively. Consequently, the error rate dropped from 0.51 to 0.40% after removing potentially mismapped SNPs with error rates greater than 5%.

The relationship between MAF and imputation error rate is illustrated in Figure 6 for the Montbéliarde breed. SNPs were divided in two groups based on their imputation error rates. No relationship with MAF was detected for SNPs with error rates less than 0.1, whereas error rates increased with MAF for SNPs with high error rates. Assuming that the LD between a mismapped SNP and its direct neighbors is low, a mismapped SNP is imputed by chance and therefore its error rate is related to its MAF.

We looked for SNPs with high error rates in each breed based on a threshold that was defined as the mean error rate plus 3 times its standard deviation for that breed. Three thousand and eighty-three, 1980, and 1146 SNPs had high error rates in at least two, three or six breeds, respectively. SNPs with high error rates in at least three breeds are likely mismapped and are listed in Additional file 1: Table S1.

Multi-breed imputation accuracy

Since imputation accuracy depends highly on the number of HD genotypes, we increased the size of the reference population by combining breeds. The results for single- and multi-breed situations are in Table 4. No difference in accuracy was found between single- and multi-breed imputation for the Abondance and Montbéliarde breeds. For the Tarentaise breed, imputation error rate was slightly higher in the multi-breed analysis than in the single-breed analysis.

Discussion

In this study, we evaluated accuracy of imputation from 50K to HD-genotypes for 16 cattle breeds and we investigated the corresponding causes of variation. We observed large differences in imputation accuracy between breeds. Several factors may explain these differences, i.e. size of the reference population and the relationship between training and validation populations (closely related to population structure). Results from the multiple linear regression performed in this study, combined with other published results, lead us to propose several hypotheses on the impact of each factor.

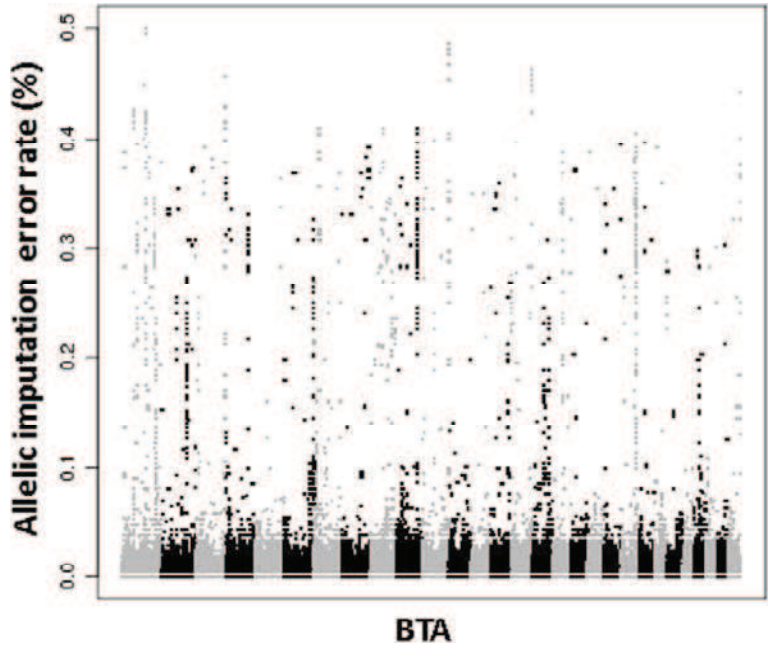


Figure 5 Allelic imputation error rate along the genome in Montbéliarde breed.

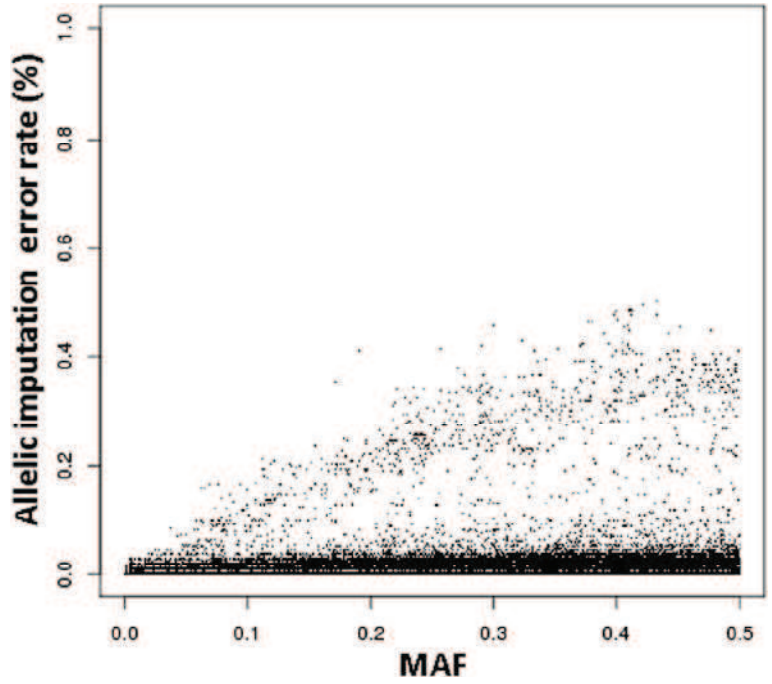


Figure 6 Relationship between allelic imputation error rate and minor allele frequency in Montbéliarde breed.

The number of HD genotypes in the reference population ranged from 72 in the Bazadaise breed to 634 in the Holstein breed and is the major factor that explains differences in imputation accuracy between breeds. The imputation error rate decreased by 0.26% when 100 animals were added to the reference population. However, other studies have shown a non-linear effect of reference population size on imputation error rate. In Holstein cattle, Schrooten et al. (unpublished data) reported that imputation error rate decreased by between 0.17 and 0.04% when moving from 200 to 500 HD genotypes by steps of 100, while in sheep, the decrease was 5% when moving from 50 to 150 individuals in the reference population but only 2% when moving from 204 to 2512 individuals [28]. In our study, the effect of the size of the reference population on imputation error rate was linear. This effect would probably have been found non-linear if more breeds with a large reference population had been included. In fact, the size of the reference population had a limited effect when the number of HD genotypes was greater than a minimum threshold that was estimated at 200–400 animals.

Table 4 Imputation error rate using single-breed populations compared to a multi-breed reference population for three breeds

	Abondance	Montbéliarde	Tarentaise
Size of training / validation population	159 / 40	422 / 106	146 / 36
Single-breed imputation error rate (%)	0.755	0.487	0.763
Multi-breed imputation error rate (%)	0.753	0.485	0.824

Hayes et al. [28] reported that most of the differences in imputation accuracy are due to differences in the relationship between the reference and target populations and the genetic diversity of the breed. Schrooten et al. (unpublished data) used traceability, defined as "the expected contribution of HD-genotyped ancestors to the genotype of an animal", as a measure of the relationship between the reference population and one animal from the target population. Imputation error rates were lower for animals with higher traceability, meaning that a higher average traceability, i.e. a higher relationship between the reference and target populations, will result in lower error rates. We reached the same conclusion, i.e. the imputation error rate was decreased by 0.12% when the average relationship increased by 0.01.

Increasing the size of the reference population decreases the probability to miss a haplotype in the reference population. For a fixed reference population size, an increase in the number of effective ancestors, i.e. an increase in the number of haplotypes in the total population, increases the probability to miss a haplotype and thus increases the error rate. This explains why reference population size and number of effective ancestors had opposite effects on imputation accuracy and compensate for

each other. In our study, differences in within-breed diversity explain why the error rate was higher in beef breeds than in dairy breeds and why more than 99% accuracy was achieved with only 200 animals for dairy breeds, while 400 animals were necessary for beef breeds. The poor results obtained with the Simmental and Brown Swiss breeds were also due to lower relationships between the reference and target populations; in these breeds, key ancestors mainly originate from abroad and were not included in our data. Combining reference populations from different breeds did not improve imputation accuracy, which confirms the results on multi-breed imputation of Hayes et al. [28] and Erbe et al. [29]. In fact, multi-breed imputation is expected to improve imputation accuracy only when 50K haplotypes are conserved across breeds, which is quite unlikely given the history of the breeds and their estimated divergence time, even for the most closely related breeds.

Although observed imputation error rates were low in all breeds, 1980 SNPs had particularly high error rates (Figure 5 and Additional file 1: Table S1), which suggests that errors exist in the marker map. Erbe et al. [29] identified 1231 SNPs with genotype error rates greater than 20%. When some of these SNPs were remapped using LD, error rates dropped. In our study, no remapping was performed but removing SNPs with high error rates resulted in a 0.1% drop in error rate, which suggests that lower error rates can be achieved with a more accurate map. However, Erbe et al. [29] still found 630 poorly imputed SNPs after remapping, which means that other reasons, such as recombination hot spots or regions on the 50K panel with lower SNP density, explain the high imputation error rates for some SNPs. SNPs with high imputation error rates likely also impact the quality of genomic selection and QTL detection. This has not been specifically investigated, but some studies have focused on the impact of imputation from low-density panels on reliability of genomic selection [14,30] and concluded that an imputation error rate between 2 and 3% leads to a mean loss of reliability of 2%. For imputation from 50K to the HD chip, the mean error rate is close to 1%, which suggests that the impact on reliability of genomic selection is even lower. However, because of the large number of markers available after imputation, it is preferable to discard markers with high error rates.

Conclusions

The mean error rate for imputation from the Illumina Bovine50K® to the BovineHD® was around 1%. Differences in error rates between breeds were large and ranged from 2.41% in the Simmental breed to 0.31% in the Normande breed. These differences were mainly due to the size of the reference population and the relationship between the reference and target populations. This means that imputation accuracy could be increased by increasing the number of HD genotypes and by improving the reference population to maximize its relationship with the population to impute. Using 50K genotypes to impute HD genotypes is possible, which implies that a large HD imputed reference population can be available for genomic selection at low cost. However, new HD genotypes are likely required in the future generations in order to maintain relationship links between the reference and target populations and limit the imputation error rate.

Additional file

Additional file 1: High_Error_Rate_SNP. Csv file with header and separator ';' containing the list of markers with high imputation error rate. Columns are as follows: chromosome number, SNP name, number of breeds for which a high error rate was detected and average error rates across breeds.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

CH and MNF performed the analysis. DB, VD, FP and PC designed the study. EV and SF managed the databases. FG and RD provided computer programmes. CH, DB, VD, PC wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

GEMBAL project is funded by the Agence Nationale de la Recherche (ANR-10-GENM-0014), APISGENE, Races de France and INRA "AIP Bioressources". The EUROGENOMICS consortium provided most of the Holstein genotypes. Most 50K genotypes originated from the Cartofine-ANR-05-GENANIMAL-007 project funded by ANR (French National Research Agency) and ApisGene, from the Qualvigene project funded by Apisgene, from the EUROGENOMICS consortium, and from genomic selection activity generated by the French cattle breeding companies.

Author details

[1]INRA, UMR 1313 Génétique Animale et Biologie Intégrative, 78350 Jouy-en-Josas, France. [2]AgroParisTech, UMR1313 Génétique Animale et Biologie Intégrative, 75231 Paris 05, France. [3]Union Nationale des Coopératives agricoles d'Elevage et d'Insémination Animale, 149 rue de Bercy, 75595 Paris Cedex 12, France. [4]Institut de l'Elevage, 149 rue de Bercy, 75595 Paris Cedex 12, France.

Received: 8 February 2013 Accepted: 19 July 2013
Published: 3 September 2013

References

1. Meuwissen THE, Hayes BJ, Goddard ME: **Prediction of total genetic value using genome-wide dense marker maps.** *Genetics* 2001, **157:**1819–1829.
2. Hayes BJ, Goddard ME: **Technical note: prediction of breeding values using marker-derived relationship matrices.** *J Anim Sci* 2008, **86:**2089–2092.
3. Goddard M: **Genomic selection: prediction of accuracy and maximisation of long term response.** *Genetica* 2009, **136:**245–257.
4. de Roos APW, Hayes BJ, Spelman RJ, Goddard ME: **Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle.** *Genetics* 2008, **179:**1503–1512.
5. Scheet P, Stephens M: **A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase.** *Am J Hum Genet* 2006, **78:**629–644.

6. Browning SR, Browning BL: **Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering.** *Am J Hum Genet* 2007, **81**:1084–1097.
7. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR: **MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes.** *Genet Epidemiol* 2010, **34**:816–834.
8. Howie BN, Donnelly P, Marchini J: **A flexible and accurate genotype imputation method for the next generation of genome-wide association studies.** *PLoS Genet* 2009, **5**:e1000529.
9. Sargolzaei M, Chesnais JP, Schenkel F: **FImpute - An efficient imputation algorithm for dairy cattle populations.** *J Dairy Sci* 2011, **94**:421.
10. Druet T, Georges M: **A hidden Markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping.** *Genetics* 2010, **184**:789–798.
11. Hickey JM, Kinghorn BP, Tier B, Wilson JF, Dunstan N, van der Werf JHJ: **A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes.** *Genet Sel Evol* 2011, **43**:12.
12. VanRaden PM, O'Connell JR, Wiggans GR, Weigel KA: **Genomic evaluations with many more genotypes.** *Genet Sel Evol* 2011, **43**:10.
13. Sun C, Wu XL, Weigel KA, Rosa GJM, Bauck S, Woodward BW, Schnabel RD, Taylor JF, Gianola D: **An ensemble-based approach to imputation of moderate-density genotypes for genomic selection with application to Angus cattle.** *Genet Res (Camb)* 2012, **94**:133–150.
14. Mulder HA, Calus MPL, Druet T, Schrooten C: **Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle.** *J Dairy Sci* 2012, **95**:876–889.
15. Calus MPL, Veerkamp RF, Mulder HA: **Imputation of missing single nucleotide polymorphism genotypes using a multivariate mixed model framework.** *J Anim Sci* 2011, **89**:2042–2049.
16. Rincon G, Weber KL, Van Eenennaam AL, Golden BL, Medrano JF: **Hot topic: performance of bovine high-density genotyping platforms in Holsteins and Jerseys.** *J Dairy Sci* 2011, **94**:6116–6121.
17. Boichard D, Maignel L, Verrier E: **The value of using probabilities of gene origin to measure genetic variability in a population.** *Genet Sel Evol* 1997, **29**:5–23.
18. Boichard D: **Pedig: a fortran package for pedigree analysis suited for large populations.** In *Proceedings of the 7th World Congress on Genetics applied to Livestock Production (WCGALP):19–23 August 2002; Montpellier.* CD-ROM communication No. 28–13; 2002.
19. Lund MS, de Roos APW, de Vries AG, Druet T, Ducrocq V, Fritz S, Guillaume F, Guldbrandtsen B, Liu Z, Reents R, Schrooten C, Seefried F, Su G: **A common reference population from four European Holstein populations increases reliability of genomic predictions.** *Genet Sel Evol* 2011, **43**:43.
20. Su G, Brøndum RF, Ma P, Guldbrandtsen B, Aamand GP, Lund MS: **Comparison of genomic predictions using medium-density (approximately 54,000) and high-density (approximately 777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations.** *J Dairy Sci* 2012, **95**:4657–4665.
21. Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, O'Connell J, Moore SS, Smith TPL, Sonstegard TS, Van Tassell CP: **Development and characterization of a high density SNP genotyping assay for cattle.** *PLoS One* 2009, **4**:e5350.
22. Zhang Z, Druet T: **Marker imputation with low-density marker panels in Dutch Holstein cattle.** *J Dairy Sci* 2010, **93**:5487–5494.
23. Danchin-Burge C: *Estimation de la variabilité génétique de 19 races bovines à partir de leurs généalogies.* Paris: Institut de l'Elevage; 2009.
24. Hill WG, Robertson A: **Linkage disequilibrium in finite populations.** *Theor Appl Genet* 1968, **38**:226–231.
25. Goddard KAB, Hopkins PJ, Hall JM, Witte JS: **Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations.** *Am J Hum Genet* 2000, **66**:216–234.
26. Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES: **Linkage disequilibrium in the human genome.** *Nature* 2001, **411**:199–204.
27. Gautier M, Laloë D, Moazami-Goudarzi K: **Insights into the genetic history of French cattle from dense SNP data on 47 worldwide breeds.** *PLoS One* 2010, **5**:e0013038.
28. Hayes BJ, Bowman PJ, Daetwyler HD, Kijas JW, Van der Werf JHJ: **Accuracy of genotype imputation in sheep breeds.** *Anim Genet* 2012, **43**:72–80.
29. Erbe M, Hayes BJ, Matukumalli LK, Goswami S, Bowman PJ, Reich CM, Mason BA, Goddard ME: **Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels.** *J Dairy Sci* 2012, **95**:4114–4129.
30. Dassonneville R, Brøndum RF, Druet T, Fritz S, Guillaume F, Guldbrandtsen B, Lund MS, Ducrocq V, Su G: **Effect of imputing markers from a low-density chip on the reliability of genomic breeding values in Holstein populations.** *J Dairy Sci* 2011, **94**:3679–3686.

doi:10.1186/1297-9686-45-33
Cite this article as: Hozé *et al.*: High-density marker imputation accuracy in sixteen French cattle breeds. *Genetics Selection Evolution* 2013 **45**:33.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central

C. Erratum Article II

Les valeurs indiquées dans le texte diffèrent légèrement de celles indiquées dans le tableau 2. Les valeurs correctes sont celles présentes dans le tableau 2.

Par ailleurs, la ligne concernant la race Tarentaise était absente du tableau 2, elle a été ajoutée.

L'auteur s'excuse de cette erreur.

Table 2 Within-breed imputation error rate and others parameters affecting imputation error rate

	Training population size	Validation population size	Allelic imputation error rate (%)	LD level at 70 kb	Average $R_{T/V}$
Dairy breeds					
Abondance (ABO)	169	40	0.75	0.217	0.146
Brown Swiss (BSW)	79	20	1.92	0.255	0.074
Holstein (HOL)	634	154	0.73	0.255	0.078
Montbéliarde (MON)	424	106	0.51	0.196	0.116
Normande (HOR)	444	107	0.33	0.233	0.104
Simmental (SIM)	100	25	2.55	0.209	0.050
Tarentaise (TAR)	149	36	0.81	0.222	0.100
Beef breeds					
Aubrac (AUB)	204	50	2.03	0.177	0.028
Bazadaise (BAZ)	72	17	2.07	0.239	0.038
Blonde d'Aquitaine (BLA)	262	65	1.80	0.175	0.038
Charolais (CHA)	539	133	0.68	0.176	0.018
Gasconne (GAS)	131	32	2.26	0.174	0.026
Limousine (LIM)	370	92	1.09	0.164	0.014
Parthenaise (PAR)	245	59	1.88	0.161	0.024
Rouge des Prés (RDP)	119	30	2.39	0.206	0.028
Salers (SAL)	197	49	1.27	0.213	0.024

D. Bilan

L'étude de l'efficacité d'imputation a montré des taux d'erreur faibles (1%) dans l'ensemble des races laitières à l'exception des races Brune et Simmental. Dans les races allaitantes, le taux d'erreur est plus élevé : généralement situé autour de 2%. Ces taux d'erreurs plus élevés s'explique par une plus grande taille efficace de population conduisant à un déséquilibre de liaison moyen plus faible entre deux marqueurs successifs. Par ailleurs, une analyse du taux d'erreur d'imputation marqueur par marqueur a révélé 1 146 SNP avec des taux d'erreurs élevés dans aux moins trois races. Il est vraisemblable que ces SNP soient mal positionnés sur le génome. Ils ont donc été éliminés des fichiers de génotypages en vue des analyses futures.

Il est donc possible de poursuivre les travaux à partir de génotypes 50K imputés HD dans la majorité des races laitières. Dans les autres races laitières, en particulier la Brune et la Simmental, des génotypages HD complémentaires ou des échanges internationaux seront nécessaires pour augmenter le nombre d'individus disposant d'un génotype HD et ainsi diminuer les taux d'erreur d'imputation. Pour les races laitières à petits effectifs, le faible nombre d'individus dans la population d'apprentissage n'a pas permis de réaliser un test d'efficacité d'imputation. En revanche, à la lumière des résultats observés sur les autres races, le taux d'erreur d'imputation sera vraisemblablement supérieur à 2%.

Le taux d'erreur d'imputation est également d'une importance capitale dans la perspective d'une évaluation génomique de routine basée sur des génotypes haute densité. Il est difficilement envisageable du fait de la différence de prix, (160 € HT pour un génotypage sur la puce HD et 55€ HT sur la puce 50K (S. Barbier, Valogène, communication personnelle), que l'ensemble des candidats à la sélection soit génotypés sur la puce HD. Si le taux d'erreur d'imputation est faible, l'inclusion d'une étape d'imputation dans la chaine d'évaluation génomique devrait permettre de limiter les coûts de génotypage tout en maintenant la précision des valeurs génétiques estimées (Dassonneville et al., 2011 ; Mulder et al., 2012).

IV. Comparaison de stratégies d'évaluations génomiques pour une race disposant d'une population de référence de taille limitée

A. Introduction

L'étude décrite dans le chapitre précédent a permis de démontrer la faisabilité de l'imputation des génotypes HD d'animaux génotypés sur la puce 50K. En 2012, seules les trois principales races laitières disposaient d'un nombre significatif de génotypes 50K, l'imputation a donc uniquement été réalisée dans ces trois races. Après imputation, la population de référence HD était constituée de 28 743 taureaux laitiers dont 83% de race Holstein, 7.5% de race Montbéliarde et 6.5% de race Normande.

Les faibles effectifs disponibles rendent difficile l'estimation de l'efficacité de la sélection génomique dans les races régionales. Classiquement, l'estimation de la précision des évaluations génomiques est réalisée à l'aide d'une étude de validation. On replace les animaux les plus jeunes dans la situation d'un candidat à la sélection (c'est-à-dire sans phénotype) dont on cherche à prédire la valeur génétique. La taille de l'intervalle de confiance associé à la corrélation entre phénotypes et valeurs génétiques estimées des évaluations génomique est fonction du nombre d'animaux de la population de validation (Erbe et al., 2010). Dans les races régionales, la taille de la population de référence est inférieure à 500 individus. Le nombre d'animaux dans la population de validation est donc le plus souvent inférieur à 100 ce qui entraine une erreur d'estimation de la précision des évaluations génomiques importante (Erbe et al., 2010).

Il a donc été proposé de réaliser un premier test d'évaluation génomique sur puce haute densité en se plaçant dans la situation des trois principales races laitières. Afin de limiter les temps de calcul et d'équilibrer la représentation de chacune des races pour une évaluation multiraciale, seuls les taureaux Holstein ayant des filles en France (4989 animaux), et non pas la totalité de la population de références issues des échanges à l'intérieur du consortium EuroGenomics, ont été conservés dans la population de référence.

Plusieurs évaluations utilisant des génotypes 50K ou HD et une population de référence mono-raciale ou multiraciale ont été réalisées. Arbitrairement, la race Normande a été considérée comme race à évaluer à partir de sa population de référence et/ou des populations de référence des deux autres races. Afin de simuler différentes tailles de population de référence, la population d'apprentissage Normande a été soit considérée dans son intégralité (1597 individus) soit réduite à 198 ou 404 individus. Ainsi, la précision des différents types d'évaluations a été comparée en utilisant la corrélation entre les phénotypes et les valeurs génétiques estimées de 394 taureaux normands (population de référence de taille relativement grande).

B. Article II : Efficacité d'une sélection génomique intra-race et multi-race pour différentes tailles de population de référence

Hozé C, Fritz S, Phocas F, Boichard D, Ducrocq V, Croiseau P. 2014,

Efficiency of multi-breed genomic selection for dairy cattle breeds with different sizes of reference population

J. Dairy. Sci., 2014, 97 : 3918-3929

J. Dairy Sci. 97:3918–3929
http://dx.doi.org/10.3168/jds.2013-7761
© American Dairy Science Association®, 2014.

Efficiency of multi-breed genomic selection for dairy cattle breeds with different sizes of reference population

C. Hozé,*†[1] S. Fritz,† F. Phocas,* D. Boichard,* V. Ducrocq,* and P. Croiseau*
*Institut National de la Recherche Agronomique (INRA), UMR 1313, Génétique Animale et Biologie Intégrative (GABI), 78350 Jouy-en-Josas, France
†Union Nationales des Coopératives d'Élevages et d'Insémination Animales (UNCEIA), 149 rue de Bercy, 75012 Paris, France

ABSTRACT

Single-breed genomic selection (GS) based on medium single nucleotide polymorphism (SNP) density (~50,000; 50K) is now routinely implemented in several large cattle breeds. However, building large enough reference populations remains a challenge for many medium or small breeds. The high-density BovineHD BeadChip (HD chip; Illumina Inc., San Diego, CA) containing 777,609 SNP developed in 2010 is characterized by short-distance linkage disequilibrium expected to be maintained across breeds. Therefore, combining reference populations can be envisioned. A population of 1,869 influential ancestors from 3 dairy breeds (Holstein, Montbéliarde, and Normande) was genotyped with the HD chip. Using this sample, 50K genotypes were imputed within breed to high-density genotypes, leading to a large HD reference population. This population was used to develop a multi-breed genomic evaluation. The goal of this paper was to investigate the gain of multi-breed genomic evaluation for a small breed. The advantage of using a large breed (Normande in the present study) to mimic a small breed is the large potential validation population to compare alternative genomic selection approaches more reliably. In the Normande breed, 3 training sets were defined with 1,597, 404, and 198 bulls, and a unique validation set included the 394 youngest bulls. For each training set, estimated breeding values (EBV) were computed using pedigree-based BLUP, single-breed BayesC, or multi-breed BayesC for which the reference population was formed by any of the Normande training data sets and 4,989 Holstein and 1,788 Montbéliarde bulls. Phenotypes were standardized by within-breed genetic standard deviation, the proportion of polygenic variance was set to 30%, and the estimated number of SNP with a nonzero effect was about 7,000. The 2 genomic selection (GS) approaches were performed using either the 50K or HD genotypes. The correlations between EBV and observed daughter yield deviations (DYD) were computed for 6 traits and using the different prediction approaches. Compared with pedigree-based BLUP, the average gain in accuracy with GS in small populations was 0.057 for the single-breed and 0.086 for multi-breed approach. This gain was up to 0.193 and 0.209, respectively, with the large reference population. Improvement of EBV prediction due to the multi-breed evaluation was higher for animals not closely related to the reference population. In the case of a breed with a small reference population size, the increase in correlation due to multi-breed GS was 0.141 for bulls without their sire in reference population compared with 0.016 for bulls with their sire in reference population. These results demonstrate that multi-breed GS can contribute to increase genomic evaluation accuracy in small breeds.

Key words: multi-breed genomic selection, dairy cattle, high-density chip

INTRODUCTION

Genomic selection has been implemented in many countries. To date, more than 16 countries apply genomic information for dairy cattle breeding (Nilforooshan et al., 2010; Eggen, 2012). However, the accuracy of genomic breeding values depends mainly on the size of the reference population (Hayes and Goddard, 2008; Goddard, 2009) and therefore, genomic evaluations are mainly implemented in large breeds. In small breeds, only a small number of progeny-tested bulls is available and assembling a large reference population consisting of animals with accurate phenotypes is challenging. In this context, a potential solution is to combine reference populations from different breeds and develop multi-breed genomic selection (**GS**).

Such an approach requires conserved linkage disequilibrium (**LD**) across breeds to maintain the association between SNP and QTL. Studies on real data show that the association between marker alleles is maintained for SNP <10 kb apart (Gautier et al., 2007; de Roos et al., 2008), a condition not fulfilled with the classically used BovineSNP50 BeadChip (**50K**, ~50,000 SNP; Illumina

Received November 25, 2013.
Accepted February 25, 2014.
[1]Corresponding author: chris.hoze@jouy.inra.fr

Inc., San Diego, CA). A simulation study by de Roos et al. (2009) confirmed these results and found that the most accurate genomic prediction was achieved when all reference populations were combined and a chip containing more than 300,000 SNP was used. Therefore, the BovineHD BeadChip (**HD**; Illumina Inc.), developed in 2010 and containing ~777,000 SNP, should be sufficiently dense to allow for efficient multi-breed GS.

Given the large number of animals already genotyped on the 50K chip, regenotyping reference populations on this HD chip is not economically justified. Imputation studies reported an observed allelic imputation error rate <1% when 50K genotypes were imputed to HD (Su et al., 2012; Hozé et al., 2013; Pausch et al., 2013; Schrooten et al., 2014). These low error rates are expected to have a minor effect on the reliability of GS (Dassonneville et al., 2011; Mulder et al., 2012). Imputation of HD reference populations from 50K genotypes and investigations on the benefit of the genomic evaluation are therefore possible.

Until now, studies on the 50K and HD panels showed limited gain in accuracy when comparing multi-breed to single-breed GS (Hayes et al., 2009; Erbe et al., 2012). However, when comparing methods, these researchers confirmed the advantage of Bayesian approaches compared with genomic BLUP for EBV estimation (Hayes et al., 2009; Erbe et al., 2012); they stated that setting a large proportion of SNP effects to zero is necessary to take advantage of the density of the HD chip (Erbe et al., 2012). This conclusion is in agreement with conserved QTL-marker association at small distances only. Furthermore, adding a polygenic component avoids spurious SNP-QTL associations due to pedigree relationship (Solberg et al., 2009; Liu et al., 2011) and helps to select QTL with rare alleles, small effects, or both (Calus and Veerkamp, 2007; Goddard, 2009). Inclusion of a polygenic component also increases the accuracy of genomic EBV (**GEBV**) prediction and allows for regression slopes closer to 1 [M. Gunia (Institut National de la Recherche Agronomique Génétique Animale et Biologie Intégrative (INRA GABI), Jouy-en-Josas, France), R. Saintilan (INRA GABI; UNCEIA, Paris, France), E. Venot (INRA GABI), C. Hozé, M. N. Fouilloux (Institut de l'Élevage, Paris, France), and F. Phocas; unpublished data].

Within-breed, genomic evaluation relies on short distance QTL-SNP associations and on long-distance LD due to relationships. We assumed here that in a multi-breed situation, across-breed information is brought only by QTL-SNP associations shared across breeds; therefore, focusing on them should avoid detecting SNP associated with genetic background of the breed. BayesC (Kizilkaya et al., 2010) and BayesCπ (Habier et al., 2011) approaches have been widely used in GS programs and allow setting a proportion (π) of SNP with a zero effect. Therefore, these approaches were chosen here to compare accuracy of single-breed and multi-breed GS for a small reference population.

In small breeds, reference population size is limiting. Not only is the training set small but assessing the achieved accuracy is difficult because of the small validation population. If the validation population is enlarged, it would be at the expense of the training set and, therefore, at the expense of prediction accuracy (Erbe et al., 2010). Therefore, we chose here to use a large dairy breed to mimic a small breed and develop a multi-breed GS method. This strategy offers the opportunity to study several training population sizes to mimic either small or large breeds while using a unique reasonably large validation set. We first investigated the phenotypes used, the proportions of SNP with a nonzero effect, and the proportion of residual polygenic variance on a 50K basis. Then, we compared the predictive ability of genomic evaluation based on single-breed and multi-breed reference populations using the HD data set. Then, we assessed the benefit of using the HD chip for multi-breed analysis and investigated the effect of population structure on GS accuracy.

MATERIALS AND METHODS

Reference Population

Genotypes used in this study came from the French genomic evaluation. The biological tissues, either cryopreserved semen or blood samples, were provided by various commercial AI companies and breeder organizations in the framework of their breeding program activities. Therefore, no ethical approval was required for this study.

In total, 535, 527, and 773 influential bulls from the Normande (**NO**), Montbéliarde (**MO**), and Holstein (**HO**) breeds were genotyped with the Illumina Bovine HD BeadChip and were used to impute HD genotypes for animals of the French reference population genotyped with the 50K chip. Quality control was performed within breed on the HD and 50K genotypes, using the same criteria for both chips. Genotyped animals with a call rate <0.95 were removed from the analysis. Only markers mapped on the UMD3.1 assembly (http://bovinegenome.org/cgi-bin/gbrowse/bovine_UMD31/) covering the 29 bovine autosomes were used. Any SNP showing departure from Hardy-Weinberg equilibrium ($P < 0.0001$) or with more than 10% missing genotypes were removed. In addition, genotype consistency was checked using 1,838 animals that were genotyped on both chips, and 352 markers that were discordant for more than 1% of these animals were excluded.

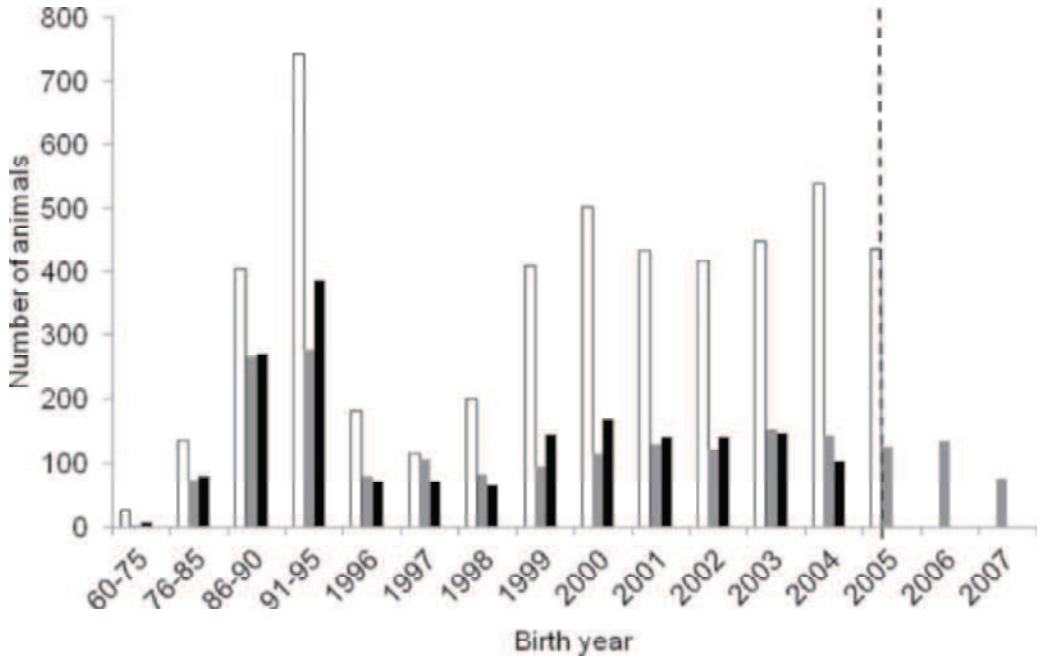


Figure 1. Birth year distribution for the training populations: Holstein = white bars; Montbéliarde = black bars; Normande = gray bars. The dashed line corresponds to the cut-off date between training and validation populations.

Imputation was performed using Beagle 3.3.0 software (Browning and Browning, 2007) with estimated allelic imputation error rates of 0.33, 0.51, and 0.71% in NO, MO, and HO, respectively (Hozé et al., 2013). The imputation study revealed 1,980 SNP with high imputation error rates in most breeds (Hozé et al., 2013). These SNP were also discarded for the analysis leading to an HD data set of 706,791 SNP. The 37,364 SNP in common between the HD and 50K panels were used to mimic 50K genotypes and compare 50K and HD panels for genomic evaluation.

Among the complete French reference population, only bulls with daughters in France and an equivalent daughter contribution (**EDC**; VanRaden and Wiggans, 1991) >20 were retained, leading to an overall reference population of 8,768 animals (Table 1). Birth year distribution per breed is shown in Figure 1. Phenotypes used were daughter yield deviation (**DYD**; VanRaden and Wiggans, 1991) for milk yield, fat yield (**FatY**), protein yield (**ProY**), fat content (**Fat%**), protein content (**Pro%**), and SCS.

Scenarios Tested

A validation scheme was used to study predictive ability of the genomic selection approaches. In a small breed, where the number of animals in the reference population is low, having a reasonably large validation population to correctly interpret accuracy is nearly impossible. To circumvent this problem, we chose here to use the Normande reference population. The youngest 20% bulls (i.e., those born after November 2004) formed the validation population and the 80% oldest animals were used to form several training populations of different sizes (Figure 1).

The first scenario (A) mimicked a large-breed situation: all 1,597 available NO bulls formed the training population. Most (96.5%) of validation bulls had their sire and maternal grandsires in training set A. In the second scenario (B), only the 404 HD-genotyped NO bulls from the 1,597 were kept in the training population. These 404 animals correspond to the major contributors of the breeds with, whenever possible,

Table 1. Reference population size, number of animals genotyped using a high-density (HD) chip, and imputation error rate in Montbéliarde, Holstein, and Normande animals

Breed	Reference population (imputed + genotyped)	HD-genotyped population	Estimated allelic imputation error rate (%)
Montbéliarde	1,788	788	0.51
Holstein	4,989	530	0.73
Normande	1,991	551	0.33

2 to 4 siblings per sire to keep a familial structure. Major contributors were selected based on their marginal contributions to the population (Boichard et al., 1997) computed using the PEDIG software (Boichard, 2002). This training set B mimicked a regional breed. In this case, 83% of validation bulls had their sire and maternal grandsires in training set. A third scenario (C), with 198 NO bulls in the training set, was used to mimic an even smaller breed. It was built by selecting the ancestors of the validation bulls among the HD-genotyped bulls. Because of the selection procedure, the proportion of validation bulls with their sire and maternal grandsires in training set was the same as in scenario B (83%). Finally, scenario D mimicked an extreme case with no animal from the NO breed in the reference population; therefore, no sires of validation animals were part of the training population.

For animals in the validation population, phenotypes were erased to mimic candidates to selection. Then, the training population was used to estimate EBV for the validation animals.

Evaluation Model

The EBV were estimated using 3 approaches: pedigree-based BLUP performed with average information (AI)-REML approach (Jensen et al., 1996) and single-breed and multi-breed genomic evaluations performed using the BayesC approach (Kizilkaya et al., 2010) implemented in GS3 software (Legarra et al., 2013).

The general Bayesian mixture model (Meuwissen, 2009) is

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Mq} + \mathbf{Zu} + \mathbf{e}, \quad [1]$$

where $\mathbf{y}$ is the vector of DYD, $\mathbf{1}$ is a vector of 1s, $\mathbf{q}$ is the vector of genotype effects, $\mathbf{M}$ is the incidence matrix for markers genotypes, $\mathbf{u}$ is the vector of polygenic effect, $\mathbf{Z}$ is the incidence matrix for the polygenic effect, and $\mathbf{e}$ is a vector of uncorrelated residuals normally distributed, whose variance is inversely proportional to the EDC of each DYD. The model for multi-breed genomic evaluation was the same as in equation [1], except that a breed effect was included.

The model assumes that a small proportion (π) of SNP has a nonzero effect and divides total additive variance into the sum of the variances due to each SNP effect and the residual polygenic variance. In our data sets, convergence was not obtained and we had to assume π known. A strong prior on the variances was used to constrain the proportion of residual polygenic component to a constant.

Phenotypes and Parameters Used in Genomic Evaluations

Total genetic variances were estimated using the average information (AI)-REML approach (Jensen et al., 1996). Genetic standard deviations (**SD**) vary widely among breeds (Table 2), in particular for fat and protein contents. Therefore, we tested 2 types of phenotypes for genomic evaluation: DYD and DYD standardized by the genetic SD of the breed.

Four values ranging from 10 to 40% were tested for the proportion of variance due to residual polygenic effect. This parameter was used to divide total genetic variance into the sum of the variance due to SNP effect and that due to the residual polygenic effect. Two values were tested for the proportion of SNP with a nonzero effect expressed as an equivalent number of expected SNP with a nonzero effect (Su et al., 2012): a value of about 700 SNP ($\pi = 0.001$ on the HD panel and $\pi = 0.02$ on the 50K) corresponding to a scenario with a small number of QTL influencing the trait, and a value of about 7,000 SNP ($\pi = 0.01$ on the HD panel and $\pi = 0.2$ on the 50K) corresponding to a scenario with a larger number of QTL.

We tested several sets of parameters using training set B and 50K genotypes as listed above and summarized in Table 3. The parameter settings for the analysis on HD data values were those resulting in the highest average weighted Pearson correlation between EBV and DYD across the 6 traits for validation animals. Because computational requirements for an HD analysis were high, we assumed that the parameters were stable among SNP panels and used them for the analysis on the HD chip. This assumption is supported by Erbe et al. (2012) and Haile-Mariam et al. (2013), where it was shown that the average number of SNP with a nonzero effect and the proportion of genetic variance unaccounted by SNP markers were relatively similar between the 50K and the HD panel.

Comparison of Evaluation Methods

After defining the parameters for genomic evaluation, we compared the accuracy of evaluation methods. The criterion to assess the accuracy of prediction was the weighted Pearson correlation (Peers, 1996) between EBV and DYD for validation animals and the weights being the EDC. Data sets A, B, and C were used to compare accuracies of EBV prediction using pedigree-based BLUP, single-breed BayesC, and multi-breed BayesC. For scenarios A and B, we also compared the achieved accuracy of GS using a 50K or an HD panel. In scenario D, no animal from the NO breed was used

Table 2. Phenotypic and genetic standard deviations of the traits analyzed for Montbéliarde (MO), Normande (NO), and Holstein (HO) breeds

Item	Milk (kg)			Fat yield (kg)			Protein yield (kg)			Fat content (g/kg)			Protein content (g/kg)			SCS (score unit)		
	MO	NO	HO	MO	NO	HO	MO	NO	HO	MO	NO	HO	MO	NO	HO	MO	NO	HO
Phenotypic SD	1,188	1,093	1,307	46.9	46	50.5	37	35	36.6	3.2	3.91	4.94	1.89	1.99	2.13	1.06	1.03	1.14
Genetic SD	651	599	716	25.7	25.2	27.6	20.3	19.2	20	2.26	2.76	3.5	1.34	1.41	1.51	0.46	0.46	0.49

in training; therefore, only multi-breed GS was used to estimate EBV. Tested scenarios and evaluation methods are summarized in Table 3.

In our scenarios, most validation animals had their sire and grandsires in the training set. We assumed that the accuracy achieved with pedigree-based BLUP was close to that achieved with a classical pedigree-based evaluation except that dam performances were not taken into account. Therefore, we compared the gain in accuracy of genomic evaluation methods to the accuracy achieved with pedigree-based BLUP.

RESULTS

Parameters for Genomic Evaluations

Several sets of parameters were tested for BayesC on the 50K data set: 2 numbers of SNP with a nonzero effect, 4 values of proportions of residual polygenic variance ranging from 10 to 40%, and 2 types of phenotypes (raw or standardized DYD) were tested (Table 3). Retained values were based on the highest average accuracy over the 6 traits for training set B (containing 404 NO bulls) and for a single-breed or multi-breed situation.

In the single-breed situation, no clear differences were found between sets of parameters and the average accuracy over the 6 traits for the 394 animals in validation ranged from 0.377 to 0.390 depending on the chosen parameters (not shown). A larger influence of type of phenotypes, proportion of polygenic variance, and proportion of SNP with a nonzero effect was observed for multi-breed genomic evaluations (Figure 2), and average achieved accuracies ranged from 0.338 to 0.398. The largest differences were observed between the 2 tested proportions of SNP: moving from an expected number of SNP with an effect of 700 to 7,000 led to an increase in correlation of 0.03. The effect of the proportion of polygenic variance depended on the proportion of SNP with a nonzero effect. We observed that when a larger polygenic component was assumed (>20% of the genetic variance), fewer SNP were required to accurately predict phenotypes. This suggests that when fewer SNP are included in the model, a smaller part of genetic variance is explained by SNP and therefore a larger variance due to residual polygenic is required. The optimal proportion of variance due to polygenic components seemed to be related to the heritability of the traits. A higher proportion of residual polygenic variances gave slightly higher results for low heritability traits (SCS), whereas lower proportions were required for fat and protein contents. A proportion of polygenic variance of 30% was the optimal trade-off across the 6 traits. Standardization of phenotypes allowed an aver-

Table 3. Overview of the main features of the scenarios tested for Montbéliarde (MO), Normande (NO), and Holstein (HO) breeds

Feature	Training A	Training B	Training C	Training D
BLUP	1,597 NO	404 NO	198 NO	
Single-breed GS[1]	1,597 NO	404 NO	198 NO	
Multi-breed GS	1,597 NO	404 NO	198 NO	4,989 HO
	4,989 HO	4,989 HO	4,989 HO	1,788 MO
	1,788 MO	1,788 MO	1,788 MO	
Phenotypes[2]	DYD/sDYD	DYD/sDYD	sDYD	sDYD
SNP panel[3]	50K/HD	50K/HD	HD	HD
Eff_SNP[4]	7,000	700/7,000	7,000	7,000
Polygenic (%)	30	10/20/30/40	30	30
Validation (with sire + mgs)[5]	394 NO (380)	39 NO (326)	394 NO (326)	394 NO (0)

[1]GS = genomic selection.

[2]DYD = daughter yield deviation; sDYD = standardized DYD.

[3]50K = ~50,000 SNP panel; HD = high-density SNP chip.

[4]Eff_SNP = expected number of markers with a nonzero effect.

[5]Number of validation bulls with sire and maternal grandsire (mgs) in the training set.

age gain of 0.01 in accuracy. Standardization had little effect on accuracy when the proportion of SNP with a nonzero effect was high. This suggests that standardization reduces the residual noise associated with raw phenotypes and avoids selecting SNP that explain differences in genetic variance between breeds.

The highest accuracy was obtained when the phenotypes were standardized, the proportion of polygenic variance was set to 30%, and an estimated number of SNP with a nonzero effect was around 7,000. Therefore, these parameters were used in the rest of this study.

Multi-Breed Versus Single-Breed GS

Multi-breed and single-breed genomic evaluations were compared for training sets A, B, and C. Results based on the HD panel are presented in Table 4. Regardless of the method, accuracies were higher in training set A than in the smaller training sets, which was at least partly due to our design. Training set A included 97% of the validation bulls sires and maternal grandsires. This proportion decreased to 83% with training sets B and C. This led to a lower accuracy of pedigree-based BLUP and a poorer estimation of the polygenic component in GS. For the same reasons, regression slopes were closer to 1 in training set A.

Comparing pedigree-based and genomic evaluations, we notice that the use of DNA information increased the accuracy of EBV prediction in both data sets (Table 4). However, the gain was 2.5 times higher in training set A (+0.193) than in training sets B (+0.077) and C (+0.057). The larger reference population allowed a better estimation of SNP effects and therefore higher accuracies. The increase in correlation due to the GS was trait-dependent (Table 4). The largest gains were observed for highly heritable traits influenced by large QTL such as milk content and the lowest were observed for SCS.

In all data sets, moving from single-breed to multi-breed GS allowed a slight increase in the quality of prediction (Table 4). The average increase in accuracy compared with pedigree-based BLUP varied from 0.193 for single-breed GS to 0.209 with multi-breed GS in

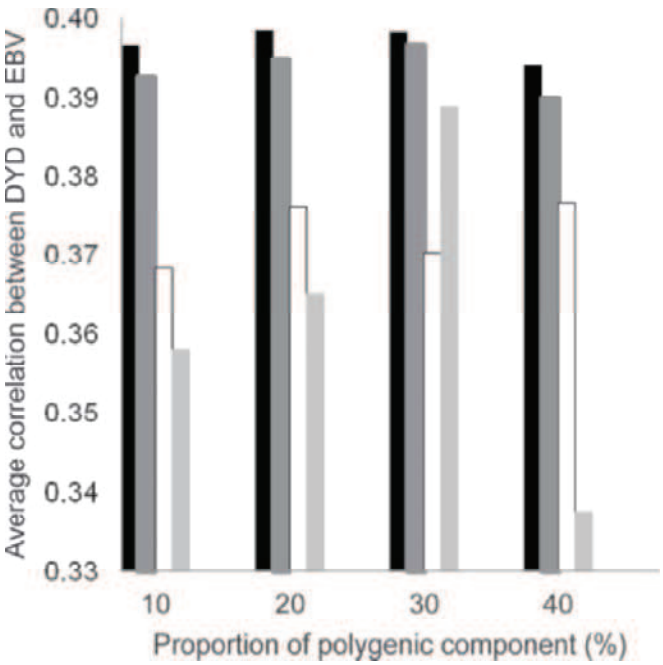


Figure 2. Average correlation between daughter yield deviation (DYD) and genomic EBV (GEBV) across the 6 traits in relation to the proportion of polygenic variance, standardization of performance, and proportion of SNP with a nonzero effect. The GEBV were estimated using multi-breed BayesC, training set B, and 50K (~50,000) genotypes. Black bars = proportion of SNP with a nonzero effect of 0.1 and standardized phenotypes; dark gray bars = proportion of SNP with a nonzero effect of 0.1 and nonstandardized phenotypes; white bars = proportion of SNP with a nonzero effect of 0.01 and standardized phenotypes; light gray bars = proportion of SNP with a nonzero effect of 0.01 and nonstandardized phenotypes.

Table 4. Weighted Pearson correlations between observed and predicted daughter yield deviations (DYD) for the 6 traits using pedigree-based BLUP, single-breed genomic selection (GS) and multi-breed GS using training sets A (1,597 Normande bulls), B (404 Normande bulls), and C (198 Normande bulls) with the high-density panel[1]

Training set	Model	Trait[2]						Average corr.[3]	Average slope[4]
		Milk	FatY	ProY	Fat%	Pro%	SCS		
A	Multi-breed GS	0.505	0.518	0.514	0.647	0.644	0.568	0.566	0.875
	Single-breed GS	0.484	0.489	0.495	0.638	0.632	0.561	0.550	0.863
	BLUP	0.316	0.346	0.305	0.350	0.397	0.430	0.357	0.760
B	Multi-breed GS	0.345	0.401	0.353	0.478	0.494	0.427	0.416	0.726
	Single-breed GS	0.312	0.389	0.333	0.398	0.470	0.418	0.387	0.710
	BLUP	0.234	0.299	0.244	0.305	0.382	0.399	0.311	0.663
C	Multi-breed GS	0.347	0.401	0.368	0.434	0.444	0.389	0.397	0.717
	Single-breed GS	0.319	0.396	0.340	0.368	0.420	0.368	0.368	0.729
	BLUP	0.243	0.306	0.249	0.305	0.373	0.393	0.311	0.685

[1]Genomic evaluations were based on a BayesC approach. Phenotypes used were standardized within breed; the proportion of residual polygenic component was set to 30% and the number of expected SNP with a nonzero effect to 7,000.

[2]Milk yield (Milk), fat yield (FatY), protein yield (ProY), fat content (Fat%), protein content (Pro%), and SCS.

[3]Average correlation between DYD and EBV over the 6 traits.

[4]Average regression slope over the 6 traits.

training set A, from 0.079 to 0.106 in training set B, and from 0.057 to 0.086 in training set C. It represents extra gain of, respectively, 8, 39, and 50% with multi-breed GS compared with single-breed GS. The use of information from MO and HO breeds allowed a gain for the 6 traits studied (Table 4) but the extra gain was trait- and data set-dependent. The largest increase was observed for Fat% in training set B (+0.08) and the lowest for SCS in training set A (+0.007). Fat content is influenced by large QTL, some of which are probably common across breeds. This could explain why the largest increases were observed for this trait. Common QTL may be more difficult to identify in traits influenced by small QTL (SCS, ProY). The benefit of multi-breed GS is also dependent on the accuracy of prediction of single-breed GS. For training set A, where the NO reference population was large, the accuracy of single-breed GS was already high and adding information from other breeds had a limited effect. Regression slopes of multi-breed GS were in the same range as those from single-breed GS, meaning that adding information from other breeds did not affect the bias.

Benefit of the HD Chip for Single- and Multi-Breed GS

The comparison between the 50K study and the HD study allowed us to assess the effect of SNP density on the accuracy of prediction. Results for single-breed and multi-breed GS are presented in Figure 3 for training sets A and B.

As seen before, accuracies were higher for training set A than for training set B and were slightly higher for multi-breed GS than for single-breed GS. However, the effect of SNP density depended on the evaluation method. Slight or no differences were observed between the 50K and HD panels for single-breed GS for training set A (+0.006) and training set B (+0.002). Larger differences were observed for multi-breed GS. The average gain of HD data set was 0.022 for training set A and 0.018 for training set B. This could be explained by the higher linkage disequilibrium on the HD chip allowing a SNP-QTL association better conserved across breeds.

Effect of Population Structure

Accuracies of prediction were compared for validation bulls whether they had their sire in the training sets B and C or not (Table 5). Differences in correlation across training sets and evaluation methods were smaller for bulls with their sire in reference population than for those without. In both training sets, the correlation was halved for bulls without their sire in the training population but the average gain in correlation with multi-breed GS was more than 10 times higher for bulls that did not have their sire in the training population (+0.16) than for those that did (+0.013): moving to a larger reference population or a multi-breed GS was more beneficial for validation animals that did not have their sire in training population. The increase in EBV prediction accuracy is likely due to a more accurate estimation of SNP effect. For bulls without their sire in training set B, the correlation between direct genomic breeding value and DYD increased from 0.218 to 0.375 when the reference population included more animals, regardless of the breed of the additional training animals. In the case of validation bulls with their sire in the reference population, accuracies were not improved.

Table 5. Weighted-average Pearson correlations between observed daughter yield deviations and EBV for the 6 traits[1] in relation to single-breed and multi-breed genomic selection (GS) in the training set, the evaluation methods on the high-density panel, and the presence of sires in the training set[2]

	Training set C (198 bulls)		Training set B (404 bulls)	
Group	Single-breed GS	Multi-breed GS	Single-breed GS	Multi-breed GS
Bulls with sire in training set (n = 355)	0.414	0.432	0.431	0.444
Bulls without sire in training set (n = 39)	0.129	0.251	0.189	0.349

[1]Milk yield, fat yield, protein yield, fat content, protein content, and SCS.

[2]Genomic evaluations were based on a BayesC approach. Phenotypes used were standardized within breed; the proportion of residual polygenic component was set to 30% and the number of expected SNP with a nonzero effect to 7,000.

When no animals from the Normande breed were in the reference population, accuracies of multi-breed GS were around 0.18 for all traits except for Pro% and SCS (Table 6). For the latter traits, the benefit of multi-breed GS compared with single-breed GS was already low in training sets B and C (Table 5). On average, the accuracy achieved was close to the one using data set C and single-breed GS for animals without their sire in the reference population. This suggests that multi-breed GS could be beneficial even for breeds with no animals at all in the reference population.

DISCUSSION

In this study, we assessed the benefit of combining reference populations from several breeds and performed multi-breed GS. A BayesC approach known to provide good results in genomic evaluation (e.g., Croiseau et al., 2012; Duchemin et al., 2012) was used.

The model used here requires some key parameters such as the proportion of SNP with a nonzero effect and the proportion of genetic variance due to a residual polygenic component. We looked for the optimal values for these parameters using data from the 50K chip and later used these for the analysis on the HD chip. When the population was small (<500 bulls) and HD genotypes were used, the average gain in accuracy compared with pedigree-based BLUP was 0.067 using single-breed GS and up to 0.105 using multi-breed GS. The benefit of multi-breed GS compared with single-breed GS was reduced when 50K genotypes were used and almost disappeared in the case of an already-large reference population.

Because of the low number of progeny-tested bulls, the current reference population for French regional breeds (Abondance, Brown Swiss, Tarentaise, and Simmental) consisted of fewer than 200 bulls (Hozé et al., 2013); therefore, using them for a validation study was questionable. The alternative implemented here was to use a national breed with a large reference population and to artificially reduce the training population size. The major advantage of this approach is to keep a reasonably large validation population, allowing for more robust conclusions and for a better estimation of the effect of reference population. This design also has some disadvantages. In particular, the choice of the animals kept in a reduced training set has a major influence on population structure and on genomic selection accuracy. Studies on simulated and real data also showed that for one given animal, the reliability of GEBV and direct genomic breeding value was linked to its relationship to reference population (Habier et al., 2010; Clark et al., 2012; Pszczola et al., 2012). Erbe et al. (2012) observed a loss of correlation between 0.05 and 0.09 when the sire of the animal considered was not in the reference population. Here, observed losses were dramatically larger (between 0.095 and 0.285) because reference population sizes considered in this study were much smaller. Accuracy of GS depends on the average relationship between individuals and LD between

Table 6. Weighted Pearson correlations between observed and predicted daughter yield deviations for the 6 traits[1] using multi-breed genomic selection (GS) with training set D (no Normande bulls) and a high-density panel[2]

Training set	Model	Milk	FatY	ProY	Fat%	Pro%	SCS	Average
D	Multi-breed GS[2]	0.172	0.168	0.182	0.174	0.097	0.045	0.140

[1]Milk yield (Milk), fat yield (FatY), protein yield (ProY), fat content (Fat%), protein content (Pro%), and SCS.

[2]Genomic evaluations were based on a BayesC approach. Phenotypes used were standardized within breed; the proportion of residual polygenic component was set to 30% and the number of expected SNP with a nonzero effect to 7,000.

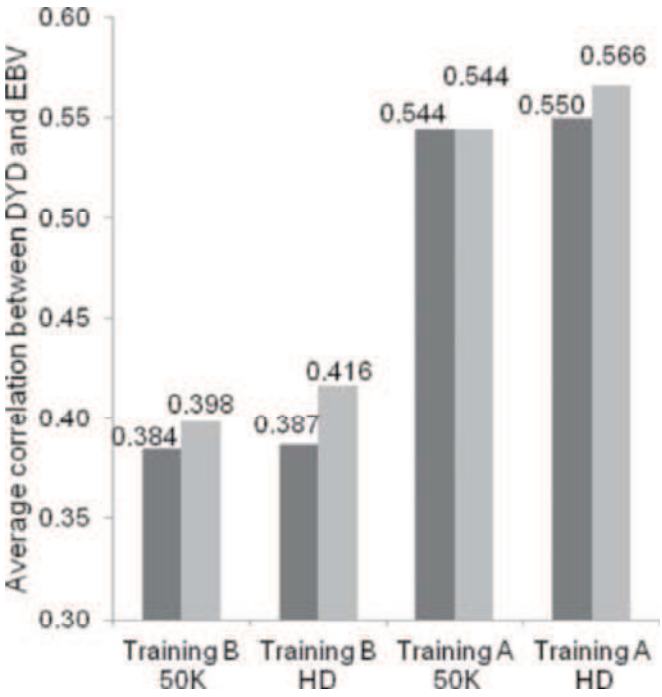


Figure 3. Average correlation across the 6 traits in relation to the training set (A = 1,597 bulls, B = 404 bulls), the genomic evaluation approach, and the SNP panel. Dark gray = results for single-breed genomic selection; light gray = results from multi-breed genomic selection. Genomic evaluations consisted in a BayesC approach. Phenotypes used were standardized within breed; the proportion of residual polygenic component was set to 30% and the number of expected SNP with a nonzero effect to 7,000. DYD = daughter yield deviation; 50K = ~50,000 SNP panel; HD = high-density SNP chip.

markers and causative genes (Habier et al., 2011). The accuracy due to LD is considered as a lower bound of the global accuracy for animals with no close relatives in the reference population and was found to increase with the reference population size (Habier et al., 2011). Our results are consistent with this statement: the average increase in accuracy for animals without their sires in the reference population was higher with larger training population or with a multi-breed evaluation.

The accuracy achieved in GS depends on the familial structure and on the accuracy obtained with pedigree-based BLUP. For this reason, comparing the gain in accuracy due to GS may be more relevant than comparing the accuracies themselves. The accuracy of prediction dramatically increased with GS in training set A (+0.193) and increased less for training sets B (+0.077) and C (+0.057). This result confirms the effect of reference population size on the accuracy of GS (Daetwyler et al., 2008; Goddard, 2009). In contrast, the extra gain with multi-breed GS was higher with training sets B (+0.030) and C (+0.029) than with training set A (+0.016). Holstein-Jersey studies (Hayes et al., 2009; Erbe et al., 2012) and a Holstein-Jersey-Brown Swiss study (Olson et al., 2012) also showed that combining reference populations is more beneficial for small populations than for large populations. It suggests that when the single-breed reference population is large enough, estimation of SNP effects and genomic evaluation are already accurate and adding information from other breeds does not improve them.

Density of markers is also known to affect the accuracy of GS (Daetwyler et al., 2008; Goddard, 2009). In single-breed evaluations, simulation studies forecast high increases in accuracy when marker density is increased (de Roos et al., 2009; Meuwissen and Goddard, 2010; VanRaden et al., 2011) but results on real data found limited gain with the HD chip (Erbe et al., 2012; Su et al., 2012; VanRaden et al., 2013). In our study, we also found little or no improvement (+0.006 for training set A and +0.003 for training set B) when moving from the 50K panel to the HD chip. One reason for this limited gain is probably the already high proportion of genetic variance captured by SNP on production traits with the 50K panel. This was estimated at 80% by Haile-Mariam et al. (2013) and was only slightly higher with the HD chip (Erbe et al., 2012). The higher number of effects to estimate may also be a factor limiting the increase in correlation. This hypothesis is supported by the lower gain in correlation due to the use of the HD chip for training set B where the reference population was small. Erbe et al. (2012) even observed a decrease in correlation with the HD chip for a Jersey population of 540 bulls. The benefit of the HD chip was clearer when multi-breed GS was used. The increase in correlation observed with the HD chip was 0.022 for training set A and 0.018 for training set B. These results are in the same range as in Erbe et al. (2012), where the use of HD chip led to a 0.03 increase of multi-breed GS accuracy observed in the Jersey population. The higher density of SNP allowed a higher persistence of LD across breeds and a more conserved QTL-SNP phase, as stated by de Roos et al. (2009).

However, the achieved accuracies observed in multi-breed studies using the HD chip are lower than those observed in simulated data sets where QTL and SNP effects are assumed to be shared across breeds (de Roos et al., 2009; Erbe et al., 2012). This may suggest that marker-QTL associations across breeds are less conserved than expected or that QTL effects may differ among breeds. Applying SNP effects estimated in one breed to another breed led to poor correlations between EBV and observed phenotypes (Erbe et al., 2012; Olson et al., 2012). This confirmed that associations between QTL and SNP are at least partially breed specific and that SNP effects cannot be directly transposed from one breed to another. In this study, breeds were pooled

together and an average SNP effect was estimated. This model may dilute the observed associations of markers with phenotypes as it does not take into account for possible differences in SNP effect among breeds or for the possible existence of breed-specific QTL. A possible improvement for multi-breed GEBV estimation would be to include a SNP × breed interaction in the model and to allow SNP effect to differ in variance, value, and sign between populations. Multi-trait models, where phenotypes of each breed are considered correlated traits, were proposed to account for the breed-QTL interaction. In Nordic Red dairy cattle where the populations are closely related, an increase in correlation of 2 to 3% was observed when comparing a breed-specific model and a genomic BLUP assuming a uniform population (Makgahlela et al., 2013). Considering Holstein, Jersey, and Brown Swiss data sets, Olson et al. (2012) showed that using phenotypes of each breed as correlated traits led to significantly higher reliabilities than just pooling the reference population over several breeds but they found limited gain compared with a single-breed genomic evaluation. In a study on the 3 main French dairy breeds, Karoui et al. (2012) observed an increase in correlation when comparing a multi-trait evaluation with each breed considered as a different trait to a single-breed genomic evaluation. However, using estimated genetic correlations between breeds did not improve EBV estimation compared with a situation where genetic correlations were set arbitrarily to 0.95. This situation is close to that where populations are pooled and suggests that for the main French breeds, using a multi-trait model would not improve genomic evaluation. An alternative to account for possible differences in SNP-QTL association was proposed by Brøndum et al. (2012). In their model, a first prediction of SNP effects is performed on the largest breeds and the posterior probability for a SNP to have a small or a large effect is used as a prior for the estimation of SNP effect in the second breed. Accuracies achieved with this approach were higher than those obtained with a single-breed GS but lower than those achieved when reference populations are pooled (Brøndum et al., 2012).

Until now, methods proposed to account for incompletely conserved LD between QTL and SNP resulted in limited gain in accuracy. This is probably explained by the higher number of effects to estimate with a breed-specific allele model, which requires a large reference population per breed (Ibáñez-Escriche et al., 2009). Use of haplotypes could also be useful to better detect conserved LD across breeds as it increases LD with the causal mutations (Hayes et al., 2007; Calus et al., 2008). With the reduction of genome resequencing cost and imputation from SNP chip to sequence data, GS based on sequence data can be envisioned (Druet et al., 2014). Sequence data are particularly appealing for multi-breed GS as the causal mutation is in the data set. Therefore, it is no longer needed to have a conserved marker-QTL association across breed. Again, the number of effects to estimate would be limiting and SNP preselection through QTL detection could probably be useful to better estimate SNP effect and take benefit of the higher density (Croiseau et al., 2011).

In this study, multi-breed genomic evaluations were used to increase reference population size and therefore to increase the accuracy of EBV predictions. A complementary solution is to genotype females and to include cow information in genomic evaluation. The benefit of inclusion of females on accuracy of genomic selection and increase in genetic gain has been assessed in simulations studies (Mc Hugh et al., 2011). On real data, adding 10,000 cows to a population of 3,000 bulls led to an increase in reliability of genomic prediction ranging from 4 to 8% (Pryce et al., 2012). For a smaller population and a highly heritable trait, Calus et al. (2013) showed that adding 1,609 cows to a population of 296 bulls increased prediction accuracy by 45%. Indeed, the increase in accuracy depends on the heritability of the traits studied, the number of cows added to the reference population, and the bull reference population size (Egger-Danner et al., 2012). Therefore, it is likely that, in a small breed where the number of progeny-tested bulls is limited, inclusion of cows in reference population is very beneficial.

CONCLUSIONS

This study compared GS approaches for small breeds of dairy cattle and 6 traits. It focused on the benefit of multi-breed GS for breeds with fewer than 500 animals in their reference population. We observed an average increase in accuracy over the 6 traits ranging from 0.057 and 0.209 with single-breed genomic evaluation compared with pedigree-based BLUP, and the lowest gain was observed for a training population of <200 animals. Compared with single-breed GS, multi-breed GS allows a gain in correlation between EBV and observed DYD that ranges between 0.016 and 0.029. The highest increase was observed for traits influenced by large QTL (e.g., fat content). Conserved linkage disequilibrium across breeds does exist and multi-breed GS may benefit breeds with a small reference population size. However, gain in correlation with multi-breed GS was low when the breed reference population was already large and when the validation animals had their sire in the reference population. Implementation of GS in small breeds is feasible provided that ancestors of selection candidates are in the reference population.

With the increase in reference population size under a genomic evaluation breeding scheme, female genotyping, and multi-breed GS, EBV prediction accuracy in small dairy cattle breeds could quickly increase in the near future.

ACKNOWLEDGMENTS

The GEMBAL project was funded by the Agence Nationale de la Recherche (Paris, France; ANR-10-GENM-0014), ApisGene (Paris, France), Races de France (Paris, France), and INRA AIP Bioressources (Paris, France). The EUROGENOMICS consortium provided most of the Holstein genotypes. Most 50K genotypes originated from the Cartofine-ANR-05-GEN-ANIMAL-007 project funded by ANR (French National Research Agency) and ApisGene, and from genomic selection activity generated by the French cattle breeding companies, with Labogena (Jouy-en-Josas, France) as main genotyping laboratory.

REFERENCES

Boichard, D. 2002. PEDIG: A Fortran package for pedigree analysis suited for large populations. CD-ROM Commun. No. 28–13 in Proc. 7th World Congr. Genet. Appl. Livest. Prod., Montpellier, France. Organizing Committee 7WCGALP, Castanet-Tolosan, France.

Boichard, D., L. Maignel, and E. Verrier. 1997. The value of using probabilities of gene origin to measure genetic variability in a population. Genet. Sel. Evol. 29:5–23.

Brøndum, R. F., G. Su, M. S. Lund, P. J. Bowman, M. E. Goddard, and B. J. Hayes. 2012. Genome position specific priors for genomic prediction. BMC Genomics 13:543.

Browning, S. R., and B. L. Browning. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 81:1084–1097.

Calus, M. P. L., Y. de Haas, and R. F. Veerkamp. 2013. Combining cow and bull reference populations to increase accuracy of genomic prediction and genome-wide association studies. J. Dairy Sci. 96:6703–6715.

Calus, M. P. L., A. P. W. De Roos, and R. F. Veerkamp. 2008. Accuracy of genomic selection using different methods to define haplotypes. Genetics 178:553–561.

Calus, M. P. L., and 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. 124:362–368.

Clark, S. A., J. M. Hickey, H. D. Daetwyler, and J. H. Van der Werf. 2012. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genet. Sel. Evol. 44:4.

Croiseau, P., F. Guillaume, and S. Fritz. 2012. Comparison of genomic selection approaches in Brown Swiss within Intergenomics. Interbull Bull. 46:127–132.

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 preselection with the Elastic-Net algorithm. Genet. Res. (Camb.) 93:409–417.

Daetwyler, H. D., B. Villanueva, and J. A. Woolliams. 2008. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS ONE 3:e3395.

Dassonneville, R., R. F. Brøndum, T. Druet, S. Fritz, F. Guillaume, B. Guldbrandtsen, M. S. Lund, V. Ducrocq, and G. Su. 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.

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

de Roos, A. P. W., B. J. Hayes, R. J. Spelman, and M. E. Goddard. 2008. Linkage disequilibrium and persistence of phase in Holstein–Friesian, Jersey and Angus cattle. Genetics 179:1503–1512.

Druet, T., I. M. Macleod, and B. J. Hayes. 2014. Toward genomic prediction from whole-genome sequence data: Impact of sequencing design on genotype imputation and accuracy of predictions. Heredity 112(Special Issue):39–47.

Duchemin, S. I., C. Colombani, A. Legarra, G. Baloche, H. Larroque, J. M. Astruc, F. Barillet, C. Robert-Granié, and E. Manfredi. 2012. Genomic selection in the French Lacaune dairy sheep breed. J. Dairy Sci. 95:2723–2733.

Eggen, A. 2012. The development and application of genomic selection as a new breeding paradigm. Anim. Front. 2:10–15.

Egger-Danner, C., H. Schwarzenbacher, and A. Willam. 2012. Genotyping of cows for genomic EBVs for direct health traits—Genetic and economic aspects. Page 84 in Book of Abstracts of the 63rd Annual Meeting of the European Federation of Animal Science, Bratislava, Slovakia. Wageningen Academic Publ., Wageningen, the Netherlands.

Erbe, M., B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman, C. M. Reich, B. A. Mason, and M. E. Goddard. 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J. Dairy Sci. 95:4114–4129.

Erbe, M., E. C. J. Pimentel, A. R. Sharifi, and H. Simianer. 2010. Assessment of cross-validation strategies for genomic prediction in cattle. Page 129 in 9th World Congr. Genet. Appl. Livest. Prod., Leipzig, Germany. Gesellschaft für Tierzuchtwissenschaften, Gießen, Germany.

Gautier, M., T. Faraut, K. Moazami-Goudarzi, V. Navratil, M. Foglio, C. Grohs, A. Boland, J. G. Garnier, D. Boichard, G. M. Lathrop, I. G. Gut, and A. Eggen. 2007. Genetic and haplotypic structure in 14 European and African cattle breeds. Genetics 177:1059–1070.

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

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick. 2011. Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics 12:186.

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. 42:5.

Haile-Mariam, M., G. J. Nieuwhof, K. T. Beard, K. V. Konstatinov, and B. J. Hayes. 2013. Comparison of heritabilities of dairy traits in Australian Holstein-Friesian cattle from genomic and pedigree data and implications for genomic evaluations. J. Anim. Breed. Genet. 130:20–31.

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

Hayes, B. J., A. J. Chamberlain, H. McPartlan, I. Macleod, L. Sethuraman, and M. E. Goddard. 2007. Accuracy of marker assisted selection with single markers and marker haplotypes in cattle. Genet. Res. 89:215–220.

Hayes, B. J., and M. E. Goddard. 2008. Technical note: Prediction of breeding values using marker-derived relationship matrices. J. Anim. Sci. 86:2089–2092.

Hozé, C., M. N. Fouilloux, E. Venot, F. Guillaume, R. Dassonneville, S. Fritz, V. Ducrocq, F. Phocas, D. Boichard, and P. Croiseau.

2013. High-density marker imputation accuracy in sixteen French cattle breeds. Genet. Sel. Evol. 45:33.

Ibáñez-Escriche, N., R. L. Fernando, A. Toosi, and J. C. Dekkers., 2009. Genomic selection of purebreds for crossbred performance. Genet. Sel. Evol. 41:12.

Jensen, J., E. A. Mantysaari, P. Madsen, and R. Thompson. 1996. Residual maximum likelihood estimation of (co)variance components in multivariate mixed linear models using average information. J. Ind. Soc. Agric. Stat. 49:215–236.

Karoui, S., M. J. Carabaño, C. Díaz, and A. Legarra. 2012. Joint genomic evaluation of French dairy cattle breeds using multiple-trait models. Genet. Sel. Evol. 44:39.

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. 88:544–551.

Legarra, A., A. Ricard, and O. Filangi. 2013. GS3—Genomic selection, Gibbs Sampling. Gauss Seidel and Bayes Cπ. Accessed Nov. 6, 2013. http://snp.toulouse.inra.fr/~alegarra.

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.

Makgahlela, M. L., E. A. Mäntysaari, I. Strandén, M. Koivula, U. S. Nielsen, M. J. Sillanpää, and J. Juga. 2013. Across breed multitrait random regression genomic predictions in the Nordic Red dairy cattle. J. Anim. Breed. Genet. 130:10–19.

Mc Hugh, N., T. H. E. Meuwissen, A. R. Cromie, and A. K. Sonesson. 2011. Use of female information in dairy cattle genomic breeding programs. J. Dairy Sci. 94:4109–4118.

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., and M. Goddard. 2010. Accurate prediction of genetic values for complex traits by whole-genome resequencing. Genetics 185:623–631.

Mulder, H. A., M. P. L. Calus, T. Druet, and C. Schrooten. 2012. Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle. J. Dairy Sci. 95:876–889.

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

Olson, K. M., P. M. VanRaden, and M. E. Tooker. 2012. Multibreed genomic evaluations using purebred Holsteins, Jerseys, and Brown Swiss. J. Dairy Sci. 95:5378–5383.

Pausch, H., B. Aigner, R. Emmerling, C. Edel, K. U. Götz, and R. Fries. 2013. Imputation of high-density genotypes in the Fleckvieh cattle population. Genet. Sel. Evol. 45:3.

Peers, I. 1996. Statistical Analysis for Education and Psychology Researchers: Tools for Researchers in Education and Psychology. The Falmer Press, London, UK.

Pryce, J. E., B. J. Hayes, and M. E. Goddard. 2012. Genotyping dairy females can improve the reliability of genomic selection for young bulls and heifers and provide farmers with new management tools. Proc. ICAR Congr., Cork, Ireland. Accessed Jan. 15, 2014. http://www.icar.org/Cork_2012/Manuscripts/Published/Pryce%202.pdf.

Pszczola, M., T. Strabel, H. A. Mulder, and M. P. L. Calus. 2012. Reliability of direct genomic values for animals with different relationships within and to the reference population. J. Dairy Sci. 95:389–400.

Schrooten, C., R. Dassonneville, V. Ducrocq, R. F. Brøndum, M. S. Lund, J. Chen, Z. Liu, O. González-Recio, J. Pena, and T. Druet. 2014. Error Rate for Imputation from Illumina BovineSNP50 to Illumina BovineHD. Genet. Sel. Evol. 46:10.

Solberg, T. R., A. K. Sonesson, J. A. Woolliams, J. Odegard, and T. H. E. Meuwissen. 2009. Persistence of accuracy of genome-wide breeding values over generations when including a polygenic effect. Genet. Sel. Evol. 41:53.

Su, G., R. F. Brøndum, P. Ma, B. Guldbrandtsen, G. P. Aamand, and M. S. Lund. 2012. Comparison of genomic predictions using medium-density (~54,000) and high-density (~777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations. J. Dairy Sci. 95:4657–4665.

VanRaden, P. M., D. J. Null, M. Sargolzaei, G. R. Wiggans, M. E. Tooker, J. B. Cole, and G. A. Doak. 2013. Genomic imputation and evaluation using high-density Holstein genotypes. J. Dairy Sci. 96:668–678.

VanRaden, P. M., J. R. O'Connell, G. R. Wiggans, and K. A. Weigel. 2011. Genomic evaluations with many more genotypes. Genet. Sel. Evol. 43:10.

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

C. Bilan

Cette étude a comparé différentes stratégies d'évaluation génomique en fonction de la taille de la population de référence.

Quelle que soit la taille de la population de référence, augmenter la densité en marqueurs n'a pas permis d'améliorer significativement la corrélation entre valeurs génétiques estimées et phénotypes dans le cas d'une évaluation génomique intra-race. Ce résultat, observé dans plusieurs autres études (Erbe et al., 2012 ; Su et al., 2012 ; VanRaden et al., 2013), confirme qu'au-delà d'une certaine densité en marqueurs, augmenter le nombre de marqueurs entraine une augmentation du déséquilibre de liaison entre QTL et marqueurs mais aussi du nombre d'effets à estimer ce qui pénalise la qualité d'estimation.

En revanche dans le cas d'une évaluation multiraciale, l'utilisation de la puce HD est recommandée puisqu'elle permet d'augmenter la précision des valeurs génétiques estimées. Ce résultat est cohérent avec une meilleure conservation du déséquilibre de liaison ainsi qu'une meilleure cohérence des phases sur la puce HD (Larmer et al., 2014). Le gain de précision lié aux évaluations multiraciales est plus important lorsque la population de référence est limitée que dans le cas d'une population de référence de grande taille. Si le nombre d'individu dans la population de référence permet déjà une estimation de relativement bonne qualité des valeurs génétiques des animaux, ajouter des animaux d'une autre race dans la population de référence ne permet pas d'améliorer la qualité d'estimation. Ce résultat est probablement en partie lié à l'effet non linéaire de l'augmentation de la taille de la population de référence sur l'efficacité de la sélection génomique. De plus, chez les animaux d'une autre race que celle dont on cherche à prédire les valeurs génomiques, le lien entre allèles aux QTL et allèles aux marqueurs peut ne pas être totalement conservé chez les candidats.

On pourra noter que la précision estimée des évaluations génomiques intra-races sur les puces 50K ou HD est relativement élevée même dans les cas où la taille de la population de référence est inférieure à 500 individus.

V. Etude du cas de deux races génétiquement proches : exemple des races Montbéliarde et Simmental

A. Introduction

Nous avons vu dans l'étude précédente que combiner les populations de référence de différentes races pouvait constituer une stratégie intéressante pour le développement d'évaluations génomiques pour les races régionales. Cependant, sauf cas particulier, elle nécessite l'utilisation de génotypes haute densité ce qui augmente considérablement la demande en ressources informatiques.

Une alternative possible est de considérer un ensemble de races ayant divergé récemment et pour lequel le déséquilibre de liaison serait mieux conservé qu'au sein de l'ensemble de l'espèce *bos taurus*. Dans les populations rouges nordiques, il a été démontré que la mutualisation des populations de référence génotypés 50K permettait déjà un fort gain de précision des évaluations génomiques (Brondum et al., 2011). La situation des races nordiques s'apparente à celle d'une population unique : certains taureaux sont utilisés dans plusieurs races ce qui assure la connexion entre les populations, les populations sont évaluées conjointement ce qui assure la cohérence des phénotypes entre populations.

Le cas des races Simmental Française et Montbéliarde se rapproche de la situation des races nordiques. Les deux races appartiennent au même groupe de races bovines et participent séparément à l'évaluation génétique internationale des races « *Fleckvieh-Simmental* ». En revanche, l'échange de taureaux entre les deux populations est inexistant. Il est cependant possible qu'une connexion entre les deux populations subsiste à travers des échanges de chacune des deux races avec les autres populations européennes (Fouilloux et al., 2006). L'utilisation de la population de référence de grande taille disponible en race Montbéliarde pour le développement d'une évaluation génomique Simmental basée sur la puce 50K apparait donc comme une approche intéressante.

Plusieurs stratégies ont été comparées ici : une évaluation polygénique en race Simmental seule, une évaluation génomique où les effets des marqueurs sont estimés en race Simmental, une estimation des effets des marqueurs en race Montbéliarde, une présélection des marqueurs en race Montbéliarde suivie d'une estimation des effets en race Simmental, et enfin une estimation des effets des marqueurs à partir d'une population multiraciale (Simmental et Montbéliarde).

B. Article III : Evaluations génomiques à partir d'une population de référence multiraciale composée d'animaux des races Montbéliarde et Simmental

Hozé C, Fritz S, Phocas F, Boichard D, Ducrocq V, Croiseau P. 2014,

Genomic evaluation using combined reference populations from Montbéliarde and French Simmental breeds

In proceedings of the 10th World Congr. Genet. Appl. Livest. , accepté

Proceedings, 10th World Congress of Genetics Applied to Livestock Production

Genomic evaluation using combined reference populations from Montbéliarde and French Simmental breeds

C. Hozé[*,†], S. Fritz[*,†], F. Phocas[*], D. Boichard[*], V. Ducrocq[*] and P. Croiseau[*].
[*]INRA UMR1313 GABI, Jouy en Josas, France, [†]UNCEIA, Paris, France

ABSTRACT

The French-Simmental and Montbéliarde breeds are currently evaluated jointly for production traits in international evaluations. Therefore, we investigated the feasibility of using the large reference population available in Montbéliarde breed to implement a genomic evaluation in French-Simmental. Data consisted in 229 Simmental and 1,758 Montbéliarde progeny tested bulls genotyped for a common set of 37,364 SNP. Genomic evaluations were performed using GBLUP and BayesCπ. SNP effects were either estimated based on Montbéliarde, Simmental or a joint reference population. Montbéliarde population was also used to preselect SNP before Simmental evaluation. Accuracy of evaluation was assessed based on the youngest Simmental bulls. Average correlations were 0.28 with SNP effects estimated in Montbéliarde, 0.39 with effects estimated in Simmental, 0.42 after pre-selection, and 0.48 with multi-breed reference population. This study demonstrates that the Montbéliarde reference population is beneficial to implement genomic evaluation in Simmental.
Keywords:
dairy cattle
genomic selection
multi-breed evaluation

Introduction

In France, a very large number of animals from the main dairy breeds have been genotyped with the Illumina Bovine SNP50 BeadChip® (50K) for genomic selection (GS) and it is now possible to predict breeding values of animals at birth with high accuracy in Holstein, Montbéliarde and Normande. For the French Simmental breed and more generally for all regional breeds, a limited number of progeny-tested bulls are available and assembling a large enough reference population is a real challenge.
One strategy proposed to implement genomic evaluations in small breeds was the use of multi-breed reference population (Hayes et al., 2009, Erbe et al., 2012). Some increase in correlation was observed when combining Holstein and Jersey populations (Erbe et al., 2012) and major French breeds (Hozé et al., 2014). However, the gain in reliability was limited, depended on the trait considered and on the size of reference populations (Hozé et al 2014).

Larger increases with multi-breed reference populations were observed when populations were closely related (Brondum et al., 2011) or when a denser chip was used (Hozé et al., 2014). One possible explanation is the better conservation of linkage disequilibrium (LD) in breeds with common origin whereas conserved LD is only observed at short distance in unrelated breeds.

In France, the French Simmental breed has a limited population and therefore cannot benefit alone from efficient genomic evaluations. In contrast, the Montbéliarde breed, which also belongs to the Simmental and Fleckvieh breed group, is a national breed and has more than 1,700 progeny-tested bulls genotyped. The two breeds take part separately to the international genetic evaluations for production traits implemented since 1996 and involving ten Simmental European populations. Such an evaluation leads to estimated breeding values on the French Simmental scale for Montbéliarde bulls. Accordingly, we can envisage using Montbéliarde bulls for French Simmental genomic evaluations.
The aim of this paper is to assess the accuracy of predictions of genomic evaluation in French Simmental using four different strategies involving Montbéliarde and Simmental reference populations separately or jointly.

Materials and Methods

Data. The French Simmental reference population consisted of 229 progeny tested-bulls genotyped on the BovineHD BeadChip® (HD). Among them, 123 were German bulls selected on their marginal contribution to the French Simmental population and whose genotype was exchanged with LfL. In the Montbéliarde breed, 1,758 progeny tested bulls genotyped on the BovineSNP50 BeadChip® (50K) were available. Most of them originate from the French genomic evaluations. A quality control procedure described in Hozé et al. (2014) was first performed on both datasets. Then, markers present on both 50K and HD chip were extracted leading to a common set of 37,364 SNP for the two breeds.
The phenotypes used were deregressed proofs (DP) and associated equivalent daughter contribution (EDC) from Interbull evaluation expressed on the French Simmental scale. The traits studied here were milk yield, fat yield, protein yield and somatic cell score (SCS).

Evaluation model. Breeding values (EBV) were estimated using three approaches : a pedigree-based BLUP, GBLUP (VanRaden, 2008) and BayesCπ approaches (Habier et al., 2010). Evaluations were performed using GS3 software (Legarra et al., 2013). BayesCπ was used for single-breed and multi-breed genomic evaluation. For this latter case, a breed effect was included in the model to account for the genetic mean of the breeds and avoid selecting SNP to estimate this difference.

Four scenarios were compared in this study: a) SNP effects were estimated in the Simmental training population and then applied to the Simmental validation population; b) SNP effects were estimated in Montbéliarde and applied to the Simmental validation population; c) SNP effects were estimated in the joint reference population of Montbéliarde and Simmental bulls; d) a preselection of SNP was performed with the Montbéliarde data and SNP effects were estimated in Simmental breed.
The latter approach, derived from the study of Croiseau et al. (2011) allows a reduction of the number of effects to estimate in the Simmental breed. SNP pre-selection was based on results of a QTL detection in the Montbéliarde breed. QTL detection was based on a Linkage Disequilibrium Linkage Analysis (LDLA, Druet et al., 2008) study. Phenotypes used were daughter yield deviation of the national Montbéliarde evaluation as they were assumed to be more reliable for the Montbéliarde breed, than Montbéliarde deregressed proofs expressed on the Simmental scale. First, we defined a QTL as the SNP with a LRT higher than 5 and with the highest value within a window of 50 SNP. Then, we retained this SNP and the 50 SNP around the LRT peak for the Simmental analysis.

Accuracy of evaluation. The accuracy of evaluation was assessed using a validation study. The 46 youngest bulls of the Simmental reference population formed the validation set. Since phenotypes used here were deregressed proof expressed on the Simmental scale, we did not assess the accuracy of evaluations in Montbéliarde. For animals in the validation set, phenotypes were erased to mimic candidates to selection. SNP effects were estimated on the training population and then applied to the validation population to compute GEBV (Genomic Enhanced Breeding Value). Then, the accuracy of genomic evaluation was measured through the weighted correlation between GEBV and DP of validation bulls.
Among the 46 bulls in the validation population, 28 had their sire in the reference population and 18 did not. We also computed the correlation separately for the two subpopulations to assess the impact of connectedness to reference population on the accuracy of genomic evaluation.

Results and Discussion

French Simmental breed considered alone. Accuracies of pedigree-based BLUP, GBLUP and BayesCπ were computed for the four traits using Simmental reference population alone. Results are presented in Table 1.
It can be observed that even for a breed with fewer than 200 animals in the reference population, using genomic information allowed a gain in correlation of 0.06 with GBLUP and 0.08 with BayesCπ. Increases in correlation were trait dependent: a large gain was observed for protein yield (+0.23) while a decrease (-0.04) was observed for fat yield.
Whatever the method, correlations were clearly lower for somatic cell score (SCS) than for other traits. SCS is also the only traits where GBLUP performed better than BayesCπ. This is at least partially due to the lower heritability and a more polygenic genetic determinism of this trait.
When computing correlations for bulls that did or did not have their sire in the training population, we observed a mean difference between the two groups of bulls of 0.18. Though the limited number of individuals and probably large sampling errors associated with these values, part of the difference between groups could be explained by the lack of polygenic component in our model and spurious association between SNP and causative mutations due to pedigree relationship (Solberg et al., 2009, Liu et al., 2011).

Using Montbéliarde breed to estimate SNP effect. SNP effects estimated in the Montbéliarde breed were used to predict GEBV in Simmental. Results are presented in Table 2. The mean correlation achieved with this strategy (0.28) was lower than the one achieved with pedigree-based BLUP (0.32) and with effects estimated in Simmental for bulls without their sire in the reference population (0.31). The Montbéliarde breed has been highly selected for milk production, therefore large QTL influencing production traits which may have been fixed in this breed may still have an influence in Simmental. For SCS, which has been less selected than production, estimating SNP effects in Montbéliarde improved accuracy compared to the single-breed situation.

Using the Montbéliarde breed to preselect SNP. When the number of effects to estimate was reduced by preselection based on Montbéliarde data, correlations achieved were up to 0.06 higher than without preselection (Table 2). However, for somatic cell score the preselection led to a decrease in correlation (-0.03). This result could be explained by the relatively low number of SNP preselected for this trait compared to production traits (Table 2).

Pooling Montbéliarde and Simmental reference populations. Combining Simmental and Montbéliarde reference populations led to a large increase in accuracy compared to single-breed genomic evaluation (Table 3). The gain in correlation was on average 0.09 and ranged from 0.05 to 0.13 depending on the trait. Larger increases in correlation were observed for SCS and protein yield. It is consistent with the relatively high correlation observed for these two traits using SNP effects estimated in Montbéliarde breed.

The benefit of multi-breed genomic evaluation was higher for bulls that did have their sire in reference population (+0.14) than for those that did not (+0.05). The larger reference population in the multi-breed GS allows a better estimation of SNP effect and avoids detecting SNP associated with pedigree relationship (Hozé et al., 2014).

Conclusion

The Montbéliarde and French Simmental breeds belong to the same breed family and are jointly evaluated in international evaluations. Despite the lack of connectedness observed between the two breeds (Fouilloux et al., 2006), using both Montbéliarde and French Simmental reference populations for the French Simmental genomic selection allowed a 0.09 increase in the accuracy of prediction compared to the use of the French Simmental breed alone. This study was performed with a small validation population and results need to be confirmed on a larger population. However, accuracies achieved with multi-breed GS are promising and encourage the development of a joint routine genomic evaluation.

A noticeable limit for this implementation is the use of deregressed proofs from Interbull evaluation as phenotypes. Indeed, up to now, production traits are the only traits evaluated by Interbull for Simmental. The extension of international evaluations to other traits, in particular to type traits which has been shown to be feasible (Regaldo et al., 2006), is required for the development of genomic selection in French Simmental. An alternative is to use daughter yield deviation from national evaluation as phenotypes for genomic evaluation but its consequences need to be investigated further.

Acknowledgment

GEMBAL project is funded by the Agence Nationale de la Recherche (ANR-10-GENM-0014), APISGENE, Races de France and INRA "AIP Bioressources". Montbéliarde genotypes originated from the Cartofine-ANR-05-GENANIMAL-007 project funded by ANR (French National Research Agency) and ApisGene, and from genomic selection activity generated by the French cattle breeding companies, with LABOGENA as main genotyping lab. The Bayerische Landesanstalt für Landwirtschaft (LfL) 123 Simmental genotypes.

Table 1: Correlations between deregressed proofs and estimated breeding values for the four traits using pedigree-based BLUP, GBLUP and BayesCπ based on a training population of 181 Simmental bulls

	Milk	Fat Yield	Protein Yield	SCS[1]	**Mean**
BLUP	0.46	0.41	0.31	0.12	**0.33**
GBLUP	0.47	0.34	0.54	0.21	**0.39**
BayesCπ	0.48	0.36	0.54	0.17	**0.39**

[1] Somatic Cell Score

Table 2: Correlations between deregressed proofs and estimated breeding values for the four traits using BayesCπ with SNP effects estimated in Montbéliarde breed or SNP effects estimated in the French Simmental breed after a SNP pre-selection in the Montbéliarde breed

	Milk	Fat Yield	Protein Yield	SCS	**Mean**
EstMO[1]	0.28	0.36	0.28	0.19	**0.28**
PreselMO[2]	0.54	0.38	0.58	0.16	**0.42**
Nb SNP[3]	11,627	12,690	12,698	7,927	**11,235**

[1] SNP effect estimated inMontbéliarde breed
[2] SNP preselected in Montbéliarde breed
[3]Number of SNP conserved after preselection

Table 3: Correlations between deregressed proofs and estimated breeding values for the four traits using BayesCπ and a joint Montbéliade/Simmental reference population

	Milk	Fat Yield	Protein Yield	SCS	**Mean**
BayesCπ	0.53	0.49	0.61	0.27	**0.48**

Literature cited

Brondum, R. F., Rius-Vilarrasa, E.,. Stranden, I. et al.. (2011). J Dairy Sci 94:4700-4707.

Croiseau, P., Legarra, A., Guillaume, F. et al.. (2011). Genet Res 93:409-417.

Druet, T., Fritz, S., Boussaha, M. et al. (2008). Genetics 178:2227-2235.

Erbe, M., Hayes, B. J., Matukumalli L. K. et al. (2012). J. Dairy Sci 95:4114-4129.

Fouilloux, M.N., Minery, S., Mattalia, S., and Laloë D. (2006). Interbull Bulletin 35: 129-135

Habier, D., Fernando, R. L., Kizilkaya, K., and Garrick, D. J.. (2010). BMC bioinformatics 12:186.

Hayes, B., Bowman, P., Chamberlain, A. et al. (2009). Genet Sel Evol 41:51.

Hozé, C., Fritz, S., Phocas, F. et al. (2014) J Dairy Sci. (in press)

Legarra, A., Ricard, A., and O. Filangi (2013). http://snp.toulouse.inra.fr/~alegarra.

Liu, Z., Seefried, F. R., Reinhardt, F. et al. (2011). Genet Sel Evol 43(19):9.

Regaldo, D., Minery, S., and Mattalia S. (2006). Interbull Bulletin 35: 136-140

Solberg, T. R., Sonesson, A.K., Woolliams, J.A. et al. (2009). Genet Sel Evol 41:53.

VanRaden P. M., (2008). J Dairy Sci 91: 4414-4423

C. Bilan

Cette étude démontre qu'il est possible de mettre en place une évaluation génomique en race Simmental en utilisant la population de référence Montbéliarde.

La distance génétique relativement faible entre les populations Montbéliarde et Simmental permet un fort gain de précision grâce aux évaluations multiraciales à partir de génotypes moyenne densité. Le gain observé par rapport aux évaluations intra-races (+ 0,09) est trois fois supérieur à ce qui avait été observé dans l'étude précédente pour une population théorique normande de 200 animaux. Il est donc possible, même à partir de génotypes de densité moyenne d'utiliser une évaluation génomique multiraciale. Cette stratégie nécessite toutefois que les deux races soient suffisamment proches pour que le déséquilibre de liaison soit conservé entre race malgré la relativement faible densité en marqueurs.

Cette étude se concentre sur les caractères de production qui sont actuellement les seuls caractères à être évalués par Interbull pour les races « *Fleckvieh-Simmental* ». La stratégie proposée ici ne pourra donc pas être directement transposable à l'ensemble des caractères. Une stratégie alternative est l'utilisation de phénotypes issus de l'évaluation génétique nationale (après transformation ou non) pour estimer l'effet des marqueurs sur une population multiraciale.

Il est également possible à partir des informations nationales de présélectionner des marqueurs en race Montbéliarde et d'estimer leurs effets en race Simmental. Cette stratégie, moins efficace que la mutualisation des populations de référence, devrait permettre de limiter l'impact d'une éventuelle hétérogénéité de caractères entre les deux races tout en réduisant le temps de calcul nécessaire aux évaluations.

VI. Discussion et perspectives

A. Bilan et limites des études

Les différentes études réalisées ont mis en évidence qu'il était possible d'augmenter la précision des évaluations génomiques, pour les races disposant d'une petite taille de population de référence, grâce aux évaluations génomiques multiraciales.

1. Imputation des génotypes haute-densité

Dans la plupart des cas, l'évaluation multiraciale nécessite l'utilisation de génotypes HD pour tirer profit de la meilleure conservation du déséquilibre de liaison à courte distance et de la meilleure cohérence de phases entre races (Larmer et al., 2014). Afin de limiter les coûts de génotypages, il est envisagé de génotyper les candidats à la sélection sur la puce 50K et d'imputer leur génotypes haute densité.

Nous avons pu démontrer ici qu'une population de 200 à 300 individus génotypés en haute densité permettait d'atteindre, par la suite, une efficacité d'imputation supérieure à 99%. Cette taille de population est atteinte dans la plupart des races bovines françaises du projet GEMBAL ce qui autorise l'imputation de génotypes haute densité à partir de génotypes 50K. Les races pour lesquelles cette taille de population, et donc cette efficacité d'imputation, ne sont pas atteintes, devront réaliser des génotypages HD complémentaires pour augmenter la qualité de l'imputation. Il est également envisageable pour les races internationales comme la race Brune et la race Simmental de procéder à des échanges internationaux pour augmenter la taille de la population d'imputation à moindre coût.

Il est possible que les taux d'erreur estimés ici soient légèrement surestimés puisque seuls 80% des génotypes HD (à cause du retrait de 20% des génotypes pour la création de la population de validation) ont été utilisés pour l'imputation. En revanche, l'étude du taux d'erreur d'imputation a été réalisée dans une situation particulière où de forts contributeurs à la race ainsi que 2 à 4 de leurs descendants ont été génotypés. Il est donc vraisemblable qu'à terme l'efficacité d'imputation diminue si la population d'imputation n'est pas régulièrement renouvelée.

Enfin, il faut préciser que l'imputation en routine des génotypes HD peut s'avérer difficile à mettre en œuvre. Si le nombre de génotypes 50K à imputer est important, la demande en ressources informatiques peut être élevée en particulier si l'imputation est réalisée avec le logiciel Beagle (Browning et Browning, 2007). Dans le cadre de l'imputation hebdomadaire de génotypes 50K à partir de génotypes basse densité (réalisée dans le cadre des évaluations génomiques des trois grandes races laitières en France), ce problème a été contourné grâce à l'utilisation du logiciel DAGPHASE (Druet et Georges, 2010). Ce logiciel bien qu'intéressant pour l'imputation des génotypes d'une population présentant une forte structure familiale s'est révélé peu efficace dans le cadre des données GEMBAL (non publié). Ce constat avait déjà été réalisé dans le cadre de l'imputation de génotypes HD à partir des données EuroGenomics (Schrooten et al., 2014). L'utilisation d'autres logiciels d'imputation comme FindHap (VanRaden et al., 2011) ou FImpute (Sargolzaei et al., 2011) pourrait permettre de réduire la demande en ressources informatiques en maintenant (ou augmentant légèrement) le taux d'erreur d'imputation (Larmer et al., 2014 ; Pausch et al., 2013 ; Sun et al., 2012). Il faudra réaliser de nouvelles études d'imputation sur l'ensemble des races du projet en utilisant ces logiciels avant d'envisager l'utilisation en routine de génotypes haute densité. Il est également envisageable, pour limiter les temps de calcul, d'utiliser les populations de référence HD déjà imputés des races nationales et d'imputer uniquement les génotypes d'animaux des races régionales.

2. Stratégie d'évaluations génomiques en fonction de la taille de population

Afin de disposer d'une population de validation de taille relativement élevée (>200 individus), l'étude cherchant à déterminer quelle sélection génomique (intra-race, multi-race, basée sur génotypes 50K ou HD) permettrait d'obtenir la qualité de prédiction la plus élevée dans le cas des races régionales a été réalisée en se servant des populations des trois grandes races laitières. Nous avons vu, en introduction, que l'efficacité des évaluations génomiques pouvaient dépendre de la taille efficace de la population et des caractères étudiés (Daetwyler et al., 2010 ; Ducrocq et al., 2014). Il est possible qu'une étude utilisant les races régionales conduise à des résultats légèrement différents de ceux observés sur les races nationales en raison de la différence de taille efficace entre populations. Les résultats obtenus lors ce travail pourraient également varier en fonction du type de caractères étudiés.

Dans le cadre des évaluations multiraciales, le type de caractères étudiés peut avoir une influence forte sur la précision des évaluations. Selon les objectifs de sélection de chaque race, il est possible que certains QTL soient fixés dans une race et pas dans une autre. D'autre part, la définition du caractère, son héritabilité et sa mesure peuvent varier d'une race à l'autre, ce qui pourrait entrainer une hétérogénéité sous-jacente et limiter l'efficacité des évaluations multiraciales. Dans le cas des caractères laitiers, pour lesquels le phénotypage et l'héritabilité sont relativement homogènes entre races, nous avons observé qu'une standardisation des phénotypes permettait d'accroitre la précision des évaluations génomiques multiraciales (Hozé et al. 2014a). Dans le cas de caractères plus hétérogènes entre races, comme par exemple certains caractères de morphologie, il est probable qu'un réel travail d'homogénéisation des phénotypes (définition, mesure) soit nécessaire avant de pouvoir envisager une évaluation multiraciale.

Les résultats obtenus dans le cadre de cette thèse devront donc être confirmés dans le cas d'autres races et d'autres caractères.

3. Evaluations génomiques multiraciales dans le cas de deux races proches

Une meilleure précision des évaluations multiraciales a été observée dans le cas d'une population composée d'animaux Simmental et Montbéliard que dans le cas d'une population composée d'animaux Normand, Montbéliard et Holstein. L'ajout de 1750 taureaux Montbéliard à une population de référence de 200 Simmental a permis un gain de corrélation de 0,09 (Hozé et al., 2014b). L'ajout des mêmes 1750 taureaux Montbéliard et de 5000 taureaux Holstein à une population de 200 taureaux Normand n'a permis quant à lui qu'un gain de précision de 0,03 (Hozé et al. 2014a). Dans le cas d'évaluations intra-races, une population de référence génétiquement proche des candidats à la sélection permet d'améliorer la précision des évaluations génomiques (Clark et al., 2012 ; Pszczola et al., 2012). Il est probable que la distance génétique entre les races des animaux de la population de référence influence également la précision des évaluations. Une étude réalisée sur les données de six populations ovines génotypées sur la puce OvineSNP50 Beadchip© confirme cette hypothèse. Il a en effet été démontré que le regroupement des populations de deux races proches permettait d'atteindre la même précision d'évaluation que lorsque les animaux de l'ensemble des races étaient regroupés (Legarra et al., 2014).

Il semble donc plus intéressant de réaliser des évaluations multiraciales à partir de sous-groupes de races génétiquement proches plutôt que de combiner les populations de référence de l'ensemble des races. Cette stratégie peut également permettre de limiter les besoins en ressources informatiques grâce à une taille de population de référence totale réduite et l'utilisation d'une puce 50K. Le choix des races à regrouper au sein d'une même évaluation peut être réalisé à partir d'études de génétique des populations comme celle présentée sur la Figure 11. Le regroupement des populations Holstein et Pie Rouge des Plaines déjà réalisé dans les évaluations génomiques françaises est justifié par la faible distance génétique observée entre ces races. On retrouve également le fait que les populations Simmental et Montbéliarde sont beaucoup plus proches l'une de l'autre que les populations Holstein, Montbéliarde et Normande ce qui explique la différence de gain de précision permis par les évaluations multiraciales. Enfin, l'ensemble des races des montagnes (Vosgienne, Tarentaise, Abondance, Simmental et Montbéliarde) semble former un groupe relativement cohérent, ce qui encourage la mise en place d'une évaluation multiraciale dans ces populations.

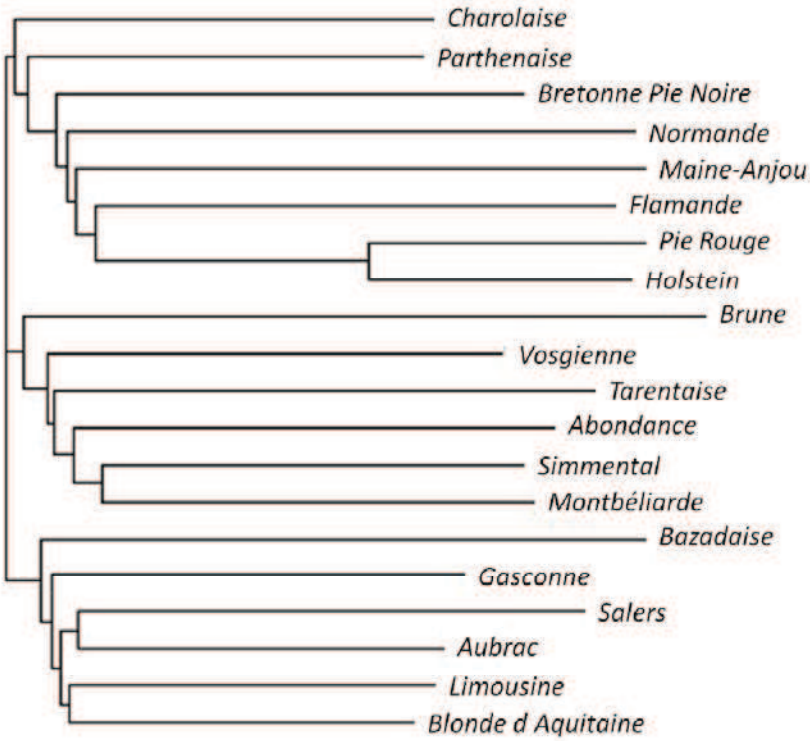


Figure 11 : Dendrogramme tracé à partir des distances génétiques (estimées à partir des génotypes haute densité) entre les populations du projet GEMBAL, d'après D. Laloë, INRA, Jouy en Josas (communication personnelle)

B. Perspectives de développement d'évaluations génomiques multiraciales

D'autres stratégies peuvent également permettre d'améliorer la précision des évaluations. Les travaux réalisés dans le cadre de cette thèse ont montré l'intérêt de la mise en place d'une évaluation génomique multiraciale pour les races laitières. En revanche, le gain observé (Erbe et al., 2012 ; Hozé et al., 2014a) est plus faible que ce qui avait été estimé sur données simulées (De Roos et al., 2009). Les hypothèses faites lors de la simulation et en particulier la présence ou non de QTL communs entre races et la conservation totale ou partielle du déséquilibre de liaison entre populations peuvent avoir une grande influence sur les capacités prédictives d'une évaluation multiraciale.

1. Conservation du déséquilibre de liaison et présence de QTL communs entre races

Larmer et al. (2014) a démontré que le niveau de déséquilibre de liaison moyen entre SNP consécutifs était deux à trois fois plus élevé sur la puce haute densité (entre 0,58 et 0,63) que sur la puce 50K (0,21 à 0,27). La cohérence entre les phases était également plus élevée sur la puce HD que sur la puce 50K (Larmer et al., 2014). Nous avons également observé sur le mélange des 11 populations bovines laitières du projet GEMBAL, que plus de 50% des couples de marqueurs distants de 4kb présentait un déséquilibre de liaison supérieur à 0,5 (Hozé et al., 2012). Ce même pourcentage n'était que de 10% pour une distance de 70kb qui correspond à deux marqueurs adjacents de la puce 50K. L'utilisation d'une puce haute densité permet donc une meilleure conservation du déséquilibre de liaison ainsi qu'une meilleure cohérence des phases et semble autoriser le regroupement des populations de référence entre races.

Cependant, indépendamment de la cohérence des phases et de la conservation du déséquilibre de liaison, une évaluation génomique multiraciale nécessite l'existence de QTL communs entre races.

Les premières études de détection de QTL influençant les caractères laitiers ont montré une faible concordance (moins de 30%) entre les QTL détectés chez les trois grandes races laitières (Guillaume, 2009). Une étude des régions chromosomiques influençant la qualité de la viande a également révélé un nombre limité de QTL communs entre races (Allais, 2011). Il est toutefois impossible de déterminer si la non-détection d'un QTL est liée à son absence au sein d'une race ou à un manque de puissance du dispositif. La puissance d'un dispositif de détection de QTL est fonction du déséquilibre de liaison entre le marqueur et le QTL, de l'effet du QTL sur le caractère, de la fréquence des allèles aux QTL et du nombre d'animaux de la population de référence (Goddard et Hayes, 2009). Il est donc possible qu'à effet et déséquilibre de liaison constants, un QTL soit détecté dans une race et non dans l'autre à cause d'une faible fréquence allélique et/ou d'une taille de population de référence limitée.

2. Développement de nouvelles approches d'évaluations génomiques multiraciales

Face au manque de précision des méthodes classiques d'évaluations génomiques utilisant une population de référence multiraciale, plusieurs autres stratégies de sélection génomique multiraciale ont été proposées. Deux grands types de stratégies coexistent : celles supposant que les QTL et leurs effets sont communs entre races et celles supposant qu'il existe des QTL spécifiques à chaque race.

a) Approches supposant que l'ensemble des QTL est commun entre races

Les méthodes supposant que l'ensemble des QTL est commun entre races reposent sur le regroupement des populations de référence pour l'estimation d'un effet marqueur unique pour l'ensemble des animaux. On cherche alors, pour augmenter l'efficacité des évaluations génomiques, à augmenter le déséquilibre de liaison entre le marqueur dont on cherche à estimer l'effet et le QTL ainsi qu'à améliorer la conservation du lien entre QTL et marqueurs entre races.

Nous avons vu que l'utilisation d'haplotypes permettait d'augmenter la qualité de prédiction dans le cas d'une évaluation génomique intra-race (Calus et al., 2008 ; Villumsen et al., 2009). Le même bénéfice pourrait être observé dans le cas d'évaluations multiraciales dont l'efficacité est très dépendante du déséquilibre de liaison moyen entre marqueurs successifs (Erbe et al., 2012 ; Hozé et al., 2014a). Afin de vérifier cette hypothèse, le développement et le test d'un logiciel permettant d'estimer à partir d'une approche bayésienne (BayesCπ) les effets haplotypiques font l'objet de travaux en cours (Croiseau et al., 2014).

L'augmentation de la densité en marqueurs et en particulier l'utilisation de données de séquences complètes pourrait également permettre d'améliorer l'efficacité des évaluations génomiques multiraciales. Van den Berg et al. (2014) ont montré à partir d'une population Normande-Jersiaise, que le coefficient de régression entre parentés génomiques calculées aux marqueurs et celles calculées aux mutations causales simulées était de 0,14 en utilisant des marqueurs distants de 25Kb des mutations causales mais de 0,36 avec des marqueurs distants de 1Kb de la mutation causale. Dans le cas de données de séquences complètes, la mutation causale est dans le jeu de données, ce qui lève la contrainte de la conservation du déséquilibre de liaison entre races et pourrait augmenter la précision des évaluations génomiques multiraciales. Cependant, la mise en place d'évaluations génomiques basées sur une puce haute densité entraine déjà de nombreuses difficultés en termes de besoins informatiques, celles-ci seront amplifiées avec l'utilisation de données de séquences complètes. Elles nécessiteront également un fort investissement financier lié au re-séquençage d'un nombre important d'animaux issus de plusieurs races. L'évaluation génomique multiraciale utilisant des données de séquences n'est donc pas envisageable à court terme.

b) Approches supposant qu'il existe une variabilité des QTL entre races

La faible précision des valeurs génétiques obtenues dans une race lors de l'application d'effets de marqueurs estimés dans une autre race laisse à penser qu'il existe des différences entre les QTL des différentes races (Brondum et al., 2011 ; Hozé et al., 2014b ; Olson et al., 2012). Le développement de méthodes permettant de prendre en compte cette variabilité de QTL reste un défi. Des modèles multi-caractères considérant les phénotypes de chacune des races comme des caractères corrélés ont été proposés pour prendre en compte l'interaction race-QTL (Karoui et al., 2012 ; Makgahlela et al., 2013 ; Olson et al., 2012). Des approches en deux étapes ont également été suggérées. Elles permettent d'utiliser les résultats d'estimation des effets des marqueurs ou de détection des effets QTL obtenus dans une race soit directement pour présélectionner les marqueurs à utiliser dans une autre race (Hozé et al., 2014b), ou bien indirectement comme une information *a priori* pour l'estimation des effets des marqueurs dans une autre race (Brondum et al., 2012). Ces différentes approches n'ont toutefois pas permis d'augmenter la précision des valeurs génétiques estimées par rapport à une situation où les populations de référence des différentes races sont regroupées et/ou celle ou une évaluation intra-race est utilisée.

L'introduction de sous-groupes de QTL (ceux communs à l'ensemble des races et ceux spécifiques à une race) au sein de l'approche SAMg actuellement utilisée dans les évaluations génomiques françaises est aussi envisageable. Cette approche peut s'avérer particulièrement intéressante dans le cas d'évaluations génomiques multiraciales puisqu'elle utilise des haplotypes et est moins coûteuse en ressources informatiques que les méthodes de sélection génomique classique. Cette stratégie nécessite cependant de pouvoir détecter et distinguer les QTL spécifiques et communs à chacune des races. Cette détection de QTL peut être réalisée avec des approches LDLA, Elastic Net ou BayesCπ (sur marqueurs ou sur haplotypes) qui ont déjà fait leur preuve intra-race (Colombani et al., 2011 ; Druet et al., 2008 ; Van den Berg et al., 2013). A moyen terme, l'utilisation de données de séquences pourrait elle aussi permettre d'augmenter la puissance du dispositif de détection de QTL et en particulier d'améliorer la détection des QTL dont la fréquence est faible dans la population (Druet et al., 2013).

C. Perspectives de développement d'évaluations génomiques intra-races

Une approche multiraciale permet d'améliorer la précision des évaluations génomiques mais nécessite des développements méthodologiques pour une bonne harmonisation des caractères et une meilleure prise en compte de la variabilité des QTL entre races. Nos travaux ont montré que la précision d'une évaluation génomique intra-race réalisée à partir d'une population de référence de 200 à 500 animaux pourrait être correcte. Il est donc possible de mettre en place dans un premier temps, une sélection génomique intra-race utilisant des méthodes d'évaluation génomique classiques. On peut également imaginer des approches spécifiques, plus adaptées à la taille de la population de référence comme une sélection assistée par marqueurs ou sur mutations, l'inclusion de femelles dans la population de référence, ou une méthode permettant de prendre en compte l'ensemble des individus (génotypés ou non) avec performances. Ce chapitre présente les avantages et inconvénients de chacune de ses alternatives.

1. Une sélection assistée par marqueurs ou sur mutations

La sélection assistée par marqueurs utilisée en France permet de fortement réduire le nombre d'effets à estimer ce qui peut être avantageux dans les cas où le nombre d'animaux phénotypés et génotypés est limité. La mise en place d'un programme de sélection assistée par marqueurs dans les races laitières régionales semble donc intéressante.

Deux stratégies sont possibles : soit utiliser une approche de type SAM2 (Fritz et al., 2008) intégrant par exemple une quarantaine de QTL par caractère détecté par LDLA, soit une approche de type SAMg (Boichard et al., 2012b) intégrant des centaines de QTL par caractère détectés grâce à une méthode de sélection génomique (Elastic Net, BayesCπ, etc.). Il est également envisageable d'utiliser le LDLA pour présélectionner les SNP avant l'étape de sélection génomique dans le but d'orienter le choix des SNP et d'augmenter la qualité d'estimation de leurs effets en limitant le nombre de QTL retenus pour la SAM (Croiseau et al., 2011).

A moyen terme, l'utilisation de données de séquences complètes et l'identification d'un nombre croissant de mutations causales peuvent également permettre la mise en place d'une sélection sur mutations. Une fois une mutation causale identifiée dans une race disposant d'une grande population de référence, Il est possible d'estimer son effet pour des races à plus petits effectifs. Cette stratégie est facilitée par le développement de puces basse densité personnalisables dites « *custom* » permettant l'ajout de marqueurs spécifiques. Une fois une mutation causale (ou supposée causale) identifiée, il est possible de l'ajouter sur cette puce afin de disposer d'un génotypage exact à la mutation, valider son existence à moindre coût, puis l'inclure dans une sélection assistée sur mutations pour les races régionales. Cette stratégie suppose qu'un QTL commun à plusieurs races soit associé à une même mutation causale ce qui n'est pas toujours le cas. Ainsi, Gautier et al. (2007b) ont par exemple montré qu'il existait plusieurs polymorphismes du gène DGAT1 influençant le taux butyreux chez les bovins laitiers.

2. Une sélection génomique utilisant uniquement les mâles génotypés

Dans le cas des populations de référence Normande ou Simmental composées respectivement de 194 et 183 animaux, la mise en place d'évaluations génomiques à partir d'un BayesCπ intra-race a permis un gain de corrélation de 0,07 et 0,06 par rapport à un modèle polygénique (Hozé et al. 2014a ; 2014b). Ce gain était similaire en utilisant la méthode du GBLUP et significativement inférieur pour les animaux dont le père était absent de la population de référence. Dans les populations laitières, le nombre de pères de candidats à la sélection est généralement très réduit. Il est donc possible de mettre en place une première approche de sélection génomique qui imposerait le génotypage du père d'un animal pour officialiser la publication d'un index.

3. Une sélection génomique utilisant à la fois les mâles et les femelles génotypés

Le facteur limitant la précision des évaluations génomiques des races régionales est la taille de la population de référence et donc le nombre de taureaux testés sur descendance. Or, plusieurs centaines à plusieurs milliers de femelles avec performances sont disponibles et pourraient être génotypées afin de compléter les populations de référence des races régionales. L'inclusion de ces femelles dans les populations de référence peut permettre une (forte) augmentation de la taille de la population de référence et de la précision des évaluations génomiques.

Dans les races nationales, l'inclusion de femelles dans les populations de référence n'a pas encore été mise en pratique car les femelles disposent de phénotypes nettement moins précis (très généralement uniquement des performances propres) que les taureaux confirmés. Cette plus faible précision des phénotypes nécessite de disposer d'un plus grand nombre d'animaux pour atteindre la même précision d'évaluation (Figure6). Calus et al. (2013b) ont ainsi démontré que la précision obtenue à partir d'une population de référence de 1600 femelles n'était que légèrement supérieure à celle obtenue à partir d'une population de 300 taureaux. Le gain de précision permis par l'inclusion de femelles est donc limité lorsqu'un relativement grand nombre de mâles génotypés est déjà disponible. Il peut cependant être plus élevé lorsque le nombre de taureaux testés sur descendance est faible.

Ainsi, alors qu'ajouter 10 000 femelles à une population de référence de 3 000 mâles ne permet qu'un gain de précision de 4 à 8% (Pryce et al., 2012), ajouter 1609 femelles à une population de 296 taureaux permet un gain de précision de 45% (Calus et al., 2013b). Le même constat a été réalisé sur données caprines où 1985 chèvres ont été ajoutées à une population de référence de 67 ou 677 boucs (Carillier et al., 2013). Dans cette étude, Carillier et al. (2013) ont également observé que l'ajout de femelles permettait d'augmenter fortement la précision des valeurs génétiques estimées de leurs demi-frères. Bapst et al. (2013) ont quant à eux démontré que l'inclusion de femelles était particulièrement utile lorsque les femelles étaient peu apparentées aux taureaux déjà présents dans la population de référence.

Dans le cas des races régionales, l'inclusion de femelles dans la population de référence devrait donc permettre une augmentation significative de l'efficacité de la sélection génomique. Elle peut, en particulier, améliorer la prédiction des valeurs génétiques des animaux peu apparentés à la population de référence et dont le père n'est pas génotypé.

L'inclusion de femelles dans la population de référence présente cependant un risque lié au traitement préférentiel. Le traitement préférentiel se définit comme les pratiques d'élevage qui modifient la production laitière et biaisent les estimations des valeurs génétiques d'une ou plusieurs vaches du troupeau (Kuhn et al., 1994). Le plus souvent ces pratiques se traduisent sur le terrain par des animaux (généralement les meilleures femelles) bénéficiant d'un meilleur logement et/ou d'une meilleure alimentation que le reste du troupeau. La prise en compte de phénotypes biaisés dans les évaluations génomiques pourrait donc introduire une hétérogénéité dans la qualité des phénotypes analysés plutôt qu'améliorer l'estimation des effets des marqueurs.

Dans le cas de l'évaluation française des grandes races laitières, aucune différence de précision n'a été observée entre une évaluation incluant uniquement les DYD des taureaux ou une incluant les DYD des taureaux et les phénotypes des femelles. Cette absence de différence suggère que le biais lié au traitement préférentiel compense l'augmentation de précision attendue par l'agrandissement de la population de référence (Dassonneville et al., 2012). Toutefois, cette étude a été réalisée dans une situation où la majorité des femelles génotypées étaient constituées de femelles élites qui sont les plus susceptibles d'être sujettes à un traitement préférentiel. L'inclusion de performances de femelles génotypées dans le cadre de projets de recherche (supposées sélectionnées aléatoirement dans la population « commerciale ») n'introduit, quant à elle, pas de biais dans l'évaluation génomique (Dassonneville et al., 2012).

Il faut donc veiller avant d'inclure des femelles génotypées dans la population de référence à ce qu'elles soient représentatives de l'ensemble de la population et non uniquement de femelles élites. Une correction peut également être apportée aux phénotypes des femelles afin de limiter l'impact du traitement préférentiel sur les évaluations génomiques (Wiggans et al., 2011).

4. Une sélection génomique utilisant l'ensemble des individus, génotypés ou non

La différence de précision observée entre les animaux ayant ou non leur père dans la population de référence est liée à l'absence du génotype de ce père qui ne permet pas de mettre à profit le déséquilibre de liaison familial, mais également, à l'absence de son phénotype qui empêche l'utilisation de cette information pour l'estimation des valeurs génétiques. En effet, dans le cas des évaluations polygéniques, l'ensemble des performances des animaux (génotypés ou non) est utilisé pour prédire les valeurs génétiques des animaux. Dans le cas d'évaluations génomiques, la quantité d'information est réduite puisque la plupart des méthodes considère uniquement les phénotypes des animaux génotypés.

Une nouvelle approche, dite « Single Step Genomic BLUP (ssGBLUP) » proposé par Aguilar et al. (2010), Christensen et Lund, (2010), Legarra et al. (2009) et Misztal et al. (2009) permet l'évaluation conjointe de l'ensemble des individus (génotypés ou non) et donc, l'utilisation de l'ensemble de l'information disponible pour prédire les valeurs génétiques des animaux. Cette approche permet de transmettre l'information génomique à l'ensemble des animaux qu'ils soient génotypés ou non. Elle repose sur un système d'équations similaire à celui du modèle mixte en intégrant une modification de la matrice de parenté pour combiner les coefficients de parenté calculés à partir des marqueurs et ceux calculés à partir de la généalogie. La matrice intégrant information génomique et polygénique est appelée matrice **H.** Elle s'obtient à partir de l'expression suivante (Misztal et al., 2009) :

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{A_{11}} & \boldsymbol{A_{12}} \\ \boldsymbol{A_{21}} & \boldsymbol{G} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A_{11}} & \boldsymbol{A_{12}} \\ \boldsymbol{A_{21}} & \boldsymbol{A_{21} + (G - A_{22})} \end{bmatrix}$$

Où :

- $\mathbf{A_{11}}$, $\mathbf{A_{12}}$ ' $\mathbf{A_{11}}$ et $\mathbf{A_{22}}$ sont les sous-matrices de parenté calculées à partir de la généalogie pour les animaux non-génotypés (1) et génotypés (2)
- $\mathbf{G}$ est la matrice de parenté génomique
- $\boldsymbol{G - A_{22}}$ représente l'écart entre la parenté mesurée à partir des marqueurs et la parenté attendue à partir de la connaissance de la généalogie

Les équations du modèle mixte font intervenir non pas directement la matrice de parenté mais son inverse. Dans le cas des évaluations polygéniques, l'inversion de la matrice **A** pouvait être évitée grâce au calcul direct de la matrice $\mathbf{A^{-1}}$ (Henderson, 1976). Dans le cas du ssGBLUP, le calcul explicite de la matrice $\mathbf{H^{-1}}$ est impossible. En revanche, l'inversion de la matrice peut être simplifiée grâce à l'égalité suivante :

$$\boldsymbol{H^{-1}} = \boldsymbol{A^{-1}} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{(G^{-1} - A_{22}^{-1})} \end{bmatrix}$$

Aguilar et al., (2010) et Christensen et al. (2012) ont démontré une meilleure précision des valeurs génétiques estimées avec cette approche que celle observée avec un GBLUP classique. Sur les populations ovines disposant d'une taille de population de référence limitée (moins de 500 béliers), Legarra et al. (2014) ont observé que, contrairement à l'utilisation du GBLUP, l'utilisation du ssGBLUP peut permettre d'améliorer la précision des valeurs génétiques estimées par rapport à une évaluation polygénique. Un second avantage du ssGBLUP est qu'il permet d'inclure les phénotypes des femelles qu'elles soient génotypées ou non. Baloche et al., (2014) ont trouvé qu'à partir d'une population de 3 822 000 femelles et 7497 béliers dont 1593 génotypés, les évaluations génomiques ont permis un gain de 0,15 de corrélation par rapport aux évaluations polygéniques mais uniquement de 0,10 lorsque les performances des femelles n'étaient pas prises en compte (i.e. prises en compte uniquement à travers le DYD de leur père). Toutefois, aucune différence de précision n'a été observée entre les deux modèles dans le cas d'un caractère peu héritable comme les comptages cellulaires (Baloche et al., 2014).

Une approche de type ssGBLUP pourrait donc être intéressante pour la mise en place d'évaluations génomiques dans les races régionales. Toutefois, dans le cas des races laitières, pour lesquelles la quasi-totalité des pères des candidats à la sélection sont génotypés et la taille efficace de population est limitée, la différence de précision entre les modèles d'évaluation classiques et le ssGBLUP pourrait être plus faible.

Par ailleurs, deux difficultés majeures limitent la mise en œuvre d'un ssGBLUP pour les évaluations génomiques nationales. D'une part, contrairement aux méthodes d'évaluations génomiques classiques qui s'intégrent à la suite des évaluations polygéniques, l'utilisation du ssGBLUP nécessite une refonte complète les chaines d'évaluations génétiques nationales. D'autre part, le calcul , l'inversion et le stockage des matrices de parenté peuvent s'avérer coûteuse en ressources informatiques. Cette seconde difficulté est cependant moins limitante dans le cas des races régionales où le d'animaux évalués est réduit. Par ailleurs, de nombreux travaux sont actuellement en cours pour rendre le ssGBLUP compatible avec les évaluations génétiques nationales en cherchant à optimiser l'inversion des matrices de parenté (Meyer et al., 2013) ou en utilisant un système de résolution itératif alternant une partie « évaluation nationale » et une autre « évaluation génomique » (Legarra et Ducrocq, 2012)

En conclusion, plusieurs stratégies, qui sont résumées dans le Tableau 7, sont envisageables pour la mise en œuvre d'évaluations génomiques dans les races régionales. L'efficacité de ces différentes approches devra être étudiée. Il est probable que la stratégie finalement adoptée diffère en fonction des situations de chacune des races.

Tableau 7 : Récapitulatif des stratégies envisageables pour la mise en place d'évaluations génomiques dans les races régionales

	Stratégie	Avantages	Inconvénients
Intra-race	**Sélection assistée sur marqueurs et/ou mutations**	• Temps de calcul réduit • Intégration de l'information biologique • Ne nécessite pas d'imputation • Ne nécessite pas une harmonisation des phénotypes	• Nécessite d'identifier des QTL et des mutations dans les grandes races • Nécessite la présence de QTL et mutations communs entre races
	Sélection génomique utilisant uniquement les mâles génotypés	• Facilité de mise en œuvre • Ne nécessite pas d'imputation • Ne nécessite pas une harmonisation des phénotypes	• Peu précis en particulier pour les candidats peu apparentés à la population de référence
	Sélection génomique utilisant les mâles et femelles génotypés	• Facilité de mise en œuvre • Ne nécessite pas d'imputation • Ne nécessite pas une harmonisation des phénotypes	• Risque de biais lié au traitement préférentiel
	Sélection génomique utilisant l'ensemble des individus génotypés ou non	• Une seule étape (polygénique et génomique) • Ne nécessite pas d'imputation • Ne nécessite pas une harmonisation des phénotypes	• Temps de calcul • Développement méthodologique
Multi-race	**Sélection génomique utilisant la puce moyenne densité**	• Facilité de mise en œuvre • Ne nécessite pas d'imputation	• Uniquement dans le cas de races proches • Nécessite une harmonisation des phénotypes
	Sélection génomique utilisant la puce haute densité	• Conservation du déséquilibre de liaison • Possible pour l'ensemble des races	• Temps de calcul important • Nécessite une harmonisation des phénotypes • Nécessite des développements méthodologiques
	Sélection génomique utilisant les données de séquences	• Mutations causales dans les données génomiques • Ne nécessite pas de conservation du déséquilibre de liaison entre races	• Temps de calcul important • Coût du séquençage • Nécessite une harmonisation des phénotypes

D. Conséquences de la mise en place d'évaluations génomiques pour les races régionales.

Les travaux sur l'optimisation des schémas de sélection comparent généralement trois scénarios : un schéma utilisant uniquement des taureaux testés sur descendance, un schéma basé uniquement sur l'utilisation de jeunes taureaux, un schéma combinant l'utilisation de taureaux confirmés sur descendance et celle de jeunes taureaux. Dans le cas des races nationales, la majorité des études concluent que l'utilisation d'un schéma génomique exclusif permet à la fois la plus forte augmentation de progrès génétique, la meilleure gestion de la variabilité génétique et la meilleure rentabilité (Bouquet et Juga, 2013 ; Colleau et al., 2009 ; Pryce et Daetwyler, 2012).

A notre connaissance, aucune étude d'optimisation des schémas n'a été conduite sur des populations de taille similaire à celle des races régionales françaises. Dans le cas des races laitières régionales, la précision des évaluations génomiques est inférieure à celle observée pour les races nationales. Le gain de corrélation entre phénotypes et valeurs génétiques estimées permis par la sélection génomique est d'environ 0,05 en race Jersiaise Danoise (Thomasen et al., 2012), Simmental (Hozé et al., 2014b) et Normande réduite à 200 ou 500 animaux (Hozé et al., 2014a). Le gain estimé de CD permis par les évaluations génomiques dans ces races est donc inférieur à 0,10 alors qu'il se situait autour de 0,20 en race Holstein au lancement des programmes de sélections génomiques (Lund et al., 2011).

Dans ces conditions, il est possible que les conclusions des études réalisées dans le cas des races nationales ne soient pas directement transposables à celui des races régionales. Nous allons essayer d'analyser dans ce chapitre quelle utilisation de la sélection génomique est la plus adaptée à une situation où le gain de précision permis par les évaluations génomiques est limité.

1. Utilisation des évaluations génomiques et conséquences sur le progrès génétique annuel et l'évolution de la variabilité génétique

Dans le cas d'une précision des évaluations génomiques de 0,60 et une situation comparable à celle de la race Montbéliarde, Colleau et al. (2009) ont démontré que l'utilisation massive de 80 jeunes taureaux permettait de doubler le progrès génétique annuel tout en maintenant le niveau d'accroissement de la consanguinité par rapport à un schéma conventionnel (30 taureaux testés sur descendance et 15 utilisés massivement). Une situation intermédiaire, correspondant à l'utilisation conjointe de jeunes taureaux et de taureaux confirmés (réutilisés après une première diffusion en tant que jeunes taureaux) permet, elle aussi, un doublement du progrès génétique par rapport à la situation de référence mais elle entraine un doublement de l'accroissement annuel de la consanguinité (Colleau et al., 2009). Le déclin plus rapide de la variabilité génétique dans le cas d'un schéma combinant jeunes taureaux et taureaux confirmés s'explique par l'abandon du testage sur descendance et la réutilisation de façon massive, en tant que taureaux confirmés, d'animaux déjà diffusés une première fois en tant que jeunes taureaux. L'accélération du rythme d'accroissement de la consanguinité s'explique donc d'une part, par une utilisation sur une plus longue période d'un même animal que dans le schéma conventionnel et d'autre part, par la potentielle utilisation simultanée d'un animal en tant que taureau confirmé et de ses fils en tant que jeunes taureaux. L'augmentation de la consanguinité peut donc être réduite par une limitation de la durée totale d'utilisation (ou du nombre total de doses produites) d'un taureau et une contrainte imposant l'arrêt de la diffusion d'un animal lorsque ses fils sont utilisés en tant que jeunes taureaux.

Pour une précision des valeurs génomiques estimées plus faible (0,46), Lillehammer et al. (2011) ont étudié à partir d'un schéma de sélection composé de 750 candidats à la sélection génotypés, 125 taureaux testés sur descendance (Figure 12), l'évolution de la consanguinité et du progrès génétique annuel permis par les évaluations génomiques pour différents scénarios. Ils considèrent soit que les évaluations génomiques permettent de présélectionner les taureaux à mettre en testage sur descendance, soit qu'elles sont directement utilisées pour le choix des taureaux à utiliser massivement. Les résultats obtenus sont présentés dans le Tableau 8.

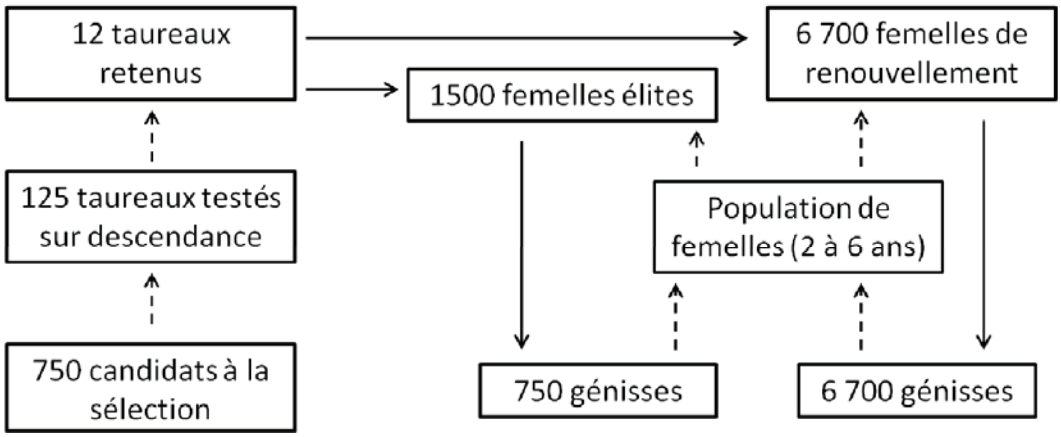


Figure 12 : Schéma de sélection simulé dans le cas de l'étude de Lillehammer et al. (2011)

Tableau 8 : Evolution de la consanguinité et du progrès génétique annuel en fonction du schéma de sélection utilisé, d'après Lillehammer et al. (2011)
PS_xx_12 : correspond à un schéma de sélection utilisant les évaluations génomiques pour présélectionner xx taureaux à tester sur descendance parmi lesquels 12 sont sélectionnés pour une utilisation sur l'ensemble de la population
GS_xx : correspond à un schéma de sélection utilisant les évaluations génomiques pour sélectionner xx jeunes taureaux à utiliser sur l'ensemble de la population

Schéma de sélection	Evolution de la consanguinité	Progrès génétique annuel
CONV	1,00	1,00
PS_125_12	0,67	1,13
PS_80_12	0,70	1,13
PS_60_12	0,63	1,11
GS_12	0,98	1,33
GS_20	0,70	1,28
GS_30	0,47	1,25
GS_40	0,36	1,20

La présélection des taureaux à mettre en testage sur descendance permet une augmentation de progrès génétique comprise entre 11 et 14% et une diminution d'environ 30% de l'augmentation de la consanguinité (Lillehammer et al., 2011). L'utilisation massive de jeunes taureaux permet, quant à elle, un gain de progrès génétique compris entre 20 et 33% et une diminution de 2 à 64% de l'accroissement annuel de la consanguinité (Lillehammer et al., 2011). Ces résultats sont cohérents avec les travaux de Colleau et al. (2009). Le gain de progrès génétique plus faible observé dans cette étude s'explique par la plus faible précision des valeurs génomiques estimées.

Dans cette étude l'utilisation limitée (260 inséminations) des jeunes taureaux pour l'étape de testage sur descendance avant leur utilisation massive en tant que taureaux confirmés, permet d'éviter l'augmentation du rythme d'accroissement de la consanguinité observée par Colleau et al. (2009). Cependant une augmentation du nombre d'inséminations réalisées (et donc du nombre de descendants) pendant le testage sur descendance par un jeune taureau, réduit le bénéfice, pour la gestion de la variabilité génétique, de la présélection des taureaux à mettre en testage sur descendance (Lillehammer et al. 2011).

Thomasen et al. (2014) a cherché à optimiser par simulation le gain de progrès génétique lié à l'intégration d'une étape de sélection génomique pour différents gains de précision permis par les évaluations génomiques. Trois grands types de schémas de sélection ont été simulés en faisant varier les proportions de jeunes taureaux utilisés sur les mères à taureaux et sur les femelles de la population commerciale :

- un schéma conventionnel utilisant uniquement des taureaux testés sur descendance
- un schéma de sélection « hybride » combinant utilisation de jeunes taureaux et taureaux testés sur descendance
- un schéma basé uniquement sur l'utilisation de jeunes taureaux.

Les effectifs utilisés pour les simulations du schéma de sélection sont représentés sur la Figure 13.

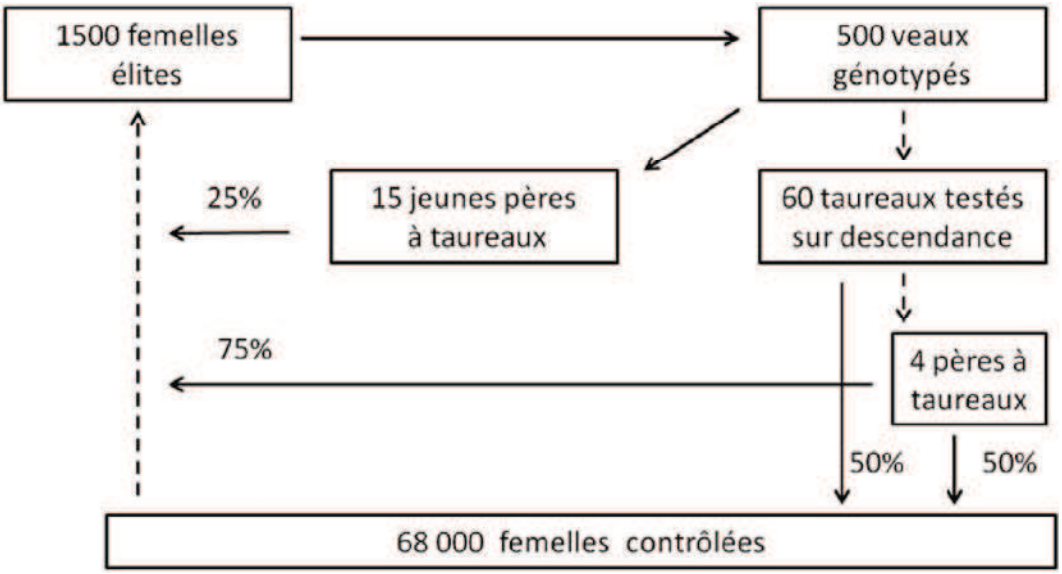


Figure 13 : Schéma de sélection « hybride » simulé dans le cas de l'étude de Thomasen et al., (2014)

Cette étude a démontré que le gain de progrès génétique annuel permis par la sélection génomique est très dépendant de l'écart de précision entre valeurs génomiques et celui des valeurs génétiques estimées sur ascendance (Thomasen et al., 2014).

Plusieurs schémas de sélection présentant différentes proportions de mères à taureaux et de femelles commerciales inséminées par de jeunes taureaux ont été comparés dans le cas où le gain de précision permis par les évaluations génomiques est de 0,05 (Figure 14). Dans toutes les situations étudiées, l'utilisation des évaluations génomiques permet un progrès génétique annuel supérieur à celui d'un schéma conventionnel. En revanche, lorsque la proportion totale de femelles inséminées avec des jeunes taureaux est élevée, il est préférable de ne pas utiliser de jeunes taureaux comme pères à taureaux. A l'opposé, lorsque l'utilisation de jeunes taureaux sur la population totale est faible (inférieure à 50%), il semble intéressant d'inséminer une large proportion de mères à taureaux avec des jeunes taureaux.

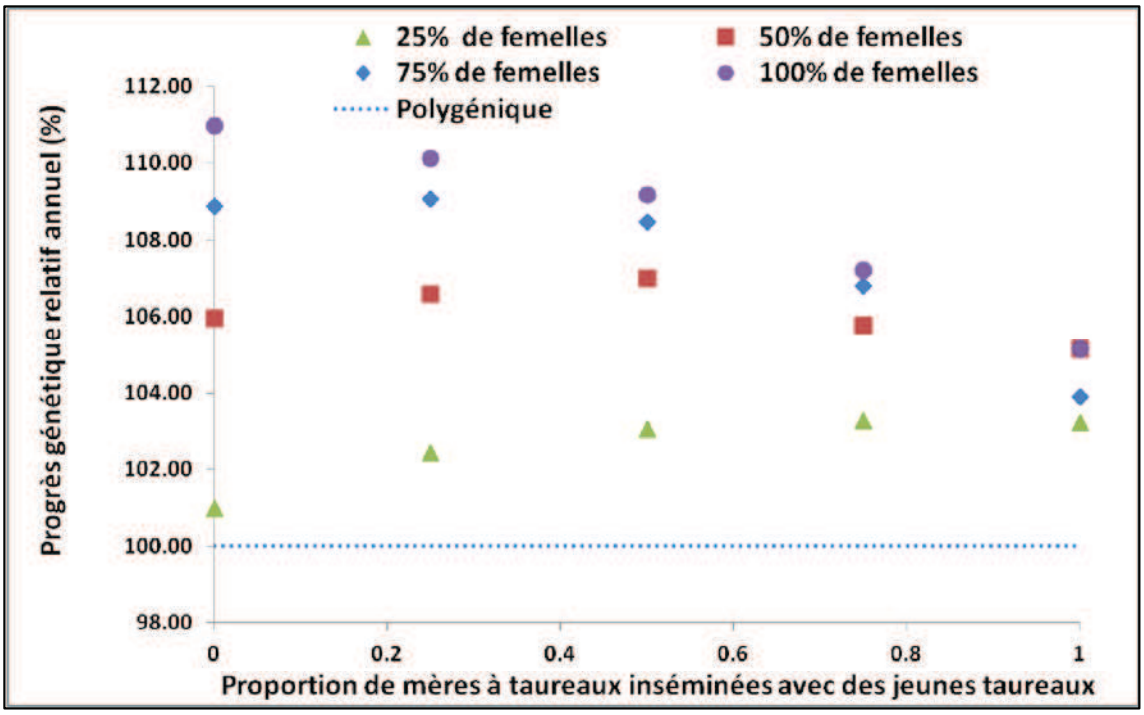


Figure 14 : Progrès génétique relatif annuel en fonction de la proportion de mères à taureaux et la proportion de femelles commerciales inséminées avec des jeunes taureaux d'après Thomasen et al. (2014). On considère ici un gain de précision permis par la sélection génomique de 0,05. Le 100 correspond à un schéma de sélection conventionnel basé sur les évaluations polygéniques

Un schéma hybride combinant utilisation de jeunes taureaux pour l'insémination des femelles de la population « commerciale » et utilisation de taureaux confirmés en tant que père à taureaux permet donc la plus forte augmentation de progrès génétique (Thomasen et al., 2014). Ces résultats sont cohérents avec les études de Colleau et al. (2009) et Lillehammer et al. (2011). Cependant la précision des valeurs génomiques plus faible simulée dans cette étude, nécessite une utilisation plus modérée des jeunes taureaux non testés sur descendance. Par ailleurs, l'impact des différents schémas de sélection sur l'évolution de la consanguinité n'a pas été pris en compte.

Les travaux de Lillehammer et al. (2011) prennent en compte l'augmentation de la précision des évaluations génomiques permises par l'augmentation de la taille de la population de référence (liée à l'arrivée des performances des filles des taureaux). Le nombre relativement important de taureaux utilisés chaque année permet en effet, une augmentation de la précision des évaluations génomiques. Cette augmentation de précision explique probablement pourquoi, dans son étude, un schéma de sélection basé uniquement sur l'utilisation de jeunes taureaux non testés sur descendance est privilégié. En effet, avec un gain de précision des évaluations génomiques légèrement plus élevée (0,10), Thomasen et al. (2014) ont également montré que l'utilisation exclusive de jeunes taureaux pour l'insémination des mères à taureaux et des femelles de la population commerciale permet la plus forte augmentation du progrès génétique.

D'après ces trois études, il semble donc plus favorable, pour concilier augmentation du progrès génétique et gestion de la variabilité génétique, d'utiliser directement les jeunes taureaux sur la base de leur information moléculaire.

2. Utilisation des évaluations génomiques, coût des schémas de sélection et précision des évaluations

Dans les paragraphes précédents, nous avons mis en avant l'intérêt d'utiliser de la sélection génomique pour augmenter le progrès génétique annuel et améliorer la gestion de la variabilité génétique. Cependant, l'intérêt économique global de chacune des stratégies (schéma conventionnel, hybride ou génomique) n'a pas été évalué. En particulier, l'investissement nécessaire à l'établissement d'une première population de référence et au génotypage des candidats à la sélection n'a pas été pris en compte.

Dans les races nationales, les coûts de génotypage ont rapidement été compensés par une forte réduction et même un abandon du testage organisé sur descendance. Dans le cas des races régionales, lorsque seul le coût économique associé à l'augmentation du progrès génétique est pris en compte, l'utilisation d'un schéma de sélection basé uniquement sur l'information moléculaire correspond à la meilleure stratégie (Thomasen et al., 2014). Cependant, sa mise œuvre peut s'avérer difficile compte tenu de la précision limitée des évaluations génomiques.

Dans les trois principales races laitières françaises, la proportion d'inséminations bovines réalisées avec de la semence de jeunes taureaux est en constante augmentation et est aujourd'hui comprise entre 50 et 70% (Tableau 4). Cette proportion (inférieure à 100%) s'explique par une disponibilité temporaire en bons taureaux issus du testage sur descendance mais également une méfiance de certains éleveurs vis-à-vis des évaluations génomiques. En effet, une précision plus faible des valeurs génomiques estimées implique que les index des jeunes taureaux peuvent subir de fortes variations (à la hausse ou à la baisse) lors de l'arrivée des performances de leurs filles. Comme dans le cas des races nationales, il est probable que dans les races régionales la proportion de femelles inséminées avec des jeunes taureaux soit d'abord faible et qu'elle augmente progressivement avec l'amélioration de la précision des évaluations génomiques.

Il faut également noter que la précision des évaluations génomiques dans les races régionales est de l'ordre de grandeur de celle qui avait été observée lors des premières études de sélection assistée par marqueurs dans les trois races nationales. La sélection génomique peut donc être, dans un premier temps, uniquement utilisée au niveau des entreprises de sélection pour la présélection des veaux à mettre en testage et le choix des pères et des mères à taureaux. Cette stratégie permet déjà d'augmenter le progrès génétique annuel réalisé et de limiter l'accroissement annuel de la consanguinité (Lillehammer et al., 2011). De plus si elle est associée à une réduction du nombre de taureaux testés sur descendance, les coûts du génotypage peuvent être, au moins en partie, compensés par la réduction du coût total du testage.

.

3. Spécificités liés aux situations des races régionales françaises

Les trois études citées précédemment démontrent qu'il est intéressant, même dans le cas où le gain de précision permis par les évaluations génomiques est limité, d'utiliser l'information moléculaire dans les schémas de sélection. Cependant, les schémas de sélection simulés sont très différents de ceux mis en place par les races régionales françaises. Au niveau international, les travaux s'intéressant à de « petites races » correspondent généralement à des populations de la taille des races Normande ou Montbéliarde. En race Abondance (actuellement 4ème race laitière française), seuls une quinzaine de taureaux sont aujourd'hui testés sur descendance chaque année parmi lesquels trois à cinq sont conservés pour une utilisation sur l'ensemble de la population (Figure 15, Tableau 2). Les grands principes d'utilisation des évaluations génomiques restent valides dans le cas des races régionales. Cependant, il existe certaines spécificités liées à la taille de leur population.

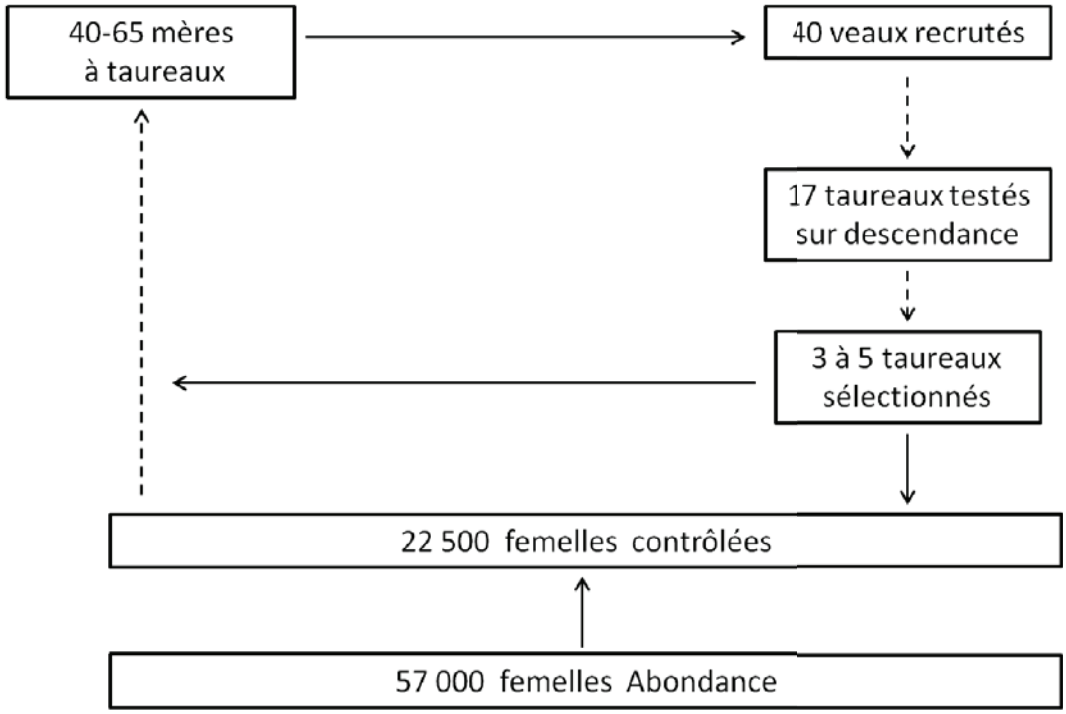


Figure 15 : Schéma de sélection en race Abondance, d'après N. Bloc, UCEAR (communication personnelle)

Etant donné le faible nombre de taureaux testés sur descendance, il est d'autant plus important de les sélectionner à partir d'une information précise afin de plus précisément s'assurer de leur niveau génétique tout en limitant l'accroissement de la consanguinité. L'utilisation des évaluations génomiques permet également de mieux sélectionner les mères et pères à taureaux afin d'améliorer la gestion de la variabilité génétique mais aussi le progrès génétique annuel. En effet, augmenter le nombre de candidats à la sélection permet d'augmenter la pression de sélection avant testage sur descendance et d'augmenter la probabilité qu'un taureau soit bon sur tout ou partie des caractères évalués. Il devient donc possible de sélectionner une dizaine de taureaux (et non plus 5) à utiliser sur l'ensemble de la population.

Par ailleurs, la précision des valeurs génomiques estimées est moins sensible à l'héritabilité du caractère que celles des valeurs génétiques estimées sur ascendance. L'utilisation de l'information moléculaire pour la présélection de taureaux à mettre en testage ou l'utilisation directe de jeunes taureaux permet donc de mieux équilibrer l'effort de sélection entre caractères et en particulier d'améliorer la sélection sur les caractères peu héritables (Lillehammer et al.,2011 ; Thomasen et al., 2014). Dans le cas des races régionales, où les taureaux sont testés sur une trentaine de filles (au lieu de 50 à 100 pour les races nationales françaises) l'information disponible sur les caractères peu héritables pour les taureaux en sortie de testage est très peu précise. Les évaluations génomiques devraient donc être particulièrement utiles pour améliorer la sélection sur les caractères fonctionnels.

Enfin, en fonction des stratégies adoptées, le nombre de taureaux génotypés avec des filles (et donc des phénotypes) susceptibles d'augmenter la taille de la population de référence varie, ce qui influence l'évolution de la précision des valeurs génomiques (Lillehammer et al.,2011). Dans le cas des races régionales, la possibilité d'augmenter la taille de référence et la précision des évaluations génomiques doivent aussi être prises en considération lors de l'établissement d'un schéma de sélection. Il peut donc sembler souhaitable de maintenir, dans un premier temps, le testage sur descendance afin d'augmenter rapidement la taille de la population de référence. Cependant, le maintien du testage nécessite d'augmenter le coût total du schéma de sélection pour financer les génotypages.

Une autre stratégie consiste à utiliser un nombre important de jeunes taureaux sur un nombre fixé (par exemple compris entre 200 et 500) de femelles afin qu'à l'arrivée des performances de leurs filles, ils intègrent la population de référence. En effet, il faut rappeler que la mise en place d'évaluations génomiques n'est surtout pas synonyme de l'abandon du système de contrôle des performances. Le maintien du contrôle de performances et le renouvellement régulier de la population de référence permettent d'augmenter la précision les évaluations. Ils sont également nécessaires pour conserver un lien fort entre population de référence et candidats à la sélection ce qui assure la persistance de l'efficacité de la sélection génomique au fil des générations (Bastiaansen et al., 2012). L'utilisation modérée d'un grand nombre de jeunes taureaux peut donc permettre de limiter la prise de risque liée à la faible précision des valeurs génomiques et d'augmenter la population de référence à moindre coût (génotypages financés par l'abandon du testage sur descendance).

4. Utilisation de la sélection génomique et gestion du troupeau

En élevage, le génotypage et les évaluations génomiques des femelles peuvent être utilisés comme des outils de gestion du troupeau (choix des génisses de renouvellement, gestion des accouplements visant à corriger les défauts de la femelle, etc.). Cette utilisation bien qu'intéressante d'un point de vue zootechnique ne l'est pas toujours d'un point de vue économique. En effet, cette stratégie n'est rentable que dans le cas où le coût associé au génotypage est compensé par le gain de progrès génétique (Dassonneville, 2012).

Pour les races nationales, cette rentabilité est actuellement permise par un coût de génotypage réduit (grâce à l'utilisation de la puce basse densité et l'imputation de génotypes 50K) et une forte précision des valeurs génomiques estimées. Dans le cas des races régionales, la différence de précision d'estimation entre évaluations génomiques et évaluations polygéniques est trop faible pour compenser le coût d'un génotypage (y compris basse densité). De plus il peut être préférable dans un premier temps, de privilégier le génotypage des femelle avec la puce 50K, afin de les inclure dans la population de référence et d'améliorer la précision des évaluations génomiques.

Une fois la précision des évaluations génomiques satisfaisante, il sera possible d'envisager l'utilisation de la puce basse densité et le génotypage femelle comme un outil de gestion du troupeau.

En conclusion, la mise en place d'évaluations génomiques pour les races régionales et l'utilisation raisonnée de l'information moléculaire dans les schémas de sélection (sélection des mères et pères à taureaux, présélection des veaux à mettre en testage) peuvent permettre d'augmenter, et mieux équilibrer entre caractères, le progrès génétique annuel réalisé tout en limitant l'accroissement de la consanguinité. En revanche, la précision des évaluations génomiques est, pour l'instant, peut-être insuffisante pour envisager l'utilisation exclusive de taureaux non testés sur descendance en tant que pères à taureaux et pères des femelles de la population commerciale. A terme, l'augmentation de la taille de population de référence permise par la mise en place d'une évaluation génomique et la poursuite des développements méthodologiques autour des évaluations génomiques multiraciales devraient permettre de rapidement augmenter la précision des évaluations. Dès lors que la précision des évaluations génomiques sera satisfaisante, il sera possible de basculer vers un schéma de sélection uniquement basé sur l'information moléculaire et l'utilisation de jeunes taureaux.

VII. Conclusion

Depuis la fin du vingtième siècle, le développement des techniques de séquençage et de génotypage à haut débit a bouleversé le monde de la génétique animale. Il est maintenant possible d'accéder à l'information moléculaire à un coût modéré, d'identifier de nombreuses régions influençant un caractère d'intérêt (QTL) et de les utiliser en sélection animale. En 2008, le développement de la puce pangénomique bovine a considérablement accéléré l'avancée des connaissances sur le génome bovin et révolutionné le monde de la sélection bovine.

Les schémas de sélection, qui s'appuyaient jusqu'alors sur les index sur ascendance et le testage sur descendance, ont aujourd'hui également accès aux index génomiques. Ces index sont obtenus grâce à la sélection génomique consistant à mieux évaluer les candidats à la sélection en utilisant les effets des marqueurs estimés sur une population génotypée et phénotypée appelée population de référence. En France, les populations de référence disponibles dans les principales races laitières (Holstein, Montbéliarde et Normande) ont permis de disposer de valeurs génomiques estimées précises à partir d'un simple échantillon biologique.

Il a alors été possible de sélectionner les animaux dès leur plus jeune âge et de supprimer l'étape de testage organisé sur descendance. Cette suppression a conduit à une forte réduction de l'intervalle de génération et à l'augmentation du nombre de candidats à la sélection. Les schémas de sélection ont alors été revisités afin de permettre à coût total constant (voire moindre) une augmentation du progrès génétique annuel, une meilleure gestion de la variabilité génétique des animaux et un effort de sélection plus équilibré entre les différents caractères.

La précision des valeurs génomiques estimées est essentiellement fonction du nombre d'individus génotypés avec performances qui ont été utilisés pour l'estimation des effets des marqueurs. Dans les races régionales, le nombre réduit d'individus testés sur descendance limite la taille de la population de référence et la précision des évaluations génomiques.

Cependant, la mise en place d'évaluations génomiques dans les races régionales est nécessaire pour maintenir la diversité des races bovines françaises mise à mal par un fossé croissant entre les races disposant de « l'outil génomique » et celles n'en disposant pas. Ainsi, seules les races nationales bénéficient actuellement des avantages de la sélection génomique ce qui creuse l'écart de niveau génétique entre races nationales et races régionales. Cet écart pourrait conduire certains éleveurs à s'orienter vers une race nationale pour augmenter leurs productivités. Il a alors été suggéré, pour le développement d'évaluations génomiques dans les races régionales, de ne plus considérer une population de référence constituée d'animaux de la même race que les candidats à la sélection mais de regrouper les populations de référence entre races.

Utiliser une population de référence multiraciale nécessite de disposer d'une densité en marqueurs suffisante pour que le lien entre QTL et marqueurs soit conservé au niveau de l'espèce et non plus au niveau de la race. Pour répondre à ce besoin, 5072 individus issus de 20 races laitières et allaitantes ont été génotypés avec la puce haute densité contenant 777 962 marqueurs soit environ un marqueur toutes les 4000 paires de bases.

Dans notre première étude, nous avons démontré que le taux d'erreur d'imputation de génotypes haute densité à partir de génotypages réalisés sur la puce 50K (contentant 54 602 marqueurs) était faible (généralement inférieur à 1%) dans la majorité des races. Il est donc possible de génotyper les animaux de la future population de référence sur une puce moyenne densité puis d'imputer leurs génotypes haute densité. Cette stratégie permet à la fois de limiter le coût de génotypage et de ne pas génotyper à nouveau les animaux pour lesquels un génotype moyenne densité est déjà disponible. L'imputation des génotypes haute densité de plus de 30 000 individus des populations de référence des trois grandes races laitières a été réalisée afin de permettre les études d'estimation de la précision des évaluations génomiques multiraciales.

L'objectif de la seconde étude était de déterminer pour différentes tailles de populations de référence quelle évaluation génomique (intra-race ou multi-race à partir de génotypes moyenne ou haute densité) devait être mise en œuvre. Nous avons démontré que dans le cas d'une population de référence de taille moyenne (1500 animaux), le passage aux évaluations génomiques multiraciales et l'utilisation de la puce haute densité ne permettent qu'un gain de précision limité. En revanche, les évaluations multiraciales sont intéressantes dans le cas d'une population de référence de 500 individus ou moins.

Nous avons également observé que la puce haute densité ne permettait pas d'améliorer la précision des évaluations génomiques intra-races mais qu'elle était utile dans le cas des évaluations génomiques multiraciales. Ce résultat est cohérent avec un déséquilibre de liaison suffisant sur la puce moyenne densité pour une utilisation intra-race et une meilleure conservation du déséquilibre de liaison entre populations sur la puce haute densité. Enfin, il faut noter que la précision des évaluations génomiques intra-races est relativement élevée, même dans le cas d'une population de référence réduite, à condition que les pères des candidats à la sélection soient inclus dans la population de référence.

Deux grandes stratégies sont donc envisageables pour la mise en place d'évaluations génomiques dans les races régionales : des évaluations intra-races utilisant la puce moyenne densité et imposant le génotypage des pères des candidats ou une évaluation multiraciale utilisant la puce haute densité. L'utilisation de la puce haute densité engendre de nombreuses contraintes notamment en termes de ressources informatiques. En effet, l'imputation régulière avec le logiciel Beagle (utilisé pour l'étude du taux d'erreur d'imputation) de l'ensemble des candidats à la sélection est difficilement envisageable. L'estimation des effets de l'ensemble des marqueurs de la puce haute densité est également très coûteuse en temps de calcul et en besoins en mémoire.

Une troisième stratégie a alors été étudiée consistant à utiliser un groupe de races proches, dans notre cas la Montbéliarde et la Simmental, afin de pouvoir détecter un déséquilibre de liaison conservé entre race sur la puce moyenne densité. Le gain de précision permis par les évaluations multiraciales était, à taille de population de référence similaire, trois fois supérieur à ce qui avait été observé pour une évaluation regroupant les races Holstein, Montbéliarde et Normande. En revanche, il n'est pas apparu intéressant d'appliquer directement les effets des marqueurs estimés en race Montbéliarde pour l'estimation des valeurs génomiques des candidats Simmental.

Les travaux réalisés lors de cette thèse ont donc pu montrer que la mise en place d'évaluations génomiques dans les races régionales est possible. Elle impose cependant plus de contraintes que le développement d'évaluations génomiques dans les races nationales. De plus, le gain de précision reste relativement faible en comparaison de celui observé dans les races nationales ce qui nécessite une utilisation raisonnée des évaluations génomiques et en particulier un abandon plus tardif de l'étape de testage organisé sur descendance. En revanche, l'information moléculaire peut être utilisée dans les schémas de sélection pour augmenter le progrès génétique annuel, limiter l'accroissement de la consanguinité et mieux équilibrer l'effort de sélection entre caractères.

A terme, de nombreuses pistes peuvent être étudiées pour améliorer la précision des évaluations génomiques. A court terme, l'inclusion de femelles dans la population de référence semble être la stratégie la meilleure pour améliorer la précision des évaluations génomiques. Il est également possible d'utiliser un modèle permettant l'inclusion des animaux non génotypés pour l'estimation des valeurs génomiques des candidats comme la Sélection Assistée par Marqueurs ou l'approche dite « Single Step ». Les travaux de recherche autour de l'utilisation des données de séquençage complet et l'intégration de mutations causales dans les modèles d'évaluations devraient eux aussi permettre d'améliorer la précision des évaluations en levant la contrainte de la conservation du déséquilibre de liaison entre races.

Il faut également prendre en compte qu'une fois un premier système d'évaluations génomiques mis en place, les animaux génotypés en tant que candidats à la sélection et finalement retenus pourront, à l'arrivée des performances de leur filles, être intégrés à la population de référence. Le maintien du contrôle de performances est donc plus que jamais nécessaire pour augmenter la taille de population de référence mais également pour maintenir le lien, essentiel à la précision des évaluations génomiques, entre population de référence et candidats à la sélection.

Les études visant à améliorer la précision des évaluations génomiques des races régionales (mais également des races nationales) doivent donc être poursuivis afin de faciliter l'utilisation de la sélection génomique en élevage et dans les schémas de sélection.

Bibliographie

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

Allais, S. 2011. **Détection et validation de marqueurs génétiques impliqués dans la qualité de la viande bovine**. Thèse de doctorat, AgroParisTech-ABIES. 257p.

Andersson, L. 2001**. Genetic dissection of phenotypic diversity in farm animals.** Nat Rev Genet 2(2):130-138.

Baloche, G., A. Legarra, G. Salle, H. Larroque, J. M. Astruc, C. Robert-Granie, et F. Barillet. 2014. **Assessment of accuracy of genomic prediction for French Lacaune dairy sheep**. J. Dairy Sci. 97:1107-1116.

Bapst, B., C. Baes, F. R. Seefried, A. Bieber, H. Simianer, et B. Gredler. 2013. **Effect of cows in the reference population: First results in Swiss Brown Swiss**. Interbull Bull. 47:187-191.

Bastiaansen, J. W., A. Coster, M. P. Calus, J. A. van Arendonk, 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.

Bennewitz, J., N. Reinsch, F. Reinhardt, Z. Liu, et E. Kalm. 2004. **Top down preselection using marker assisted estimates of breeding values in dairy cattle.** J. Anim. Breed. Genet. 121(5):307-318.

Bennewitz, J., N. Reinsch, H. Thomsen, J. Szyda, F. Reinhart, C. Kuhn, M. Schwerin, G. Erhardt, C. Weimann, et E. Kalm. 2003. **Marker assisted selection in German Holstein dairy cattle breeding: outline of the program and marker assisted breeding value estimation.** in Book of Abstracts of the 54th Annual Meeting of the EAAP. p.118

Boichard, D., H. Chung, R. Dassonneville, X. David, A. Eggen, S. Fritz, K. J. Gietzen, B. J. Hayes, C. T. Lawley, T. S. Sonstegard, C. P. Van Tassell, P. M. VanRaden, K. A. Viaud-Martinez, et G. R. Wiggans. 2012a. **Design of a bovine low-density SNP array optimized for imputation**. PLoS One 7(3):e34130.

Boichard, D., F. Guillaume, A. Baur, P. Croiseau, M. N. Rossignol, M. Y. Boscher, T. Druet, L. Genestout, J. J. Colleau, L. Journaux, V. Ducrocq, et S. Fritz. 2012b. **Genomic selection in French dairy cattle**. Anim. Prod. Sci. 52:115-120.

Boichard, D., S. Fritz, M. N. Rossignol, M. Y. Boscher, A. Malafosse, et J. J. Colleau. 2002. **Implementation of marker-assisted selection in French dairy cattle**. in Proc. of the 7th World Congr. Genet. Appl. Livest. , Montpellier, France, 19-23 Août 2002. Comm. n° 28-13.

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

Boichard, D., L. Maignel, et E. Verrier. 1997. **The value of using probabilities of gene origin to measure genetic variability in a population**. Genet. Sel. Evol. 29:5–23.

Boichard, D., L. Maignel, et E. Verrier. 1996. **Analyse généalogique des races bovines laitières françaises**. Productions Animales 9:323-334.

Bouquet, A. et J. Juga. 2013. **Integrating genomic selection into dairy cattle breeding programmes: a review**. Animal 7:705-713.

Brondum, R. F., G. Su, M. S. Lund, P. J. Bowman, M. E. Goddard, et B. J. Hayes. 2012. **Genome position specific priors for genomic prediction**. BMC Genomics 13:543.

Brondum, R. F., E. Rius-Vilarrasa, I. Stranden, G. Su, B. Guldbrandtsen, W. F. Fikse, et M. S. Lund. 2011. **Reliabilities of genomic prediction using combined reference data of the Nordic Red dairy cattle populations**. J. Dairy Sci. 94:4700-4707.

Browning, S. R. et B. L. Browning. 2007. **Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering**. Am. J. Hum. Genet. 81:1084-1097.

Calus, M. P., Y. de Haas, M. Pszczola, et R. F. Veerkamp. 2013a. **Predicted accuracy and response to genomic selection for new traits in dairy cattle**. Animal 7:183-191.

Calus, M. P., Y. de Haas, et R. F. Veerkamp. 2013b. **Combining cow and bull reference populations to increase accuracy of genomic prediction and genome-wide association studies**. J. Dairy Sci. 96:6703-6715.

Calus, M.P., R. F. Veerkamp, et H.A. Mulder. 2011. **Imputation of missing single nucleotide polymorphism genotypes using a multivariate mixed model framework**. J. Anim. Sci., 89:2042–2049.

Calus, M. P., T. H. Meuwissen, J. J. Windig, E. F. Knol, C. Schrooten, A. L. Vereijken, et R. F. Veerkamp. 2009. **Effects of the number of markers per haplotype and clustering of haplotypes on the accuracy of QTL mapping and prediction of genomic breeding values**. Genet. Sel. Evol. 41:11.

Calus, M. P., T. H. Meuwissen, A. P. de Roos, et R. F. Veerkamp. 2008**. Accuracy of genomic selection using different methods to define haplotypes**. Genetics 178:553-561.

Calus, M. P., 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. 124:362–368.

Carillier, C., H. Larroque, I. Palhiere, V. Clement, R. Rupp, et C. Robert-Granie. 2013. **A first step toward genomic selection in the multi-breed French dairy goat population**. J. Dairy Sci. 96:7294-7305.

Christensen, O. F. et M. S. Lund. 2010. **Genomic prediction when some animals are not genotyped.** Genet. Sel. Evol. 42:2.

Christensen, O. F., P. Madsen, B. Nielsen, T. Ostersen, et G. Su. 2012. **Single-step methods for genomic evaluation in pigs**. Animal 6:1565-1571.

Clark, S. A., J. M. Hickey, H. D. Daetwyler, et J. H. van der Werf. 2012. **The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes**. Genet. Sel. Evol. 44:4.

Colleau, J. J., S. Fritz, F. Guillaume, A. Baur, D. Dupassieux, M. Y. Boscher, L. Journaux, A. Eggen, et D. Boichard. 2009. **Simulating the potential of genomic selection in dairy cattle breeding**. 16èmes Renc. Rech. Rum. 419.

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

Croiseau, P., M. N. Fouilloux, D. Jonas, S. Fritz, A. Baur, V. Ducrocq, F. Phocas, et D. Boichard. 2014. **Extension of algorithms to the use of haplotypes in genomic selection**. in Proc. 10th World Congr. Genet. Appl. Livest., 17-22 Août 2014, Vancouver, (Canada), (soumis).

Croiseau, P., F. Guillaume, et S. Fritz. 2012. **Comparison of genomic selection approaches in Brown Swiss within Intergenomics**. Interbull Bull. 46:127-132.

Croiseau, P., A. Legarra, F. Guillaume, S. Fritz, A. Baur, C. Colombani, C. Robert-Granie, 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., M. P. Calus, R. Pong-Wong, G. de Los Campos, et J. M. Hickey. 2013. **Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking**. Genetics 193:347-365.

Daetwyler, H. D., R. Pong-Wong, B. Villanueva, et J. A. Woolliams. 2010. **The impact of genetic architecture on genome-wide evaluation methods**. Genetics 185:1021-1031.

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.

Danchin-Burge, C., G. Leroy, M. Brochard, S. Moureaux, et E. Verrier. 2012. **Evolution of the genetic variability of eight French dairy cattle breeds assessed by pedigree analysis**. J. Anim. Breed. Genet. 129:206-217.

Danchin-Burge, C. 2009. **Estimation de la variabilité génétique de 19 races bovines à partir de leurs généalogies**. Institut de l elevage, 149 rue de Bercy 75012 Paris (France). Disponible en ligne : http://idele.fr/recherche/publication/idelesolr/recommends/estimation-de-la-variabilite-genetique-de-19-races-bovines-a-partir-de-leurs-genealogies.html [17Avril 2014]

Dassonneville, R. 2012. **Sélection génomique des vaches laitières**. Thèse de doctorat, AgroParisTech-ABIES. 163p.

Dassonneville, R., A. Baur, S. Fritz, D. Boichard, et V. Ducrocq. 2012. **Inclusion of cow records in genomic evaluations and impact on bias due to preferential treatment**. Genet. Sel. Evol. 44:40.

Dassonneville, R., R. F. Brondum, T. Druet, S. Fritz, F. Guillaume, B. Guldbrandtsen, M. S. Lund, V. Ducrocq, et G. Su. 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.

De Los Campos, G., J. M. Hickey, R. Pong-Wong, H. D. Daetwyler, et M. P. Calus. 2013. **Whole-genome regression and prediction methods applied to plant and animal breeding**. Genetics 193:327-345.

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

De Roos, A. P. W., B. J. Hayes, R. J. Spelman, et M. E. Goddard. 2008. **Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle**. Genetics 179:1503-1512.

Dekkers, J. C. 2004. **Commercial application of marker- and gene-assisted selection in livestock: strategies and lessons**. J Anim Sci 82:313-328.

Druet, T., I. M. Macleod, et B. J. Hayes. 2014. **Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions**. Heredity 112, 39–47

Druet, T. et 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.

Druet, T., C. Schrooten, et A. P. de Roos. 2010. **Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle**. J. Dairy Sci. 93:5443-5454.

Druet, T., S. Fritz, M. Boussaha, S. Ben-Jemaa, F. Guillaume, D. Derbala, D. Zelenika, D. Lechner, C. Charon, D. Boichard, I. G. Gut, A. Eggen, et M. Gautier. 2008. **Fine mapping of quantitative trait loci affecting female fertility in dairy cattle on BTA03 using a dense single-nucleotide polymorphism map**. Genetics 178:2227-2235.

Duchemin, S. I., C. Colombani, A. Legarra, G. Baloche, H. Larroque, J. M. Astruc, F. Barillet, C. Robert-Granié, et E. Manfredi. 2012. **Genomic selection in the French Lacaune dairy sheep breed**. J. Dairy Sci. 95:2723–2733.

Ducrocq, V., P. Croiseau, A. Baur, R. Saintilan, S. Fritz, et D. Boichard. 2014. **Genomic evaluation using QTL information**. in Proc. 10th World Congr. Genet. Appl. Livest. 17-22 Août 2014, Vancouver (Canada), (soumis)

Dürr, J. et J. Philipsson. 2012. **International cooperation: The pathway for cattle genomics.** Animal Frontiers 2:16-21.

Egger-Danner, C., H. Schwarzenbacher, et A. Willam. 2012. **Genotyping of cows for genomic EBVs for direct health traits-Genetic and economic aspects**. in Book of Abstracts of the 63rd Annual Meeting of the EAAP, Bratislava (Slovaquie). p.163

Eggen, A. 2012. **The development and application of genomic selection as a new breeding paradigm**. Anim. Front. 2:10–15.

Erbe, M., B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman, C. M. Reich, B. A. Mason, et M. E. Goddard. 2012. **Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels**. J. Dairy Sci. 95:4114-4129.

Erbe, M., E. C. J. Pimentel, A. R. Sharifi, et H. Simianer. 2010. **Assessment of cross-validation strategies for genomic prediction in cattle**. in Proc. 9th World Congr. Genet. Appl. Livest., 1er-6 Août 2010, Leipzig (Allemagne), p. 129–132

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.

Fouilloux, M. N., S. Minery, S. Mattalia, et D. Laloë. 2006. **Assessment of Connectedness in the International Genetic Evaluation of Simmental and Montbéliard Breeds**. Interbull Bull. 35:129-135.

France Génétique Elevage. 2014. **Dispositif Génétique Chiffres clés Ruminants.** France Génétique Elevage, 149 rue de Bercy 75012 Paris (France). Disponible en ligne: http://fr.france-genetique-elevage.org/IMG/pdf/fge_chiffres_des_2013_genetique_francaise.pdf [17 Avril 2014]

Fritz, S., F. Guillaume, J. Tarres, A. Baur, M. Boussaha, M. Y. Boscher, L. Journaux, A. Malafosse, M. Gautier, et J. J. Colleau. 2008. **Utilisation des résultats de cartographie fine de QTL en sélection chez les bovins laitiers**. 15èmes Renc. Rech. Rum. :423-426.

Fritz, S., T. Druet, F. Guillaume, A. Malafosse, M. Y. Boscher, A. Eggen, M. Gautier, J. J. Colleau, et D. Boichard. 2007. **Bilan du programme de Sélection Assistée par Marqueurs dans les trois principales races bovines laitières françaises et perspectives d'évolution**. 14èmes Renc. Rech. Rum:129-132.

Gautier, M., D. Laloé, et K. Moazami-Goudarzi. 2010. **Insights into the genetic history of French cattle from dense SNP data on 47 worldwide breeds**. PLoS One 5(9).

Gautier, M., T. Faraut, K. Moazami-Goudarzi, V. Navratil, M. Foglio, C. Grohs, A. Boland, J.-G. Garnier, D. Boichard, et G. M. Lathrop. 2007a. **Genetic and haplotypic structure in 14 European and African cattle breeds**. Genetics 177:1059-1070.

Gautier M., A. Capitan A, S. Fritz, A. Eggen, D. Boichard, et T. Druet. 2007b. **Characterization of the DGAT1 K232A and variable number of tandem repeat polymorphisms in French dairy cattle.** J Dairy Sci. 90:2980-2988.

GEB Institut de l'Elevage. 2013. Chiffres clés **Production bovines lait et viande**. Institut de l'Elevage., 149 rue de Bercy 75012 Paris. Disponible sur http://idele.fr/domaines-techniques/economie-des-filieres/analyse-des-filieres/publication/idelesolr/recommends/chiffres-cles 2013-des-productions-bovines-lait-viande.html [17 Avril 2014]

Gianola, D. 2013. **Priors in whole-genome regression: the bayesian alphabet returns**. Genetics 194:573-596.

Gianola, D., G. de los Campos, W. G. Hill, E. Manfredi, et R. Fernando. 2009. **Additive genetic variability et the Bayesian alphabet**. Genetics 183(1):347-363.

Goddard, K. A. B., P. J. Hopkins, J. M. Hall, et J. S. Witte. 2000. **Linkage disequilibrium et allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations**. Am J Hum Genet, 66:216–234.

Goddard, M. E. 2009. **Genomic selection: prediction of accuracy et maximisation of long term response**. Genetica 136:245-257.

Goddard, M. E. et B. J. Hayes. 2009. **Mapping genes for complex traits in domestic animals et their use in breeding programmes**. Nat Rev Genet 10:381-391.

Goddard, M. E. et B. J. Hayes. 2007. **Genomic selection**. J Anim Breed Genet 124(6):323-330.

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

Guillaume, F. 2009. **Intégration de l'information moléculaire dans l'évaluation génétique**. Thèse de doctorat, AgroParisTech-ABIES, 145p.

Guillaume, F., S. Fritz, P. Croiseau, A. Legarra, C. Robert-Granié, C. Colombani, C. Patry, D. Boichard, et V. Ducrocq. 2009. **Modèles d'évaluation génomique: application aux populations bovines laitières françaises.** 16ème Renc. Rech. Rum.:399-406.

Guillaume, F. o., S. b. Fritz, D. Boichard, et T. Druet. 2008. **Short Communication: Correlations of Marker-Assisted Breeding Values with Progeny-Test Breeding Values for Eight Hundred Ninety-Nine French Holstein Bulls**. J. Dairy. Sci. 91:2520-2522.

Habier, D., R. L. Fernando, et D. J. Garrick. 2013. **Genomic BLUP decoded: a look into the black box of genomic prediction.** Genetics 194:597-607.

Habier, D., R. L. Fernando, K. Kizilkaya, et D. J. Garrick. 2011. **Extension of the Bayesian alphabet for genomic selection**. BMC Bioinformatics 12:186.

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

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

Haile-Mariam, M., G. J. Nieuwhof, K. T. Beard, K. V. Konstatinov, et B. J. Hayes. 2013. **Comparison of heritabilities of dairy traits in Australian Holstein-Friesian cattle from genomic et pedigree data et implications for genomic evaluations**. J. Anim. Breed. Genet. 130:20-31.

Haley, C. S. et P. M. Visscher. 1998. **Strategies to Utilize Marker-Quantitative Trait Loci Associations**. J. Dairy. Sci. 81:85.

Hardy, G. H. 1908. **Mendelian proportions in a mixed population**. Science 28:49-50.

Hayes, B. J., P. J. Bowman, H. D. Daetwyler, J. W. Kijas, et J. H. J. Van der Werf. **Accuracy of genotype imputation in sheep breeds**. 2012. Anim. Genet. 43:72–80.

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

Hayes, B. J., H. D. Daetwyler, P. Bowman, G. Moser, B. Tier, R. Crump, M. Khatkar, H. W. Raadsma, et M. E. Goddard. 2009b. **Accuracy of genomic selection: comparing theory et results**. Proc. Assoc. Advmt. Anim. Breed. Genet. 2009;18:34–37

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

Hayes, B. J., et M. E. Goddard. 2008. **Technical note: Prediction of breeding values using marker-derived relationship matrices**. J. Anim. Sci. 86:2089–2092.

Hayes, B. J., A. J. Chamberlain, H. McPartlan, I. Macleod, L. Sethuraman, et M. E. Goddard. 2007. **Accuracy of marker assisted selection with single markers and marker haplotypes in cattle**. Genet. Res. 89:215–220.

Hayes, B. et M. E. Goddard. 2001. **The distribution of the effects of genes affecting quantitative traits in livestock.** Genet. Sel. Evol.33:209-230.

Henderson, C. R. 1975. **Best linear unbiased estimation et prediction under a selection model**. Biometrics 31:423-447.

Henderson, C. R. 1976. **Simple method for computing inverse of a numerator relationship matrix used in prediction of breeding values**. Biometrics 32:69-83.

Hickey J.M., B. P. Kinghorn, B. Tier, J. F. Wilson, N. Dunstan, J. H. J. Van der Werf: **A combined long-range phasing et long haplotype imputation method to impute phase for SNP genotypes**. 2011. Genet. Sel. Evol. 43:12.

Hill, W. G. et A. Robertson. 1968. **Linkage disequilibrium in finite populations**. Theor. Appl. Genet. 38:226-231.

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

Hosking, L., S. Lumsden, K. Lewis, A. Yeo, L. McCarthy, A. Bansal, J. Riley, I. Purvis, et C. F. Xu. 2004. **Detection of genotyping errors by Hardy-Weinberg equilibrium testing**. Eur J Hum Genet 12:395-399.

Howie B.N., P. Donnelly, et J. Marchini. **A flexible et accurate genotypeimputation method for the next generation of genome-wide association studies.** PLoS Genet 2009, 5:e1000529.

Hozé, C., S. Fritz, F. Phocas, D. Boichard, V. Ducrocq, et P. Croiseau. 2014a. **Efficiency of multi-breed genomic selection for dairy cattle breeds with different sizes of reference population**. J. Dairy. Sci. 97: 3918-3929

Hozé, C., S. Fritz, F. Phocas, D. Boichard, V. Ducrocq, et P. Croiseau. 2014b. **Genomic evaluation using combined reference populations from Montbéliarde et French Simmental breeds**. in Proc. 10th World Congr. Genet. Appl. Livest. 17-22 Août 2014, Vancouver (Canada), (soumis)

Hozé, C., M. N. Fouilloux, E. Venot, F. Guillaume, L. Journaux, D. Boichard, F. Phocas, S. Fritz, V. Ducrocq, et P. Croiseau. 2013. **High-density marker imputation accuracy in sixteen French cattle breeds**. Genet. Sel. Evol. 45:33.

Hozé, C., M. N. Fouilloux, E. Venot, F. Guillaume, L. Journaux, D. Boichard, F. Phocas, S. Fritz, V. Ducrocq, et P. Croiseau. 2012. **High density chip brings new opportunity for multibreed genomic evaluations in dairy cattle**. in 16th QTL MAS Workshop. 24-25 Mai 2012, Alghero, (Sardaigne, Italie). p.6

Interbull, **Résultats officiels des tests de validation des évaluations génomiques**. Disponible en ligne : http://www.interbull.org/web/web/article/gebvtest [16 Avril 2014]

Ibánez-Escriche, N., R. L. Fernando, A. Toosi, et J. C. Dekkers., 2009. **Genomic selection of purebreds for crossbred performance.** Genet. Sel. Evol. 41:12.

Ihara, N., A. Takasuga, K. Mizoshita, H. Takeda, M. Sugimoto, Y. Mizoguchi, T. Hirano, T. Itoh, T. Watanabe, et K. M. Reed. 2004. **A comprehensive genetic map of the cattle genome based on 3802 microsatellites**. Genome Research 14:1987-1998.

Jensen, J., E. A. Mantysaari, P. Madsen, et R. Thompson. 1996. **Residual maximum likelihood estimation of (co)variance components in multivariate mixed linear models using average information**. J. Ind. Soc. Agric. Stat. 49:215–236.

Karoui, S., M. J. Carabano, C. Diaz, et A. Legarra. 2012. **Joint genomic evaluation of French dairy cattle breeds using multiple-trait models**. Genet. Sel. Evol. 44:39

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

Khatkar, M. S., P. C. Thomson, I. Tammen, et H. W. Raadsma. 2004. **Quantitative trait loci mapping in dairy cattle: review et meta-analysis**. Genet. Sel. Evol. 36:163-190.

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

Kuhn, M. T., P. J. Boettcher, et A. E. Freeman. 1994. **Potential Biases in Predicted Transmitting Abilities of Females from Preferential Treatment**. J. Dairy. Sci. 77:2428-2437.

Lande, R. et R. Thompson. 1990. **Efficiency of marker-assisted selection in the improvement of quantitative traits**. Genetics 124:743-756.

Larmer, S. G., M. Sargolzaei, et F. S. Schenkel. 2014. **Extent of linkage disequilibrium, consistency of gametic phase, et imputation accuracy within et across Canadian dairy breeds**. J. Dairy Sci. (in press).

Le 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. 7:35.

Legarra, A., G. Baloche, F. Barillet, J. M. Astruc, C. Soulas, X. Aguerre, F. Arrese, L. Mintegi, M. Lasarte, F. Maeztu, I. Beltran de Heredia, et E. Ugarte. 2014. **Within- et across-breed genomic predictions et genomic relationships for Western Pyrenees dairy sheep breeds Latxa, Manech, et Basco-Bearnaise.** J. Dairy Sci. (in press).

Legarra, A., A. Ricard, et O. Filangi. 2013. **GS3-Genomic selection, Gibbs Sampling, Gauss Seidel et Bayes Cπ.** Disponible en ligne http://snp.toulouse.inra.fr/~alegarra [06 Novembre 2013]

Legarra, A. et V. Ducrocq. 2012. **Computational strategies for national integration of phenotypic, genomic, et pedigree data in a single-step best linear unbiased prediction**. J. Dairy Sci. 95:4629-4645.

Legarra, A., I. Aguilar, et I. Misztal. 2009**. A relationship matrix including full pedigree et genomic information**. J. Dairy Sci. 92:4656-4663.

Li Y, C. J. Willer, J. Ding, P. Scheet, et G. R. Abecasis, **MaCH: using sequence et genotype data to estimate haplotypes et unobserved genotypes**. Genet. Epidemiol. 2010, 34:816–834.

Lillehammer, M., T. H. Meuwissen, et A. K. Sonesson. 2011. **A comparison of dairy cattle breeding designs that use genomic selection**. J. Dairy Sci. 94:493-500.

Liu, Z., G. P. Aamand, S. Fritz, et C. Schrooten. 2013. **Comparison of national genomic prediction of EuroGenomics exchanged young bulls**. Interbull Bull. 47:38-42.

Liu, Z., F. R. Seefried, F. Reinhardt, S. Rensing, G. Thaller, et R. Reents. 2011. **Impacts of both reference population size et inclusion of a residual polygenic effect on the accuracy of genomic prediction**. Genet. Sel. Evol. 43:9.

Loberg, A. et J. W. Dürr. 2009. **Interbull survey on the use of genomic information**. Interbull Bull. 39:3-13.

Loberg, A., H. Jorjani, et J. W. Dürr. 2011. **Validation of genomic national evaluations**. Interbull Bull. 44:62-66.

Lund, M. S., A. P. Roos, A. G. Vries, T. Druet, V. Ducrocq, S. Fritz, F. Guillaume, B. Guldbrandtsen, Z. Liu, R. Reents, C. Schrooten, F. Seefried, et G. Su. 2011. **A common reference population from four European Holstein populations increases reliability of genomic predictions**. Genet. Sel. Evol. 43:43.

Makgahlela, M. L., E. A. Mantysaari, I. Strandén, M. Koivula, U. S. Nielsen, M. J. Sillanpaa, et J. Juga. 2013. **Across breed multi-trait random regression genomic predictions in the Nordic Red dairy cattle**. J. Anim. Breed. Genet. 130:10-19.

Malécot, G. 1948. **Les mathématiques de l'hérédité**. Masson & Cie, Paris. 60p.

Marchini, J. et B. Howie. 2010. **Genotype imputation for genome-wide association studies**. Nat Rev Genet 11:499-511.

Matukumalli, L. K., C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan, M. P. Heaton, J. O'Connell, S. S. Moore, T. P. L. Smith, et T. S. Sonstegard. 2009. **Development et characterization of a high density SNP genotyping assay for cattle**. PloS one 4(4):e5350.

Mc Hugh, N., T. H. E. Meuwissen, A. R. Cromie, et A. K. Sonesson.2011. **Use of female information in dairy cattle genomic breeding programs**. J. Dairy Sci. 94:4109–4118.

Meuwissen, T. H. E., et M. E. Goddard. 2010. **Accurate prediction of genetic values for complex traits by whole-genome resequencing**. Genetics 185:623–631.

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., A. Karlsen, S. Lien, I. Olsaker, et M. E. Goddard. 2002. **Fine mapping of a quantitative trait locus for twinning rate using combined linkage et linkage disequilibrium mapping**. Genetics 161:373-379.

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

Meuwissen, T. H. et M. E. Goddard. 1997. **Estimation of effects of quantitative trait loci in large complex pedigrees**. Genetics 146:409-416.

Meyer, K., B. Tier, et H. U. Graser. 2013. **Technical note: updating the inverse of the genomic relationship matrix**. J. Anim. Sci. 91:2583-2586.

Minvielle, F. 1990. **Principes d'amélioration génétique des animaux domestiques. Institut national de la recherche agronomique**; Québec: Presses de l'Université Laval. 211p.

Misztal, I., A. Legarra, et I. Aguilar. 2009. **Computing procedures for genetic evaluation including phenotypic, full pedigree, et genomic information**. J. Dairy Sci. 92:4648-4655.

Moreau, L., A. Charcosset, F. Hospital, et A. Gallais. 1998. **Marker-assisted selection efficiency in populations of finite size**. Genetics 148:1353-1365.

Muir, B., B. Van Doormaal, et G. Kistemaker. **International genomic co-operation—North American perspective**. Interbull Bull. 41:71-76.

Mulder, H. A., M. P. L. Calus, T. Druet, et C. Schrooten. 2012. **Imputation of genotypes with low-density chips et its effect on reliability of direct genomic values in Dutch Holstein cattle**. J. Dairy. Sci. 95:876-889.

Nilforooshan, M. A., B. Zumbach, J. Jakobsen, A. Loberg, H. Jorjani, et J. Dürr. 2010. **Validation of national genomic evaluations**. Interbull Bull. 42:56.

Olson, K. M., P. M. VanRaden, et M. E. Tooker. 2012. **Multibreed genomic evaluations using purebred Holsteins, Jerseys, et Brown Swiss**. J. Dairy. Sci. 95:5378-5383.

Pausch, H., B. Aigner, R. Emmerling, C. Edel, K.-U. Götz, et R. Fries. 2013. **Imputation of high-density genotypes in the Fleckvieh cattle population**. Genet. Sel. Evol.45:3.

Peers, I. 1996. **Statistical Analysis for Education and Psychology Researchers: Tools for Researchers in Education and Psychology**. The Falmer Press, Londres, Angleterre. 411p.

Pryce, J. et H. Daetwyler. 2012. **Designing dairy cattle breeding schemes under genomic selection: a review of international research**. Anim. Prod. Sci. 52:107-114.

Pryce, J. et B. Hayes. 2012. **A review of how dairy farmers can use et profit from genomic technologies**. Anim. Prod. Sci. 52:180-184.

Pryce, J., B. Hayes, et M. E. Goddard. 2012. **Genotyping dairy females can improve the reliability of genomic selection for young bull s et heifers et provide farmers with new management tools**. in 38th ICAR Conference. Cork, Ireland. Disponible en ligne http:// www.icar.org/Cork_2012/Manuscripts/Published/Pryce%202.pdf. [15 Janvier 2014]

Pryce, J. E., M. E. Goddard, H. W. Raadsma, et B. J. Hayes. 2010. **Deterministic models of breeding scheme designs that incorporate genomic selection**. J. Dairy Sci. 93:5455-5466.

Pszczola, M., T. Strabel, H. A. Mulder, et M. P. L. Calus. 2012. **Reliability of direct genomic values for animals with different relationships within et to the reference population**. J. Dairy. Sci. 95:389-400.

Raboisson, D. **Evolution raciale du cheptel bovin français des années 1970 aux années 2000: analyse à partir des données des Recensements Généraux agricoles de 1979, 1988 et 2000.** Thèse d'exercice, Ecole Nationale Vétérinaire de Toulouse - ENVT, 2004, 172 p.

Reich, D. E., M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti, D. J. Richter, T. Lavery, R. Kouyoumjian, S. F. Farhadian, R. Ward, et E. S. Lander. **Linkage disequilibrium in the human genome**. Nature 2001, 411:199–204.

Regaldo, D., Minery, S., et Mattalia S. 2006. **Harmonisation of type traits in the Simmental/Montbéliard breeds**. Interbull Bull. 35: 136-140

Rincent, R., D. Laloe, S. Nicolas, T. Altmann, D. Brunel, P. Revilla, V. M. Rodriguez, J. Moreno-Gonzalez, A. Melchinger, E. Bauer, C. C. Schoen, N. Meyer, C. Giauffret, C. Bauland, P. Jamin, J. Laborde, H. Monod, P. Flament, A. Charcosset, et L. Moreau. 2012. **Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds**. Genetics 192:715-728.

Rincon G., K. L. Weber, A. L. Van Eenennaam, B. L. Golden, J. F. Medrano. 2011. **Hot topic: performance of bovine high-density genotyping platforms in Holsteins and Jerseys**. J. Dairy Sci. 94:6116–6121.

Robert-Granie, C., A. Legarra, et V. Ducrocq. 2011. **Basic principles of genomic selection**. INRA Prod. Anim. 24:331-340.

Ruane, J. et J. J. Colleau. 1995. **Marker assisted selection for genetic improvement of animal populations when a single QTL is marked**. Genet. Res. 66:71-83.

Sargolzaei, M., J. P. Chesnais, et F. S. Schenkel. 2011. **FImpute-an efficient imputation algorithm for dairy cattle populations**. J. Dairy Sci 94:421.

Schaeffer, L. R. 2006. **Strategy for applying genome-wide selection in dairy cattle**. J Anim Breed Genet 123:218-223.

Scheet, P. et M. Stephens. 2006. **A fast et flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes et haplotypic phase**. Am. J. Hum. Genet. 78:629-644.

Schrooten, C., R. Dassonneville, V. Ducrocq, R. F. Brondum, M. S. Lund, J. Chen, Z. Liu, O. Gonzalez-Recio, J. Pena, et T. Druet. 2014. **Error rate for imputation from the Illumina BovineSNP50 chip to the Illumina BovineHD chip**. Genet. Sel. Evol. 46:10.

Schrooten, C., H. Bovenhuis, J. A. van Arendonk, et P. Bijma. 2005. **Genetic progress in multistage dairy cattle breeding schemes using genetic markers**. J. Dairy Sci. 88:1569-1581.

Sigurdsson, A. et G. Banos. 1995. **Dependent Variables in International Sire Evaluations**. Anim. Sci. 45:209-217.

Solberg, T. R., A. K. Sonesson, J. A. Woolliams, J. Odegard, et T. H. Meuwissen. 2009. **Persistence of accuracy of genome-wide breeding values over generations when including a polygenic effect**. Genet. Sel. Evol. 41:53.

Spelman, R. et D. Garrick. 1997. **Utilisation of marker assisted selection in a commercial dairy cow population**. Livest. Prod. Sci. 47:139-147.

Spelman, R. J., D. J. Garrick, et J. A. M. van Arendonk. 1999. **Utilisation of genetic variation by marker assisted selection in commercial dairy cattle populations**. Livest. Prod. Sci. 59:51-60.

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

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. 85:2358–2367

Stephens, M., N. J. Smith, et P. Donnelly. 2001. **A new statistical method for haplotype reconstruction from population data.** Am J Hum Genet 68:978-989.

Su, G., R. F. Brondum, P. Ma, B. Guldbrandtsen, G. P. Aamand, et M. S. Lund. 2012. **Comparison of genomic predictions using medium-density (~ 54,000) et high-density (~ 777,000) single nucleotide polymorphism marker panels in Nordic Holstein et Red Dairy Cattle populations**. J. Dairy. Sci. 95:4657-4665.

Sun, C., X. L. Wu, K. A. Weigel, G. J. Rosa, S. Bauck, B. W. Woodward, R. D. Schnabel, J. F. Taylor, et D. Gianola. 2012. **An ensemble-based approach to imputation of moderate-density genotypes for genomic selection with application to Angus cattle**. Genet. Res. 94:133-150.

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

The Bovine HapMap, C. 2009. **Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds**. Science 324:528-532.

Thomasen, J. R., C. Egger-Danner, A. Willam, B. Guldbrandtsen, M. S. Lund, et A. C. Sørensen. 2014. **Genomic selection strategies in a small dairy cattle population evaluated for genetic gain et profit**. J. Dairy. Sci. 97:458-470.

Thomasen, J. R., B. Guldbrandtsen, G. Su, R. F. Brondum, et M. S. Lund. 2012. **Reliabilities of genomic estimated breeding values in Danish Jersey**. Animal 6:789-796.

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

Tibshirani R. 1996. **Regression shrinkage et selection via the Lasso**. J. R. Stat. Soc. Ser.B-Methodol. 58:267-288.

Toosi, A., R. L. Fernando, et J. C. Dekkers. 2010. **Genomic selection in admixed et crossbred populations**. J Anim Sci 88:32-46.

Van den Berg, I., B. Guldbrandtsen, C. Hozé, R. F. Brøndum, D. Boichard, et M. S. Lund. 2014. Across Breed QTL **Detection et Genomic Prediction in French et Danish Dairy Cattle Breeds**. in Proc. 10th World Congr. Genet.Appl. Livest. 17-22 Août 2014, Vancouver (Canada), (soumis)

Van den Berg, I., S. Fritz, et D. Boichard, 2013, **QTL fine mapping with Bayes C(π): a simulation study**. Genet. Sel. Evol.45:19

VanRaden, P. M., D. J. Null, M. Sargolzaei, G. R. Wiggans, M. E. Tooker, J. B. Cole, T. S. Sonstegard, E. E. Connor, M. Winters, J. B. van Kaam, A. Valentini, B. J. Van Doormaal, M. A. Faust, et G. A. Doak. 2013. **Genomic imputation et evaluation using high-density Holstein genotypes**. J Dairy Sci 96:668-678.

VanRaden, P. M., J. R. O'Connell, G. R. Wiggans, et K. A. Weigel. 2011. **Genomic evaluations with many more genotypes.** Genet. Sel. Evol 43:10.

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

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

Villumsen, T. M., L. Janss, et M. S. Lund. 2009. **The importance of haplotype length et heritability using genomic selection in dairy cattle.** J Anim Breed Genet 126:3-13.

Wiggans, G. R., T. A. Cooper, P. M. Vanraden, K. M. Olson, et M. E. Tooker. 2012**. Use of the Illumina Bovine3K BeadChip in dairy genomic evaluation**. J Dairy Sci 95:1552-1558.

Wiggans, G. R., T. A. Cooper, P. M. Vanraden, et J. B. Cole. 2011. **Technical note: adjustment of traditional cow evaluations to improve accuracy of genomic predictions**. J Dairy Sci 94:6188-6193.

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

Wright, S. 1921. **Correlation and causation**. Journal of agricultural research 20:557-585.

Wright, S. 1931. **Evolution in mendelian population**. Genetics 16:97-159.

Zhang, Z., et T. Druet. 2010. **Marker imputation with low-density marker panels in Dutch Holstein cattle**. J Dairy Sci 2010, 93:5487–5494.

Zou, H. et T. Hastie. 2005. **Regularization et variable selection via the elastic net**. J. R. Stat. Soc. 67:301-320.

Liste des tableaux

Tableau 1 : Les composantes du coefficient de détermination en évaluation génétique d'après (Minvielle, 1990)..... 17

Tableau 2 : Nombre de taureaux testés sur descendance par année d'après Boichard et al. (1996) et France Génétique Elevage..... 20

Tableau 3 : Corrélations entre valeurs génomiques estimées selon différentes approches et déviations moyennes des filles (DYD) en race Holstein pour les animaux de la population de validation, d'après Boichard et al. (2012b)..... 58

Tableau 4 : Utilisation des jeunes taureaux et des taureaux testés sur descendance en 2013 dans les trois grandes races laitières françaises, d'après S. Moureaux (Institut de l'Elevage, Jouy en Josas, communication personnelle)..... 59

Tableau 5 : Liste des pays et des races disposant d'évaluations génomiques. Les évaluations génomiques validées par Interbull sont indiquées en souligné..... 60

Tableau 6 : Répartition des génotypages réalisés dans le cadre du projet GEMBAL en fonction de la race et du type (laitier ou allaitant) de l'animal..... 64

Tableau 7 : Récapitulatif des stratégies envisageables pour la mise en place d'évaluations génomiques dans les races régionales..... 114

Tableau 8 : Evolution de la consanguinité et du progrès génétique annuel en fonction du schéma de sélection utilisé, d'après Lillehammer et al. (2011)..... 117

Liste des figures

Figure 1 : Schéma de sélection avec testage sur descendance d'après S. Fritz, UNCEIA (communication personnelle) 20

Figure 2 : Transmission d'une association entre une mutation et un marqueur d'un individu fondateur à ses descendants. 24

Figure 3 : Déséquilibre de liaison (r^2) moyen en fonction de la distance entre loci dans les principales races bovines françaises, d'après Hozé et al. (2012) 25

Figure 4 : Schématisation des méthodes d'imputation d'après Marchini et Howie (2010) 37

Figure 5 : Précision des évaluations génomiques en fonction de la méthode et de la proportion de segments chromosomiques ayant un effet sur le caractère, d'après Daetwyler et al. (2010) 45

Figure 6 : Impact de la taille de la population de référence, de l'héritabilité (h^2) et de la taille efficace (Ne) de la population sur la précision des évaluations génomiques, d'après Goddard et Hayes (2009). 50

Figure 7 : Evolution de la précision des valeurs génomiques estimées en fonction du nombre de générations séparant la population de référence et les candidats à la sélection, d'après Bastiaansen et al. (2012) 53

Figure 8 : Evolution de la précision des valeurs génomiques estimées en fonction du nombre de générations séparant la population de référence et les candidats à la sélection pour des densités en marqueurs égale à 8, 4, 2 ou 1 fois la taille efficace de la population (ici Ne=100) d'après Solberg et al., (2009) 54

Figure 9 : Evolution de la précision des valeurs génomiques estimées en fonction du nombre de générations séparant la population de référence et les candidats à la sélection et de la longueur de l'haplotype (ici constitué de 1, 2, 5, 10, 20 ou 40 marqueurs) d'après Villumsen et al. (2009) 55

Figure 10 : Corrélations de Pearson entre déséquilibres de liaison (r) à différentes distances. Les barres verticales correspondent à la distance moyenne entre deux marqueurs sur la puce 50K (trait plein) ou HD (trait pointillé), d'après Larmer et al. (2014). 63

Figure 11 : Dendrogramme tracé à partir des distances génétiques (estimées à partir des génotypes haute densité) entre les populations du projet GEMBAL, d'après D. Laloë, INRA, Jouy en Josas (communication personnelle) 104

Figure 12 : Schéma de sélection simulé dans le cas de l'étude de Lillehammer et al. (2011) 117

Figure 13 : Schéma de sélection « hybride » simulé dans le cas de l'étude de Thomasen et al., (2014) 118

Figure 14 : Progrès génétique relatif annuel en fonction de la proportion de mères à taureaux et la proportion de femelles commerciales inséminées avec des jeunes taureaux d'après Thomasen et al. (2014)..... 119

Figure 15 : Schéma de sélection en race Abondance, d'après N. Bloc, UCEAR (communication personnelle)..... 122

Liste des publications

A. Articles scientifiques

C. Hozé, M. N. Fouilloux, E. Venot, F. Guillaume, R. Dassonneville, S. Fritz, V. Ducrocq, F. Phocas, D. Boichard, et P. Croiseau, 2013, **High-density marker imputation accuracy in sixteen French cattle breeds.** Genet. Sel. Evol. 45:33

C. Hozé, S. Fritz, F. Phocas, D. Boichard, V. Ducrocq, et P. Croiseau, 2014, **Efficiency of multi-breed genomic selection for dairy cattle breeds with different sizes of reference population.** J. Dairy Sci. 97 : 3918-3929

B. Communications à des congrès internationaux

C. Hozé, M. N. Fouilloux, E. Venot, F. Guillaume, R. Dassonneville, L. Journaux, D. Boichard, F. Phocas, S. Fritz, V. Ducrocq, et P. Croiseau, 2012, **From the 50K chip to the HD: Imputation efficiency in 9 French dairy cattle breeds.** In 63rd Annual EAAP Meeting. 27–31 Août-2012, Bratislava (Slovénie), Comm. n° 42.8

C. Hozé, M. N. Fouilloux, E. Venot, F. Guillaume, R. Dassonneville, L. Journaux, D. Boichard, F. Phocas, S. Fritz, V. Ducrocq, et P. Croiseau, 2012, **High density chip brings new opportunities for multi-breed genomic evaluations in dairy cattle.** In 16th QTL MAS Workshop. 24-25 Mai 2012, Alghero, (Sardaigne, Italie), p.6

C. Hozé, S. Fritz, D. Boichard, V. Ducrocq, et P. Croiseau, 2013, **Comparison of genomic selection approaches for small breeds**, In 30th Interbull meeting 24-26 Août 2013, Nantes (France), Interbull Bull. 47, 90-94.

C. Hozé, S. Fritz, D. Boichard, V. Ducrocq, et P. Croiseau, 2014, **Genomic evaluation using combined reference populations from Montbéliarde et French Simmental breeds**, In 10thWorld Congr. Genet. Appl. Livest. 17-22 Août 2014, Vancouver (Canada), (soumis).