# Evaluation of haplotype-based genomic selection methods with focus on their performances in a multi-breed context in dairy cattle

David Jonas

NNT : 2016SACLA026

# THESE DE DOCTORAT
DE
L'UNIVERSITE PARIS-SACLAY
PREPAREE A
AGROPARISTECH (L'INSTITUT DES SCIENCES ET INDUSTRIES DU VIVANT ET DE L'ENVIRONNEMENT)

ÉCOLE DOCTORALE N°581
Agriculture, alimentation, biologie, environnement et santé

Spécialité de doctorat : Génétique Animale

Par

**M. Dávid Jónás**

Evaluation of haplotype-based genomic selection methods with focus on their performances in a multi-breed context in dairy cattle

**Thèse présentée et soutenue à Paris, le 12 décembre 2016 :**

**Composition du Jury :**

| | | |
|---|---|---|
| M. Etienne VERRIER | Professeur, AgroParisTech, Paris, FR | Président |
| M. Esa MÄNTYSAARI | Professor, Natural Resources Institute, Helsinki, FI | Rapporteur |
| M. Antonio REVERTER-GÓMEZ | Res. scientist, QLD Biosci. Precinct, AUS | Rapporteur |
| M. Hervé CHAPUIS | Ingénieur de recherche, INRA-GenPhySE, Toulouse, FR | Examinateur |
| M. Laurent SCHIBLER | Responsable dév. et innov., ALLICE, Paris, FR | Examinateur |
| M. Vincent DUCROCQ | Directeur de recherche, INRA GABI, Jouy en Josas, FR | Directeur de thèse |

**Doctoral School**

| | |
|---|---|
| **Name:** | **ABIES Doctoral School** |
| **University:** | **Paris Institute of technology for life, food and environmental sciences** |
| **Discipline:** | **Animal genetics** |


**Thesis supervisors:**

**Dr. Vincent Ducrocq**
Director of research
French National Institute for Agricultural Research
Animal Genetics and Integrative Biology Unit
Bovine Genetics & Genomics Research Group

**Dr. Pascal Croiseau**
Research scientist
French National Institute for Agricultural Research
Animal Genetics and Integrative Biology Unit
Bovine Genetics & Genomics Research Group

# Abstract

Genomic evaluation exploits DNA marker information for selection purposes in breeds with agricultural importance. The majority of the available genomic evaluation methods today rely on SNP information, although it is hypothesized that haplotypes would perform better due to their higher polymorphism. Genomic evaluation was not implemented in regional dairy cattle breeds as of 2014, resulting in serious economical disadvantages for these breeds, urging breeders and scientists to address the issue. Our main aim was to evaluate haplotypes in genomic evaluation with focus on their performance in combination with multi-breed reference populations, which is an appealing way to enlarge the otherwise small reference populations of regional breeds.

The performance of haplotypes compared to SNP was assessed in a large dairy cattle breed. The higher performance of haplotypes was confirmed and haplotypes outperformed the SNP-based analyses in all scenarios. Furthermore, we also tested the hypotheses that information on allele frequency and on linkage pattern along the chromosomes are both relevant in marker selection for genomic evaluation purposes. After the development and assessment of two haplotype selection criteria capable of incorporating these information, we could prove that these hypotheses are valid and the efficiency of genomic evaluation methods can be improved using haplotypes. In addition, the developed haplotype selection criteria also allowed the reduction of the number of markers used in the prediction process by a significant proportion.

Out of these two criteria, the higher performing one was incorporated in the French routine genomic evaluation in 2015. The performance of this evaluation in the regional breeds was assessed and possible ways of improvements were implemented and evaluated. As a result of the sufficiently high performance of the French routine evaluation in the regional breeds, genomic selection was officially implemented in these breeds in 2016. The use of the bovine high-density SNP-chip did not improve the performance of genomic evaluation in these breeds, while multi-

breed training populations were only partially beneficial. On the other hand, genotyping females led to notable increases in selection accuracies. Inclusion of candidate mutations identified in large breeds also led to a small improvement in these breeds.

**Keywords**: dairy cattle, genomic evaluation, multi-breed, haplotype, haploblock

# Résumé

En sélection génomique, des marqueurs de l'ADN sont utilisés pour l'estimation des valeurs génétiques. La sélection génomique a été mise en placedans les trois grandes races (inter)nationales (Montbéliarde, Normande et Holstein) en 2014 en utilisant les données SNP de la puce 50K et elle a entraîné une augmentation significative (~2 fois plus) du progrès génétique annuel dans les caractères sélectionnés. Pour les races dites régionales, le nombre de taureaux testés est trop restreint pour permettre la constitution d'une population de référence suffisamment grande. Le manque d'évaluation génomique chez les races régionales – étant donné qu'elle a été mise en pratique dans les grandes races – place les races régionales dans un sérieux désavantage économique.

La plupart des méthodes d'évaluation génomique utilisées depuis 2014utilisent les SNP comme marqueurs de l'ADN, bien que les haplotypes (combinaisons de N SNP) soient plus informatifs en raison de leur polymorphisme plus élevé. En outre, unepuceHaute Densité (HD) est disponible chez les bovins depuis 2011 en plus de la puce 50K. Malgré les attentes initiales, aucune amélioration significative n'a été pas observée avec la puce HD par rapport au puce 50K.

Dans une première étude, nous avons évalué les avantages de l'utilisation des haplotypes dans l'évaluation génomique. Nous avons également évalué l'utilisation des haplotypes en combinaison avec lapuce HD dans l'évaluation génomique. Toutefois, le nombre d'effets de marqueur à estimer dans le modèle rend cette analyse difficile. En effet, en utilisant la puce HD, entre 1 et 2,3 millions d'effets sont à estimer avec des haplotypes de 2 à 5 SNP ce qui est bien trop complexe pour un modèle d'évaluation génomique. Par conséquent, nous avons également dû réduire le nombre des haplotypes utilisés dans les modèles.

De plus, nous avons également contribué à la mise en place d'une méthode d'évaluation génomique efficace pour les races régionales. Afin d'augmenter la taille de la population de référence et donc de maximiser la performance d'évaluation

génomique dans ces races, les vaches avecdes performances enregistrées ont été génotypées en plus des taureaux testés. Avec ces populations de référence mixtes, nous avons évalué la performance des méthodes d'évaluation génomique disponibles dans les races régionales. En outre, nous avons également évalué plusieurs façons prometteuses d'améliorer la performance des évaluations génomiques dans les races régionales.Ainsi, l'utilisation la puce HD, lespopulations de référence multi-raciales (c'est-à-dire des populations de référence comprenant des animaux de plus d'une seul race), l'utilisation d'information de mutation candidate ou d'information de haploblock (c'est-à-dire exploitant l'information de déséquilibre de liaison entre des SNP) ont été évaluées.

Pour cette analyse, cinq races ont été utilisées : Une grande race bovine laitière française (la Montbéliarde) a été utilisée pour l'évaluation des nouvelles méthodes qui utilise des haplotypes (voir ci-dessous). La population de référence de cette race incluait 2235 taureaux testés. Par ailleurs, les quatre races laitières régionales suivantes étaient disponibles également: Abondance, Tarentaise, Simmental et Vosgienne. La population de référence de ces races incluait des mâles et des femelles. La taille de la population de référence– en nombre des taureaux testés – variait entre 348 et 767 en 2015. Ces effectifs ont été revus à la hausse en2016, ce qui a porté la population de référence à 575-1593 animaux. En fonction de la race, entre 34 et 40 caractères sont disponibles dont 5 caractères de production laitières (quantité du lait, matière grasse, matière protéine, taux butyreux et taux protéique). Les observations de performance disponibles ont été converties en 'daughter yield deviations' (DYD) pour les mâles et en 'yield deviations' (YD) pour les femelles avant les analyses. Les animaux intégrés à cette analyse ont tous été génotypés soit en LD, 50K ou HD. Des travaux d'imputation (prédiction des génotypes) ont été menés et ont permis d'avoir un génotype HD (imputé ou réel) pour l'ensemble des animaux disponibles. Ainsi, les tests d'évaluation génomique ont pu être réalisé avec différentes densités de puce. Environ 3000 mutations candidates ont été génotypées dans les races Abondance, Tarentaise et Vosgienne et ont donc pu être également exploitées.

Tous les tests ont été réalisés dans le cadre d'études de validation classiques avec les 20% plus jeunes animaux dans la population de validation et les 80% restant dans la population d'apprentissage. Dans le cas des races régionales, les animaux de la population de validation étaient exclusivement des femelles. Mesurée sur la population de validation, les coefficients de corrélation entre (D)YD et GEBV ainsi que les pentes de régression de (D)YD sur GEBV ont été utilisés pour évaluer la performance de chaque densité de puce et de chaque méthode .

Une application de BayesC-π capable d'utiliser des haplotypes au lieu des SNP individuels a été développée et évaluée. Deux critères légèrement différents ont été également développés afin de réduire le nombre de marqueurs utilisés dans lesévaluations génomiques. Ces critères ont pour but de sélectionner l'haplotype avec les meilleures propriétés de fréquence alléliqueau sein d'une région donnée. Ces deux critères comptent uniquement sur l'information de fréquence allélique: le premier (que nous appelons Critère-A) maximise le nombre d'allèles dont la fréquence allèlique est supérieure à un seuil défini par l'utilisateur, tandis que le deuxième critère (Critère-B) met plus d'accent sur l'équilibre entre les fréquences alléliqueet le nombre d'allèle afin de maximiser le nombre d'allèles avec une fréquence suffisamment élevée pour pouvoir permettre l'estimation d'effet d'allélique.

Une des faiblesses de la méthode précédemment décrite est l'exigence de la connaissance préalable de laposition desrégions QTL. Afin de contourner cette condition, nous avons découpé le génome en régions au sein desquelles le déséquilibre de liaison est élevé (haploblock). Au sein de ces régions, tous les marqueurssont en fort LD avec tous les autres SNP de la même région ce qui signifie que ces régions sont héritées de génération en génération. La sélection d'un haplotype pour représenter chacun de ces haploblock ne nécessite pas une étape de détection QTL antérieure. L'utilisation de ces haploblocks avec les critères de sélection d'haplotype décrits précédemment permet de (1) réduire davantage le nombre d'haplotypes dans le modèle et (2) d'améliorer la précision de la sélection.

La performance de l'évaluation génomique de routine française a été évaluée chez les races régionales qui –depuis 2015 – incorporaient la méthode de sélection

Criterion-B. En outre, les avantages possibles en raison d'addition des mutations candidates ont été également évalués avec BayesC et BayesR en même temps.

Des évaluations multi-raciales ont été réalisées en fusionnant la population d'apprentissage des races régionales. L'étape de validation de ces études a été maintenues dans un contexte intra-race, parce qu'ilnous a permis une comparaison facile entre des résultats multi-raciauxet des résultats intra-race. Les populations d'apprentissage multi-raciales ont été formées en incluant les 4 races régionales ou la combinaison de 2 ou 3 races seulement. Au total, 11 scénarios multi-raciaux différents ont été testésavec l'utilisation de la puce 50K et HD.

Nous avons pu démontrer que les haplotypes étaient plus performant que les SNP ensélection génomique (+ 2% en coefficients de corrélation en moyenne pour les 5 caractères de production). Nous avons également pu montrer que l'information de fréquence alléliques et l'étendu du déséquilibre de liaison sont importants pour une construction optimale des haplotypes. Les deux critères nous avons proposé pour la sélection des haplotypes ont permis d'augmenter la précision de sélection de 0,7-0,9% en moyenne sur les 5 caractères de production. Lorsque la sélection d'haplotypes a été conjointement utilisée avec l'information de blocs haplotypiquesbasée sur le LD, une augmentation supplémentaire de 1,5% est observée. Dans nos analyses, le Critère-B s'est montré plus performant que le Critère-A. En outre, par rapport aunombre total d'haplotypes consécutifs, le nombre d'haplotypes pourrait être réduit de ~26% et ~90% respectivement avec les puces 50K et HD, lorsque les haploblocks et les critères de sélection sont utilisés simultanément.

Le Critère-B a été includans les évaluations génomiquesofficielles en France en 2015. La performance de cette évaluation a été ensuite évaluée dans les quatre races régionales. Cesanalyses ontabouti, pour les taureaux testés sur descendance, à des précisions au moins semblable à celles obtenus sous un modèle polygénique (sans information de génotypage). Par conséquent, une évaluation génomique a été mise en pratique dans ces races en 2016. En comparant les résultats obtenus en 2015 et 2016, on pourrait conclure que le génotypage d'individus supplémentaires

(principalement des femelles) était avantageux dans les races régionales (augmentation de 4 à 7% des coefficients de corrélation entre les valeurs de YD et de GEBV dans la population de validation).

L'addition de l'information de mutation candidate aux données ordinaires de 50K n'a pas permis d'améliorer notre modèle. En termes de précisions de la sélection, BayesC a généré une augmentation moyenne de 0,5% (moyenne sur les 5 traits de production), tout comme leBayesR(+0,3%). En termes de biais de sélection, aucune amélioration significative n'a pas été observée avec l'inclusion des mutations candidates.

L'utilisation de génotypes haute densité n'a pas amélioré la performance de l'évaluation génomique dans les racesévaluées, alors que la formation des populations multi-raciales ne sontbénéfiques que pour certaines d'entre elles.

L'utilisation d'une population multi-raciale a été avantageuse dans les races Abondance (+5,8% en corrélation entre YD et GEBV en moyenne pour les 5 traits de production) et Simmental (+ 5,4%), mais a été désavantageusepour la Tarentaise (-3%) et la Vosgienne (-2,5%). Plusieurs auteurs ont suggéré que la puce HD seraient nécessaires pour les évaluations multi-raciales, en raison de la diminution du déséquilibre de liaison (LD) entre les marqueurs et QTL, lorsqu'on utilise une population de référence multi-raciale. Cependant, ces populations de référence sonttoujours génétiquement plus distante que lapopulationde référence d'une seule race et, dans notre cas, l'utilisation de la puce HD dans un contexte multi-racial n'a pas amélioré l'efficacité de l'évaluation.

Au vu de ces résultats, une évaluation génomique officielle a été mise en placedans trois races régionales : Abondance, Tarentaise et Vosgienne. Pour la Simmental, une population de référence internationale, plus grande, est aussi disponible. Ainsi, cela permet une plus grande précision de sélection et un biais plus faible par rapport à ceux que nous pouvons fournir.

L'arrivée desévaluations génomiques dans ces races devrait également avoir un impact positif sur la biodiversité : auparavant ~5-20 taureaux étaient testés sur

descendance (en fonction la race) et seulement une fraction de ces taureaux devenait reproducteur. Toutefois, à partir de 2016, les organismes de sélection visent à évaluer entre 50-150 taureaux avec une utilisation de ces taureaux plus homogène. Plus le nombre de taureaux reproducteurs augmentera et plus la taille efficace de la population de ces races augmentera également, ce qui facilitera la gestion de la population et la préservation des races.

Lescoefficient de déterminationobtenus avec la sélection génomique dans ces races est similaire à ceux obtenus sous un modèle polygénique. Toutefois, les GEBV sont disponiblespour un plus grand nombre d'animaux et à la fois pour les mâle et les femmes. Cela facilite et accélère le processus de sélection pour ces races. Ainsi, d'après nos estimations, on s'attend à ce que le gain génétique annuel soit multiplié par 3 dans les races régionales, comparativement au programme de testage sur descendants. Cependant, il sera toujours inférieur par rapport au progrès génétique annuel observé chez les grandes races laitières.

Nous avons également fourni des preuves empiriques de la supériorité des haplotypes sur les SNP individuels dans les modèles d'évaluation génomique. En outre, nous avons prouvé qu'il est avantageux de considérer l'information de fréquence allélique et de LD lors de la sélection des marqueurs pour former les haplotypespour les évaluations génomiques. Notre méthode est particulièrement intéressante pour améliorer la précision de la sélection génomique, car elle n'a besoind'aucune information supplémentaire. Ces méthodes permettent une exploitation des données disponibles plus pertinente.

# Acknowledgement

I would like to express my gratitude to my supervisors, Pascal Croiseau and Vincent Ducrocq, without whom this work could not have been done. I also thank the help of everyone from the $G^2B$ group of INRA/GABI, who helped me at some stage during my studies and especially to Sébastien Fritz, Didier Boichard and Marie-Pierre Sanchez. I also would like to acknowledge the help provided by the external members of my thesis committee meeting: Andres Legarra, Tristan Mary-Huard and Etienne Verrier. I am especially grateful to my friends (from France and abroad) and family, who supported me during the past 3 years. I thank you all for your help and encouraging that guided me during my PhD studies.

Finally, I would like to thank for the financial support of INRA, ALLICE and ANRT, which organizations guaranteed funding for my PhD studies.

# List of abbreviations

**50K (SNP-chip)**: Bovine 50K SNP panel

**(%)p. a.**: in/for each year (from latin per annum)

**AEGIS**: European genebank integrated system

**AFT**: Allele frequency threshold

**AI**: Artificial insemination

**BLP**: Best linear prediction

**BLUE**: Best linear unbiased estimate

**BLUP**: Best linear unbiased prediction

**bp**: Base pair

**CD**: Coefficient of determination

**cM**: centiMorgan

**DGAT1**: Diacylglycerol O-acyltransferase-1

**DNA**: Deoxyribonucleic acid

**DYD**: Daughter yield deviation

**EDC**: Equivalent daughter contributions

**EBV**: Estimated breeding value

**EFABIS**: European farm animal biodiversity information system

**GBLUP**: Genomic BLUP

**GEBV**: Genomic estimated breeding value

**HD (SNP-chip)**: Bovine high-density (777K) SNP panel

**$h^2$**: Heritability

**HS**: Haplotype size

**HWE**: Hardy-Weinberg equilibrium

**K**: Thousand

**Kb**: Thousand base pairs

**LD**: Linkage disequilibrium

**LD (SNP-chip)**: Bovine low-density (10-20K) SNP panel

**MA-BLUP**: Marker-assisted BLUP

**MAF**: Minor allele frequency

**Mb**: Million base pairs

**MD**: Maximum deviation

**QTL**: Quantitative trait loci

**QTL-SNP**: SNP in strong LD with QTL

**$R^2$**: Coefficient of determination

**RE**: Record equivalent

**RNA**: Ribonucleic acid

**SNP**: Single nucleotide polymorphism

**WGS**: Whole-genome sequencing

**WS**: Window size

**YD**: Yield deviation

# Table of contents

**Abstract** ..... **5**

**Résumé** ..... **7**

**Acknowledgement** ..... **13**

**List of abbreviations** ..... **14**

**Table of contents** ..... **16**

**List of Tables** ..... **19**

**List of Figures** ..... **22**

**Chapter 1 Introduction** ..... **27**

**Chapter 2 Background** ..... **30**

2.1 Characteristics of dairy cattle breeding ..... 30

2.2 Pedigree-based selection methods ..... 32

2.2.1 Best linear unbiased prediction ..... 33

2.2.2 Implementation in our study ..... 35

2.3 Genetic background of quantitative traits and genetic markers ..... 35

2.3.1 Quantitative trait loci ..... 35

2.3.2 Genetic markers ..... 36

2.3.2.1. Microsatellite ..... 36

2.3.2.2. Single nucleotide polymorphism ..... 36

2.3.3 Haplotype ..... 39

2.3.4 Imputation and phase reconstruction ..... 40

2.4 Genomic evaluation ..... 41

2.4.1 Marker-assisted BLUP ..... 42

2.4.2 Genomic-BLUP ..... 43

2.4.3 Bayesian methods ..... 44

2.4.4 Genomic evaluation methods with haplotype markers ..... 47

2.5 French routine genomic evaluation of dairy cattle ..... 50

2.6 Consequences of genomic selection ..... 53

2.6.1 Advantages of genomic selection ..... 53

2.6.2 Drawbacks of genomic evaluation ..... 56

2.7 Assessment of genomic evaluation studies ..... 57

2.7.1 Principles of validation in genomic evaluation studies ..... 57
2.7.2 Measured parameters ..... 58
2.8 Analyzed breeds and traits ..... 59
2.9 Single-breed and multi-breed genomic evaluation ..... 61
2.9.1 Review of the recent multi-breed genomic evaluation studies ..... 63
2.10 Problem statement and motivation ..... 66

**Chapter 3 Haplotype construction for genomic evaluation purposes ..... 68**
3.1 The Montbéliarde dataset ..... 69
3.2 Haplotypic BayesC-π results ..... 70
3.3 Influence of allele frequency on genomic evaluation ..... 73
3.3.1 Introduction ..... 73
3.3.2 Alternative haplotype construction methods for genomic evaluation 75
3.3.3 Discussion ..... 94
3.4 Genomic evaluation with HD data ..... 95
3.5 Inclusion of linkage disequilibrium information ..... 101
3.5.1 Introduction ..... 101
3.5.2 Combining LD and allele frequency information to improve selection accuracy ..... 102
3.5.3 Discussion ..... 118

**Chapter 4 Genomic evaluation in regional breeds ..... 121**
4.1 Datasets ..... 122
4.1.1 Genotyping and imputation ..... 123
4.2 LD-pattern in the regional breeds ..... 125
4.3 Genomic evaluation with 50K data ..... 127
4.3.1 Introduction ..... 127
4.3.2 Single-breed and multi-breed genomic evaluation with 50K data ..... 128
4.3.3 BayesC results ..... 144
4.3.4 Discussion ..... 145
4.4 Genomic evaluation with high-density data ..... 147
4.4.1 Introduction ..... 147
4.4.2 Materials and methods ..... 147
4.4.3 Results ..... 149
4.4.4 Conclusions ..... 150
4.5 Genomic evaluation with causative mutations ..... 151
4.5.1 Introduction ..... 151

4.5.2 Materials and Methods ..... 151
4.5.3 Results and discussion ..... 154
4.5.4 Conclusions ..... 157

**Chapter 5 General discussion ..... 159**
5.1 Introduction ..... 159
5.2 Biodiversity ..... 160
5.3 Effects of the slower genetic progress ..... 162
5.4 Perspectives for the regional breeds ..... 163
5.5 Genomic evaluation in the regional breeds ..... 168
5.6 Financial considerations ..... 175
5.7 Genomic evaluation with haplotypes ..... 177
5.8 Future perspectives ..... 179

**Chapter 6 Concluding remarks ..... 183**

**References ..... 186**

**Appendix A ..... 199**

**Appendix B ..... 200**

**Appendix C ..... 205**

**Appendix D ..... 207**

**Appendix E ..... 208**

**Publications and trainings ..... 214**

# List of Tables

**Table 1**: Number of progeny-tested bulls and number of cows under performance recording in the 5 breeds used through this Thesis. ..... 60

**Table 2**: Average standard 305-day production level of the 5 breeds used through this Thesis (data from 2015). ..... 61

**Table 3**: Number of consecutive, non-overlapping haplotypes that can be built with data from either the 50K or the HD SNP-chips and the number of allele effects to be estimated. ..... 71

**Table 4**: Correlation coefficients and regression slopes of DYD on GEBV values measured on the validation set with *haplotypic-GS3* (Croiseau et al., 2014)..... 73

**Table 5**: Average number of alleles per haplotype observed with the 3 different haplotype construction methods, as function of haplotype size and number of QTL-SNP in the model. Window size: 80 SNP. ..... 96

**Table 6**: Observed correlations in the validation set between DYD and GEBV values using either only the QTL-SNP or the flanking haplotypes as genomic markers. Average correlations over the 5 traits. ..... 98

**Table 7**: Average correlations calculated between DYD and GEBV of the validation set for 5 production traits (Criterion-B)..... 99

**Table 8**: Regression slopes with the 2 different haplotype construction methods and when only QTL-SNP were used as genetic markers. Values measured on the validation set and averaged over 5 traits. ..... 100

**Table 9**: Correlation coefficients and regression slopes of DYD on GEBV values of the validation population with a D' threshold of 45% or 90%. ..... 119

**Table 10**: Total number of genotyped or imputed males and females in the 4 regional breeds, as of either August 2015 or August 2016..... 122

**Table 11**: Number of monomorphic SNP on the different SNP-chips in the four regional breeds. ..... 124

**Table 12**: Average correlation coefficients and regression slopes (expressed as deviations from 1) of the 5 traits measured on the validation set from a BayesC and from the routine genomic evaluation. ..... 144

**Table 13**: Correlations and regression slopes between the DYD and GEBV in the 4 regional breeds. Average single-breed (SB) and multi-breed (MB) results with 50K are also added. ..... 150

**Table 14**: Number of imputed SNP and number of SNP retained from the LD SNP-chip after quality control. ..... 152

**Table 15**: Summary of the QTL groups used with BayesR. ..... 153

**Table 16**: Correlation coefficients obtained in the validation population with either BayesC or with BayesR (π=9%) using 50K SNP-chip information. ..... 155

**Table 17**: Number of genotyped young candidates and selected bull sires and bull dams during the first year after the implementation of genomic selection in the regional breeds. ..... 165

**Table 18**: Asymptotic annual genetic gain and different parameters affecting it in large breeds with genomic selection (GS) or in regional breeds with or without genomic selection (indicative values). ..... 166

**Table 19**: Number of females with one individual phenotype required to bring information equivalent to one male, according to heritability and male estimated breeding value (EBV) reliability based on progeny information only (Table 1 from Boichard et al., 2015) ..... 169

**Table 20**: Estimated reliabilities of selection candidates with the French routine evaluation (from Sanchez et al., 2016). ..... 172

**S. table 1**: Correlation coefficients and regression slopes of DYD on GEBV values obtained with the GBLUP analysis (Montbéliarde breed). ..... 199

**S. table 2**: Correlations between genomic estimated breeding values and DYD in the validation population for the scenario with an optimal number of QTL are presented. Window size: 80 SNP; Montbéliarde breed. ..... 203

**S. table 3**: Regression slopes of DYD on GEBV in the validation population for the scenario with an optimal number of QTL are presented. Window size: 80 SNP; Montbéliarde breed..... 204

# List of Figures

**Figure 1**: Average number of alleles when using consecutive haplotypes from either the 50K or from the HD SNP-chip with 4 different haplotype sizes (the theoretical maximum number of alleles (i.e. $2^N$) is also plotted). ..... 40

**Figure 2**: Probability density distributions of QTL effects in dairy cattle (after Hayes and Goddard, 2001; axis labels were removed since they are trait-dependent). ..... 47

**Figure 3**: Tree representing the genetic distances between 20 French cattle breeds. Genetic distances were estimated from allele frequencies using the bovine 50K SNP-chip (from Gautier et al., 2010). Breed name abbreviations: CHA – Charolais; PAR – Parthenaise; BPN – Bretonne Pie Noire; Noire – Normande; MAI – Maine Anjou (Rouge des prês); FLA – Flamande; PRP – Pie Rouge des Plaines [→Red Holstein]; HOL – Holstein; BRU – Brune; VOS – Vosgienne; TAR – Tarentaise; ABO – Abondance; PRE – Pie Rouge de l'Est (French Simmental); MON – Montbéliarde; BAZ – Bazadaise; GAS – Gasconne; SAL – Salers; AUB – Aubrac; LIM – Limousin; BLA – Blonde d'aquitaine. ..... 63

**Figure 4**: Convergence plots obtained with haplotypes of 4 SNP. Proportion of haplotypes without an effect (π), residual variance (vare), variance of a single locus (vara) and residual polygenic variance (varg) are plotted. The thinning value was 1000..... 72

**Figure 5**: Overall distribution of haplotype allele frequencies with either flanking or with Criterion-B selected haplotypes (haplotype size: 4 SNP; window size: 80 SNP; 6,000 QTL-SNP). The 0-10% region is also depicted with more detailed scale on the x-axis. ..... 97

**Figure 6**: Average observed correlations between DYD and GEBV values for 5 production traits with different haplotype selection methods and haplotype sizes. Solid lines indicate the correlations for the haplotype-based tests while dashed lines show the correlations observed when the same SNP were used but as single-SNP markers (Criterion-B; validation set). ..... 99

**Figure 7**: Distribution of the minor allele frequency in the regional breeds (MAF resolution: 1%)..... 124

**Figure 8**: Linkage disequilibrium decay in the single-breed contexts. ..... 126

**Figure 9**: Linkage disequilibrium decay in the multi-breed (MB) context (average of the 11 different multi-breed combinations (solid, black line); minimum/maximum of these combinations (dashed, black lines) and average of the four single-breed (SB) scenarios). ..... 127

**Figure 10**: Frequency distribution of the number of SNP from the HD SNP-chip overlapping with the 10 SNP-wide windows from the 50K SNP-chip (Montbéliarde breed). Trait name abbreviations: MY – milk yield; FY – fat yield; PY – protein yield; FC – fat content; PC – protein content..... 149

**Figure 11**: Effect of the inclusion of candidate mutations on the correlation between YD and GEBV measured on the validation population (BayesC). ..... 156

**Figure 12**: Average absolute deviation of regression slopes from 1 with either BayesC or BayesR and with the 50K and 50K+custom SNP-chip data..... 157

**Figure 13**: Illustration of the long-term effect of genomic selection (GS) on the production level of the regional and large breeds. ..... 168

**S. figure 1**: Frequency distribution of the distances between neighboring SNP from the (A) 50K and (B) HD SNP panels. Frequencies are calculated for every bins of 100 bp and 2500 bp for the HD and 50K SNP panels, respectively. ..... 201

**S. figure 2**: Overall distribution of haplotype allele frequencies according to the haplotype construction approach (haplotype size: 3 SNP; 6,000 QTL-SNP). The 0-10% region is also depicted with a more detailed scale on the x-axis. ..... 202

**S. figure 3**: Linkage disequilibrium decay in the multi-breed (2-breed) scenarios. Breed name abbreviations: A – Abondance; T – Tarentaise; S – Simmental ; V – Vosgienne..... 205

**S. figure 4**: Linkage disequilibrium decay in the multi-breed (3-breed) scenarios. Breed name abbreviations: A – Abondance; T – Tarentaise; S – Simmental ; V – Vosgienne..... 206

**S. figure 5**: Linkage disequilibrium decay in the multi-breed (4-breed) scenario. Breed name abbreviations: A – Abondance; T – Tarentaise; S – Simmental ; V – Vosgienne..... 206

**S. figure 6**: Effect of the inclusion of candidate mutations on the correlation between YD and GEBV measured on the validation population (BayesR). ..... 207

*The beginning of knowledge is the discovery of something we do not understand.*

**Frank Herbert**

# Chapter 1

# Introduction

Some of the most important challenges modern agriculture faces today are the fast human population growth (projected World population in 2050: 9.7 billion; current increase: +83 million/year; FAO, 2015), the expected freshwater shortage and the continuing decline of arable land in use per person (Alexandratos and Bruinsma, 2012). Livestock production is especially affected by these challenges, because it directly (for pastures) or indirectly (for feedcrop production) uses 70% of the World's agricultural lands (FAO, 2006). Furthermore, especially in Western countries, a shift can be observed in consumer expectations towards, for example, healthier products or higher animal welfare (e.g. Støier et al., 2016; Thaxton et al., 2016). Proper adaptation of animals to the technological conditions in modern farming systems (e.g. to milking machines in dairy cattle) as well as secondary traits with significant effects on animal production, such as stress resistance or resistance against infections and diseases are also of interest. Therefore, it is of great importance to develop sustainable and more efficient production systems in all fields of agriculture and especially in animal breeding.

The phenotypic characteristics of animals are determined by two major components: the genetic background (i.e. the DNA) of the animals and the environment in which

they produce. In order to successfully cope with the challenges agriculture must face in the foreseeable future, genetic improvement of livestock is crucial because it focuses on maximizing genetic gain in the long-term and therefore all future generations benefit from it. Genetic improvement in agronomically important species/breeds is obtained through artificial selection on economically important traits, such as milk production and udder health in dairy cattle, growth rate and stress resistance in pigs or number of eggs produced by laying hens. Traditional selection methods use phenotypic observations combined with pedigree information to estimate the genetic merit of selection candidates. However, recent biotechnological advances in molecular genetics and genomics (e.g. Bentley, 2006; Shen et al., 2005; applications in cattle: Matukumalli et al., 2009; Liu et al., 2009) allowed the development of genomic selection (e.g. Meuwissen et al., 2001) and its implementation in practice, particularly in dairy cattle breeding (for example in France: Croiseau et al., 2015b). These modern selection tools permit the direct utilization of information on DNA sequence variations in the selection process, leading to significant increases in annual genetic gain in the selected traits.

Genetic diversity is a key element of population management. Without genetic diversity, there is no chance for genetic improvement of animal populations. With a declining genetic diversity, populations (breeds or even whole species) can become endangered and in extreme cases might ultimately face extinction. For the same considerations, it is crucial to maintain the genetic diversity in agriculturally relevant species and breeds. Furthermore, preservation of regional breeds (see the definition in the next paragraph) is important as well because future production environments are unknown and therefore it is unknown which breeds could produce efficiently in the future. To support the preservation of regional breeds, their competitiveness has to be maintained. However, due to their smaller population size and to the less available funding, breeding programs are usually less efficient in these breeds.

Through this manuscript the term "regional breed" is used to denominate breeds, which are raised in a limited area, much smaller than the whole territory of France. A first category of regional breeds comprises native breeds with a small (e.g. the

Vosgienne with ~5,000 cows) to moderate (e.g. the Abondance with ~50,000 cows) current population size. A second category comprises breeds of foreign origin with a small-moderate population size in France, such as the Simmental Française or the Brown Swiss breeds (both with about 25000 cows).

Currently available genomic selection methods require large animal populations with both phenotype and genotype data in order to achieve high prediction accuracy (Goddard, 2009), which is a prerequisite for successful selection. However, these so called “reference populations” are limited for regional cattle breeds, which are characterized by a small population size and are bred only by a limited number of breeders. Breeders and breeding organizations of regional breeds are therefore in disadvantage with regard to genomic selection with the serious risk of increasing the gap between the genetic potential of these regional breeds compared to larger (inter)national breeds, in which genomic selection has already been implemented.

Currently there are numerous projects in our research group aiming to improve the efficiency of genomic selection in dairy cattle. One of these projects focuses on the development of efficient genomic selection methods for regional breeds in collaboration with breeding organizations representing four such French dairy cattle breeds. The primary aim of my PhD within this framework was to investigate the performance of state of the art genomic evaluation procedures in regional breeds and to develop new methods to improve the genomic selection efficiency in these breeds.

In particular, testing the efficiency of new tools such as haplotype markers, the BovineHD BeadChip® (HD; manufactured by Illumina Inc., San Diego, CA) and putative causative mutations in genomic selection were among our aims. Our long-term objective was to contribute to a new genomic evaluation procedure which is efficient in breeds with small reference populations. Practical implementation of the newly developed methods is made possible by the collaborations with breeding organizations.

# Chapter 2

# Background

The main objective of animal breeding is to genetically improve animal populations for economically important traits. The phenotypic performance of animals is affected by both genetic and environmental factors. Although the existence of genotype-by-environment interactions is currently actively studied – e.g. in Rauw and Gomez-Raya, 2015 – they are most often not taken into account as its removal simplifies the models without compromising the selection efficiency. In modern farming systems, both of the other two factors (i.e. the environmental conditions and the genetic background of the animals) are improved – independently from each other – in order to increase the production level of the animals. Genetic improvement of livestock is done by means of selection. In the following sections, we will introduce the main characteristics of selection in dairy cattle breeding as well as the fundamental basics of both classical and genomic selection procedures.

## 2.1 Characteristics of dairy cattle breeding

There are several key features of the dairy cattle industry which have major impacts on the applied breeding system. Firstly, all the production traits (e.g. milk yield, milk fat and protein content) and many other traits (e.g. udder health, milking speed, somatic cell count) can be measured only on females. Hence, own performances do

not exist in males for most of the economically important traits and selection of males must rely on information from female relatives. Secondly, a much larger proportion of the young female animals are required in order to keep the population size constant compared to the required proportion of males. Therefore, in dairy cattle (similarly to most animal species) much larger selection pressure can be applied on males than on females. In addition, most of the traits of interest have low (e.g. functional traits, such as fertility, resistance to mastitis or ease of calving) to moderate heritabilities (e.g. production traits, such as milk yield) in dairy cattle, although some exceptions exist, for example milk fat content, which has a heritability of about 0.7 in certain breeds.

Due to the extensive use of artificial insemination in dairy cattle breeding, bulls may have several hundreds of thousands of daughters and therefore a huge contribution to the gene pool of the next generation. In order to ensure that only the best bulls will have such a strong contribution, an accurate breeding value estimation for male selection candidates is inevitable in dairy cattle breeding.

As a consequence of the mainly low-moderate heritabilities and the lack of own performance in males, progeny testing had to be implemented in order to achieve reasonably high accuracy of breeding value estimations in males. Due to progeny testing, the precision of the available performance information is much higher for progeny-tested males than for females; however, this come at the cost of a lengthened generation interval, which is usually more than 6 years when measures of males and their offspring can be gathered (Schaeffer, 2006).

Furthermore, an important characteristic of dairy cattle breeding is the high per animal costs (e.g. raising, housing or feeding). These costs are much higher in the dairy cattle industry than – for example – in the pig or poultry industry. These unit costs in dairy cattle are also considerably higher than they are in case of small ruminants (goat, sheep), which species can be considered as competitors of dairy cattle.

Due to the low prolificacy, the applied breeding programs in dairy cattle are aiming to maximize the gain in the additive genetic effects, i.e. the heritable part of the genetic effect and other types of breeding (e.g. cross-breeding) is not widespread. In the following, I will discuss genomic evaluation methods, which are frequently used either in practice or in research for breeding value estimation in dairy cattle. However, before reviewing these, pedigree-based selection methods will be discussed, because one of these (BLUP) will be used to obtain a baseline for comparison purposes.

## 2.2 Pedigree-based selection methods

Pedigree-based selection methods assume that genetic relationships between animals are known and that phenotype data is available for a significant part of the population. The traits of interest are most often quantitative traits with a continuous (normal) distribution. These traits are assumed to be influenced by a very large (in theory by an *infinite*) number of loci, each having an (*infinitesimally*) small effect on the phenotype under study.

An individual's phenotypic performance ($P_i$) is influenced by multiple factors, including an additive genetic effect ($A_i$), a dominance effect ($D_i$), epistatic effects ($I_i$) and environmental effects ($E_i$):

$$P_i = \mu + A_i + D_i + I_i + E_i \quad (1)$$

where μ is the population mean. Other effects, such as genotype-environment interactions or maternal effects can be included as well, but are usually assumed to be negligible. $D_i$ and $I_i$ are also ignored, because they are not directly transmitted to the next generation.

Additive genetic effects "$A_i$" (also called breeding values) are estimated using linear regression models. **Best linear predictions** (or BLP) of the breeding values are obtained by constructing optimal linear combinations of performances of each animal and close relatives (progeny, parents, sibs) expressed as deviation from a general mean. However, such procedures assume that breeding values do not differ

systematically within any of the environmental effects, an assumption which usually does not hold in practical animal breeding. Therefore these estimates are usually biased.

### 2.2.1 Best linear unbiased prediction

Best linear unbiased prediction (BLUP) can be used to estimate the environmental effects and genetic effects simultaneously using mixed models. These models include the identifiable environmental effects as fixed effects and the breeding values as random effects. Since all effects are estimated at the same time and under the same assumptions, BLUP results in unbiased estimations for both types of effects. Using matrix notations, a statistical model including both types of explanatory variables can be written as:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Za} + \boldsymbol{e} \qquad (2)$$

where **y** is a vector of phenotypic observations (dimension: n × 1, where n is the number of phenotypes), **b** is a vector of fixed effects (dimension: p × 1, where p is the total number of levels of fixed effects), **a** is a vector of random additive genetic effects of all animals (dimension: q × 1, where q is the number of such “animal” effects), **X** is an incidence matrix of dimension n × p relating the levels of fixed effects to the observations, **Z** is an incidence matrix of dimension n × q relating the animal effects to the observations and **e** is a vector of random errors (dimension: n × 1).

With (univariate) evaluation models, BLUP usually assumes that random error terms (**e**) are normally distributed, have a mean equal to zero and a variance equal to $\boldsymbol{R} = \boldsymbol{I}\sigma_e^2$ (where **I** is an n × n identity matrix): $\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{R})$. The additive genetic effects are also assumed to follow a normal distribution with a vector mean of **0** and a variance-covariance matrix of $\boldsymbol{G} = \boldsymbol{A}\sigma_a^2$: $\boldsymbol{a} \sim N(\boldsymbol{0}, \boldsymbol{G})$, where **A** is the additive genetic relationship matrix built from pedigree information. It follows, that the performances (**y**) are assumed to have a mean of **X*****b** and a variance equal to $\sigma_p^2 = \sigma_a^2 + \sigma_e^2$: $y \sim N(\boldsymbol{Xb}, \boldsymbol{ZGZ'} + \boldsymbol{R})$. All explanatory variables are assumed to be independent from the random error term.

In dairy cattle breeding, a contemporary group effect is used most often as a fixed effect, in order to integrate information from both the calendar (year/season/…)- and herd effects. For the model presented above, the mixed model equations leading to BLUE (for fixed effects) and BLUP (for random effects) solutions can be written as:

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{a} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \quad (3a)$$

Best linear unbiased estimates (BLUE) of fixed effects are distinguished from best linear unbiased predictions (BLUP) of random effects, because they are calculated differently: for fixed effects only point *estimates* of the specific effect levels present in the model (i.e. the contemporary groups) are of interest. On the other hand, in case of the random effects first parameters of the underlying distribution (i.e. for the animal population) are estimated and then the realized levels of this distribution (i.e. animal effects) are predicted. Equation 3a can be simplified in case of a univariate animal model (Henderson, 1984; Lynch and Walsh, 1998):

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \alpha\mathbf{A}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{a} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix} \quad (3b)$$

where $\alpha = \frac{\sigma_e^2}{\sigma_a^2} = \frac{1-h^2}{h^2}$, $h^2$ is the heritability of the trait, $\mathbf{A}^{-1}$ is the inverse of the additive genetic relationship matrix and all other terms are as described previously. The heritability (more precisely, the *narrow-sense* heritability; $h^2$) of a trait is defined as the proportion of the phenotypic variance that is due to the additive genetic variance. Therefore, heritabilities are trait-dependent and they can be different for different breeds as well as for different populations of the same breed. Solving the mixed model equations for **b** and **a** will give BLUE & BLUP estimates for the fixed and random effects, respectively.

The theoretical accuracy of the estimated breeding values is often measured by the *reliability*, which is the square of the correlation coefficient between the estimated and true breeding values.

### 2.2.2 Implementation in our study

The BLUP analyses were carried out using the BLUPF90 software (Misztal, 1999, after Misztal, 2016) and the results constituted a baseline for comparisons. On several occasions the performance of different genomic evaluation methods will be compared to those obtained with a pedigree-based BLUP model. The models used for breeding value estimation were the ones currently implemented for all dairy cattle breeds in France – including the regional breeds – for the traits we were interested in (discussed later).

Traits were analyzed in a single-trait context. Multiple-trait models also exist and they can result in higher accuracies when the genetic correlations between the analyzed traits are not zero. These methods assume knowledge on genetic correlations and are computationally more demanding than single-trait analyses (Lynch and Walsh, 1998). Because these genetic correlations were not always available and also because the French routine genomic evaluation is conducted in a single-breed context, multiple-trait models were not used and they will not be further discussed.

## 2.3 Genetic background of quantitative traits and genetic markers

Genomic selection procedures differ from pedigree-based selection methods in their use of genetic markers during the breeding value estimation process. In this section first a brief introduction is given on quantitative traits, which is followed by the presentation and characterization of the most frequently used markers and by the detailed description of the genomic evaluation procedures.

### 2.3.1 Quantitative trait loci

Quantitative trait loci (QTL) are the loci (e.g. genes, non-cooding RNA, etc.) affecting the expression of a quantitative trait. The ultimate aim of animal breeders is to identify through genomic evaluation all QTL as well as to accurately estimate the size of their effects. If such information would be available together with the genotypes of animals at all QTL, selection could be done purely on observed genotype data and phenotype recording would be dispensable. However, the identification of all QTL is

currently not possible and therefore in nearly all cases breeders have to rely on genetic markers “linked” to the QTL.

### 2.3.2 Genetic markers

Genetic markers are DNA variations generated by mutations that occurred during the evolution of the species and of the breeds. We will see in section 2.4 that such DNA sequence information can be exploited for selection purposes in animal breeding: in genomic selection, genetic markers are used to trace the inheritance of chromosome segments carrying quantitative trait loci. Unless the QTL is/are known, these marker effects are used as proxies of the QTL effects. Since the exact locations of the QTL are unknown, denser marker maps increase the probability that at least one marker will be “linked” to each QTL. Several types of genetic markers are used for genomic evaluation purposes.

#### 2.3.2.1. Microsatellite

Historically, the first markers used were microsatellites, which are defined as "simple sequence repeats with a repeat length of up to 13 bases" (Gibson and Muse, 2009). These markers have a high mutation rate and therefore are highly polymorphic with an average of at least 10 alleles per locus in human (Gibson and Muse, 2009). However, due to their sparse distribution along the genome, the observed gain in terms of accuracy of genomic evaluation was very limited (Boichard et al., 2012b, Guillaume et al. 2008a; Guillaume et al., 2008b) and genotyping costs of microsatellites were substantial.

#### 2.3.2.2. Single nucleotide polymorphism

The key biotechnological breakthrough that led to significant improvements in selection accuracy (as compared to the pedigree-based selection methods) was the development of the first commercial SNP arrays (in cattle: Matukumalli et al., 2009). Single nucleotide polymorphisms (SNP) are mutations affecting a single locus on the genome. Due to the nature of these mutations, multi-allelic SNP are extraordinarily rare and the vast majority of them are bi-allelic. Furthermore, SNP are the most frequent type of markers on the genome and per-marker genotyping costs are

constantly decreasing (e.g. Holland et al., 1991; Shen et al., 2005; Tobler et al., 2005).

In cattle, three main types of SNP-chips were developed: first the *Bovine SNP50 BeadChip* with approximately 54,000 SNP (50K; Illumina Inc., San Diego, CA, USA; Matukumalli et al., 2009) followed by the *BovineHD BeadChip®* with ~777,000 SNP (Illumina Inc., San Diego, CA, USA; Matukumalli et al., 2011 after Rincon et al., 2011) and finally the *Illumina Infinium BovineLD Genotyping BeadChip* hosting 3-18 thousand SNP, depending on the version of the SNP-chip (LD; Illumina Inc., San Diego, CA, USA). The bovine 50K chip was developed as an initial tool to allow both researchers and industry members to genotype a large number of animals and to enable them to evaluate the performance of the previously proposed genomic evaluation procedures (e.g. Meuwissen et al., 2001) on real data. The HD SNP-chip was developed to grant very fine mapping resolution to scientists, because it was envisioned that this would further improve the resolution and performance of QTL detections, genomic evaluations and other studies. Finally, the LD chip was specifically designed to include a relatively small number of SNP (~3-18 thousand) so the chip could be efficiently used to genotype a large number of animals at a low cost. The first LD SNP-chip contained only ~3,000 SNP and was specifically developed for the request of the United States Department of Agriculture by Illumina and to be used in the US Holstein population (SNP on the chip were selected accordingly). This chip was however quickly replaced by a larger one (~7,000 SNP), which was done for the request of the Bovine LD consortium (Boichard et al., 2012a). The chip then went through an evolution, during which the number of SNP increased to ~18,000; meanwhile several SNP were also replaced by others of larger importance. The larger versions of the LD SNP-chip were also more appropriate to be used in breeds other than the Holstein.

The development of these SNP arrays allowed breeding organizations in various countries in collaboration with research centers to genotype cost-effectively large numbers of SNP for thousands of individuals.

Genetic markers are said to be linked, when the co-occurrence of their different alleles is more frequent than it is expected from their allele frequencies under the assumption that the markers are segregating independently from each other. In other words, linkage is the non-random association between markers (Gibson and Muse, 2009). The stronger the linkage between a marker and a QTL is, the better the QTL effect can be "captured" with the marker alleles and therefore the more appropriate the marker is to trace the transmission of the QTL alleles from one generation to the other. Consequently, it is of interest to have genetic markers closely located to the QTL in order to be able to accurately estimate the marker effects. The strength of the linkage can be characterized by the level of linkage disequilibrium (LD). There are two commonly used measures of LD: D' (the normalized) form of a linkage disequilibrium measure D and $r^2$ (the square of a correlation coefficient between the frequencies of loci). Consider two biallelic markers SNP-A (with alleles $A_1$ and $A_2$) and SNP-B (with alleles $B_1$ and $B_2$), the allele frequencies $p_{A_1}$, $p_{A_2}$, $p_{B_1}$ and $p_{B_2}$ and the frequency of the $A_1B_1$ genotype ($p_{A_1B_1}$), $r^2$ and D' are calculated as shown in equations (4) and (5), respectively:

$$r_{AB}^2 = \frac{\left(p_{A_1B_1} - p_{A_1}p_{B_1}\right)^2}{p_{A_1}p_{A_2}p_{B_1}p_{B_2}} \tag{4}$$

$$D'_{AB} = \begin{cases} \dfrac{p_{A_1B_1} - p_{A_1}p_{B_1}}{\max\left(-p_{A_1}p_{B_1}, -p_{A_2}p_{B_2}\right)}, & if\ p_{A_1B_1} - p_{A_1}p_{B_1} < 0 \\ \dfrac{p_{A_1B_1} - p_{A_1}p_{B_1}}{\min\left(p_{A_2}p_{B_1}, p_{A_1}p_{B_2}\right)}, & if\ p_{A_1B_1} - p_{A_1}p_{B_1} > 0 \end{cases} \tag{5}$$

The most important disadvantage of the $r^2$ parameter is that it depends much on the (marginal) allele frequencies and is sensitive to low allele frequencies (e.g. Devlin and Risch, 1995). In contrast, D' is less dependent on allele frequencies, although it is still influenced by it if a rare allele is present. D' estimates are also inflated in small samples, which is a serious disadvantage of this parameter.

Linkage breaks down with increasing distance between markers due to a higher probability of recombination events between more distinct markers. This phenomenon is known as LD-decay (Baird, 2015).

### 2.3.3 Haplotype

A notable disadvantage of SNP compared to microsatellites is that SNP are bi-allelic and therefore a single SNP carries less information than a single microsatellite. A possible solution to circumvent this issue is the use of *combinations of SNP* instead of individual SNP markers. Haplotypes can be defined in at least two different ways:

- haplotypes are the sets of alleles of markers or genes of an organism, which were inherited together by the individual on one of the ancestral chromosomes (e.g.: The International HapMap Consortium, 2005; Gibson and Muse, 2009; Stephens et al., 2001)
- More simply, haplotypes are combinations of N SNP markers (e.g.: Hayes et al., 2007; Villumsen et al., 2009; Garrick et al., 2014)

In this study, the term “haplotype” refers to the second definition, while the term “phase” will be used to cover the first definition. The term “alleles” or “haplotype alleles” will be used to refer to the alternative forms of the haplotypes (similarly to the case of SNP). Given this definition of a haplotype, it can be shown that a haplotype can carry a maximum of $2^N$ different alleles, where N is the number of bi-allelic SNP forming the haplotype. Due to the multi-allelic nature of haplotypes, there is an increased chance – as compared to individual SNP – that at least one of these alleles will be in LD with the (ungenotyped) causative mutation at a QTL, if one is present. In addition, LD between haplotype and QTL alleles are more stable over time as well, because if a whole haplotype allele is passed to the next generation, it is very unlikely that two recombinations took place within the chromosome segment it represents.

Before haplotypes can be built, phases must be reconstructed from genotype data, since these are not readily available with the genotyping tools available today. Phase-reconstruction will be discussed in detail in the next section. Although haplotypes can increase the LD between the genomic markers and QTL, as it was proven by Croiseau et al. (2015b) and as we will see later, the number of alleles increases exponentially with the haplotype size (when the latter is measured in number of SNP), leading to a rapid increase in the number of allele effects that need

to be estimated. **Figure 1** shows the average number of segregating haplotype alleles in a Montbéliarde population either with the 50K or with the HD chip as well as the maximum possible number of alleles for 4 different haplotype sizes (this Montbéliarde population will be described in section 2.8 below). It can be seen that the number of segregating alleles is close to its theoretical maximum only with short haplotypes (2 or 3 SNP/haplotype). With haplotypes of 4 SNP, the deviation from the theoretical maximum is ~23.0% and 33.5% with the 50K- and HD data, respectively. This deviation shows a substantial increase with haplotypes of 5 SNP. **Figure 1** also illustrates that haplotypes built from consecutive SNP have less segregating alleles when the HD panel is used compared to the 50K SNP-chip. This phenomenon can be explained by the fact that markers are less dense on the 50K array and therefore there is a higher chance for recombinations to occur between markers from this chip than between those from the HD array. This in turn leads to a larger number of segregating haplotype alleles.

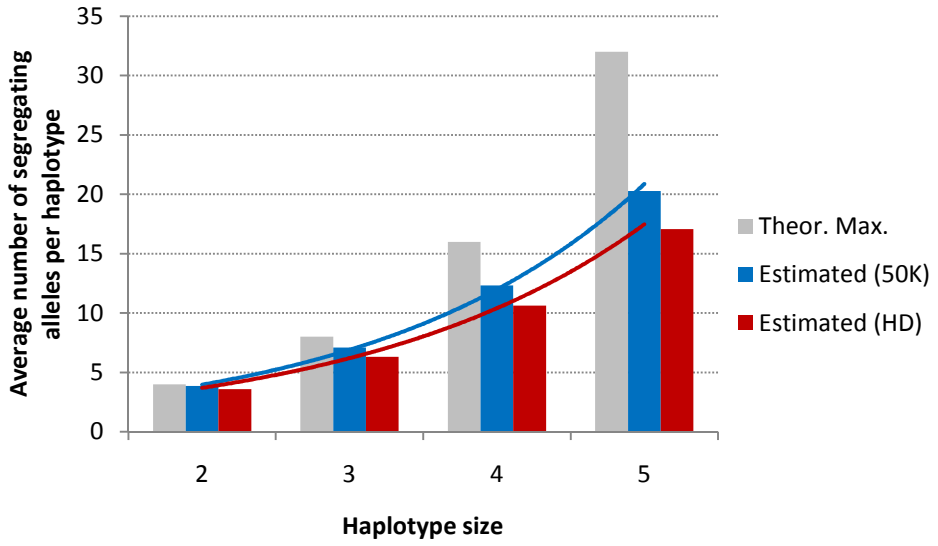



**Figure 1**: Average number of alleles when using consecutive haplotypes from either the 50K or from the HD SNP-chip with 4 different haplotype sizes (the theoretical maximum number of alleles (i.e. $2^N$) is also plotted).

### 2.3.4 Imputation and phase reconstruction

Imputation is the prediction of ungenotyped SNP from genotypes of linked SNP and/or with the use of pedigree information (Li et al., 2009; more generally, any type of marker can be imputed). Phasing is the process in which the parental phases – i.e.

the ordered sequence of SNP alleles which are located either on the paternal or on the maternal chromosome inherited by an individual (see the definition in section 2.3.3) – are reconstructed from genotype data by exploiting pedigree information (Fallin and Schork, 2000). Through the intensive use of imputation, breeders and breeding organizations were able to genotype animals for a decreased number of SNP (for reduced costs), because imputation allowed them to predict the ungenotyped markers with a high accuracy (e.g. Saintilan et al., 2015; prediction error (as concordance rate) was less than 1%). This resulted in substantial savings. Furthermore, determination of parental phases is a prerequisite for haplotype construction. Therefore, both imputation and phase reconstruction (if haplotypes are used) are of great importance with a large impact on every downstream step of a genomic evaluation pipeline. The imputation and phasing methods used in our study will be described later.

## 2.4 Genomic evaluation

The availability of genetic marker information allows us to trace the transmitted marker alleles from ancestors to descendants. Genomic evaluation methods require both phenotype and genotype data (although pedigree data is not a prerequisite, it can improve the performance of genomic evaluation). Most of the genomic evaluation methods estimate allele effects of markers (microsatellite, SNP, haplotype or any other type of marker) using a reference population of animals, i.e. a population of animals with both phenotype and genotype data. Once estimated allele effects are available, they are used in combination with genotype data on the selection candidates to calculate their genomic estimated breeding values (GEBV). Furthermore, availability of marker information also enables QTL detection studies as well, which aim to identify causative mutations, i.e. those genetic markers that are responsible for the observed genetic diversity (e.g. Grisart et al., 2002). This information might be important to improve the performance of genomic evaluation in the future.

Whether or not genomic selection is efficient in any animal population depends both on the characteristics of the species and on those of the production system. Genomic evaluation was quickly introduced in dairy cattle breeding because it allowed

breeding organizations to stop progeny testing, leading to substantial savings (although these were then invested in further genotyping). The genetic gain obtained annually increased significantly as well (see section 2.6 below).

### 2.4.1 Marker-assisted BLUP

In marker-assisted BLUP (MA-BLUP) selection, a limited number of markers are added as random covariable effects to the pedigree-based BLUP model (Fernando and Grossman, 1989). These markers are assumed to be the proxies of causative mutations (i.e. the QTL). A pedigree-based residual polygenic effect is retained in the model in order to account for the additive genetic effect of those QTL which were not identified previously and therefore are not represented in the model by any marker. A general MA-BLUP model can be written as:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \sum_{i=1}^{N}\sum_{j=1}^{2} m_{ij} + \mathbf{e} \qquad (6)$$

where **y**, **X**, **b**, **Z** and **e** are defined as previously for equation 2, **u** is the residual polygenic effect, N is the number of markers included in the model and $m_{ij}$ is the effect of allele j of marker i. A major difference between MA-BLUP and pedigree-based BLUP is the increased number of explanatory variables. Meuwissen and Goddard (1996) showed that substantial gain can be obtained with MA-BLUP compared to BLUP results using microsatellites. Marker-assisted BLUP was first implemented in practice in France (Boichard et al., 2002), followed by Germany (Bennewitz et al., 2003).

In theory, if all QTL would be known and the model would be purely additive, MA-BLUP methods would result in 100% accuracy. However, the identification of all QTL as well as the accurate estimation of each of their effects in any breed is currently not feasible. The two main disadvantages of the MA-BLUP procedure is that all QTL detection methods include false positives and that the QTL linked to the selected markers explain only a fraction of the total genetic variance (de Roos et al., 2009a). For example, if a single marker for each of the ~20,000 genes from the bovine

genome (data from ENSEMBL, 2016) is used, the number of marker effects would exceed the number of phenotypes in most of the breeds.

### 2.4.2 Genomic-BLUP

The most straightforward genomic selection procedure is an extension of the BLUP methodology (equation 3) with a "genomic relationship matrix" (**G**) replacing the pedigree relationship matrix (**A**). This is called genomic-BLUP (GBLUP). This genomic relationship matrix can be constructed in at least 3 different ways (VanRaden, 2008), which are outlined here:

The first one is calculated as $\boldsymbol{G} = \frac{\boldsymbol{ZZ'}}{2*\sum_{n=1}^{N} p_n*(1-p_n)}$, with N being the number of bi-allelic SNP, $p_n$ the minor allele frequency (i.e. the frequency of the less frequent allele of a SNP; MAF) of SNP n and **Z** being an incidence matrix of markers calculated as $\boldsymbol{Z} = \boldsymbol{M} - \boldsymbol{P}$ with one row per animal. In the calculation of the **Z** matrix, each row of **M** contains values (-1), 0 and 1 for the homozygous, heterozygous and the other homozygous genotypes for each animal × SNP combination and any value of column *i* of matrix **P** is calculated as $\boldsymbol{P_i} = 2(\boldsymbol{p} - 0.5)$, where **p** is the vector of minor allele frequencies of the SNP. Matrices **M**, **P** and **Z** have as many rows as the number of genotyped individuals in the population and as many columns as the number of SNP genotyped.

The second one, using the same notations is Calculated as $\boldsymbol{G} = \boldsymbol{ZDZ'}$, where $D_{ii} = \frac{1}{m[2p_i*(1-p_i)]}$. This formula weights the different SNP separately based on their expected variance in contrast with the previous one, which weighted all SNP with the sum of variances of all the SNP.

The last method includes a regression on the pedigree relationship matrix ($\boldsymbol{MM'} = g_0\mathbf{11'} + g_1\boldsymbol{A} + \boldsymbol{E}$, where $g_0$ and $g_1$ are the intercept and regression slopes, respectively) and is calculated as: $\boldsymbol{G} = \frac{\boldsymbol{MM'} - g_0\mathbf{11'}}{g_1}$.

The inverse of the genomic relationship matrix, $\mathbf{G}^{-1}$ is then used to replace the inverse of the additive genetic relationship matrix in BLUP. The **G** matrix is supposed

to reflect the relationship between genotyped animals more accurately than the pedigree-based **A** matrix, because it relies on *observed* genotype data. In contrast, the **A** matrix is based on probabilities and *expected* levels of similarities between relatives, which can be considered less accurate. That is because in case of the **A** matrix all individuals that have the same relationship to each other (e.g. half-sibs) receive the same genetic relationships based on pedigree. However, in the case of the **G** matrix, genetic relationships are estimated from *observed* genotype data, which can deviate from their expected values, based on the number of SNP alleles in common between the animals (e.g. between the half-sibs).

Meuwissen et al. (2001) described a GBLUP applied to a model including marker effects as random variables drawn from a single normal distribution (their model also included a contemporary group effect as fixed effect). This model is equivalent to the GBLUP model described in the previous paragraph, because the breeding values (vector **a** in equation 3) equal to the sum of the allele effects, as it was shown by (VanRaden, 2008). This implies that breeding values can be estimated indirectly, by first estimating the allele effects and then calculating the breeding values of individuals from the estimated allele effects and from their observed genotypes.

The problem with the **G** matrix is that it measures the relationship between animals by the average number of shared alleles, i.e. it considers the alleles *identity in state* rather than those *identity by descent*. Furthermore, usually the same weights are given to all SNP irrespective of the trait, although it is reasonable to assume that not all genotyped SNP are linked to QTL for all the traits (and also that their relative importance also differ from trait to trait). However, there are some studies to circumvent this issue and Zhang et al. (2010) for example proposed the use of a trait-specific relationship matrix instead of a regular G-matrix.

### 2.4.3 Bayesian methods

To cope with the mentioned issues of MA-BLUP, Meuwissen et al. (2001) proposed using *all* SNP in genomic evaluation and not a subset of them. Bayesian methods were originally suggested to be used for genomic evaluation purposes because they are computationally efficient and because they can successfully deal with the

problem of estimating many more effects than the number of dependent variables available for the analysis (the $p \gg n$ problem). Furthermore, the use of the Gibbs sampler algorithm was also suggested to generate samples from the posterior distribution of each effect. This was a convenient choice because it allowed the sampling of allele effects from their posterior distribution conditional on all other effects, but not on the effect being sampled, which is relatively straightforward to implement.

The methods proposed by Meuwissen et al. (2001) became known as BayesA (when all SNP is assumed to have a larger-than-zero effect) and BayesB (when a predefined proportion of the SNP are assumed to have an effect of 0 and only the rest of the SNP to have an effect >0). We mainly worked with an extension of BayesB which will be described in detail below. In BayesA, each marker is assumed to explain a different proportion of the genetic variance ($\sigma_a^2$). The prior distribution of the marker variances is modeled with a scaled inverted chi-square distribution. As it is reasonable to assume that most of the SNP from any SNP panel are neither a causative mutation nor linked to any of those, the BayesB method has a fixed prior probability (π) that a given marker has no effect on the analyzed trait (in Meuwissen et al. (2001) π varied between 78.8% and 94.7%, depending on the marker density). For technical reasons (it is impossible to directly sample an effect from a “simple” distribution), marker variances were sampled with the Metropolis-Hastings sampling procedure with BayesB, instead of sampling with the Gibbs sampler. A serious problem arising with both BayesA and BayesB methods is shrinkage (i.e. the risk of shrinking allele effects when the estimates are applied on a dataset other than the one used to calculate them), which was shown to depend on the initial value of the scale parameter S of the scaled inverse chi-square distribution (Gianola et al., 2009).

The BayesC method was proposed as an extension to the BayesA and BayesB methods (Habier et al., 2011). In contrast to BayesA and BayesB, the BayesC model assumes a single marker-effect variance for all markers. This modification was shown to decrease the chance of shrinking.

A modification of BayesC is the so called BayesC-π, where the proportion "π" (i.e. the proportion of markers without an effect on the trait) is allowed to vary during the analysis and is estimated from the data. In our work, we used the GS3 software with an implementation of the BayesC and BayesC-π methods (Legarra et al., 2013). In the original paper in which the BayesC-π method was introduced (Habier et al., 2011) π was defined as the proportion of SNP *without* an effect on the analyzed trait (in accordance with the definition of π in BayesB in Meuwissen et al., 2001). However, in the GS3 implementation, π refers to the opposite proportion, that is the fraction of SNP *with* an effect on the trait of interest. In order to avoid ambiguities, π will be defined here according to the original definition given by Meuwissen et al. (2001) and by Habier et al. (2011).

BayesC(-π) distinguishes only 2 groups of SNP: those with an effect (from a distribution with a unique variance) and those without an effect on the analyzed trait. However, it is known from previous studies that the size of SNP effects can differ substantially. The distribution of the marker effects (after standardization) was shown to follow a gamma distribution (Hayes and Goddard, 2001; also see **Figure 2**), i.e. there is a small number of QTL with large effects in addition to a large number of QTL with small effects. However, it is reasonable to assume that the parameter estimates (scale and shape parameters were estimated by Hayes and Goddard (2001) to be 5.4 and 0.42, respectively) are dependent both on the analyzed population and trait.

Erbe et al. (2012) proposed a method termed BayesR which can distribute the SNP into more than 2 groups, i.e. the distinction of small, medium and large QTL becomes possible in addition to a group of SNP with no effect. In this method, each group is defined by the proportion of genetic variance that any SNP from that group is expected to explain.

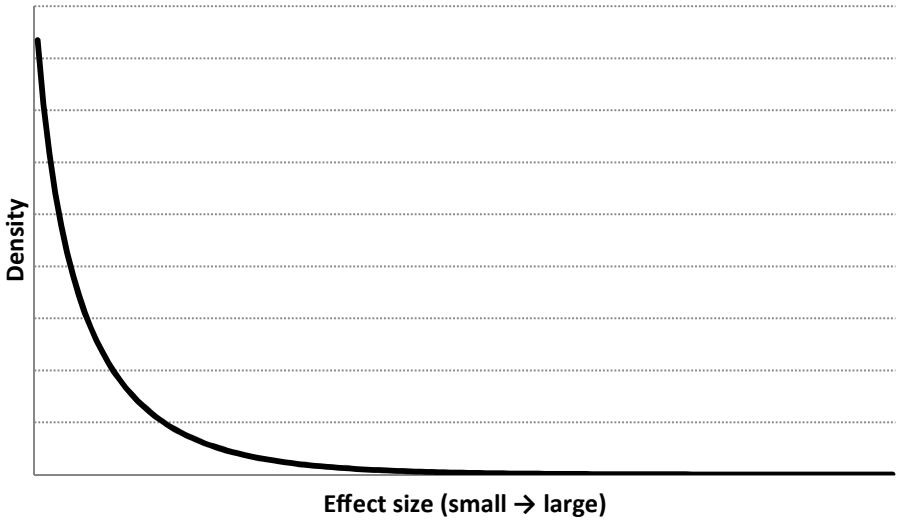



**Figure 2**: Probability density distributions of QTL effects in dairy cattle (after Hayes and Goddard, 2001; axis labels were removed since they are trait-dependent).

Other Bayesian methods include the BayesD(-π) (Habier et al., 2011), Bayesian Lasso (Park and Casella, 2008; de los Campos et al., 2009; Weigel et al., 2009, Legarra et al., 2011), emBayesR (Wang et al., 2015) or the BayesSSVS (Verbyla et al., 2009). The latter method is very similar to BayesC-π (SNP effects are assumed to follow a normal distribution and a proportion (π) of the SNP are assumed to have a negligible effect on the trait of interest; Lukić et al., 2015). These methods will not be further discussed as they are not used in routine genetic evaluation.

A serious drawback of the presented Bayesian methods compared to the other methods presented (GBLUP, MA-BLUP) is that they are not suitable to evaluate large datasets in routine due to long running times. However, they are still adequate for QTL detection for scientific purposes and this information can then be exploited for routine evaluations (for example, see the French routine genomic evaluation pipeline in section 2.5).

### 2.4.4 Genomic evaluation methods with haplotype markers

In our studies, two haplotype-based genomic evaluation methods were implemented. The first one, the marker-assisted BLUP model on haplotypes is a straightforward extension of equation (6). In this model, SNP effects are simply replaced with haplotype effects as follows:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \sum_{i=1}^{N}\sum_{j=1}^{H_i} h_{ij} + \mathbf{e} \quad (7)$$

where **y**, **X**, **b**, **Z**, **u**, **N** and **e** are defined as previously for equation 6, $H_i$ is the number of alleles carried by haplotype i and $h_{ij}$ is the allele effect of allele j of haplotype i.

In 2013, before our studies there was no software available for the implementation of a Bayesian genomic evaluation procedure using haplotypes. Therefore the GS3 software by Legarra et al. (2013) was modified by P. Croiseau and M-N. Fouilloux in our group to be able to handle multi-allelic haplotypes instead of bi-allelic SNP in a BayesC-π approach. This version of the software will be referred as *haplotypic GS3* hereafter. I used this software to assess the performance of two criteria to define optimal haplotypes. In this section the most important aspects of the method will be described as well as the differences compared to the regular, SNP-based BayesC-π. A typical model with haplotype effects is:

$$y_i = cge_i + u_i + \sum_{j=1}^{N} \delta_j \left(h_{ij}^{p} + h_{ij}^{m}\right) + e_i \quad (8)$$

where $y_i$ is the performance value of individual i, $cge_i$ is the contemporary group effect of animal i (fixed effect; additional fixed effects can be included as well), $u_i$ is the residual polygenic effect of animal i ($u \sim MVN(0, \boldsymbol{A}\sigma_u^2)$), N is the total number of haplotypes in the model, $h_{ij}^{p}$ and $h_{ij}^{m}$ are the random effects of the maternal and paternal alleles of haplotype j of animal i, $\delta_j$ is a 0/1 variable indicating whether or not marker j is assumed to have an effect ($\delta_j$ is zero with a probability of π; when it is zero, all alleles of the given haplotype are assumed to have no effect on the trait) and $e_i$ is a random error term for animal i.

In this implementation, haplotype size is a user-specified parameter (usually between 1 and 5, with 1 corresponding to the SNP-based BayesC-π model; this parameter will be abbreviated as $N_h$ in this section). The software then creates every consecutive,

non-overlapping haplotypes of $N_h$ SNP from the genotype files. The last haplotypes were truncated if a complete haplotype of $N_h$ SNP could not be built from them. In order to avoid haplotypes spreading across multiple chromosomes, separate genotype files must be provided for each chromosome. Similarly to the SNP-based BayesC-π, a common variance (sampled from an inverted chi-square distribution) is used for all haplotypes in the model.

In certain cases, it is desirable to exclude certain SNP from the analysis, therefore an important question is how one can simply remove SNP from the dataset. The solution I proposed to this issue was not to address it within the software (i.e. making both the code and the software input file more complex) but to simply adjust the genotype files prior to running the software. On the one hand, this did not require further programming and additional input files and parameters, which is convenient from the perspective of both the programmer and the user. On the other hand, it made necessary that the user creates a new set of genotype files each time (s)he wants to test a different set of haplotypes, which can be – depending on the density of the SNP-chip and on the number of different genotype sets to be tested – very demanding in terms of data-storage.

This work was presented at the *World Congress on Genetics Applied to Livestock Production* in Vancouver, Canada (Croiseau et al., 2014).

An important question that immediately arises when haplotypes are used instead of SNP in genomic evaluation is: what is the optimal haplotype size for genomic selection? Too long haplotypes would result in increasingly large number of segregating alleles and therefore in a rapid decrease in the average number of available observations per allele, leading to a quick decrease in estimation accuracy of allele effects. To overcome this difficulty, an efficient technique is needed to reduce the number of haplotypes used in the prediction models as much as possible without risking the loss of relevant genotype information.

In conclusion, the use of haplotype markers in genomic prediction is intuitively a promising way to increase the selection accuracy, because they are much more

polymorphic. However, they pose serious risks as well. On the one hand, it is desirable to increase the number of marker alleles (i.e., the number of effects to estimate in the genetic model) in order to increase the probability of capturing the QTL effects. On the other hand, the increase of the number of effects in the model is detrimental to the accuracy of parameter estimates. These issues will be addressed in Chapter 3.

## 2.5 French routine genomic evaluation of dairy cattle

In France, marker-assisted evaluation was first introduced in 2001 (Boichard et al., 2002) based on microsatellites, but quickly evolved into a real genomic evaluation and went through several steps of evolution (Ducrocq et al., 2009, Boichard et al., 2012b, Croiseau et al., 2015b) with the last major changes implemented in April 2015 (Croiseau et al., 2015a). At the present time, the routine genomic evaluation consists of 4 steps (see below) and incorporates part of my PhD work. In France, genomic evaluation is officially applied to (i) the 3 major dairy cattle breeds, namely the Holstein, Montbéliarde and Normande breeds (since 2009), (ii) to the Brown Swiss (since 2014) and (iii) to 3 local breeds, namely the Abondance, Tarentaise and Vosgienne (since 2016). In the case of five breeds (the 3 regional breeds, Montbéliarde and Normande) both males and females are included in the training population in contrast with the two international breeds (Holstein and Brown Swiss), for which only males are used. It is worth mentioning, that the French Brown Swiss population is small, but within the framework of the Intergenomics project (http://www.brown-swiss.org/genetics), a large international reference population was assembled for this breed from smaller national populations (contributing countries included – among others – Germany, USA, Canada and France). Genomic evaluation is carried out on 34-46 traits, depending on the breed. The four steps of the evaluation pipeline are:

1 QTL detection
2 Haplotype construction
3 Estimation of (haplotype) allele effects
4 GEBV calculation for selection candidates

The first two steps were done in the research phase only and are not repeated at each routine evaluation. In contrast, the last 2 steps are routinely done 3 times a year in order to obtain estimates of marker effects using all available data to compute GEBV values for selection candidates. Genomic evaluation is carried out on 34-46 traits per breed (traits analyzed independently).

For the research phase (steps 1 and 2), phenotypes were first converted into 'daughter yield deviations' (DYD) for progeny-tested bulls and into 'yield deviations' (YD) for females with own performance recording only. (D)YD values are calculated by correcting the observed phenotypes to all fixed and random effects except of the effect of the animal (Liu et al., 2004; Szyda et al., 2008); at the end of each BLUP genetic evaluation. (D)YD values are the most accurate indicators of the true breeding values calculated from the available data.

Genotype data from both the 50K and LD SNP-chips are currently used. Genotype sets are standardized for each breed: a set of 43,801 SNP are retained from the 50K and a set of 8,218 SNP from the LD chip for genomic evaluations.

In the first step of the pipeline, SNP effects are estimated for all SNP from the 50K SNP-chip using a BayesC-π procedure with the following model:

$$y_i = \mu_{si} + p_i + \sum_{j=1}^{N} z_{ij} m_j \delta_j + e_i \qquad (9)$$

where $y_i$ is the performance value of individual i, $\mu_{si}$ is an overall mean effect (calculated separately for males (s=1) and females (s=2), when applicable) of animal i, $p_i$ is the residual polygenic effect of animal i (**p** ~ MVN(0, $\boldsymbol{A}\sigma_a^2$), with MVN refering to a multivariate normal distribution, **A** to the additive relationship matrix and $\sigma_a^2$ to the genetic variance), N is the total number of SNP in the model, $z_{ij}$ is an indicator variable representing the number of copies of one of the alleles at marker j in animal i, $m_j$ is the allele effect for marker j, $\delta_j$ is a 0/1 variable indicating whether or not marker j has an effect and $e_i$ is the random error term for animal i. The proportion of the genetic variance attributed to the residual polygenic effect in the BayesC-π model

(equation 9) is determined empirically for each trait separately and the most optimal value is used.

Once marker effects are available for all the 43,801 SNP, those with the highest probability of inclusion (i.e. the highest probability to have an effect different from zero) are identified to trace the QTL with moderate-high effects. The analyses with 1,000 and 3,000 SNP included in the model were compared and the most optimal value is used for each trait. This is done in order to properly adapt the models to the genetic background of the traits. In practice 3,000 SNP was found to be optimal for most of the traits. Probability of inclusion is used preferably to the estimated allele effects because it was found to give slightly better results (S. Fritz, 2014, personal communication).

It is reasonable to assume that the markers selected from the 50K SNP-chip are not the causative mutations but are merely linked to them: this is because the 43,801 SNP from the chip represent only ~0.16% of all the ~28 million known SNP on the bovine genome (Boussaha et al., 2016). Therefore SNP from the 50K chip likely indicate only the approximate location of the causative mutations on the chromosomes. In order to better capture the QTL effects, haplotypes are built around each of the selected SNP for the routine evaluation. Haplotypes are built using the method proposed in Chapter 3. This method exploits information on haplotype allele frequencies. In this method, a short (10 SNP-wide), symmetric window is created around the selected SNP and from all possible haplotypes of 4 SNP within the window, one is selected to represent the given region based on observed allele frequencies. The main goal of this method is to balance between allele frequencies and number of segregating alleles when a haplotype is selected. Different haplotype sizes between 2 and 5 SNP were compared. Haplotype size of 4 SNP was found to be optimal and therefore was applied in the routine evaluation in France.

Once the haplotypes are available, their allele effects are estimated using a marker-assisted BLUP model:

$$y_i = \mu_{si} + \sum_{j=1}^{8218} z_{ij} m_j + \sum_{k=1}^{N_h} \left( \sum_{l=1}^{N_{ka}} \beta_{kl} \varepsilon_{ikl} \right) + e_i \qquad (10)$$

where $N_h$ is the number of haplotypes (i.e. 1,000 or – most often – 3,000), $N_{ka}$ is the number of segregating alleles at haplotype k, $\beta_{kl}$ is the estimated allele effect of allele i at haplotype k and $\varepsilon_{ikl}$ is an indicator variable indicating how many copies (0, 1 or 2) of allele l at haplotype k individual i carries. The other terms are defined as in equation (9). The polygenic effect from equation (9) is replaced by the combined effect of the 8,218 SNP from the LD SNP-chip in the MA-BLUP model. This modification was done because the combined effect of the 8,218 SNP from the LD SNP-chip can be considered as equivalent to a residual polygenic effect with a genomic relationship matrix (see section 2.4.2) and therefore is expected to perform better than the pedigree-based residual polygenic effect.

Following the allele effect estimation of the haplotypes, these estimates are applied to the genotypes of the selection candidates to estimate their GEBV.

To adapt the routine evaluation procedure to the regional breeds (most importantly to the lower amount of available performance records), there were 2 important changes. First, the number of QTL traced was reduced to 1000 from the original 1,000-3,000. Secondly, due to convergence problems in the first step, π had to be fixed to 80%.

## 2.6 Consequences of genomic selection

### 2.6.1 Advantages of genomic selection

The technological advances previously presented and the theoretical developments achieved since the early 2000s led to the practical implementation of genomic evaluation in dairy and beef cattle in at least 16 countries by 2016 (e.g. for Holstein in the USA: Wiggans et al., 2011; in France: Boichard et al., 2012b and Croiseau et al., 2015b; in the Netherlands and in New Zealand: de Roos et al., 2009b; the Eurogenomics initiative: Lund et al., 2011). Genomic evaluation also led to the elimination of the expensive progeny-testing phase of the previous breeding program in several countries (e.g. France, United States).

Genomic evaluation has an effect on the annual genetic gain. When calculating the annual genetic gain, four different paths have to be distinguished in dairy cattle breeding, because multiple parameters affect genetic gain (namely: the generation interval, selection accuracy and selection intensity) differ significantly for these paths. Generation interval is the average age of the breeding animals when their offspring, which are kept for breeding are born. Selection accuracy is the correlation between the true and estimated breeding values, while the selection intensity is the performance of breeding animals expressed as a deviation from the population mean and as a proportion of phenotypic standard deviation. The aforementioned four paths differ mainly due to progeny testing in males and because a much larger selection pressure can be applied on males. The paths are distinguished based on whether bulls or cows are selected and whether they are selected to contribute to the next generation of bulls or cows:

- males to produce females (denoted “mf” in the subscripts in equation 11)
- males to produce males (denoted as “mm”)
- females to produce females (denoted as “ff”)
- females to produce males (denoted as “fm”)

The annual genetic gain obtained with any breeding program can be calculated using the following formula (Rendel and Robertson, 1950):

$$\Delta G = \frac{\left(i_{mf} * r_{IH,mf} + i_{mm} * r_{IH,mm} + i_{ff} * r_{IH,ff} + i_{fm} * r_{IH,fm}\right) * \sigma_a}{L_{mf} + L_{mm} + L_{fm} + L_{ff}} \qquad (11)$$

where ΔG is the annual genetic gain, $i_{..}$ is the selection intensity calculated for the four different paths, $r_{IH,..}$ is the selection accuracy calculated for the four paths, $\sigma_a$ is the standard deviation of the additive genetic effect of the trait under selection and $L_{..}$ are the generation intervals (expressed in years) again for the four paths. Genomic selection affects the following factors in the above equation:

1. Selection accuracy ($r_{IH,..}$): For males, selection accuracy of genomic selection is usually inferior compared to the selection accuracy of progeny-tested bulls given that a large number of progeny is evaluated for the bulls (this was

typically done in large breeds). However, selection accuracy is higher for females with genomic evaluation compared to the BLUP selection accuracy based on own performance only (Boichard et al., 2015). Furthermore, genomic evaluation increases the selection accuracy in case of males without a large number of progeny as well.

2. Selection intensity ($i_{..}$): Selection intensity can be increased for females (Boichard et al., 2015). This is due to the increasing use of sexed semen as well as due to the introduction of genomic evaluation. The former biotechnological development leads to a larger number of selection candidates for females while the latter results in more accurate breeding values for females, which enables the selection of the best females. Sexed semen accounted for 37% of all inseminations in dairy cattle in France (Institut de l'Elevage, 2016).
3. Generation interval ($L_{..}$): Due to the availability of DNA sample of selection candidates immediately after birth, generation interval is greatly reduced for progeny-tested bulls. Schaeffer (2006) assumed the generation interval of progeny-tested bulls between 6 and 6.5 years, while in the same study he predicted that the generation interval with genomic selection could be ~1.75 years. García-Ruiz et al. (2016) observed such trends and values in the US Holstein population, although the decrease was more moderate (~25-50%); in this population, the generation interval was ~6.8 years with progeny testing vs. 3-5 years with genomic selection. Le Mézec et al. (2015) observed similar results in the French dairy cattle breeds, however, the generation interval was slightly shorter in the French case (5.6 years before genomic evaluation; Institut de l'Elevage, 2015c). Generation interval of dams of cows is largely unaffected by genomic evaluation, because they were used for reproduction at an early age previously as well, which could not be further decreased by the introduction of genomic evaluation.

Overall, after combining all these changes, the introduction of genomic selection is extremely advantageous in dairy cattle. Schaeffer (2006) estimated that the annual genetic gain would be approximately doubled with genomic selection compared to the previous *state of the art* breeding programs (such gains were observed in

practice in France: Le Mézec et al., 2015). Furthermore, because progeny-testing became unnecessary, significant savings were accumulated in the dairy cattle industry.

### 2.6.2 Drawbacks of genomic evaluation

Most of the currently available genomic evaluation procedures use bi-allelic SNP markers to trace QTL on the genome, with the notable exception of the French routine genomic evaluation procedure, which uses haplotype markers. A major drawback of the SNP markers lies in their bi-allelic nature: because of it, SNP in strong linkage disequilibrium with the causative mutations are required to efficiently capture their effects. Such SNP are not always available, especially when SNP-chips of low or moderate density are used. Yang et al. (2010) showed that even with ~300,000 SNP, part of the additive genetic variance could not be explained by SNP due to low linkage disequilibrium between the markers and QTL. Although it is desirable to have a high SNP density along the genome to maximize the probability that there is a SNP linked to every imortant QTL, the abundance of SNP across the genome can be considered as a disadvantage as well. This is because a majority of them are not relevant for the analyzed trait(s) and these SNP make it more difficult to identify the significant SNP as well as to obtain accurate allele effect estimates for them.

Therefore, a major difficulty that needs to be addressed in genomic evaluation is the balance between the number of effects that needs to be estimated and the estimation accuracy. Due to the dense SNP assays available and efficient imputation methods, the amount of phenotype data available is at least one order of magnitude lower than the amount of genotype data. Therefore, the main limiting factor in genomic selection is the size of the reference population, i.e. the number of animals with both phenotype and genotype information available (Hayes et al., 2009a). This limitation is more stringent in populations with a limited number of recorded animals (for example in regional breeds) or in cases when (multi-allelic) haplotypes are used as genetic markers. Due to the insufficient amount of phenotype data in these breeds, it is difficult to identify all the markers with a significant effect on the analyzed trait.

Furthermore – especially when markers are linked to small QTL – accurate estimation of the allele effects is also challenging (Wientjes et al., 2015).

## 2.7 Assessment of genomic evaluation studies

### 2.7.1 Principles of validation in genomic evaluation studies

The performance of the genetic/statistical models must be assessed before they can be applied in practical animal breeding. Validation studies have been often used to assess the performance of genomic evaluation models since they were first proposed (Meuwissen et al., 2001). These studies first split the available dataset into a training set and a validation set. The model is then fitted to the training set and the quality of genomic prediction is evaluated on the validation set, from which data was not used for model fitting. The evaluation on the validation set incorporates two sub-steps: first the dependent variable (that is the breeding value in a genomic evaluation experiment) is estimated for all individuals in the validation population either using the estimated allele effects from the training dataset (when marker effects were estimated) or exploiting the genomic relationship information between animals (e.g. in GBLUP). In the second step, measures of accuracy such as the correlation coefficient between the GEBV and (D)YD are calculated.

In genomic evaluation studies, the division of the datasets into training- and validation sets is adapted to the main target population, which is the set of young animals, usually without any performance observations for which we want estimated breeding values. Therefore, the validation population typically consists of the youngest individuals (usually the 20-30% youngest animals) in order to objectively simulate real-life conditions, where performance values are available only on the older individuals of the populations but not on the youngest ones.

From this point on, the “training population” and “validation population” terms will be used according to their definitions above, while the term “reference population” will be used to refer to these two populations combined.

### 2.7.2 Measured parameters

The performances of different (genomic) evaluation procedures are compared based on 2 parameters: the accuracy and bias of the (genomic) estimated breeding values. In the following these parameters are discussed with DYD used as measure of performance. However, it can be replaced with other measures, such as deregressed proofs or simulated true breeding values in a simulation study. Furthermore, observations are weighted, using equivalent daughter contributions (EDC) in case of males and number of record equivalents (RE) in case of females.

#### Reliability of selection candidates

The accuracy of an EBV is the correlation between the estimated and true breeding values. The reliability is the accuracy squared. The higher the reliability of the selection candidates, the more accurate the breeding values are. Reliability is bounded between 0 and 1.

In a validation study, the accuracy is measured by the weighted correlation coefficient between DYD and GEBV in the validation population. This is calculated as:

$$\rho_{wt} = \frac{\sum_{i=1}^{n} w_i (DYD_i - \overline{DYD})(GEBV_i - \overline{GEBV})}{\sqrt{\sum_{i=1}^{n} w_i (DYD_i - \overline{DYD})^2 \sum_{i=1}^{n} w_i (GEBV_i - \overline{GEBV})^2}} \quad (12)$$

where $\rho_{wt}$ is the weighted correlation coefficient between DYD and GEBV, $w_i$ is the weighting factor of animal i, $DYD_i$ and $GEBV_i$ are the DYD and GEBV of animal i; $\overline{DYD}$ and $\overline{GEBV}$ are the weighted means of DYD and GEBV, respectively. The corresponding reliability is $\rho_{wt}^2$.

#### Regression slope of DYD on GEBV

In addition to be accurate, breeding values are also expected to be unbiased. In other words, we want that the average (genomic) estimated breeding values of particular groups of animals (in particular the youngest ones) is nearly the same as their average (unknown) true breeding values. The regression slope of DYD on GEBV indicates a bias: the optimal value of this parameter is 1 (indicating no bias).

When the regression slope is less than 1 it indicates that the young animals are overestimated, while a slope higher than 1 indicates the opposite (i.e. underestimated young selection candidates). Regression slopes are estimated using the following equation:

$$DYD = \beta_0 + \beta_1 GEBV + e \quad (13)$$

where $\beta_0$ is the intercept, $\beta_1$ is the regression slope and e is the random error term ($e \sim N(0, \mathbf{D}\sigma_e^2)$, where **D** is a diagonal matrix with diagonal elements equal to 1/EDC and 1/RE for males and females, respectively). Although there is no theoretical lower or upper limit of the regression slope in terms of statistics, in the context of breeding value estimation they are never lower than zero and not frequently higher than 1. A large bias (say, a regression slope significantly lower than one) results in "inflation" of GEBV of the young candidates. This is undesirable, because this leads to the overestimation of the genetic merit of the young candidates. When young AI sires are considered, this means that their progeny performances will be disappointing, generating some distrust of the quality of genomic evaluation.

## 2.8 Analyzed breeds and traits

Five breeds were included in this work: one of them is Montbéliarde, the second largest French dairy cattle breed with genomic evaluation. The Montbéliarde population is currently of approximately 648,000 cows (with ~68% of them under performance recording), which represents more than ~18% of the dairy cattle population of France (Institut de l'Elevage, 2015a). The Montbéliarde breed was selected to test the new methods, because of the availability of the large reference population of progeny-tested bulls ($n = 2{,}235$).

Multi-breed tests were carried our using the following four regional French dairy breeds (abbreviations of the breed names are given in parenthesis): Abondance (A), Tarentaise (T), Simmental (S) and Vosgienne (V).

**Table 1** shows the number of bulls progeny tested every year as well as the number of females under performance recording, as of 2015. **Table 2** shows the average performance records of these breeds for production traits.

**Table 1**: Number of progeny-tested bulls and number of cows under performance recording in the 5 breeds used through this Thesis.

| Breed | Number of progeny-tested males[1] | Number of cows under performance recording |
|---|---|---|
| Montbéliarde | 164 | 439,609 |
| Abondance | 18 | 23,412 |
| Tarentaise | 11 | 7,816 |
| Simmental | 10 | 16,938 |
| Vosgienne | 5 | 1,372 |

1: Before the implementation of genomic evaluation. Data from Institut de l'Elevage, 2014 and 2015b.

Phenotype data were available in the form of daughter yield deviations in case of progeny tested bulls and as yield deviations in case of females with own performance information only. In case of all the 5 presented breeds, both male and female animals were genotyped. However, while only the progeny tested bulls were used from the Montbéliarde breed, all genotyped males and females were used in case of the regional breeds. This decision was made because the Montbéliarde was specifically selected due to the available large number of progeny tested bulls, which allowed an efficient within-breed evaluation for this breed. In contrast, the lack of such a male reference population in the regional breeds required all animals – irrespective to its gender – to be included in the reference population to enable genomic evaluation. Furthermore, one of the main aims was to maximize the selection efficiency in the regional breeds, therefore it made no sense to remove animals from the reference population of these breeds. Majority of this work was done on 5 dairy cattle production traits (these are: milk yield, fat yield, protein yield, fat content and protein content), which are moderately heritable traits (**Table 2**). Although, some of the developed methods (mainly those that were later included in the French routine genomic evaluation) were tested on a wider range of traits including some with lower or higher heritabilities.

**Table 2**: Average standard 305-day production level of the 5 breeds used through this Thesis (data from 2015).

| Breed | Milk yield (kg) | Fat yield (kg) | Fat content (%) | Protein yield (kg) | Protein content (%) |
|---|---|---|---|---|---|
| Heritability | 0.3 | 0.3 | 0.5 | 0.3 | 0.5 |
| Montbéliarde | 6515 | 250 | 3.83 | 212 | 3.25 |
| Abondance | 5085 | 186 | 3.66 | 168 | 3.30 |
| Tarentaise | 4045 | 147 | 3.64 | 130 | 3.22 |
| Simmental | 5751 | 228 | 3.96 | 192 | 3.34 |
| Vosgienne | 3963 | 149 | 3.75 | 125 | 3.15 |

Data from Institut de l'Elevage, 2015a

The Simmental and Vosgienne breeds were particular among the 4 regional breeds. The number of imported breeding animals was relatively large in the Simmental breed and the available pedigree information on these animals (in France) was very limited. Therefore the BLUP analysis is expected to be less accurate than it would be in another breed with similar characteristics but more pedigree data. On the other hand, in Vosgienne the average age of the breeding animals was higher than it was in the other breeds and therefore more phenotype data was available on these individuals. In consequence, the pedigree-based BLUP is expected to perform well in this breed.

## 2.9 Single-breed and multi-breed genomic evaluation

As mentioned earlier, current genomic evaluation methods require reference populations because neither the QTL nor their relative effects are known. Genomic evaluation studies can be split into 2 groups based on the composition of the reference population: the reference population consists of individuals either from a single breed or from multiple breeds. The main difference between these two scenarios is that when several breeds are considered together, either artificial or natural barriers (or both) prevented gene flow from one population to another. Therefore different QTL might exist in the different populations, the same QTL might have a different relative effect (compared to the other QTL), LD phases might differ

across breeds or the linkage phases between the QTL and markers in the different breeds might be also different (de Roos et al., 2009a). In the cases of these QTL, multi-breed genomic evaluations can be expected to be less efficient, which can counterbalance the impact of having a larger reference population (e. g. this was discussed in Hayes et al., 2009). This is because the multi-breed training population introduces mainly noise to the allele effect estimation process of these markers. Whether or not multi-breed genomic evaluation in specific breeds is advantageous or not depends on the relative frequency and importance of the shared QTL. Both the emergence of new (i.e. breed-specific) QTL and the break-down of QTL-marker phases in the breeds depend on the evolutionary distance from the most recent common ancestors (de Roos et al., 2009a). Therefore breeds that are closer to each other from an evolutionary perspective can be expected to benefit more from a multi-breed genomic evaluation, while for breeds that diverged earlier in time (time measured in number of generations) a multi-breed reference population is expected to be detrimental.

These remarks can be generalized to a “single-subpopulation” – “multi-subpopulation” case, because natural barriers might prevent gene flow from one population to another even among two populations of the same breed.

Gautier et al. (2010) estimated genetic distances between 47 cattle breeds using 50K SNP-chip data, including the five breeds presented here. All of these breeds were clustered very closely together based on this study (**Figure 3**).

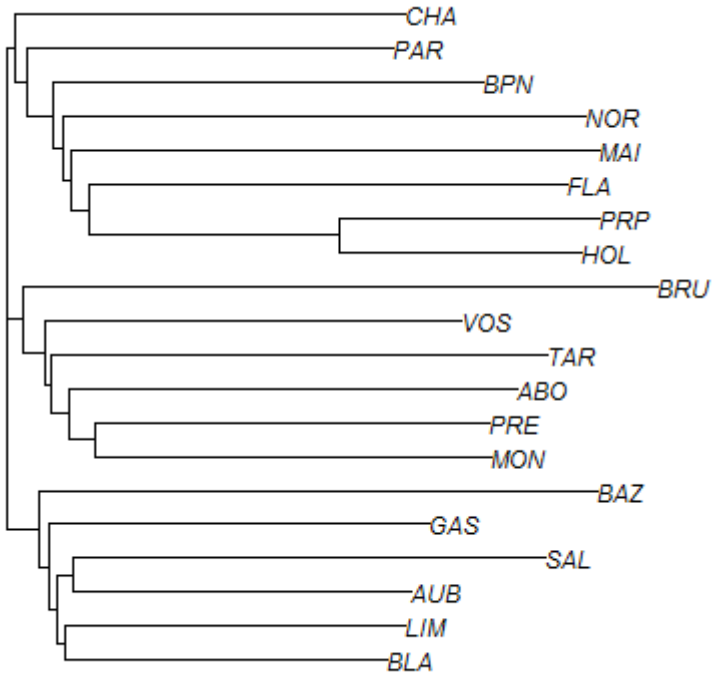



**Figure 3**: Tree representing the genetic distances between 20 French cattle breeds. Genetic distances were estimated from allele frequencies using the bovine 50K SNP-chip (from Gautier et al., 2010). Breed name abbreviations: CHA – Charolais; PAR – Parthenaise; BPN – Bretonne Pie Noire; Noire – Normande; MAI – Maine Anjou (Rouge des prês); FLA – Flamande; PRP – Pie Rouge des Plaines [→Red Holstein]; HOL – Holstein; BRU – Brune; VOS – Vosgienne; TAR – Tarentaise; ABO – Abondance; PRE – Pie Rouge de l'Est (French Simmental); MON – Montbéliarde; BAZ – Bazadaise; GAS – Gasconne; SAL – Salers; AUB – Aubrac; LIM – Limousin; BLA – Blonde d'aquitaine.

### 2.9.1 Review of the recent multi-breed genomic evaluation studies

It was shown that allele effects estimated in one breed cannot be used for genomic valuation in another breed to obtain accurate estimated breeding values (e.g. Hayes et al., 2009b; Brøndum et al., 2011; Olson et al., 2012).

The most widely used multi-breed genomic evaluation method is when the training populations of different breeds are merged into a single training population, which is then used to estimate allele effects (e.g. Hozé et al., 2014). Other proposed multi-breed methods include a multi-task Bayesian approach (Chen et al., 2014) or a multi-trait model in which the same trait from different breeds are handled as different correlated traits (Olson et al., 2012).

**Simulation studies**

In a simulation study Calus et al. (2008) simulated genotype data of different SNP densities and used them to estimate breeding values. They concluded that for a trait with moderate heritability ($h^2 = 0.5$), LD with $r^2 = 20\%$ is sufficient between neighboring SNP and that stronger LD does not increase selection accuracy. They obtained a somewhat lower value (15%) for haplotypes for the same, moderately heritable trait. For a lowly heritable trait ($h^2 = 0.1$), the optimal value was 20% for SNP and haplotypes likewise. In a very similar experimental setup, VanRaden et al. (2009a) arrived to similar conclusions. Using real data from five populations of three breeds (Angus, Jersey and Holstein), de Roos et al. (2008) estimated that in a within-breed context to obtain an $r^2 \geq 0.20$ between adjacent markers, approximately ~45-75K SNP would be needed across the genome, depending on the population structure. In order to obtain a similar level of LD between adjacent markers, ~300K SNP would be needed in a multi-breed context (de Roos et al., 2008).

Using a simulated 50K SNP-chip data, de Roos et al. (2009a) demonstrated that depending on the simulated genetic distance between the breeds, on the marker density and on the heritability of the trait, genomic evaluation can be efficient even in a multi-breed context. It was also hypothesized that HD data is necessary only if the training population consists of animals from different breeds (de Roos et al., 2009a). That is because breeds are genetically more distant from each other than populations of the same breed. Due to the longer genetic distance, the linkage between adjacent markers (or between markers and QTL) broke down to a greater extent and therefore to capture the effect of a common QTL, SNP that are located closer to the QTL are required. Harris and Johnson (2010b) showed in a simulation study that in order to efficiently exploit the larger marker density from a high-density SNP-chip, a large reference population is required. This is in contradiction with the characteristics of regional breeds, but fits well the concept of multi-breed genomic evaluation (given that the multi-breed training population is large).

**Results based on real data**

Using 50K data, analyses of real dataset including Holstein and Jersey led to the conclusions that multi-breed genomic evaluation can be efficient, but efficiency depends on parameters such as marker density or genetic distance between the breeds (Hayes et al., 2009b; Harris and Johnson, 2010a; Erbe et al., 2012). Similar results, but lower differences were observed when 3 closely related Nordic breeds (Danish Red, Swedish Red and Finnish Red) were analyzed simultaneously (Brøndum et al., 2011) as well as when a mixed population of Holsteins, Jerseys and Fleckvieh was analyzed (Pryce et al., 2011). Analysis of a joint Holstein, Jersey and Brown Swiss population resulted in similar conclusions (Olson et al., 2012).

The genetic gain obtained with multi-breed training population was however limited in the previously mentioned studies. Hayes et al. (2009) and de Roos et al. (2009a) concluded that the inclusion of individuals from a different breed was beneficial if the included breeds diverged more recently or when reference populations included crossbred animals (Lourenco et al., 2016). Larger gains were observed for more heritable traits and/or with a higher marker density.

Also, Bayesian methods were found to perform generally better in a multi-breed context than a GBLUP (e.g. Hayes et al., 2009b; Pryce et al., 2011).

The use of HD data was initially expected to outperform the 50K (Brøndum et al., 2011), especially in small breeds (Hozé et al., 2014; Khansefid et al., 2014). Khansefid et al. (2014) divided the SNP effects into an overall- and a breed-specific component. With such a model, they obtained a limited gain for prediction of residual feed intake using a mixed dairy- and beef cattle population. On a combined Holstein and Ayrshire multi-breed dataset, only a limited increase in selection accuracy was observed with a Bayesian approach compared to a within-breed evaluation (Chen et al., 2014). When analyzing a combined Holstein-Jersey population, Erbe et al. (2012) obtained inferior accuracies with the HD compared to the 50K. Hozé et al. (2014) showed that the potential gain due to a multi-breed training population (with HD data)

is limited when sires of selection candidates are genotyped, which is the case in the four regional breeds presented earlier.

Most of these studies could not show any improvement in selection accuracy for the larger breed contributing to the reference population (usually the Holstein) compared to a within-breed evaluation (e.g. Chen et al., 2014; Erbe et al., 2012). Gains in smaller breeds were often larger, but did not reach expectations. The main challenge in using HD data in genomic evaluation is the ~14-fold increase in the number of allele effects compared to the 50K SNP-chip. Accurate estimation of this many alleles require much more phenotype data. This problem can equally affect single- and multi-breed evaluations.

## 2.10 Problem statement and motivation

In the large dairy cattle breeds, genomic selection led to higher annual genetic gains, drastically decreased costs of selection and selection for a wider range of traits also became possible (e.g. García-Ruiz et al., 2016). These advantages cannot be reached by the means of traditional (i.e. pedigree-based) selection methods, resulting in substantial disadvantages (including economical drawbacks) for regional breeds, where sufficient funding is more difficult to obtain and large reference populations are not available for the implementation of genomic selection in practice.

In our research group, there are several ongoing projects aiming at successfully addressing these challenges. Within the framework of one of these projects, our main aim was to develop new methods and analysis tools for the breeders and breeding organizations of regional breeds (first and foremost the Abondance, Tarentaise and Vosgienne breeds), which would allow them to implement genomic evaluation in practice.

Our primary focus was initially on the use of haplotype markers in combination with the HD SNP-chip in a multi-breed context. Indeed, because of the relatively short genetic distance between these breeds, a multi-breed reference population seemed a good way to increase the reference population size for these breeds. Haplotype markers seemed necessary to maximize the probability of capturing the QTL effects

and the HD SNP-chip was also required to assure a sufficiently high LD between markers and QTL (following the suggestion of, for example, de Roos et al., 2008).

The performance of the methods developed was first evaluated in a single-breed context using a large breed (Montbéliarde) and then in the 4 regional breeds (the previously mentioned 3 breeds together with the Simmental breed). Once the performance of these methods was verified in a within-breed context, they were applied in several multi-breed scenarios using the four regional breeds.

Our long-term aim was to provide an efficient genomic evaluation to breeding organizations of regional breeds and to contribute to the future development of genomic selection in these breeds.

# Chapter 3

# Haplotype construction for genomic evaluation purposes

---

The use of haplotypes is expected to increase the probability of identifying markers linked to QTL affecting the analyzed trait. Furthermore, It was hypothesized that for a multi-breed genomic evaluation to be efficient, the use of HD SNP-chip data is a prerequisite (de Roos et al., 2008). However, the combined use of the HD SNP-chip and haplotypes is currently not realistic, because the number of allele effects to be estimated dramatically increases and in parallel the estimation accuracy of every allele decreases. Overall, this leads to decreased selection accuracy, especially in regional breeds where the amount of phenotypic information is already scarce. To overcome these difficulties, we intended to develop a new haplotype selection procedure that on the one hand allows a more accurate allele effect estimation and on the other hand reduces the number of allele effects to be predicted.

This haplotype selection procedure is presented in detail in this chapter. The chapter is divided into five sections and it starts with the presentation of the dataset used for evaluating the method as well as the first analyses with haplotypes. Then, the

haplotype selection method is presented and evaluated on both 50K and HD data. Finally, possible improvements of the method are presented and discussed.

## 3.1 The Montbéliarde dataset

The Montbéliarde breed was used to test the performance of the developed methods, which breed is one of the large French dairy cattle breeds. The choice of this breed was convenient because for this breed a large reference population of progeny-tested bulls is available, and allows the validation of our results using accurate DYD measures and to compare the performance of different genomic evaluation methods to the performance of a reasonably accurate BLUP analysis.

A population of 2,235 progeny-tested bulls was available for testing. Phenotypes, in the form of DYD were available for 5 production traits: milk yield, protein yield, protein content, fat yield and fat content. Individuals were genotyped either for the 50K or for both the 50K and high-density SNP-chips. Individuals genotyped only on the 50K were imputed to the HD. Multi-allelic markers were removed prior to imputation. Imputation was done by Hozé et al. (2013) using the BEAGLE software (Browning and Browning, 2007). The default parameter values of the software were used for imputation. Imputation accuracy – measured as concordance rate – was ~0.5% with this software. For linkage phasing, the DAGPHASE software (Druet and Georges, 2009) was used, again with the default parameters.

Following imputation, a quality control step was implemented to remove SNP of poor quality. At this step, SNP were removed if at least one of the following conditions was not met:

a) Minor allele frequency higher than 5%
b) Minimum call rate higher than 90%
c) Hardy-Weinberg equilibrium test with $p - value > 10^{-4}$)

After quality control, 43,801 SNP were retained from the 50K SNP-chip panel and 706,791 SNP from the HD-panel. In addition to the phenotype and genotype data, pedigree information was also available.

## 3.2 Haplotypic BayesC-π results

One of our main goals was to assess the benefits of haplotpye-based genomic evaluation methods, particularly in regional breeds. The performance of the developed *haplotypic BayesC-π* (Croiseau et al., 2014; also see section 2.4.4) procedure was first assessed in the Montbéliarde breed. The *haplotypic BayesC-π* was run with all consecutive haplotypes of N SNP used as explanatory variables in the genetic model. Only the 50K SNP-chip was used in this analysis, because the number of allele effects from the HD chip would have been excessively large (this is discussed in detail later). Traits were evaluated independently from each other in a classical validation study, where 20% of the youngest bulls were in the validation population. In practice, 4 different analyses were run for each trait, depending on the value of N (i.e. the number of SNP per haplotype), which ranged from 2 to 5. Performance values ($y_i$) were DYD and the proportion of π was estimated from the data. The following model was used for these tests:

$$y_i = cge_i + u_i + \sum_{j=1}^{N} \delta_j \left( h_{ij}^{p} + h_{ij}^{m} \right) + e_i \qquad (14)$$

where all parameters are as in equation 8 (section 2.4.4). The residual polygenic effect was assumed to account for 20% of the total genetic variance, while the rest of the genetic variance was attributed to the markers.

Running times of the haplotypic BayesC-π ranged from ~16 hours with haplotypes of 2 SNP to ~56 hours with haplotypes of 5 SNP.

**Table 3** gives both the number of haplotypes and the number of allele effects to be estimated during each genomic evaluation procedure with 4 different sizes of haplotype and for both the 50K- and HD-chips (number of alleles per haplotype are taken from **Figure 1**). To create **Table 3**, all consecutive, non-overlapping haplotypes of N SNP (N=2, 3, 4 or 5) were built across all chromosomes; the last markers from every chromosome were truncated if a complete haplotype could not be created. Note that the number of allele effects to be estimated is the total number of alleles

minus the total number of haplotypes, because – as with SNP – for each marker, the effect of one of the alleles (the "reference allele") can be considered to be equal to zero.

**Table 3**: Number of consecutive, non-overlapping haplotypes that can be built with data from either the 50K or the HD SNP-chips and the number of allele effects to be estimated.

| Haplotype size | Number of haplotypes | | Number of allele effects to be estimated | |
|---|---|---|---|---|
| | 50K | HD | 50K | HD |
| 2 | 21 892 | 353 388 | 62 341 | 915 617 |
| 3 | 14 592 | 235 588 | 88 745 | 1 253 312 |
| 4 | 10 936 | 176 688 | 123 886 | 1 702 330 |
| 5 | 8 746 | 141 349 | 168 494 | 2 270 150 |

Based on the Montbéliarde breed

It is clear from **Table 3** that the number of allele effects to be estimated with data from the HD SNP-chip is unreasonably large even with the shortest haplotypes. The number of allele effects to be estimated with the HD chip is close to 1 million with haplotypes of 2 SNP and it rapidly increases to ~2.3 million with haplotypes of 5 SNP. Therefore, it is essential to reduce the number of haplotypes before they can be used in combination with data from the HD SNP-chip for genomic evaluation.

Ideally, the average of samples drawn for each parameter converges to their true values. Lack of convergence of any parameter prevents the estimation of that parameter and therefore convergence is critically important. **Figure 4** gives typical examples of convergence plots for the proportion of haplotypes without an effect (π), the residual variance (vare), the variance attributed to a single haplotype (vara) and the residual polygenic variance (varg). Convergence in case of all these parameters could be observed (visually). In case of all the tests done with the *haplotypic GS3* software, the first 20,000 iterations are discarded as burn-in.

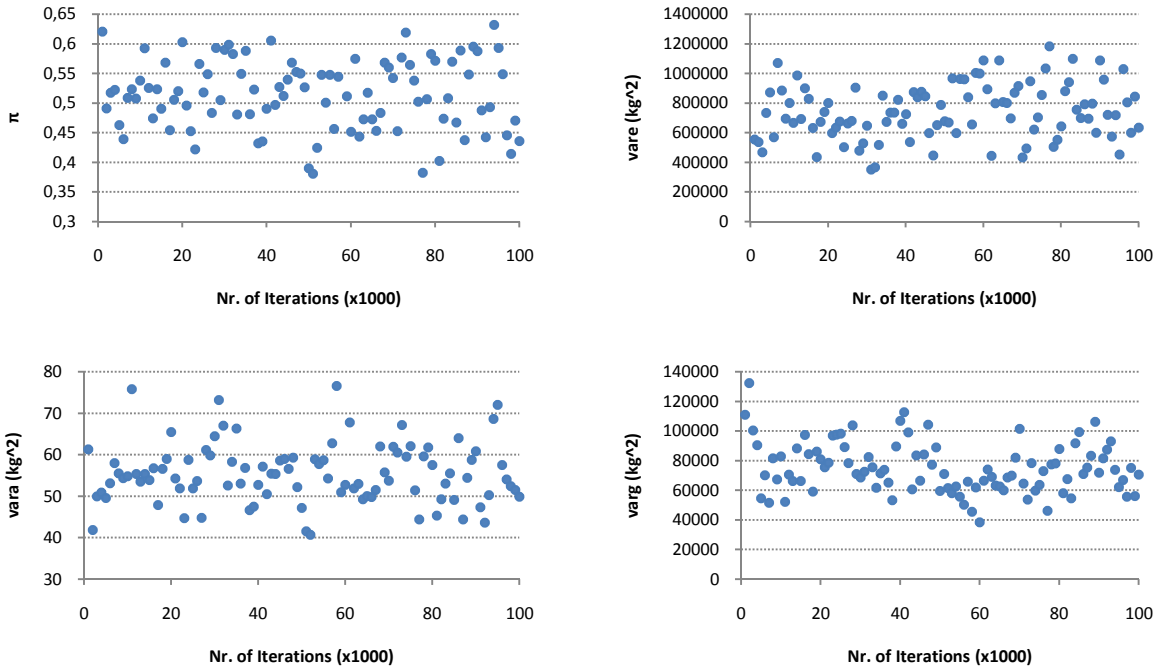



**Figure 4**: Convergence plots obtained with haplotypes of 4 SNP. Proportion of haplotypes without an effect (π), residual variance (vare), variance of a single locus (vara) and residual polygenic variance (varg) are plotted. The thinning value was 1000.

Plots on **Figure 4** indicate that convergence was reached as neither the variation nor the mean of the values change with the number of iterations (x-axis). Since the plots presented in **Figure 4** can be considered as typical ones obtained with the *haplotypic BayesC*(*-π*), no further convergence plots will be presented.

**Table 4** shows the correlation coefficients and regression slopes of DYD on GEBV values obtained in the validation population. Based on these results, the selection accuracy did not vary to a large extent from haplotypes of 2 to 4 SNP. Haplotype size 4 was slightly better than either haplotypes of 2 or 3. The correlation coefficient started declining with haplotypes of 5 SNP, probably due to over-parameterization of the model. Similar trends were observed for the regression slope (on average). Although the haplotype size of 4 SNP slightly outperformed the other haplotype sizes, this advantage was minor and the best performing haplotype size could not be clearly identified based on these results.

**Table 4**: Correlation coefficients and regression slopes of DYD on GEBV values measured on the validation set with *haplotypic-GS3* (Croiseau et al., 2014).

| Trait name[1] | Correlation coefficient | | | | Regression slope | | | |
|---|---|---|---|---|---|---|---|---|
| | HS[2]: 2 | HS: 3 | HS: 4 | HS: 5 | HS: 2 | HS: 3 | HS: 4 | HS: 5 |
| MY | 0.502 | 0.497 | 0.507 | 0.500 | 0.863 | 0.869 | 0.885 | 0.895 |
| FY | 0.557 | 0.557 | 0.563 | 0.559 | 0.863 | 0.871 | 0.912 | 0.905 |
| PY | 0.490 | 0.491 | 0.497 | 0.491 | 0.763 | 0.779 | 0.799 | 0.792 |
| FC | 0.576 | 0.572 | 0.571 | 0.559 | 0.868 | 0.874 | 0.894 | 0.894 |
| PC | 0.596 | 0.589 | 0.593 | 0.581 | 1.055 | 1.052 | 1.090 | 1.094 |
| Average[3] | 0.544 | 0.541 | 0.546 | 0.538 | 0.140 | 0.132 | 0.120 | 0.122 |

1: Trait name abbreviations: MY – milk yield; FY – fat yield; PY – protein yield; FC – fat content; PC – protein content

2: Haplotype size

3: Average deviations from 1 are indicated for regression slopes

The results obtained with the *haplotypic BayesC-π* slightly outperformed the corresponding GBLUP analysis with the **G** matrix constructed from 50K SNP markers (results of the GBLUP analysis are presented in **S. table 1** in Appendix A on page 199). The results presented in **Table 4** were also better than those of a regular, SNP-based BayesC-π (Croiseau et al., 2014).

## 3.3 Influence of allele frequency on genomic evaluation

### 3.3.1 Introduction

In the previous study we used an intuitive way to form the haplotypes by simply merging the adjacent SNP creating the so called flanking haplotypes. This choice (i.e. the flanking haplotypes) is intuitive from a biological point of view, because haplotypes are used to represent specific genomic regions and neighboring SNP necessarily represent the same regions. Therefore, if a QTL is segregating within any region, flanking haplotypes can be expected to be linked to the QTL in the same region. However, from a statistical point of view, flanking haplotypes do not have ideal allele properties: due to the relatively short distance between these markers (see **S. figure 1** in Appendix B on page 201), there is a lower chance for historical recombination events to occur between them. This is particularly the case when data

from the HD-chip is used because LD between consecutive SNP is higher. Therefore, flanking haplotypes are likely to carry a large number of under-represented (rare) alleles for which allele effect estimation is difficult and a small number of largely over-represented alleles. To circumvent these issues, instead of merging the adjacent SNP one can select SNP that result in more appropriate allele properties (i.e. number of alleles and allele frequency distribution), with the expectation that it would enhance the performance of genomic evaluation based on haplotypes. Therefore, the question is: which SNP should be used to create haplotypes with better properties?

In this study, we aimed to develop a procedure to identify haplotypes that can be expected to outperform flanking haplotypes in genomic evaluation studies. Our goal was to maximize the number of haplotype alleles, while taking into account the allele frequency distribution of the haplotypes, i.e., trying to maximize the number of well-represented alleles (alleles with a reasonably high allele frequency) and to minimize the number of rare alleles. In addition, we tried to reduce the overall number of haplotypes used for genomic evaluation, as this was a prerequisite for the combined use of haplotype markers and HD-chip data in genomic evaluation. That is because if haplotypes are used in combination with the HD SNP-chip, the number of allele effects that needs to be estimated would increase to several million (**Table 3**), which is excessive even for the largest breeds. Furthermore, the possible benefits of haplotypes compared to SNP markers were also assessed in this study.

The expected prediction accuracy of the allele effects is also influenced by the size of the effect of the linked QTL: estimated allele effects play an important role even for rare alleles if the linked QTL has a large effect. However, due to lack of *prior* information on the effect size of the QTL , this cannot be directly taken into account to select haplotypes for genomic evaluation purposes *before* the evaluation, in contrast with allele frequencies, which are available *prior* genomic evaluation.

We developed and tested two criteria to select a single haplotype from a set of potential haplotypes based on allele frequency information.

The performance of the selected haplotypes was compared to the results obtained earlier as well as to a regular GBLUP analysis and other SNP- and haplotype-based genomic evaluations. Testing was done using data from both the 50K and HD SNP panels.

### 3.3.2 Alternative haplotype construction methods for genomic evaluation

The article with the haplotype selection method and the 50K SNP-chip results was published in *Journal of Dairy Science* in 2016. The results based on the HD data are presented after the article in a separate section.

Jónás, D., Ducrocq, V., Fouilloux, M-N. and Croiseau, P. 2016. Alternative haplotype construction methods for genomic evaluation. J. Dairy. Sci. 99: 4537-4546.

J. Dairy Sci. 99:4537–4546
http://dx.doi.org/10.3168/jds.2015-10433


# Alternative haplotype construction methods for genomic evaluation

**Dávid Jónás,*†‡[1] Vincent Ducrocq,* Marie-Noëlle Fouilloux,§ and Pascal Croiseau***
*INRA, UMR1313 Génétique animale et biologie intégrative, 78350 Jouy-en-Josas, France
†AgroParisTech, 16 rue Claude Bernard, 75231 Paris 05, France
‡ALLICE, 149 rue de Bercy, 75012 Paris, France
§Idele, UMR1313 GABI, 78352 Jouy-en-Josas Cedex, France

## ABSTRACT

Genomic evaluation methods today use single nucleotide polymorphism (SNP) as genomic markers to trace quantitative trait loci (QTL). Today most genomic prediction procedures use biallelic SNP markers. However, SNP can be combined into short, multiallelic haplotypes that can improve genomic prediction due to higher linkage disequilibrium between the haplotypes and the linked QTL. The aim of this study was to develop a method to identify the haplotypes, which can be expected to be superior in genomic evaluation, as compared with either SNP or other haplotypes of the same size. We first identified the SNP (termed as QTL-SNP) from the bovine 50K SNP chip that had the largest effect on the analyzed trait. It was assumed that these SNP were not the causative mutations and they merely indicated the approximate location of the QTL. Haplotypes of 3, 4, or 5 SNP were selected from short genomic windows surrounding these markers to capture the effect of the QTL. Two methods described in this paper aim at selecting the most optimal haplotype for genomic evaluation. They assumed that if an allele has a high frequency, its allele effect can be accurately predicted. These methods were tested in a classical validation study using a dairy cattle population of 2,235 bulls with genotypes from the bovine 50K SNP chip and daughter yield deviations (DYD) on 5 dairy cattle production traits. Combining the SNP into haplotypes was beneficial with all tested haplotypes, leading to an average increase of 2% in terms of correlations between DYD and genomic breeding value estimates compared with the analysis when the same SNP were used individually. Compared with haplotypes built by merging the QTL-SNP with its flanking SNP, the haplotypes selected with the proposed criteria carried less under- and over-represented alleles: the proportion of alleles with frequencies <1 or >40% decreased, on average, by 17.4 and 43.4%, respectively. The correlations between DYD and genomic breeding value estimates increased by 0.7 to 0.9 percentage points when the haplotypes were selected using any of the proposed methods compared with using the haplotypes built from the QTL-SNP and its flanking markers. We showed that the efficiency of genomic prediction could be improved at no extra costs, only by selecting the proper markers or combinations of markers for genomic prediction. One of the presented approaches was implemented in the new genomic evaluation procedure applied in dairy cattle in France in April 2015.

**Key words:** single nucleotide polymorphism, haplotype, genomic evaluation, dairy cattle

## INTRODUCTION

Virtually all current genomic prediction methods use information from SNP markers (e.g., Meuwissen et al., 2001; Habier et al., 2011), which are abundant all over the genome. However, a major limitation of individual SNP markers as explanatory variables is that each significant causal mutation should be in high linkage disequilibrium (**LD**), with at least 1 SNP to ensure a good prediction. Given the fact that SNP on the commercial SNP chips were selected to have a high minor allele frequency, this requirement is not necessarily fulfilled when the mutated alleles are rare. For example, the development of high-density SNP chips in cattle was expected to overcome this limitation and increase genomic prediction accuracy, but recent studies could show only a limited gain (e.g., Erbe et al., 2012; VanRaden et al., 2013; Ma et al., 2014). Furthermore, the accurate separation and estimation of the effects of closely linked QTL with SNP is not feasible either.

Haplotypes (defined as combinations of 2 or more SNP as in Hayes et al., 2007; Villumsen et al., 2009; Garrick and Fernando, 2014) are multiallelic genomic markers that hold the promise of improving genomic prediction due to higher expected LD between the haplotype and the QTL alleles (e.g., Hayes et al., 2007). Indeed, haplotype information has been used in practical genomic selection in France since 2008, leading to an increased correlation between estimated breeding

Received September 23, 2015.
Accepted February 8, 2016.
[1]Corresponding author: david.jonas@jouy.inra.fr

values and performances as compared with genomic prediction methods based on SNP (Boichard et al., 2012).

Several methods have been used to construct haplotypes for genomic evaluation (Calus et al., 2008, 2009; Boichard et al., 2012; Cuyabano et al., 2014). Allele effect predictability can be defined as the expected prediction accuracy of the effect of haplotype alleles, and it is expected to have a significant effect on the performance of genomic prediction. However, none of the previously mentioned methods take into account any information on this predictability. The construction of haplotypes at a particular SNP position by merging this SNP with the flanking markers is straightforward. However, because of the short distance between the markers, the resulting haplotypes most frequently include a small number of over-represented alleles together with a large number of alleles with low frequencies within the population. An accurate estimation of allele effects for the haplotype alleles that are greatly under-represented is difficult, whereas the abundant information on over-represented alleles does not contribute efficiently to the improvement of genomic estimated breeding value (**GEBV**). The complexity of the statistical model cannot be increased to the range of hundreds of thousands of effects to be estimated, as would happen if all possible nonoverlapping haplotypes of 4 to 5 SNP were considered. Therefore, an efficient haplotype selection procedure is required to identify the haplotypes most suitable for genomic evaluation purposes. In addition, the estimated effects of rare alleles would be generally inaccurate. Hence, the selection of haplotypes with fewer rare alleles would also be beneficial.

For QTL fine mapping, Grapes et al. (2006) showed that it is beneficial to use a selected subset of markers instead of all available markers within a genomic region to build haplotypes, especially when markers are densely distributed. The main objective of the present study was to develop a method to, a priori, construct the most appropriate haplotype for genomic prediction, given a set of SNP previously detected to be in LD with QTL influencing the trait of interest. These SNP will be called QTL-SNP hereafter. Two haplotype selection methods are proposed to select the best haplotype within a window of $N$ SNP around the QTL-SNP based on observed allele frequencies. The goal is to reduce the number of under-represented alleles and to maximize the number of alleles properly represented in the population under study. The predictability of an allele effect also depends on the effect size of the linked QTL (Meuwissen et al., 2001), but this information is not available at the haplotype selection step. The effect on genomic prediction of haplotypes from the 2 haplotype selection methods versus haplotypes built from flanking markers around the QTL-SNP was compared on a real data set.

## MATERIALS AND METHODS

### General Notation

The term "QTL-SNP" refers to SNP in strong LD with causative mutations affecting a trait of interest. These SNP were identified using a Bayes-Cπ procedure (see details below). Haplotypes are defined as combinations of $N$ SNP along a chromosome (similar to the definitions of Hayes et al., 2007; Villumsen et al., 2009; Garrick and Fernando, 2014). The term "allele" refers to the alternative forms of a genetic marker present in a population; considering SNP, 2 alleles are present per marker, whereas haplotypes can be composed of $2^N$ different alleles, where $N$ is the haplotype size in number of SNP. "Flanking SNP" of a QTL-SNP are the nearest SNP surrounding the QTL-SNP. "Flanking haplotypes" are the haplotypes that are built by merging the QTL-SNP and the flanking SNP into a single haplotype. A short genomic segment around the QTL-SNP defined in number of SNP is referred to as a "QTL window," or simply as a "window."

In this study, the QTL-SNP were considered as markers indicating the approximate positions of the QTL affecting the trait of interest. A short, symmetric genomic window was constructed around each QTL-SNP and these genomic segments were assumed to contain the linked QTL. Our aim was to select a single haplotype of $N$ SNP per window to represent the QTL within that window in genomic prediction. Once haplotypes were selected around each QTL-SNP, all of them were used in genomic prediction to predict breeding values for the individuals in the validation population.

### Data and QTL Detection Methods

Performance values in the form of average daughter yield deviations (**DYD**) for 5 dairy cattle production traits (milk quantity, fat content, fat yield, protein content, and protein yield) were available for 2,235 Montbéliarde bulls genotyped with the Bovine SNP50 BeadChip (50K; Illumina Inc., San Diego, CA). Only autosomal chromosomes were used. After quality control, 43,801 SNP were retained from the 50K chip. In a first step, a QTL detection was undertaken using a Bayes-Cπ approach as implemented in the GS3 software by Legarra et al. (2013). The model used in this SNP-based Bayes-C analysis was:

$$y_i = \mu + u_i + \sum_{j=1}^{N} z_{ij} a_j \delta_j + e_i,$$

where $y_i$ is the performance value of individual $i$, μ is an overall mean effect, $u_i$ is the residual polygenic effect of animal $i\left[u \sim MVN\left(0, \mathbf{A}\sigma_u^2\right)\right]$, where $MVN$ is multivariate normal distribution, $\mathbf{A}$ is the additive relationship matrix, and $\sigma_u^2$ is 0.2 times the genetic variance, $N$ is the total number of SNP in the model, $z_{ij}$ is an indicator variable representing the number of copies of one of the alleles at marker $j$ in animal $i$, $a_j$ is the substitution effect of marker $j$, $\delta_j$ is a 0/1 variable indicating whether or not marker $j$ is assumed to have an effect, and $e_i$ is a random error term for animal $i$. The residual polygenic effect was assumed to account for 20% of the total genetic variance, whereas the rest of the genetic variance was attributed to the selected markers. Following the Bayes-Cπ analysis, the $k$ SNP with the largest probability of inclusion in the model were considered to be QTL. These SNP will be called QTL-SNP. This step was done within the framework of a classical validation study, using the same training and validation populations as for the haplotype-based tests (see in detail below). In practice, the first 1,000, 3,000, and 6,000 QTL-SNP were selected for each trait (denoted as 1K, 3K, and 6K, respectively). Due to this selection procedure, for each trait, every smaller set is a subset of the larger set(s). It is expected that these QTL-SNP were in strongest LD with the causative mutations.

The original GS3 software by Legarra et al. (2013) was extended to deal with haplotypes (Croiseau et al., 2014). This haplotypic Bayes-C was used for genomic evaluation and for testing the performance of the different haplotype construction methods. Haplotypes were modeled as class variables, with one effect predicted for each haplotype allele. The proportion π of haplotypes with no effect was fixed because of practical considerations: the haplotypic Bayes C was very time-consuming due to the increased number of effects to estimate. Fixing π allowed us to perform a large number of tests within a reasonable time, without sacrificing accuracy. Moreover, preliminary tests showed that fixing π led to validation correlations slightly higher as compared with a scenario where π was estimated during the analysis due to poor mixing in the latter case (data not shown). A constant value of π (90%) was selected because it gave a number of marker effects to be estimated similar to the number of individuals in the training population. The same model was used for the haplotype-based Bayes-C analyses as for the SNP-based tests, with the SNP effects being replaced by the haplotype effects.

Out of the 2,235 bulls with both phenotype and genotype information, the youngest 20% of individuals were selected as the validation population. Allele effects were estimated using the training population (that is, the oldest 80% of animals) and GEBV were estimated for the individuals in the validation population using only genomic information of that population and the estimated allele effects. Accuracy of the breeding value estimation was measured by the correlation coefficient between GEBV and DYD values of the validation population. The performance of the different haplotype construction methods was evaluated based on this parameter. In addition, the slopes of the regression of DYD on GEBV were calculated and compared.

### Haplotype Selection

Haplotypes were constructed within each QTL window. The most desirable one was supposed to maximize the number of alleles with an allele frequency higher than a given threshold. As previously mentioned, it is advantageous in genomic prediction to avoid both under- and over-represented alleles.

Once a window of window size (**WS**; the size in number of markers) SNP was defined around each QTL position, every possible haplotype of haplotype size (**HS**; the size in number of markers) SNP was constructed. Three different methods with different criteria were used, and each of these methods resulted in a haplotype within each window. The performances of these haplotypes (methods) in genomic evaluation were compared. These criteria are described in detail below. Considering that the QTL-SNP had the strongest LD within a window with the linked QTL, this SNP was always forced to be part of the final haplotype. The number of haplotypes that can be built within the window is therefore

$$\binom{\mathrm{WS}-1}{\mathrm{HS}-1} = \frac{(\mathrm{WS}-1)!}{(\mathrm{HS}-1)! \times (\mathrm{WS}-\mathrm{HS})!}.$$

One haplotype was selected from each window to be used in genomic evaluation based on 3 different approaches. These approaches were termed as flanking markers, criterion-A, and criterion-B and their performances were compared. To test the effect of the WS and HS on genomic prediction, windows of size WS = 10, 15, and 20 SNP, as well as haplotypes of size HS = 3, 4, and 5 SNP were constructed. All WS and HS combinations were tested.

***Flanking Markers.*** The QTL-SNP and its flanking markers were grouped into a haplotype. Haplotype allele frequency was not considered. Flanking markers were always considered symmetrically around the QTL-SNP: the flanking haplotype built from 5 SNP included the QTL-SNP and 2–2 flanking SNP on both sides of the QTL-SNP. When HS was an even number (i.e., an odd number of SNP had to be selected on the 2

sides of the QTL-SNP), a symmetric haplotype of (HS + 1) SNP was created around the QTL-SNP and the marker that was the farthest from the QTL-SNP was excluded from the haplotype. The same principle was used when asymmetric windows had to be constructed around the QTL-SNP.

***Criterion-A.*** A threshold level denoted as allele frequency threshold (**AFT**) was used to determine which alleles are considered predictable (i.e., which allele effects can be predicted with satisfactory accuracy). The following AFT values were tested: 1, 3, 5, and 8%.

With criterion-A, a 2-step approach was implemented. First, for each haplotype $i$ within a specific window, the number of predictable alleles (i.e., with a frequency higher than AFT) was determined. Then for the haplotypes carrying the maximum number ($N_{\max}$) of predictable alleles within the window, a score ($SD_{hi}$) was calculated as the squared deviation of observed allele frequencies from the ideally balanced allele frequency, where the latter was equal to $1/N_{\max}$. The score can be written as

$$SD_{hi} = \sum_{k=1}^{N_i} \left( OF_{i,k} - \frac{1}{N_i} \right)^2,$$

where $h_i$ is haplotype $i$, $N_i$ ($=N_{\max}$) is the number of predictable alleles of haplotype $i$, and $OF_{i,k}$ is the observed frequency of allele $k$ of haplotype $i$. Retaining the haplotype with the lowest squared deviation score guarantees that the observed allele frequencies are as balanced as possible.

***Criterion-B.*** A drawback of criterion-A is that the allele frequencies can still be unbalanced to a high degree, because haplotypes with more predictable alleles are always preferred over haplotypes with fewer predictable alleles. This is true even if, for example, many alleles of a certain haplotype have a frequency that barely exceeds the threshold level, whereas a small number of alleles are greatly over-represented in the population. Criterion-B consists of 2 parts, from which the first part is a modified version of the SD score calculated for criterion-A. The difference is that $1/N_i$ is replaced by $1/2^{\mathrm{HS}}$ to ensure that this part is, assuming similar variations in the allele frequencies, smaller for haplotypes with a higher number of predictable alleles. This is guaranteed because the observed frequencies of the predictable alleles will on average get closer to $1/2^{\mathrm{HS}}$ as their number is increasing. The second part is a weighted number of predictable alleles. It ensures that out of haplotypes that carry the same number of alleles, the haplotype(s) that include more predictable alleles have a lower score. A parameter that we call maximum deviation (**MD**) was introduced in the computation of the weight (see Supplemental Materials for details; http://dx.doi.org/10.3168/jds.2015-10433). It is defined as the average acceptable deviation of ($n$ − 1) alleles from the ideal frequency $\left(\frac{1}{2^{HS}}\right)$, expressed as a proportion of the ideal frequency. The $n$th allele must have a frequency equal to or larger than AFT. The MD parameter can be interpreted as follows: the smaller its value is, the less the allele frequencies are allowed to deviate from their mean. For example, if MD is set to a relatively strict value of 10%, haplotypes with fewer predictable alleles are favored when their allele frequencies are more balanced against haplotypes with more predictable alleles, but with a larger variation among the frequencies of those alleles.

In practice, criterion-B is calculated as

$$\text{Criterion-B}_{hi} = \sum_{k=1}^{N_i} \left( OF_{i,k} - \frac{1}{2^{\mathrm{HS}}} \right)^2 - w \times N_i,$$

where $w$ is the weighing factor of the number of predictable alleles. The second term of criterion-B is negative to be consistent with the first term, which is optimal when it takes the smallest value.

Table 1 illustrates the difference between criterion-A and -B. Criterion-A would prefer the second haplotype over the fourth despite of its highly unbalanced allele frequencies. This preference is reversed with criterion-B, assuming appropriate AFT and MD values.

An analysis using only the QTL-SNP as genomic markers was conducted to obtain a basis for comparisons. This analysis was conducted on all sets of QTL-SNP (1K, 3K, and 6K) and the optimal number of QTL-SNP was selected for each trait. The benefit of haplotypes versus SNP was judged by analyzing the same SNP selected by each method in a Bayes C model utilizing them as single-SNP information. The observed correlations between DYD and GEBV from these analyses were compared with those obtained with their haplotype counterparts. A genomic BLUP analysis with all retained SNP markers was also performed to complete the tests.

## RESULTS AND DISCUSSION

Table 2 shows the number of haplotypes that can be built for several different WS and HS values. The windows have a reasonably small number of combinations. Haplotype selection was performed on a single processor and running time was less than 1 min for windows of 10 SNP, haplotypes of 4 SNP and 3,000 QTL-SNP, where the total number of evaluated haplotypes was 252,000.

**Table 1**. Allele frequencies for 4 haplotypes; the selection order with both criterion-A and -B is also shown

| | | Allele frequencies | | | | | |
|---|---|---|---|---|---|---|---|
| Criterion-A | Criterion-B | A1 | A2 | A3 | A4 | A5 | A6 |
| 1 | 1 | 0.167 | 0.167 | 0.167 | 0.167 | 0.167 | 0.165 |
| 2 | 4 | 0.70 | 0.06 | 0.06 | 0.06 | 0.06 | 0.06 |
| —[1] | 3 | 0.2 | 0.2 | 0.2 | 0.19 | 0.19 | 0.02 |
| —[1] | 2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | — |

[1]As the first 2 haplotypes have 6 predictable alleles (assuming a threshold of allele frequency threshold = 5%), these haplotypes are not considered in the second step of criterion-A.

### *Distribution of Allele Frequencies*

The number of alleles with very low allele frequencies (<1%) decreased with criterion-A and -B compared with the flanking markers approach. With flanking markers and 6K QTL-SNP in the model, 2,660 alleles (i.e., 3.6% of the alleles in the population had frequency >40%) were termed as over-represented alleles; almost half of the flanking haplotypes included one such allele. The proportion of over-represented alleles with the haplotypes selected by either criterion-A or criterion-B was approximately half of this value: 2.1 and 1.56%, respectively. In case of haplotypes of 4 and 5 SNP, criterion-B tended to select haplotypes with slightly fewer rare and over-represented alleles than criterion-A.

Figure 1 shows the distribution of alleles present in the population according to their allele frequency for HS = 4, WS = 10 SNP, and 6,000 QTL-SNP. The use of criterion-A and -B led to a higher proportion of haplotype alleles in the 5 to 30% frequency range, but also to a lower proportion of over-represented alleles. These trends were observed whatever the haplotype size. The difference between the haplotypes built from the flanking markers and from the selected markers decreased when the haplotype size increased (data not shown).

Table 3 shows the average number of alleles per haplotype for different haplotype selection methods, haplotype sizes, and number of QTL-SNP. As expected, with the increase of the haplotype size, the number of segregating alleles increased rapidly. However, it was close to its theoretical maximum value ($2^{HS}$; i.e., 8, 16, or 32 for HS = 3, 4, or 5) only when HS = 3. This is not surprising, given the relatively dense SNP chips available and the corresponding high LD.

Interestingly, the average number of segregating alleles per haplotype was decreasing as the number of QTL was increasing from 1,000 to 6,000 (Table 3). One interpretation is that QTL with smaller effects (i.e., those QTL-SNP added when moving from 1,000 to 6,000 QTL in the model) are segregating in less polymorphic regions of the genome compared with QTL with larger effects. The reduced number of haplotype alleles might also slightly affect the prediction accuracy, as the probability of having at least 1 allele in strong LD with the QTL is reduced. This trend was apparent with all marker construction methods; however, the magnitude of the decrease is larger with criterion-A and -B than it is with the flanking marker haplotypes.

The number of rare and over-represented alleles was lower with criterion-B. The frequencies of these alleles were also more favorable with criterion-B than with criterion-A; rare alleles had a higher average frequency with criterion-B, whereas the average frequency of the over-represented alleles decreased when compared with criterion-A (data not shown). All of these are beneficial features for genomic prediction, which can be attributed to the changes made in criterion-B. These are the additional constraint on the allele frequency equilibrium and the replacement of $1/N_i$ by $1/2^{HS}$ in the equation of the SD. The total number of segregat-

**Table 2**. Number of possible haplotypes with different window and haplotype sizes

| Window size | Without forcing the QTL-SNP[1] | | | With forcing the QTL-SNP[2] | | |
|---|---|---|---|---|---|---|
| | HS[3] = 3 | HS = 4 | HS = 5 | HS = 3 | HS = 4 | HS = 5 |
| 10 | 120 | 210 | 252 | 36 | 84 | 126 |
| 15 | 455 | 1,365 | 3,003 | 91 | 364 | 1,001 |
| 20 | 1,140 | 4,845 | $1.55 \times 10^4$ | 171 | 969 | 3,876 |

[1]All possible haplotypes within the window are considered, whether they include the QTL-SNP or not.

[2]Within a window, only haplotypes that include the QTL-SNP are considered. Good candidate QTL-SNP are required.

[3]HS = haplotype size.

**Table 3**. Average number of alleles per haplotype observed with the 3 different haplotype construction methods, as function of haplotype size and number of QTL-SNP in the model[1]

| | Number of QTL-SNP | | |
|---|---|---|---|
| Item | 1,000 | 3,000 | 6,000 |
| HS[2] = 3 | | | |
| Flanking markers | 7.40 | 7.22 | 7.08 |
| Criterion-A | 7.51 | 7.21 | 6.87 |
| Criterion-B | 7.56 | 7.23 | 6.86 |
| HS = 4 | | | |
| Flanking markers | 13.42 | 12.80 | 12.33 |
| Criterion-A | 13.84 | 12.89 | 11.90 |
| Criterion-B | 14.41 | 13.43 | 12.43 |
| HS = 5 | | | |
| Flanking markers | 23.16 | 21.43 | 20.27 |
| Criterion-A | 22.70 | 20.74 | 18.81 |
| Criterion-B | 26.62 | 24.04 | 21.78 |

[1]Window size: 10 SNP

[2]HS = haplotype size.

ing alleles with criterion-B did not change as the AFT threshold increased, in contrast with criterion-A (see Supplemental Table S1; http://dx.doi.org/10.3168/jds.2015-10433). The number of alleles with very low (<1%) allele frequencies tended to increase with increasing AFT, whereas the number of the moderately frequent alleles (1–10%) systematically decreased (data not shown).

Although the proposed methods favor haplotypes with intermediate allele frequencies, rare alleles are inevitable. For example, with haplotypes of 4 SNP and AFT of 8%, the proportion of alleles with frequency less than 8% was 63 to 64% with the haplotypes selected by criterion-A or -B instead of ~69% with the flanking markers.

### Correlations Between DYD and GEBV Values

Genomic prediction of a set of dairy cattle production traits was implemented to investigate the performance of the haplotypes selected by the different methods.

***AFT Tests.*** The optimal AFT for the studied population with both criterion-A and -B was 8% (see Supplemental Table S2; http://dx.doi.org/10.3168/jds.2015-10433). The effect of the choice of AFT on correlations decreased when the number of QTL increased (data not shown). This may be related to the fact that the smaller QTL were segregating in less polymorphic parts of the genome, where fewer but more frequent alleles were segregating. The AFT parameter had only a minor effect on the prediction accuracy; it also had a smaller effect on the results of criterion-B than on those of criterion-A (Supplemental Table S2). The AFT was fixed to 8% for the rest of the analysis.

***MD Tests.*** Several values were tested for the MD parameter of criterion-B, which were chosen to cover

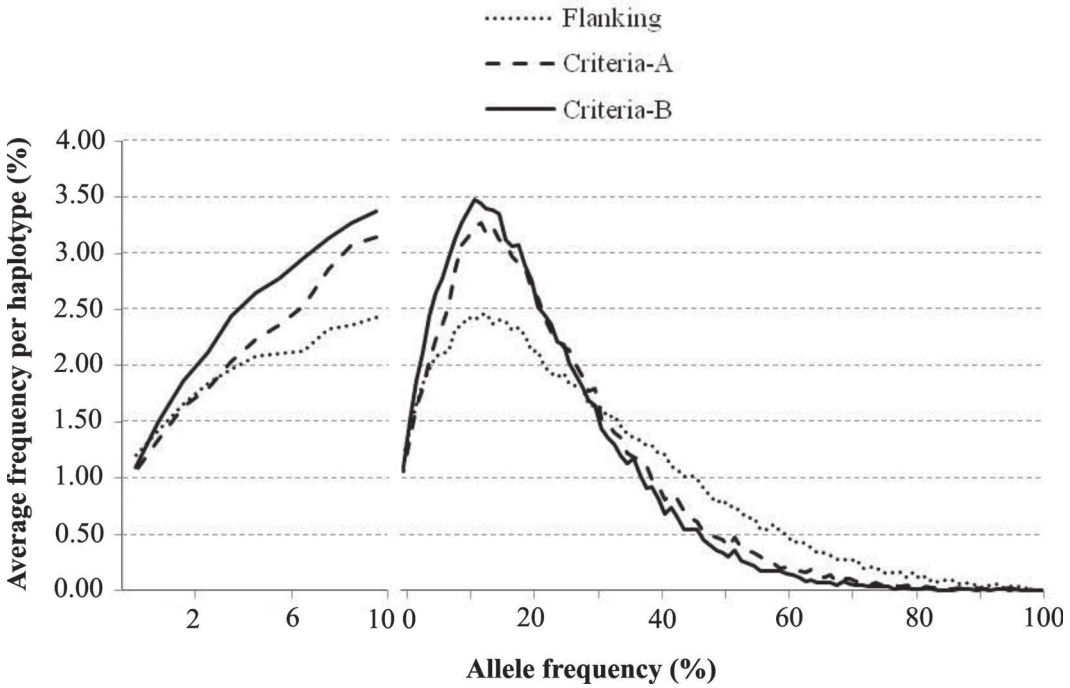



**Figure 1**. Overall distribution of haplotype allele frequencies according to the haplotype construction approach (haplotype size: 4 SNP; window size: 10 SNP; 6,000 QTL-SNP). The 0 to 10% region is also depicted with more detailed scale on the x-axis.

the whole range between 0 and 1. No large differences were observed in correlations with regard to this parameter (see Supplemental Table S3; http://dx.doi.org/10.3168/jds.2015-10433). As the MD value had only a marginal effect on the results, its value was fixed to 10% (i.e., more strongly favoring more balanced allele frequencies over a higher number of predictable alleles).

***Comparison of the Haplotype Construction Methods.*** Table 4 shows the correlations between DYD and GEBV in the validation population obtained with the analysis using either only the QTL-SNP as genomic markers or the haplotypes built from the flanking markers. Hereafter, all correlations and differences in correlations are reported in percentage points. Flanking markers outperformed the analyses, which solely used the QTL-SNP in all scenarios. The observed gain ranged between 0.8 and 2.9%, and it was larger with longer haplotypes and with a higher number of QTL-SNP in the model. The optimal number of QTL-SNP was 6,000 for most of the traits. The average gain observed for the 5 traits was 2.1 to 2.9% with flanking markers, again increasing with haplotype size. Similar results were found with criterion-A and -B, except that haplotype size 5 did not result in higher correlations than haplotypes of 4 SNP (see Supplemental Table S4; http://dx.doi.org/10.3168/jds.2015-10433).

Figure 2 shows the obtained correlations between DYD and GEBV values of the validation population with the different haplotype sizes and haplotype selection methods after selecting the optimal number of QTL-SNP for each trait. The solid lines represent the analyses using the selected SNP as haplotypes and the dashed lines correspond to the analyses using the same SNP as individual SNP information sources in genomic prediction. Average correlations of the 5 production traits are shown (for the individual results, see Supplemental Table S5; http://dx.doi.org/10.3168/jds.2015-10433). Merging the SNP into haplotypes was beneficial in all cases, leading to an increase of 1.4% in correlations when the obtained gain was averaged across the 3 haplotype construction methods. This increase in correlation was 2% when only the highest correlation for each trait was considered from those observed with 1K, 3K, and 6K haplotypes in the model. This gain was positively correlated with the increase of number of haplotypes in the model, showing an increase of 0.7, 1.6, and 1.9% with 1,000, 3,000, and 6,000 QTL modeled, respectively. No large differences were observed between the haplotype selection methods in this aspect. With the presented criteria in general, haplotypes of 5 SNP performed worse than the shorter haplotypes; on average for the 5 production traits, no additional gain was observed with criterion-A and HS = 5, compared with its flanking haplotypes counterpart (see Supplemental Table S5). The poor performance of haplotypes of 5 SNP might be a result of over-parameterization of the model. The average gains with criterion-A compared with the flanking marker haplotypes were 1.3 and 0.6% with haplotypes of 3 and 4 SNP, respectively. Haplotypes selected by criterion-B outperformed those selected by criterion-A by 0.3% on average. The observed gain compared with the flanking haplotypes with both criterion-A and criterion-B was decreasing as the haplotype size increased. This can be attributed to the diminishing differences in total number of alleles between the haplotype construction methods with increasing haplotype size (data not shown). Finally, the average correlation of the 5 production traits with genomic BLUP was 0.535; the correlations between DYD and GEBV were 1.1% higher with haplotypes built with criterion-A or -B than with a standard genomic BLUP analysis.

***WS Tests.*** The effect of window size used for haplotype construction on genomic prediction results was also investigated. Windows of 10, 15, and 20 SNP were constructed and haplotypes were selected from these windows for genomic prediction, using a value of 8% for AFT and 10% for MD. Table 5 shows the results obtained with the different window sizes for both criterion-A and criterion-B and for the 3 tested haplotype sizes. It was expected that wider windows would result in lower correlation due to a decreasing LD between QTL and haplotypes. This was indeed observed for

**Table 4**. Observed correlations between daughter-yield deviations and genetic EBV values using either only the QTL-SNP or the flanking haplotypes as genomic markers (average correlations over the 5 traits)

| Number of QTL-SNP | QTL-SNP | Flanking markers | | |
|---|---|---|---|---|
| | | HS[1] = 3 | HS = 4 | HS = 5 |
| 1K | 0.480 | 0.491 | 0.492 | 0.488 |
| 3K | 0.499 | 0.523 | 0.526 | 0.528 |
| 6K | 0.512 | 0.534 | 0.538 | 0.541 |
| Optimal[2] | 0.512 | 0.534 | 0.538 | 0.542 |

[1]HS = haplotype size.

[2]For each trait separately, the number of QTL-SNP/haplotypes is the one leading to the highest correlation.

**Table 5**. Correlations between the daughter-yield deviations and genetic EBV values for the tested window sizes[1]

| Haplotype selection method | Window size | Haplotype size | | |
|---|---|---|---|---|
| | | 3 | 4 | 5 |
| Criterion-A | 10 | 0.537 | 0.541 | 0.547 |
| | 15 | 0.542 | 0.550 | 0.543 |
| | 20 | 0.538 | 0.548 | 0.547 |
| Criterion-B | 10 | 0.548 | 0.548 | 0.546 |
| | 15 | 0.541 | 0.543 | 0.549 |
| | 20 | 0.540 | 0.546 | 0.545 |

[1]Average correlations over the 5 traits are shown (allele frequency threshold: 8%; maximum deviation: 10%). The optimal number of QTL-SNP was selected, as described in the manuscript. Allele frequency threshold = only alleles with a frequency higher than this threshold are assumed to be sufficiently predictable; maximum deviation = controls the acceptable level of variation among allele frequencies.

the correlations obtained with haplotypes constructed using criterion-B. However, the results obtained with criterion-A showed a small increase in correlations with the increase of window sizes. The apparent inconsistency in the results with respect to the effect of window size might be a result of different LD patterns around the different QTL-SNP in the model, for which the same window size was applied in our study. This might have resulted in windows that overlap with recombination sites or hotspots, greatly reducing the LD between the selected haplotypes and the linked QTL. Undoubtedly, the frequency of such windows increases with the increase of the window size. Therefore, in practical applications, it might be beneficial to take into account additional information for the definition of the windows, such as recombination hotspots or the LD pattern of the SNP along the genome. However, the testing of the effect of this information was outside the scope of our study.

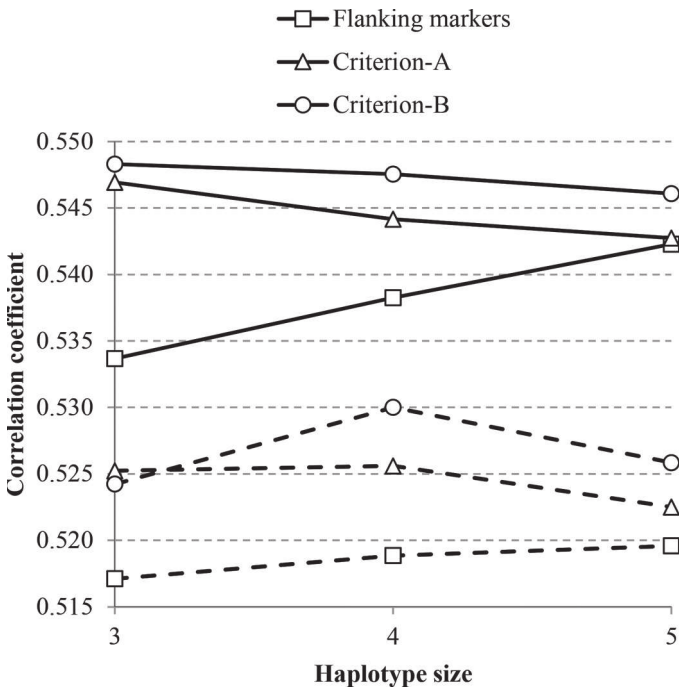



**Figure 2**. Observed correlations between daughter yield deviation (DYD) and genetic EBV (GEBV) values in the validation population with the different haplotype selection methods and haplotype sizes after selecting the optimal number of QTL-SNP for each trait. Average correlations of the 5 production traits are shown. Solid lines show the correlations for the haplotype-based analyses, whereas dashed lines show the correlations observed when the same SNP were used as single-SNP markers. Windows of 10 SNP were used for criterion-A and criterion-B.

Obviously, it is desirable to adjust parameter values for the model to the studied population. For example, population size has a major effect on the optimal AFT value; in larger populations, lower AFT values can be used. However, the presented criteria (especially criterion-B) appear to be robust to the choice of parameter values within the tested limits. With criterion-B, an increased risk of over-parameterization was noted with haplotypes of 5 SNP (compared with the flanking haplotype situation) due to the higher number of segregating alleles per haplotype (11.5% larger, on average).

### *Slope of Regression*

The average slope of regression of DYD on GEBV with haplotypes of 4 SNP over the 5 traits were 0.80, 0.80, and 0.83 with the flanking, criterion-A, and criterion-B haplotypes, respectively. When the same markers were used as single-SNP information, the slopes of regression were in the same order, 0.71, 0.73, and 0.75, respectively. The regression slope was 0.83 with the genomic BLUP model. In all cases, these values are relatively far from the desirable value of 1. Higher values were obtained when the fraction of the total genetic variance allocated to the residual polygenic effect was increased (data not shown); however, optimization of this slope was outside the scope of this paper.

### Statistical Analysis

The average differences between the correlations of criterion-B and those of the flanking markers (short horizontal lines), as well as the calculated lower confidence bounds of the tests (triangles), are shown on Figure 3. Criterion-B led to a small increase in correlation in almost all of the cases (see also Supplemental Table S5; http://dx.doi.org/10.3168/jds.2015-10433). The significance of the observed increase in correlation between DYD and GEBV was tested using Fisher's Z-transform, as implemented in the "cocor" R package (Diedenhofen and Musch, 2015) based on the work of Zou (2007). As the results of criterion-B were slightly better than those with criterion-A, these were compared with the flanking haplotypes. To test whether haplotypes selected with criterion-B outperform flanking haplotypes, a one-tailed test with $\alpha = 0.05$ and the null hypothesis that the 2 correlations are equal was performed. Out of the 15 correlations (5 traits × 3 haplotype sizes; the correlation coefficients are present in Supplemental Table S5), 3 were found to be significantly better with criterion-B than with the flanking haplotypes.

A Wilcoxon signed-rank test was performed to assess whether criterion-B, compared with the flanking markers, led globally to increased correlations. The Wilcoxon signed-rank test was chosen because normality could not be assumed due to the low sample size (n = 15) and because the available data were paired; for every HS or trait combination, a correlation coefficient was available in both the flanking marker and criterion-B cases.

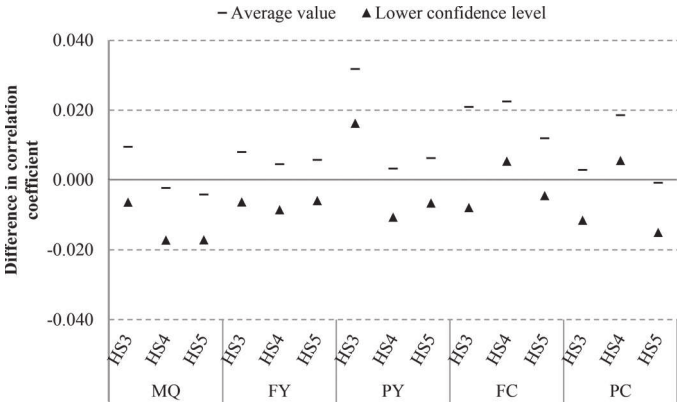



**Figure 3**. Average differences between the correlation coefficients (correlations calculated with criterion-B (using windows of 10 SNP) minus those calculated using the flanking markers) are represented by the short horizontal lines. The lower confidence intervals for the differences based on Fisher's Z-transform are also shown (black triangles). HS3, HS4, and HS5 = haplotype sizes 3, 4 and 5, respectively; MQ = milk quantity; FY = fat yield; PY = protein yield; FC = fat content; PC = protein content.

To account for the wide range of correlations for the different traits, they were first standardized by calculating their deviation from the correlation coefficients observed when only the QTL-SNP were used:

$$\text{gain}_{z,t} = \left(p_{z.hap,t} / p_{\text{QTL-SNP},t}\right) - 1,$$

where $z$ refers to one of the haplotype selection scenarios (flanking marker, criterion-A, or criterion-B), $p_{z.hap,t}$ is the observed correlation coefficient with the haplotype-based analysis using scenario $z$ for trait $t$, $p_{\text{QTL-SNP},t}$ is the observed correlation coefficient with the analysis using only the QTL-SNP as genetic markers for trait $t$, and $\text{gain}_{z,t}$ is the observed relative gain in correlation between the 2. The Wilcoxon signed-rank test was performed using $\alpha = 0.05$ (one-tailed test). The test results (W = 111 and $P = 0.001$) indicate that the haplotypes selected by criterion-B significantly increased the correlations between DYD and GEBV compared with the flanking haplotypes. The test with criterion-A was also significant (W = 76, $P = 0.02$).

### Final Remarks

The alleles that are considered predictable based solely on their allele frequencies and those that are actually well predicted in genomic selection are not equivalent because the predictability of an allele also depends on the effect size of the linked QTL. Therefore, whereas alleles carried by a sufficiently large number of individuals in the population are always predictable, effects of rare alleles can be also accurately predicted if those alleles are in strong LD with large QTL. Hence, the efficiency of haplotype selection procedures can be further improved in the future, once objective measures of QTL effect sizes will be available.

At present, interest is increased in using haplotypes as genomic markers in genomic evaluation procedures. The efficiency of the methods presented in our study might be further improved by, for example, identifying window boundaries in a more precise way [for examples, see Cuyabano et al. (2014) and Beissinger et al. (2015)].

Criterion-B is part of the new genomic evaluation procedure, which was implemented for the 4 dairy cattle breeds (Holstein, Montbéliarde, Normande, and Brown Swiss) in France in April 2015 (Croiseau et al., 2015).

## CONCLUSIONS

Two methods to improve haplotype allele predictability based on observed allele frequencies were presented and compared with haplotypes created from the

flanking markers. The obtained results indicate that an a priori selection of haplotypes from a small genomic region around each QTL-SNP can improve the correlations between DYD and GEBV at no extra costs. In addition, the proposed methods are data-independent and require neither large computing power nor excessive running time. The inclusion of additional constraints on the allele frequency equilibrium in the haplotype selection procedure was beneficial, further increasing the correlations between DYD and GEBV by 0.3% on average over 5 production traits.

## ACKNOWLEDGMENTS

The GEMBAL (Génomique Multi-race des Bovins Allaitants et Laitiers) project is funded by the Agence Nationale de la Recherche (ANR-10-GENM-0014), Apisgene, Races de France, and INRA "AIP Bioressources."

## REFERENCES

Beissinger, T. M., G. J. Rosa, S. M. Kaeppler, D. Gianola, and N. de Leon. 2015. Defining window-boundaries for genomic analyses using smoothing spline techniques. Genet. Sel. Evol. 47:30. http://dx.doi.org/10.1186/s12711-015-0105-9.

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, and S. Fritz. 2012. Genomic selection in French dairy cattle. Anim. Prod. Sci. 52:115–120. http://dx.doi.org/10.1071/AN11119.

Calus, M. P., T. H. E. Meuwissen, J. J. Windig, E. F. Knol, C. Schrooten, A. L. Vereijken, and 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. http://dx.doi.org/10.1186/1297-9686-41-11.

Calus, M. P. L., T. H. E. Meuwissen, A. P. W. de Roos, and R. F. Veerkamp. 2008. Accuracy of genomic selection using different methods to define haplotypes. Genetics 178:553–561. http://dx.doi.org/10.1534/genetics.107.080838.

Croiseau, P., A. Baur, D. Jónás, C. Hozé, J. Promp, D. Boichard, S. Fritz, and V. Ducrocq. 2015. Comparison of different Marker-Assisted BLUP models for a new French genomic evaluation. Page 248 in Book of Abstracts of the 66th Annual Meeting of the European Federation of Animal Science, Warsaw University of Life Sciences, Poland. Wageningen Academic Publisher, Wageningen, the Netherlands.

Croiseau, P., M. N. Fouilloux, D. Jónás, S. Fritz, A. Baur, V. Ducrocq, F. Phocas, and D. Boichard. 2014. Extension to haplotypes of genomic evaluation algorithms. Abstract 708 in Proc. 10th World Congress of Genetics Applied to Livestock Production, Vancouver, Canada. Am. Soc. Anim. Sci., Champaign, IL. https://asas.org/docs/default-source/wcgalp-posters/708_paper_10043_manuscript_1181_0bFD602C6D9AD3.pdf?sfvrsn=2.

Cuyabano, B. C. D., G. Su, and M. S. Lund. 2014. Genomic prediction of genetic merit using LD-based haplotypes in the Nordic Holstein population. BMC Genomics 15:1171. http://dx.doi.org/10.1186/1471-2164-15-1171.

Diedenhofen, B., and J. Musch. 2015. cocor: A comprehensive solution for the statistical comparison of correlations. PLoS ONE 10:e0121945. http://dx.doi.org/10.1371/journal.pone.0121945.

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. http://dx.doi.org/10.3168/jds.2011-5019.

Garrick, D. J., and R. Fernando. 2014. Genomic prediction and genome-wide association studies in beef and dairy cattle. Pages 474–501 in: The Genetics of Cattle. 2nd ed. D. J. Garrick and A. Ruvinsky, ed. CABI, Wallingford, UK.

Grapes, L., M. Z. Firat, J. C. Dekkers, M. F. Rothschild, and R. L. Fernando. 2006. Optimal haplotype structure for linkage disequilibrium-based fine mapping of quantitative trait loci using identity by descent. Genetics 172:1955–1965.

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick. 2011. Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics 12:186. http://dx.doi.org/10.1186/1471-2105-12-186.

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. http://dx.doi.org/10.1017/S0016672307008865.

Legarra, A., A. Ricard, and O. Filangi. 2013. GS3 software package and documentation. Accessed Jan. 1, 2013. http://snp.toulouse.inra.fr/~alegarra.

Ma, P., M. S. Lund, X. Ding, Q. Zhang, and G. Su. 2014. Increasing imputation and prediction accuracy for Chinese Holsteins using joint Chinese-Nordic reference population. J. Anim. Breed. Genet. 131:462–472. http://dx.doi.org/10.1111/jbg.12111.

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

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, and G. A. Doak. 2013. Genomic imputation and evaluation using high-density Holstein genotypes. J. Dairy Sci. 96:668–678. http://dx.doi.org/10.3168/jds.2012-5702.

Villumsen, T. M., L. Janss, and M. S. Lund. 2009. The importance of haplotype length and heritability using genomic selection in dairy cattle. J. Anim. Breed. Genet. 126:3–13. http://dx.doi.org/10.1111/j.1439-0388.2008.00747.x.

Zou, G. Y. 2007. Toward using confidence intervals to compare correlations. Psychol. Methods 12:399–413. http://dx.doi.org/10.1037/1082-989X.12.4.399.

**J. Dairy Sci. 99:4537–4546**
**http://dx.doi.org/10.3168/jds.2015-10433**


## Supplementary tables

**Supplementary Table S1.** Average number of alleles per haplotype with haplotypes of 4 SNP and AFT of 1-8%. Window size: 10 SNP.

| Criterion | Nr. of QTL | AFT[1] (%) | | | |
|---|---|---|---|---|---|
| | | 1 | 3 | 5 | 8 |
| Criterion-A | 1K[2] | 14.55 | 14.35 | 14.29 | 13.84 |
| | 3K | 13.55 | 13.41 | 13.26 | 12.89 |
| | 6K | 12.54 | 12.41 | 12.25 | 11.90 |
| Criterion-B | 1K | 14.42 | 14.41 | 14.42 | 14.41 |
| | 3K | 13.46 | 13.45 | 13.44 | 13.43 |
| | 6K | 12.44 | 12.43 | 12.42 | 12.43 |

1: AFT=Allele frequency threshold (alleles with a frequency higher than this threshold are assumed to be predictable)

2: Thousand

**J. Dairy Sci. 99:4537–4546**
**http://dx.doi.org/10.3168/jds.2015-10433**


**Supplementary Table S2.** Correlations between GEBV and DYD in the validation population for different allele frequency thresholds using Criterion-A and -B to select the haplotypes. Average values over the 5 production traits are shown. Window size: 10 SNP.

| Haplotype size | AFT[1] | Criterion-A | Criterion-B |
|---|---|---|---|
| 3 | 1 | 0.541 | 0.546 |
| | 5 | 0.537 | 0.547 |
| | 8 | 0.547 | 0.548 |
| 4 | 1 | 0.542 | 0.546 |
| | 5 | 0.541 | 0.546 |
| | 8 | 0.544 | 0.548 |
| 5 | 1 | 0.542 | 0.545 |
| | 5 | 0.547 | 0.545 |
| | 8 | 0.543 | 0.546 |

1: AFT=Allele frequency threshold (alleles with a frequency higher than this threshold are assumed to be predictable)

**J. Dairy Sci. 99:4537–4546**
**http://dx.doi.org/10.3168/jds.2015-10433**


**Supplementary Table S3.** Average DYD-GEBV correlations of the 5 production traits using different MD values with Criterion-B. AFT was set to 8% and windows of WS = 10 SNP were used. For every trait separately, the highest correlation was considered from those observed with 1K, 3K and 6K QTL-SNP in the model.

| Haplotype size | Maximum Deviation (MD)[1] | | | |
|---|---|---|---|---|
| | 10% | 30% | 50% | 80% |
| 3 | 0.548 | 0.546 | 0.547 | 0.548 |
| 4 | 0.548 | 0.548 | 0.546 | 0.545 |
| 5 | 0.546 | 0.546 | 0.547 | 0.546 |
| Average | 0.547 | 0.547 | 0.547 | 0.546 |

1: This parameter reflects the acceptable level of variation among allele frequencies.

**J. Dairy Sci. 99:4537–4546**
**http://dx.doi.org/10.3168/jds.2015-10433**


**Supplementary Table S4.** Correlation coefficients calculated between DYD and GEBV for the haplotype-based (Criterion-A/Criterion-B; window size: 10 SNP) methods as function of number of haplotypes in the model. Average correlations over the 5 production traits are shown.

| #QTL-SNP | Criterion-A | | | Criterion-B | | |
|---|---|---|---|---|---|---|
| | HS[1]=3 | HS=4 | HS=5 | HS=3 | HS=4 | HS=5 |
| 1K[2] | 0.506 | 0.509 | 0.494 | 0.505 | 0.516 | 0.501 |
| 3K | 0.525 | 0.529 | 0.525 | 0.524 | 0.538 | 0.529 |
| 6K | 0.546 | 0.544 | 0.543 | 0.544 | 0.544 | 0.544 |
| Optimal | 0.547 | 0.544 | 0.543 | 0.548 | 0.548 | 0.546 |

1: HS=Haplotype size

2: Thousand

**J. Dairy Sci. 99:4537–4546**
**http://dx.doi.org/10.3168/jds.2015-10433**


**Supplementary Table S5.** Correlations between genomic estimated breeding values and DYD in the validation population. Correlations for the optimal number of QTL are presented. Average values of the 5 production traits; window size: 10 SNP.

| Haplotype selection method | Marker type | Haplotype size | Milk quantity | Fat yield | Protein yield | Fat content | Protein content | Average |
|---|---|---|---|---|---|---|---|---|
| QTL-SNP | SNP | 1 | 0.473 | 0.509 | 0.431 | 0.567 | 0.581 | 0.512 |
| Flanking markers | SNP[1] | 3 | 0.475 | 0.525 | 0.437 | 0.568 | 0.581 | 0.517 |
| | | 4 | 0.477 | 0.523 | 0.439 | 0.575 | 0.581 | 0.519 |
| | | 5 | 0.475 | 0.522 | 0.443 | 0.572 | 0.586 | 0.520 |
| | haplotype | 3 | 0.496 | 0.546 | 0.455 | 0.570 | 0.601 | 0.534 |
| | | 4 | 0.498 | 0.558 | 0.473 | 0.571 | 0.591 | 0.538 |
| | | 5 | 0.503 | 0.556 | 0.476 | 0.567 | 0.609 | 0.542 |
| Criterion-A | SNP[1] | 3 | 0.484 | 0.521 | 0.454 | 0.581 | 0.586 | 0.525 |
| | | 4 | 0.487 | 0.530 | 0.453 | 0.578 | 0.580 | 0.526 |
| | | 5 | 0.476 | 0.527 | 0.454 | 0.572 | 0.577 | 0.521 |
| | haplotype | 3 | 0.503 | 0.558 | 0.479 | 0.584 | 0.611 | 0.547 |
| | | 4 | 0.502 | 0.558 | 0.473 | 0.582 | 0.606 | 0.544 |
| | | 5 | 0.485 | 0.562 | 0.487 | 0.577 | 0.602 | 0.543 |
| Criterion-B | SNP[1] | 3 | 0.481 | 0.522 | 0.456 | 0.575 | 0.588 | 0.524 |
| | | 4 | 0.486 | 0.528 | 0.459 | 0.586 | 0.591 | 0.530 |
| | | 5 | 0.483 | 0.530 | 0.456 | 0.578 | 0.584 | 0.526 |
| | haplotype | 3 | 0.506 | 0.554 | 0.487 | 0.591 | 0.604 | 0.548 |
| | | 4 | 0.496 | 0.562 | 0.476 | 0.594 | 0.609 | 0.548 |
| | | 5 | 0.499 | 0.561 | 0.482 | 0.579 | 0.608 | 0.546 |

1:All the SNP included in the haplotypes are included in the Bayes C analysis but they are used as independent explanatory variables.

**APPENDIX I.**

# Supplementary methods

### *Calculation of the weighing factor for Criterion-B*

In the calculation of the weighing factor, two principles need to be taken into account: on the one hand, it is desired to maximize the number of predictable alleles of the selected haplotype while on the other hand, it is also expected from Criterion-B that the allele frequencies of the predictable alleles (which were identified the same way as with Criterion-A, i.e. using the AFT parameter) do not differ extremely from each other, or in other words, their differences do not exceed certain limits. Similarly to Criterion-A, selection of the optimal haplotype with Criterion-B will be accomplished through the minimization of a function, which is expected to reflect both aims.

In order to maximize the number of predictable alleles as with Criterion-A, it must be guaranteed that any haplotype that includes a larger number of predictable alleles has a lower score than the scores calculated for haplotypes with less predictable alleles. Therefore, the *least optimal* scenario with N predictable alleles is expected to get a lower score, than the *most optimal* scenario for any $N' < N$ predictable alleles. Hence:

$$CriterionB^{+}_{N'} > CriterionB^{-}_{N} \quad (1)$$

where

N and $N'$ are the number of predictable alleles (assuming $N' < N$)
$\mathrm{CriterionB}^{+}_{N'}$ is the *most optimal* case with $N'$ predictable alleles
$\mathrm{CriterionB}^{-}_{N}$ is the *least optimal* case with N predictable alleles

The *most optimal* case with $N'$predictable alleles corresponds to the situation when the Criterion-B gives the smallest possible value, which is the case when $N'$ takes its largest value. Within the domain of $N'$ $(0 < N' < N)$, this is $N' = (N - 1)$. Therefore in the rest of the derivation, this value is used instead of $N'$ (proof not shown).

The general form of Criterion-B (without subscripts for simplicity) is:

$$CriterionB = SD - w * N \quad (2a)$$

$$SD = \sum_{k=1}^{N} (OF_k - \frac{1}{2^{HS}})^2 \quad (2b)$$

where

$OF_k$: observed frequency of allele k
w: the weighing factor of the number of predictable alleles
HS: haplotype size

Using equation (2a) in equation (1) leads to equation (3), which in turn (after simple algebraic transformations) can be written as equation (4), defining a lower limit for the weighing factor:

$$SD^{+}_{(N-1)} - w * (N-1) > SD^{-}_{N} - w * N \quad (3)$$

$$w > SD^{-}_{N} - SD^{+}_{(N-1)} \quad (4)$$

Calculating this lower limit for all suitable values of N (that is, from 2 till $2^{HS}$) results in a sequence of lower limits, from which the maximum will satisfy all inequalities. In the following, the two terms on the right side of equation 4 will be defined.

### *Calculating $SD^{+}_{(N-1)}$*

Since Criterion-B is used to solve an optimization problem by minimization, the SD value of the most optimal situation corresponds to the situation where SD takes the lowest possible value.

SD is the smallest for a particular N, when all the alleles have the same frequency (1/N). In such "optimal" cases, the minimal SD can be calculated by equation (5):

$$SD = \begin{cases} 0, if\ N = 2^{HS} \\ (\frac{1}{N} - \frac{1}{2^{HS}})^2 * N, if\ N < 2^{HS} \end{cases} \quad (5)$$

Because $2^{HS}$ is an upper limit of N, (N-1) is necessarily lower than $2^{HS}$. Therefore the lowest SD for $SD^{+}_{(N-1)}$ can be obtained by replacing N by (N-1) in equation (5): $(\frac{1}{N-1} - \frac{1}{2^{HS}})^2 * (N-1)$. Note that this value depends on the number of predictable alleles (N) and on the haplotype size (HS) used in the model only.

### *Calculating $SD^{-}_{N}$*

The least optimal corresponds to a situation where the allele frequencies are as unbalanced as possible. This is the case when (N-1) alleles have an allele frequency equal to AFT and 1 allele has a frequency equal to (1-AFT*(N-1)). The SD value then can be calculated as follows:

$$SD^{-}_{N} = \left(AFT - \frac{1}{2^{HS}}\right)^2 * (N-1) + \left[\left(1 - AFT * (N-1)\right) - \frac{1}{2^{HS}}\right]^2 \quad (6)$$

At this point a new parameter was introduced to include information on the allele frequency equilibrium: the maximum deviation (MD) is defined as the average "allowed" deviation of (N-1) alleles from the ideal frequency ($\frac{1}{2^{HS}}$), expressed as a proportion of the ideal frequency. The last, $N^{th}$ allele is assumed to have an allele frequency equal to the AFT.

With the use of this parameter, the SD of the least optimal case can be calculated as:

$$SD_N^- = \left(AFT - \frac{1}{2^{HS}}\right)^2 + \left(\frac{1}{2^{HS}} * MD\right)^2 * (N-1) \qquad (7)$$

The right side of equation (4) can be calculated for all N and the weighing factor can be selected as described above. From equation (7) it can be noted that with increasing N (from 2 till $2^{HS}$), the value of $SD_N^-$ is increasing as well, while the value of $SD_{(N-1)}^+$ is decreasing (see equation (5)). Therefore to determine the proper weighing factor, the calculation of these parameters is enough for the largest possible value of N, that is for $2^{HS}$.

In summary, to calculate the weighing factor for the number of predictable alleles in Criterion-B, the following parameters are required:

- The haplotype size (HS)
- Allele frequency threshold (AFT)
- The maximum deviation (MD)

All of these parameters are tested in the results section of the article. Since these parameters are available prior to the start of the analysis of the QTL, the weighing factor can be calculated before determining any QTL-windows on the genome and the same weighing factor is generally applicable along the whole genome.

### 3.3.3 Discussion

The most important benefits of the developed methods were described in the paper. In short, we could prove that using allele frequency information to select haplotypes for genomic evaluation purposes was beneficial. Furthermore, we could also provide empirical proof for the superiority of haplotypes over SNP.

A computer program was written and optimized to implement the two criteria in practice. All important parameters (such as AFT, haplotype size, window size and – in case of Criterion-B – MD as well) can be defined in the program by the user. Other features of this software are:

- the possibility of multi-processing using a user-defined number of processors
- the handling of different window sizes for the different QTL regions
- the possibility to force a (single) SNP per window to be part of the final haplotype

The last feature is especially important when putative causative mutations are available. Due to optimization and parallel programming, several thousands of windows as wide as 200 SNP can be analyzed simultaneously in a reasonable time.

By selecting a single haplotype of 'HS' SNP from a window of 'WS' SNP, the number of haplotypes to be used in the model can be reduced by a proportion equal to $(WS - HS)/WS$ (e.g. 60% in case of WS=10 and HS=4), compared to the case when all consecutive, non-overlapping haplotypes of HS SNP are built. This is a very important feature, which alleviates the computational burden when using HD SNP-chip data with haplotype markers.

Following the presented analyses, this method was applied to the other dairy breeds (the Holstein, Normande and Brown-Swiss populations) using the French routine genomic evaluation pipeline. These analyses are not presented here but they resulted in similar gains in terms of correlation coefficients and regression slopes of DYD on GEBV, as presented above.

QTL-mapping results can also be incorporated into these haplotype selection methods in the future, which might further improve selection accuracies. Furthermore, different QTL can be identified with different degrees of accuracy, depending on the size of the QTL effect and the LD between the QTL and the neighboring SNP. Therefore, it is reasonable to assume that different window sizes should be used for the different QTL, depending on the accuracy of QTL localization. The testing and implementation of such refined methods are interesting directions for future research.

This haplotype selection procedure became part of the new French genomic selection pipeline in April 2015 and it was used in the implementation of the new genomic evaluation in the four main dairy cattle breeds in France (Holstein, Montbéliarde, Normande and Brown Swiss breeds). A longer description of this genomic evaluation pipeline was given in section 2.5 of Chapter 2.

## 3.4 Genomic evaluation with HD data

As already indicated, it was hypothesized in the past that the HD SNP-chip could significantly improve the performance of genomic evaluation (Brøndum et al., 2011), but recent studies could not verify this expectation (e.g. Chen et al., 2014; Hozé et al., 2014). Therefore we were interested whether we can observe any improvement with our haplotype construction method combined with HD data in a single-breed scenario, compared to the similar tests using the 50K SNP-chip.

For this, the performance of Criterion-B was tested on the HD SNP-chip in the Montbéliarde breed in a within-breed context. We used the exact same Montbéliarde population as for the tests with the 50K SNP-chip data. The training and validation populations were the same as well.

Due to the shorter distances between the markers on the HD-chip (see **S. figure 1** in Appendix B on page 201), a window size covering approximately the same genomic regions as the 10 SNP-wide windows on the 50K was selected for the HD data. On average, 144 SNP from the HD-chip fell under the windows of the 50K, therefore this value was evaluated together with windows of 80 and 160. Windows of 80 SNP

outperformed the other window sizes, therefore only this analysis is presented here. When the QTL-SNP were not available in the HD data, the closest SNP were used as QTL-SNP instead. Similarly to the tests conducted on the 50K SNP-chip data, the QTL-SNP were forced to be part of the selected haplotypes with the HD data as well. Only haplotypes of 3 and 4 SNP were tested in combination with data from the HD chip to avoid over-parameterization with haplotypes of 5 SNP.

**Table 5** shows the average number of alleles per haplotype with the 2 haplotype building methods and for the 2 haplotype sizes. The average number of segregating alleles with Criterion-B was larger than that with the flanking haplotypes. The difference was larger with haplotypes of 4 SNP (~30%) than with haplotypes of 3 SNP (~16%). As expected, in case of the flanking haplotypes, the number of segregating alleles was lower with the HD data than with the 50K SNP-chip data (**Table 3** from the article). It is due to the much shorter genetic distance between the SNP from the HD chip, which corresponds to a larger LD between consecutive SNP. However, the haplotypes selected by Criterion-B carried slightly more alleles, when they were selected from the HD data compared to the haplotypes selected from the 50K data. The increase in the average number of alleles was ~7% and ~11% with haplotype size of 3 and 4 SNP, respectively.

**Table 5**: Average number of alleles per haplotype observed with the 3 different haplotype construction methods, as function of haplotype size and number of QTL-SNP in the model. Window size: 80 SNP.

| | Number of QTL-SNP | | |
|---|---|---|---|
| | **1,000** | **3,000** | **6,000** |
| | HS[1]=3 | | |
| Flanking markers | 6.59 | 6.43 | 6.32 |
| Criterion-B | 7.64 | 7.49 | 7.33 |
| | HS[1]=4 | | |
| Flanking markers | 11.57 | 11.02 | 10.63 |
| Criterion-B | 14.89 | 14.34 | 13.79 |

1 : HS=Haplotype size

**Figure 5** shows the distribution of allele frequencies with haplotypes of 4 SNP (for results on haplotypes of 3 SNP, see **S. figure 2** in Appendix B on page 202).

Similarly to the 50K results, Criterion-B outperformed the flanking-haplotype case in terms of allele frequency. A larger proportion of the alleles had an intermediate allele frequency (i.e. a frequency between 10 and 40%), while the proportion of over-represented alleles (alleles with a frequency of >40%) in the population decreased by 60% and 79% with haplotypes of 3 and 4 SNP, respectively. The frequency of under-represented alleles (i.e. alleles with a frequency < 1%) decreased by 25% with haplotypes of 3 SNP and increased by 5% with haplotypes of 4 SNP. These values (with the exception of the frequency of the under-represented alleles with haplotypes of 4 SNP) were more favorable with the HD-chip than with the 50K chip.

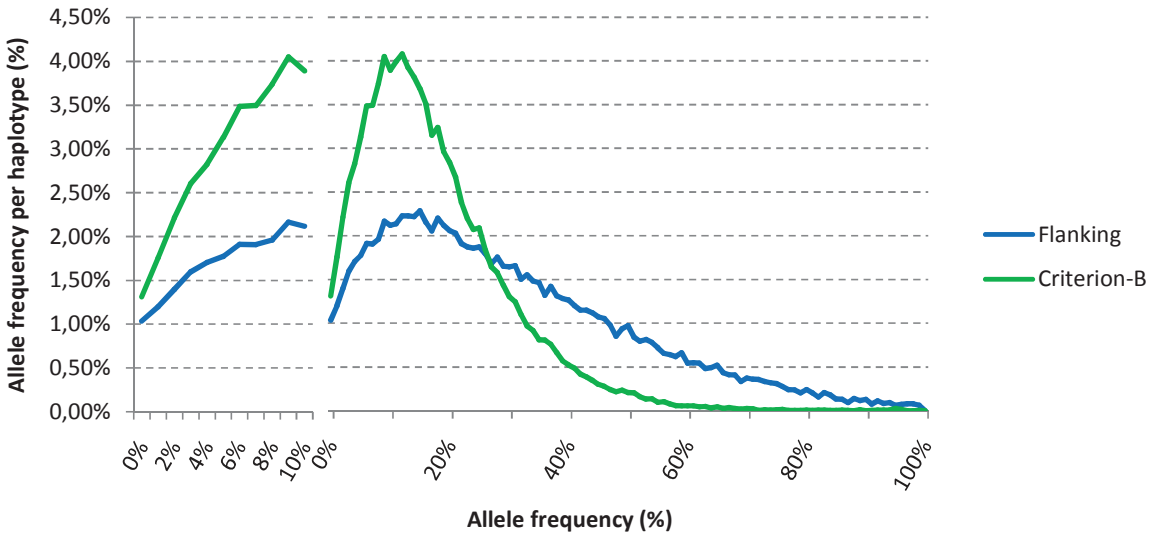



**Figure 5**: Overall distribution of haplotype allele frequencies with either flanking or with Criterion-B selected haplotypes (haplotype size: 4 SNP; window size: 80 SNP; 6,000 QTL-SNP). The 0-10% region is also depicted with more detailed scale on the x-axis.

Based on the allele numbers and allele frequency results shown earlier, Criterion-B is expected to outperform the flanking haplotypes in genomic evaluation. Like previously with the 50K data,the flanking haplotypes are expected to outperform the analysis where only the QTL-SNP are used as genetic markers due to the more informative markers.

**Table 6** shows the correlation coefficients between DYD and GEBV values in case when only the QTL-SNP are used as genetic markers and when flanking haplotypes

are built from the QTL-SNP and their neighboring markers. The flanking haplotypes outperformed the analyses with only QTL-SNP information (with the exception of the HS=3 and 1K QTL-SNP model). However, these correlations were consistently lower than their 50K SNP-chip counterparts (also see **Table 4** from the above paper).

**Table 6**: Observed correlations in the validation set between DYD and GEBV values using either only the QTL-SNP or the flanking haplotypes as genomic markers. Average correlations over the 5 traits.

| Number of QTL-SNP | QTL-SNP | Flanking haplotypes | |
|---|---|---|---|
| | | HS[1]=3 | HS[1]=4 |
| 1K | 0.459 | 0.454 | 0.463 |
| 3K | 0.484 | 0.499 | 0.516 |
| 6K | 0.498 | 0.521 | 0.528 |
| Optimal[2] | 0.498 | 0.523 | 0.529 |

1: HS=Haplotype size

2: For each trait separately, the number of QTL-SNP/haplotypes is the one leading to the highest correlation.

The comparison of the performance of the flanking markers with the selected markers is shown on **Figure 6**. This figure shows the correlation coefficients with the 2 haplotype building methods for haplotypes of 3 and 4 SNP (for individual results of each trait, see **S. table 2** in Appendix B on page 203). Combining the markers into haplotypes was beneficial, leading to an average increase of 1.8% in correlation. Criterion-B performed better than the flanking haplotypes, leading to an extra ~1% increase in correlation. These trends are similar to those with the 50K SNP-chip data. However, data from the 50K SNP-chip were superior compared to those of the HD chip. Averaged over the 2 haplotype sizes and 5 traits, HD data resulted in ~1% lower correlations either when the markers were used as single-SNP information or when they were combined into haplotype markers, as compared to the 50K data.

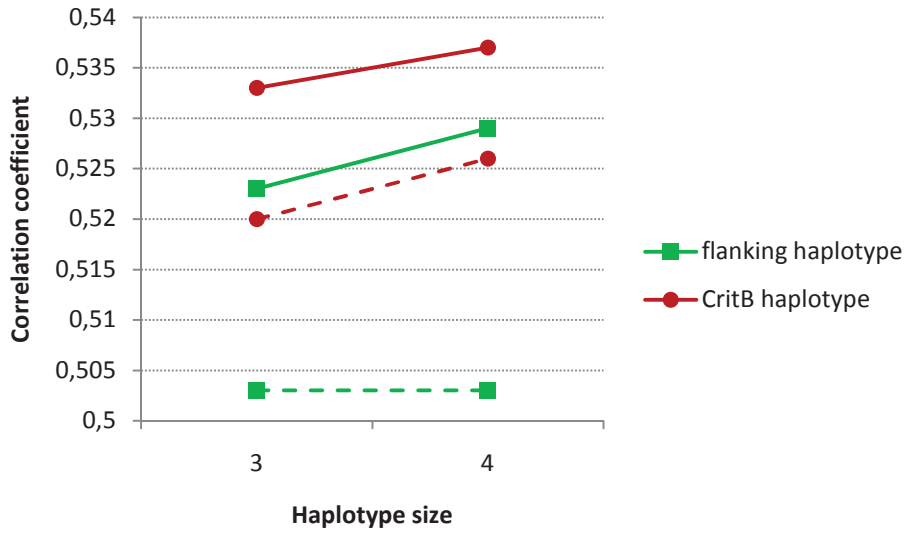



**Figure 6**: Average observed correlations between DYD and GEBV values for 5 production traits with different haplotype selection methods and haplotype sizes. Solid lines indicate the correlations for the haplotype-based tests while dashed lines show the correlations observed when the same SNP were used but as single-SNP markers (Criterion-B; validation set).

Until now, for each trait only the "optimal" (i.e. the highest) correlation coefficient was considered from among those obtained with 1,000, 3,000 and 6,000 haplotypes in the model. The performance of the different number of haplotypes are compared in **Table 7**, which presents the average correlation coefficients for Criterion-B and for the 2 haplotype sizes with either 1,000 or 3,000 or 6,000 haplotypes included in the model (for comparison purposes, the "optimal" values – i.e. those plotted on **Figure 6** – are also shown).

**Table 7**: Average correlations calculated between DYD and GEBV of the validation set for 5 production traits (Criterion-B).

| #QTL-SNP | HS[1]=3 | HS[1]=4 |
|---|---|---|
| 1K[2] | 0.494 | 0.487 |
| 3K[2] | 0.521 | 0.526 |
| 6K[2] | 0.532 | 0.536 |
| Optimal | 0.533 | 0.537 |

1: HS=Haplotype size

2: Thousand

Higher correlations were observed when more QTL were modeled. For most of the traits, 6,000 haplotypes in the model was found to be optimal. For individual results,

see **S. table 2** in Appendix B on page 203. With the exception of 1,000 QTL in the model, haplotypes of 4 SNP led to higher correlation coefficients than haplotypes of 3 SNP.

In addition to the selection accuracy, inflation of breeding values is also an important aspect that has to be considered. **Table 8** shows the estimated regression slopes of DYD on GEBV, averaged over the 5 production traits. Results of individual traits can be found in **S. table 3** in Appendix B on page 204. Criterion-B resulted in the highest regression slopes, followed by the flanking-marker scenario. Once again, the use of haplotypes instead of individual SNP markers was beneficial. These results were very similar to the regression slopes observed with the 50K SNP-chip data.

**Table 8**: Regression slopes with the 2 different haplotype construction methods and when only QTL-SNP were used as genetic markers. Values measured on the validation set and averaged over 5 traits.

| Haplotype selection method | Marker type | Haplotype size (#SNP) | |
|---|---|---|---|
| | | 3 | 4 |
| QTL-SNP | SNP | 0.656 | |
| Flanking markers | SNP | 0.685 | 0.687 |
| | haplotype | 0.742 | 0.768 |
| Criterion-B | SNP | 0.735 | 0.751 |
| | haplotype | 0.796 | 0.825 |

Similarly to the 50K SNP-chip situation, Criterion-B outperformed the flanking haplotypes when data from the HD SNP-chip was used. However, the HD SNP-chip performed worse than the 50K SNP panel. The inferior performance of the HD chip data compared to the 50K SNP panel might be because the windows used for the 2 tests differed significantly in length and in turn the LD-patterns beneath these windows were different as well. By potentially having a large effect on the selected haplotypes, this could result in different selection accuracies.

In conclusion, haplotypes can outperform individual SNP markers in genomic evaluation with the HD SNP-chip as well and the application of the haplotype selection criterion was also beneficial. However, in the studied cases the efficiency of genomic selection was lower with HD data compared to 50K data. These tests should be performed in a potentially more favorable situation for the HD data, for example in a multi-breed context, where larger differences can be expected between the performances of the HD and 50K SNP-chips in genomic selection. Indeed, it was shown earlier that in a within-breed context the resolution of the 50K SNP-chip is sufficiently high for genomic evaluation (Hozé et al., 2013; de Roos et al., 2008).

## 3.5 Inclusion of linkage disequilibrium information

### 3.5.1 Introduction

In the previous study QTL were assumed to segregate within a short (10-SNP wide) window surrounding the SNP identified in the QTL detection step. Although this window size was found to be better on average across the genome when compared to 15- and 20-SNP wide windows, this approach is not perfect and could be improved. Using a fixed window was a compromise that had to be made during the previous study. This allowed us to test a wide range of values for the different parameters. However, it is reasonable to assume that different window limits should be used along the genome as a result of adaptation to the local recombination rates (e.g. Coop et al., 2008; other drawbacks of fixed window sizes were outlined by Beissinger et al., 2015). Recombination rates can differ across chromosomes, genomic regions and populations as well (e.g. Jeffreys et al., 2005 in human or Weng et al., 2014 in beef cattle, Ma et al., 2015). Furthermore, the SNP from the SNP-chips are not equidistant (**S. figure 1** and **S. figure 2** in Appendix B on page 201), which also implies that even for a fixed window size, the different genomic regions do not have the same length. In order to remove the requirement of a preliminary QTL-detection step, one can build windows of SNP along the genome based on LD information. Haplotypes can then be selected to best represent these segments in stronger LD instead of representing the regions surrounding pre-selected SNP.

In what follows, windows are defined as a set of consecutive SNP where the LD measured with D' (after Cuyabano et al., 2014) between every pair of neighboring SNP has to exceed a pre defined limit. These windows will be called haploblocks hereafter (as in Knürr et al., 2013). Although D' is known to be more sensitive to rare alleles (McRae et al., 2002), Cuyabano et al. (2014) showed that D' performed equally well compared to the $r^2$ in creating haploblocks for genomic evaluation purposes. This can be due to the lower number of haploblocks identified with D', which leads to fewer effects to be estimated in genomic evaluation.

The definition of haploblocks based on the LD-pattern allows to account for the variable recombination rate along the genome, and in particular to avoid the inclusion of a recombination hot-spots or any historical recombination with a large impact within any window. Since haploblocks are defined using the LD-pattern along the genome, they are expected to segregate as a single unit from generation to generation (at least as long as the pre-defined D' threshold is close to its maximum).

Because in genomic evaluation the aim is to capture the *combined* effect of all the QTL affecting the trait of interest, the precise positioning of these QTL may not be essential in contrast to QTL detection studies, where the emphasis is on the identification and accurate positioning of the QTL. Therefore, in genomic evaluation the scenarios when the effects of two (or more) closely linked QTL are accurately separated and estimated independently, or when their combined effect is estimated jointly can be considered as equally good. In this context, it is sufficient to estimate a single effect for each haploblock allele, because these blocks are – by construction – closely linked chromosome segments. After determining the haploblocks, a single haplotype can be selected to represent every haploblock along the genome. Such a haplotype within each haploblock can be then selected using Criterion-B and the optimal parameter values (see section 3.3.2).

### 3.5.2 Combining LD and allele frequency information to improve selection accuracy

This article was submitted for publication to the *Journal of Dairy Science* in 2016: Jónás, D., Ducrocq, V. and Croiseau, P. Submitted. Short communication: The

combined use of LD-based haploblock and allele frequency-based haplotype selection method enhances genomic evaluation accuracy in dairy cattle. J. Dairy. Sci.

# The combined use of LD-based haploblock and allele frequency-based haplotype selection method enhances genomic evaluation accuracy in dairy cattle

| | |
|---|---|
| Journal: | *Journal of Dairy Science* |
| Manuscript ID | Draft |
| Article Type: | Short Communications |
| Date Submitted by the Author: | n/a |
| Complete List of Authors: | Jónás, Dávid; INRA, GABI; ALLICE<br>Ducrocq, Vincent; INRA,<br>Croiseau, Pascal; INRA, GABI |
| Key Words: | haplotype, haploblock, genomic evaluation |
| | |

# SHORT COMMUNICATION: HAPLOBLOCK CONSTRUCTION FOR GENOMIC EVALUATION

Exploiting simultaneously marker linkage disequilibrium- and allele frequency information improves genomic evaluation accuracy (Jónás)

Either nonrandom association between markers from dense SNP panels and marker allele frequency information has been used to reduce the number of explanatory variables in genomic evaluation and to improve its accuracy in dairy cattle. Marker allele frequency information can also reduce the number of rare alleles, which is beneficial, because their estimated effects are usually less accurate. In this paper we propose to use these information simultaneously. Our results confirm that this is a promising way to improve genomic selection efficiency.

**The combined use of LD-based haploblock and allele frequency-based haplotype selection method enhances genomic evaluation accuracy in dairy cattle**

**Dávid Jónás,*†[1] Vincent Ducrocq,* Pascal Croiseau,***

*GABI, INRA, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France

†ALLICE, 149 rue de Bercy, 75012 Paris, France

[1]Corresponding author: Dávid Jónás

INRA-GABI; bât 211
Domaine de Vilvert
78352 Jouy en Josas Cerdex
FRANCE

Phone: (33)1-34-65-29-65

david.jonas@jouy.inra.fr

The construction and use of haploblocks – i.e. adjacent SNP in strong linkage disequilibrium – for genomic evaluation purposes is advantageous, because it allows the reduction of the number of effects to be estimated in genomic prediction without the risk of discarding relevant genomic information. Furthermore, haplotypes – i.e. the combination of 2 or more SNP – can increase the probability of capturing the QTL effect compared to individual SNP markers. With regards to haplotypes, the allele frequency parameter is also of interest because as a selection criterion, it allows the reduction of the number of rare alleles, which alleles' effects are usually difficult to estimate. We propose a simple pipeline that simultaneously incorporates both linkage disequilibrium and allele-frequency information in genomic evaluation and we also present the first results we obtained with this procedure. A population of 2,235 progeny tested bulls from the Montbéliarde breed was used for the tests. Phenotype data in the form of daughter yield deviations on 5 production traits as well as genotype data from the 50K SNP-chip was available. A classical validation study was conducted by splitting the population into a training (80% oldest animals) and validation (20% youngest animals) set to emulate a real-life scenario where the selection candidates have no available phenotype data. All reported parameters were measured on the validation set.

Our results prove that the outlined method is indeed advantageous and accuracy of genomic evaluation can be improved. Correlation coefficients between true and estimated breeding values increased by 2.7% on average of the 5 traits, when results were compared to results of a GBLUP analysis. Inflation of genomic evaluation of the simulated selection candidates was significantly reduced as well. The proposed method outperformed all other SNP and haplotype-based tests we evaluated in a previous study. Therefore, the combined use of LD-based haploblocks and allele frequency-based haplotype selection methods is a promising way to improve the efficiency of genomic evaluation. Further work is still needed to optimize each step in the proposed analysis pipeline, but the first results are very promising.

Keywords: haplotype, haploblock, genomic evaluation

The development of cost-efficient SNP-chips as well as elaborate evaluation methods, such as the Bayes Alphabet: A, B, C(-π), D(-π), R (by Meuwissen et al., 2001, Habier et al., 2011 and Erbe et al., 2012) led to the practical implementation of genomic selection in dairy cattle breeding in most developed countries (e.g. in France: Boichard et al., 2012). The majority of the currently available methods use bi-allelic single nucleotide polymorphisms (SNP) as genetic markers to trace quantitative trait loci (QTL). However, haplotype markers (defined as a combination of 2 or more SNP markers, as in: Hayes et al., 2007; Villumsen et al., 2009; Garrick and Fernando, 2014) can outperform individual SNP markers in genomic evaluation (Croiseau et al., 2015 and Jónás e al., 2016). The main advantage of haplotypes lies in their multi-allelic nature: when more alleles can be tracked at a given locus, there is a higher chance that at least one of those alleles will be linked to existing QTL. However, allele effects are not always predicted more accurately with haplotypes than with SNP. The accuracy with which allele effects can be estimated is largely influenced by the alleles frequency, which determines how much phenotypic information can be directly linked to each allele. Rare haplotype alleles are more likely than with SNP, especially if the flanking (i.e., neighboring) SNP are combined into a haplotype marker, because of the short genetic distance (i.e., high LD) between SNP on medium- and high-density SNP-chips. Therefore, on one hand, it is desirable to maximize the number of haplotype alleles in genomic prediction to maximize the probability that at least one allele will be linked to the QTL (if present). But on the other hand, it is necessary to avoid rare alleles to have accurate allele effect estimation, which is essential for an efficient genomic evaluation.

Following these considerations, Jónás et al. (2016) proposed a method to select haplotype markers *prior* to genomic evaluation based on observed allele frequencies. It was shown that

such selected haplotypes outperform haplotypes of flanking SNP in genomic evaluation. However, a major drawback of the proposed method is the prerequisite that the approximate location of the QTL must be determined in a first step prior to genomic evaluation. Here we present an extension of this work aiming at removing this prerequisite by exploiting information on the linkage disequilibrium (LD) pattern along the genome.

**Dataset**

Two criteria were proposed in Jónás et al. (2016) to select haplotypes, with a small difference between their formulations. In this study, only the one with the higher performance will be considered and it will be termed as "Criterion-B" as in Jónás et al. (2016). This selection procedure selects from a set of haplotypes the one leading to the best balance between haplotype allele frequencies and number of haplotype alleles.

The exact same dataset described in Jónás et al. (2016) is used here, allowing an easy comparison between the results published earlier and the ones obtained here. The dataset included 2,235 progeny-tested bulls from the French Montbéliarde population. Phenotype data (in the form of daughter yield deviations or DYD) was available on 5 production traits, namely milk-, protein- and fat yield, protein- and fat content. Genotype data from the Bovine SNP50 BeadChip (50K; Illumina Inc., San Diego, USA) was used. After quality control, 43,801 SNP were retained for genomic evaluation.

Analyses were done in a cross-validation study with the 20% youngest animals in the validation population (as follows, the 80% oldest animals formed the training population). Haplotype allele effects were estimated using the training set; using these estimated allele effects together with genotype and pedigree information from the validation population, GEBV were estimated for all individuals within the validation set. Finally, correlations between estimated GEBV and DYD as well as regression slopes of DYD on GEBV were calculated and compared to the results published in Jónás et al. (2016), i.e. results obtained

with a GBLUP model as well as with the Criterion-B haplotype selection approach, because this approach was previously found to be optimal. In the latter procedure, SNP effects were estimated via a Bayes-Cπ analysis and the SNP with the highest probability of inclusion in the model were selected (in practice, 1000, 3000 or 6000 SNP were identified). These SNP were not assumed to be the causative mutations themselves but to merely indicate the approximate location of the QTL affecting the trait of interest. In a 10-SNP wide window symmetrically surrounding these pre-selected SNP, all possible combinations of 4 SNP were considered as a different haplotype and one haplotype was selected using Criterion-B to represent the linked QTL. These haplotypes were used to better capture the QTL effects. This procedure will be referred as "Pre-selection method" hereafter.

**Haplotype marker selection**

A fixed window size was used in Jónás et al. (2016). However, it is reasonable to assume that different window boundaries should be used along the genome, adapting to the local LD (e.g. Jeffreys et al., 2005 in human or Weng et al., 2014 in beef cattle; other drawbacks of fixed window sizes were outlined by Beissinger et al., 2015). In order to account for the different recombination rates as well as to remove the prerequisite of information on the approximate location of QTL, windows of SNP in strong LD along the genome were built and haplotypes were selected to represent these windows. Windows were defined as a set of consecutive SNP where the LD measured between every neighboring SNP exceeded a pre-defined limit. These windows will be called haploblocks following Knürr et al. (2013). In this study, D' was used as a measure of linkage disequilibrium and the threshold level was set to 45% following Cuyabano et al. (2014; a threshold of 90% was also evaluated). After determining the haploblocks, a single haplotype of 4 SNP was selected from among all possible haplotypes of 4 SNP to represent each haploblock along the genome. Haplotypes within each haploblock

were selected using Criterion-B and the optimal parameter values (i.e. haplotype size: 4 SNP, AFT: 8%; MD: 10%), as they were identified in Jónás et al. (2016).

This process also allowed to identify those haplotypes that are expected to be the most significant in genomic evaluation based on both LD and allele frequency information, before using any phenotype data. This is a notable advantage, because identification of significant markers is usually done in a prior genomic evaluation run after the training population was split into further sub-populations, which method is clearly suboptimal. This aspect is especially relevant for regional breeds, where the number of animals with both genotype and phenotype data is already scarce and their division into more sub-populations is detrimental to a greater extent.

Another advantage of this procedure is that it allowed using the same haplotypes for all the traits analyzed. This is because the haploblock construction is based on observed LD-patterns while the haplotype selection process assumes knowledge on the allele frequencies only; no information on performances were used to select the genetic markers to be used. The differences between the genetic backgrounds of the traits are expected to be reflected in the different estimated allele effects of the haplotyes.

**Genomic evaluation model**

Haplotype allele effects were estimated using a haplotypic Bayes-Cπ approach (Croiseau et al., 2014). The model included an overall mean effect and a residual polygenic effect in addition to the haplotype marker effects (as in Jónás et al., 2016). It can be written as:

$$y_i = \mu + u_i + \sum_{j=1}^{N} z_{ij} a_j \delta_j + e_i$$

where $y_i$ is the performance value (DYD) of individual i, μ is an overall mean effect, $u_i$ is the residual polygenic effect of animal i ($u \sim MVN(0, A\sigma_u^2)$), N is the total number of haplotypes in the model, $z_{ij}$ is a vector of dimension $1 \times k_j$ (where $k_j$ is the number of alleles at haplotype j)

indicating the number of each haplotype allele copies animal i carries at haplotype j for every allele of that haplotype (i.e. vector sum of $z_{ij}$ is 2), $a_j$ is a vector of substitution effects of haplotype j (of dimension $k_j \times 1$), $\delta_j$ is a 0/1 variable indicating whether or not marker j is assumed to have an effect and $e_i$ is a random error term for animal i. The proportion of genetic variance attributed to the residual polygenic effect was allowed to vary.

**Results**

Two different threshold values of the D' parameter were tested: 45% and 90%. The value of 45% was found to be optimal in Cuyabano et al. (2014) and our tests confirmed their results (data not shown). Therefore only results with a D' threshold of 45% will be presented here.

Table 1 gives a short summary of the characteristics of the haploblocks and the selected haplotypes. The 43,801 SNP were divided into 8,393 haploblocks with an average of 5.22 SNP per haploblock. This number of SNP per haploblock is relatively small due to the long distance between the markers on the 50K SNP-chip panel (on average ~57,300 bp, exceeding 100,000 bp only in 11.5% of the cases). Sometimes haploblocks were shorter than the desired haplotype size (4 SNP). In such cases, haplotypes were built using all of the SNP from the haploblock and the closest flanking SNP were added to extend the haplotypes to 4 SNP. When such short haploblocks were adjacent to each other, it was likely that the exact same haplotypes were built for them and only one of them was kept for the analysis. This is the reason why there were less haplotypes in total than haploblocks (Table 1). The average number of alleles per haplotype was higher than those observed with 6,000 haplotypes in (Jónás et al., 2016).

Table 2 presents the GBLUP results as well as the results of the pre-selection method (these results were taken from Jónás et al., 2016) together with the new results obtained using haploblock information. Both the correlation coefficients between DYD and GEBV and regression slopes of DYD on GEBV are presented. The "pre-selection method" column of the

table corresponds to the second last row of Supplementary Table S5. of (Jónás et al., 2016), displaying the best results obtained in that study.

The proportion of variance attributed to the residual polygenic effect with the haploblock based method converged to 5.7% (average of the 5 traits). The rest of the genetic variance was explained by the haplotypes. Results obtained with the combined use of LD-based haploblocks and haplotype selection based on allele frequencies outperformed the traditional GBLUP analysis by 2.7 percentage points (pp) in correlation coefficients. An average gain of 1.5pp in correlation was observed, when the basis of comparison was the best pre-selection method. Largest improvements were observed for fat content (4.3pp in correlation compared to correlations observed with the other two methods) and for protein yield (1.7pp gain in correlation). Although the observed increase in correlations was very limited for certain traits, a significant Wilcoxon signed-rank test (p-value: 0.03) between the haploblock based results and those obtained with the pre-selection method showed that an increase was always observed when haploblock information was taken into account. The large improvement with these traits is most likely because when regions are pre-selected based on a prior Bayes-Cπ analysis, multiple SNP are linked to the same major genes (such as diacylglycerol O-acyltransferase 1 or DGAT1) and as a consequence, SNP that were linked to other QTL were missed in these analyses. In contrast, they are necessarily kept when all markers from all regions are kept in the haploblock based analysis, leading to higher selection accuracies.

Regression slopes of DYD on GEBV were substantially improved as well. On average, deviation of the regression slopes from their optimal value (i.e. from 1) was 0.078 smaller when compared to either the pre-selection method or the GBLUP method.

A test using all consecutive haplotypes of 4 SNP along the genome was also implemented, resulting in inferior correlations and regression slopes compared to the haploblock based analyses (data not shown).

In conclusion, the use of information on LD-pattern along the genome in combination with allele frequency information to build haplotypes specifically for genomic evaluation purposes is a promising way to improve genomic evaluation accuracy. A very interesting feature of the proposed method is that the same haplotypes can be used to analyze all traits of interest. Further significant improvements can be expected following the refinement of the different steps of the proposed process. For example, Beissinger et al. (2015) developed a smoothing spline technique to better identify window boundaries. Application of this method can lead to a better haploblock definition, which in turn can further improve the selection efficiency. Another interesting aspect of the proposed method is that it allows the use of genotype data of the selection candidates (or that of the validation population in an experimental setup) in combination with the genotype data of the training population to build the haplotypes for genomic evaluation (that is because no phenotype data was used for the haplotype construction).

## ACKNOWLEDGEMENTS

CARTOFINE and AMASGEN projects are funded by the Agence Nationale de la Recherche (ANR-10-GENM-0014) and APISGENE.

## REFERENCES

Beissinger, T. M., G. J. M. Rosa, S. M. Kaeppler, D. Gianola and N. de Leon. 2015. Defining window-boundaries for genomic analyses using smoothing spline techniques. Genet. Sel. Evol. 47: 30.

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 and S. Fritz. 2012. Genomic selection in French dairy cattle. Anim. Prod. Sci. 52: 115-120.

Croiseau, P., A. Baur, D. Jónás, C. Hozé, J. Promp, D. Boichard, S. Fritz and V. Ducrocq. 2015. Comparison of different Marker-Assisted BLUP models for a new French genomic

evaluation. Page 248 in Book of Abstracts of the 66th Annual Meeting of the European Federation of Animal Science, Warsaw University of Life Sciences, Poland.

Croiseau, P., M. N. Fouilloux, D. Jónás, S. Fritz, A. Baur, V. Ducrocq, F. Phocas, and D. Boichard. 2014. Extension to haplotypes of genomic evaluation algorithms. AB#708 in Proc. 10th World Congress of Genetics Applied to Livestock Production. Vancouver, Canada.

Cuyabano, B. C. D., G. Su and M. S. Lund 2014. Genomic prediction of genetic merit using LD-based haplotypes in the Nordic Holstein population. BMC Genomics 15: 1171.

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.

Garrick, D. J., and R. Fernando. 2014. Genomic prediction and genome-wide association studies in beef and dairy cattle. Pages 474-501 in: The genetics of cattle. D. J. Garrick and A. Ruvinsky, ed. CABI (2nd edition), Wallingford, UK.

Habier, D., R. L. Fernando, K. Kizilkaya and D. J. Garrick. 2011. Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics. 12: 186. http://dx.doi.org/10.1186/1471-2105-12-186.

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. http://dx.doi.org/10.1017/S0016672307008865.

Jeffreys, A. J., R. Neumann, M. Panayi, S. Myers and P. Donnelly. 2005. Human recombination hot spots hidden in regions of strong marker association. Nat. Genet. 37:601-606.

Jónás, D., V. Ducrocq, M-N. Fouilloux and P. Croiseau. 2016. Alternative haplotype construction methods for genomic evaluation. J. Dairy Sci. 99: 4537-4546.

Knürr, T., I. Strandén, M. Koivula, G. P. Aamand and E. A. Mäntysaari. 2013. Haplotype-assisted genomic evaluations in Nordic red dairy cattle. Page 454 in Book of Abstracts of the 64th Annual Meeting of the European Federation of Animal Science, Nantes, France.

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

Villumsen, T. M., L. Janss, and M. S. Lund. 2009. The importance of haplotype length and heritability using genomic selection in dairy cattle. J. Anim. Breed. Genet. 126: 3-13. http://dx.doi.org/10.1111/j.1439-0388.2008.00747.x.

Weng, Z-Q., M. Saatchi, R. D. Schnabel, J. F. Taylor and D. Garrick. J. 2014. Recombination locations and rates in beef cattle assessed from parent-offspring pairs. Genet. Sel. Evol. 46:34.

256 **Tables**

**Table 1**: Descriptive statistics of the haploblocks

| Parameter name | Haploblock information[1] |
|---|---|
| Total number of markers | 43,801 |
| Number of haploblocks | 8,393 |
| Number of haplotypes built | 7,804 |
| Average number of SNP per haploblock | 5.22 |
| Average number of alleles per haplotype | 13.29 |

1: Results obtained using haploblock information with a D' threshold of 45%.

257

**Table 2**: Correlation coefficients and regression slopes of DYD on GEBV using haplotype markers. Results of GBLUP as well as those with the pre-selection and haploblock based methods are presented

| Trait name[1] | GBLUP[2] | | Pre-selection method[2] | | Haploblock information[3] | |
|---|---|---|---|---|---|---|
| | Correlation | Slope | Correlation | Slope | Correlation | Slope |
| MQ | 0.490 | 0.810 | 0.496 | 0.789 | 0.504 | 0.910 |
| FY | 0.551 | 0.850 | 0.562 | 0.806 | 0.564 | 0.943 |
| PY | 0.478 | 0.738 | 0.476 | 0.697 | 0.493 | 0.803 |
| FC | 0.570 | 0.785 | 0.594 | 0.865 | 0.637 | 0.933 |
| PC | 0.584 | 0.987 | 0.609 | 0.971 | 0.613 | 1.071 |
| Average | 0.535 | 0.166[4] | 0.547 | 0.174[4] | 0.562 | 0.096[4] |

1: Trait name abbreviations: MQ – milk quantity; FY – fat yield; PY – protein yield; FC – fat content; PC – protein content

2: Results were taken from Jónás et al. (2016).

3: Results obtained using haploblock information with a D' threshold of 45%.

4: Average deviations from 1.

258

### 3.5.3 Discussion

In the previous section we could prove that the simultaneous use of LD- and allele frequency information to pre-select genetic markers for genomic evaluation purposes is beneficial. The level of gain was comparable to the gain obtained in Jónás et al. (2016). A likely explanation is that earlier a predefined number of SNP (haplotype) was selected to represent QTL, while haploblocks cover all genomic regions (including all QTL). Also, previously there were situations where more than a single SNP was linked to a specific QTL, depending on the effect size of the QTL and on the strength of LD within the haploblock in which the QTL is located. For example, the bovine diacylglycerol O-acyltransferase-1 (DGAT1) is a known causative mutation with a major effect on milk fat content and the LD around this SNP is also known to cover a region of several centiMorgan (cM) on the bovine genome (Grisart et al., 2002). In contrast, in this second study this was efficiently avoided due to the use of haploblock information. This is desirable, because it decreases the number of haplotypes in the model without the risk of removing relevant information. This either gives space to the estimation of additional haplotype allele effects or to the better estimation of the remaining effects. In this work implicitly, additional haplotypes were included in the model (all haploblocks were added in practice). This includes those that carry undetected QTL with smaller effects as well, which were missed earlier, when the SNP in the QTL-detection step were selected based on estimated probabilities of inclusion.

We hypothesized that a larger LD threshold would result in better estimates. However, this hypothesis was not confirmed by our findings. **Table 9** shows the validation results with a D’ of 90% (for an easier comparison the results obtained with a D' threshold of 45% are also indicated). Correlation coefficients measured between DYD and GEBV of the validation population as well as the regression slopes of the same DYD on GEBV are shown. These results are inferior compared to those published with a D’ threshold of 45%, most likely because of the much larger number of haploblocks/haplotypes and therefore more allele effects (+83%) to be estimated by the model.

**Table 9**: Correlation coefficients and regression slopes of DYD on GEBV values of the validation population with a D' threshold of 45% or 90%.

| Trait name[1] | D' threshold: 45% | | D' threshold: 90% | |
|---|---|---|---|---|
| | Correlation coefficient | Regression slope | Correlation coefficient | Regression slope |
| MY | 0.504 | 0.91 | 0.497 | 0.868 |
| FY | 0.564 | 0.943 | 0.565 | 0.917 |
| PY | 0.493 | 0.803 | 0.491 | 0.786 |
| FC | 0.637 | 0.933 | 0.615 | 0.911 |
| PC | 0.613 | 1.071 | 0.603 | 1.077 |
| Average[2] | 0.562 | 0.096 | 0.554 | 0.119 |

1: Trait name abbreviations: MY – milk yield; FY – fat yield; PY – protein yield; FC – fat content; PC – protein content

2: In case of regression slopes, average deviations from 1 are shown.

In conclusion, selection accuracy could be improved with the inclusion of LD information in the haplotype selection step. This also led to a reduced inflation of the breeding value estimates of selection candidates (i.e. of the validation animals in the validation study). A major practical advantage of the presented evaluation pipeline is that it allows the use of the same haplotypes for all traits. The difference between the genetic background of the traits are expected to be reflected in the different estimated allele effects for these haplotypes: that is to say a haplotype might have an effect close to zero for a trait while for another trait, the same haplotype might have a sizeable effect.

Creating haploblocks in a more sophisticated way may further improve the efficiency of genomic evaluation. Several authors have proposed methods to define window boundaries based on the LD patterns observed within a population, including Cuyabano et al. (2014), whose definition of haploblocks was very similar to ours. However, they measured the LD between every pair of SNP instead of between every neighboring SNP. Other works include that of Gabriel et al. (2002), or Beissinger et al. (2015). The removal of markers with very rare alleles prior to haplotype construction might be also a way to improve the performance of genomic evaluation (as it was done here).

The use of different haplotype sizes for the different haploblocks (i.e. longer haplotypes for longer haploblocks) might also have a positive impact on the selection accuracy. However, this test was not feasible with the available *haplotypic BayesC-π* software, as it works only with haplotypes of identical sizes.

# Chapter 4

# Genomic evaluation in regional breeds

---

At the start of this PhD, genomic evaluation was not yet implemented in regional breeds, due to lack of a sufficiently large reference population. Since genomic selection was implemented earlier in the large breeds, the gap between the genetic potential of regional and large breeds is expected to increase. Because of these considerations, there was an increased pressure from breeders and breeding organizations of regional breeds to benefit from genomic evaluation methods relatively efficient in breeds with a reference population of limited size.

In order to address this demand, we assessed the performance of the French routine genomic evaluation pipeline in the regional breeds, which by 2015 incorporated the new methods presented in Chapter 3. Furthermore, we also investigated the possible gains of a multi-breed genomic evaluation using the 4 regional breeds available. This latter method seemed promising, because genetic distances between these breeds are relatively short (**Figure 3**; Gautier et al., 2010).

Before describing these analyses, I will briefly describe the available dataset and characterize the linkage disequilibrium within- and between the breeds, because both the quality of the available dataset and the strength of LD have a major impact on the efficiency of genomic evaluation.

## 4.1 Datasets

Regional breeds (such as the Abondance, Tarentaise, Simmental and Vosgienne) are characterized by a small population size. As follows, the reference population of these breeds consist of only a limited number of progeny tested bulls and the progeny-testing is also less accurate (in the aforementioned breeds, progeny testing is limited to ~25 recorded female offspring on average; D. Boichard, 2015, personal communication). Therefore, in order to enlarge the reference population and in turn to maximize the selection accuracy of genomic evaluation, breeding organizations invested in genotyping females from these populations in addition to the progeny-tested males. Individuals genotyped within the framework of the GEMBAL project or imputed by August 2015 were available for testing with the 50K and HD data, while those available by February 2016 were used to evaluate whether candidate mutation information from large breeds can increase selection accuracy in regional breeds or not. **Table 10** shows the number of males and females with genotype and performance records at these 2 dates. Considering all SNP-chips, Abondance had the largest reference population.

**Table 10**: Total number of genotyped or imputed males and females in the 4 regional breeds, as of either August 2015 or August 2016.

| Breed | Number of animals with genotype data (August 2015) | | Number of animals with genotype data (February 2016) | |
|---|---|---|---|---|
| | **Male** | **Female** | **Male** | **Female** |
| Abondance | 344 | 1482 | 388 | 2766 |
| Tarentaise | 297 | 1167 | 320 | 1566 |
| Simmental | 324 | 183 | 909 | 482 |
| Vosgienne | 60 | 1008 | 65 | 1167 |

In all the forthcoming validation analyses, validation sets consisted entirely of female individuals, because the 20% youngest individuals corresponded to females only. All individuals from all breeds had performance records for all the analyzed (i.e. production) traits, which were obtained in routine phenotype recording.

### 4.1.1 Genotyping and imputation

Individuals were genotyped for one or more of the low-, medium- and high-density SNP-chips and they were imputed for all SNP for which they had no genotype records. Multi-allelic markers were removed prior to imputation. Genotype imputation in the regional breeds was done in 2015 (and repeated in 2016), following the update of the French routine evaluation pipeline. This update affected the imputation and phasing steps as well. After 2015, the FImpute software (Sargolzaei et al., 2014) replaced the BEAGLE software for imputation in France. This change resulted in an increased accuracy and a 3-fold decrease in running time (Croiseau et al., 2015b). In case of FImpute, the default parameters and values were used for imputation. FImpute had a built-in phasing function, which was used for phasing, instead of the previously used DAGPHASE software (Druet and Georges, 2009).

Following imputation, the same quality control step was implemented in the regional breeds as in Montbéliarde to remove SNP of poor quality (see in section 3.1). After quality control, ~43,800 SNP were retained from the 50K SNP-chip panel, ~706,800 SNP from the HD-panel and approximately 5,000 unique SNP (i.e. SNP that are neither present on the 50K nor on the HD chip) from the LD SNP-chip.

With the FImpute software, the allelic imputation error rate (i.e. the proportion of incorrectly imputed alleles among all the imputed alleles) was lower than 1% in all of the regional breeds (S. Fritz, 2015, personal communication).

The distribution of minor allele frequencies was very similar among the regional breeds (**Figure 7**). This figure was created using HD SNP-chip data and all chromosomes. These distributions are very similar to the ones obtained in the large breeds (data not shown). In case of all breeds, 86-88% of the SNP had a MAF >5% and more than 50% of them had a MAF >25%. This is important in genomic evaluation studies, because the estimation of allele effects is difficult for rare alleles.

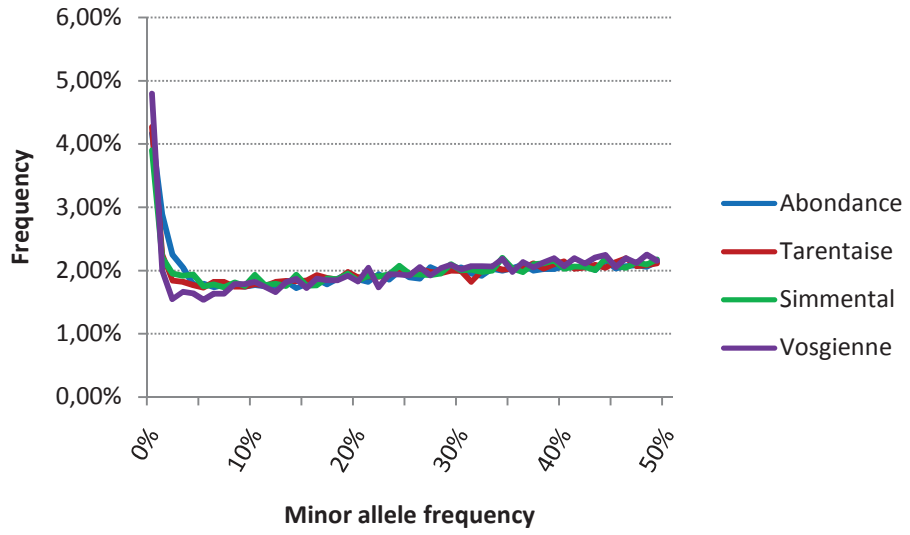



**Figure 7**: Distribution of the minor allele frequency in the regional breeds (MAF resolution: 1%).

Finally, the number and proportion of monomorphic SNP within each breed are shown in **Table 11** for all 3 SNP-chips. A much larger proportion of the custom SNP was monomorphic, because many SNP on the LD chip are candidate mutations responsible for embryo mortality and genetic disorders. Such SNP do not necessarily segregate in every breed. Furthermore, a number of problematic SNP were removed prior to imputation (e.g. because they were difficult to impute in several breeds) in case of the 50K and HD SNP-chips, but not in case of the LD chip.

**Table 11**: Number of monomorphic SNP on the different SNP-chips in the four regional breeds.

| Breed | Custom SNP-chip | | 50K SNP-chip | | HD SNP-chip | |
|---|---|---|---|---|---|---|
| | Nr. | % | Nr. | % | Nr. | % |
| Abondance | 1495 | 29.92 | 893 | 2.04 | 97649 | 13.82 |
| Tarentaise | 1788 | 37.53 | 2396 | 5.47 | 107463 | 15.20 |
| Simmental | NA[1] | NA[1] | 606 | 1.38 | 78906 | 11.16 |
| Vosgienne | 1545 | 31.04 | 813 | 1.86 | 87770 | 12.42 |

1: Custom SNP-chip data was not used from the Simmental breed, due to insufficient number of animals genotyped with this chip

## 4.2 LD-pattern in the regional breeds

The chance for recombination(s) to occur between any 2 markers is increasing with the distance between these markers, which leads to a decay in the LD between them. Since LD between markers and QTL is fundamental for an efficient genomic evaluation, it is of great importance to know the level of LD in the analyzed breeds. Furthermore, the comparison of LD-decay in the multi-breed case to the single-breed scenarios is an important indicator whether or not multi-breed genomic evaluation can be expected to outperform the single-breed tests in the analyzed breeds or not.

The $r^2$ measure of LD was used to measure the strength of LD among positions and to characterize the speed of linkage decay along the genome, because D' is known to be more sensitive to rare alleles (McRae et al., 2002). The $r^2$ measure of LD was calculated between every pair of SNP on each chromosome separately to characterize the LD-decay within each breed as well as to compare the different breeds. The average LD was calculated as a function of distance between markers. The 0-0.25 Mb region of this plot is shown on **Figure 8**. Markers with a minor allele frequency lower than 5% (including the monomorphic SNP) were removed, because it was shown that detection of LD is difficult when at least one of the SNP carries a rare allele (Goddard et al., 2000). Both the level of LD and the speed of its decay were very similar in the regional breeds and these were not different compared to the large breeds. Montbéliarde can be considered as a typical large breed in this aspect, based on Hozé et al. (2013).

The range of the values on **Figure 8** is lower from the results published by Hozé et al. (2013). This is because monomorphic SNP were removed for **Figure 8**, while they were kept in Hozé et al. (2013). The calculated $r^2$ values presented here are in a similar range than those published by de Roos et al. (2009a).

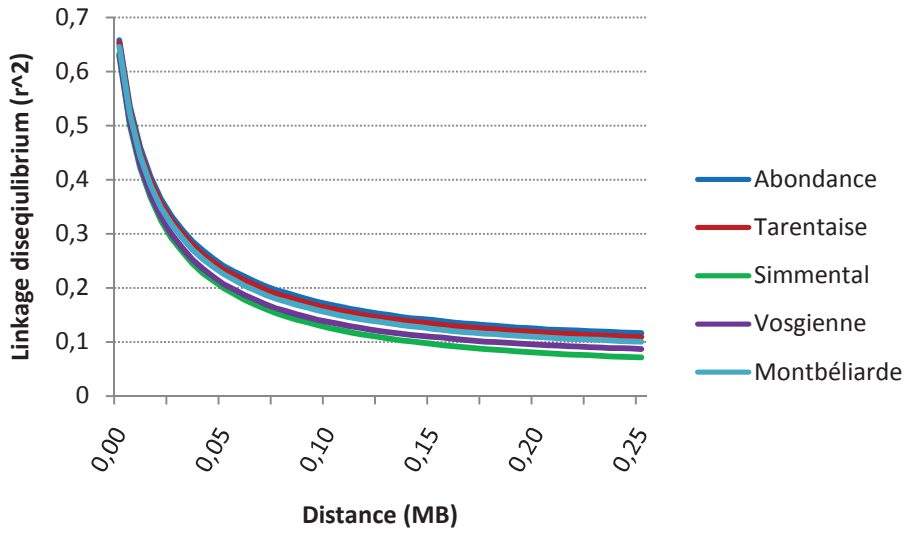



**Figure 8**: Linkage disequilibrium decay in the single-breed contexts.

The average distance between SNP on the 50K SNP-chip is 57,000 bp and it is 3,500 bp on the HD SNP-chip, suggesting that the HD chip is much more likely to have SNP in strong LD with causative mutations.

**Figure 9** shows the average LD-decay of 11 multi-breed scenarios (solid black line), which correspond to the 6 different combinations of 2 breeds out of the 4 regional breeds plus the 4 combinations of 3 breeds out of the 4 regional breeds plus the case when all 4 breeds are merged together (11 in total). The slowest and fastest LD-decays out of the 11 cases are also shown (dashed lines) as well as the average of the 4 within-breed cases (solid blue line). The 11 multi-breed scenarios are shown separately on **S. figure 3**, **S. figure 4** and **S. figure 5** in Appendix C on pages 205-206.

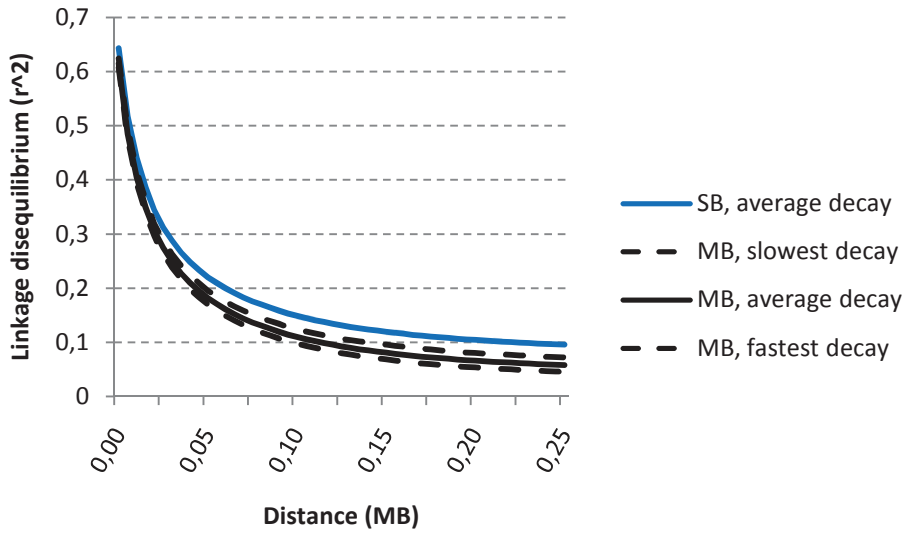



**Figure 9**: Linkage disequilibrium decay in the multi-breed (MB) context (average of the 11 different multi-breed combinations (solid, black line); minimum/maximum of these combinations (dashed, black lines) and average of the four single-breed (SB) scenarios).

Because a multi-breed population is genetically more diverse than a single-breed population, the linkage disequilibrium between adjacent markers is always weaker in multi-breed populations. Although the LD-decay in the multi-breed test is indeed faster, it is remarkably similar to the single-breed cases (**Figure 9**).

## 4.3 Genomic evaluation with 50K data

### 4.3.1 Introduction

The introduction of genomic selection drastically increased the annual genetic gain in large dairy cattle breeds (see section 2.6.1 for a summary of the advantages of genomic evaluation). The lack of sufficient phenotype data is the most important disadvantage of regional breeds as compared to large dairy cattle breeds. Because of this, genomic selection was not applied to regional breeds before 2015.

In the following, the performance of genomic evaluation methods in regional breeds both in single-breed and in multi-breed contexts is discussed. In this section the French routine genomic evaluation is applied to the 4 regional dairy cattle breeds (Abondance, Tarentaise, Simmental and Vosgienne). Afterwards, several ways to

improve the performance of genomic evaluation in these breeds will be proposed and their performances evaluated.

### 4.3.2 Single-breed and multi-breed genomic evaluation with 50K data

This article was submitted for publication to *Journal of Animal Breeding and Genetics* in 2016.

Jónás, D., Ducrocq, V., Fritz, S., Baur, A., Sanchez, M-P. and Croiseau, P. Submitted. Genomic evaluation of regional dairy cattle breeds in single-breed and multi-breed contexts. J. Anim. Breed. Genet.

Journal of
Animal Breeding and Genetics

J. Anim. Breed. Genet. ISSN 0931-2668

ORIGINAL ARTICLE

# Genomic evaluation of regional dairy cattle breeds in single-breed and multibreed contexts

D. Jónás[1,2], V. Ducrocq[1], S. Fritz[2], A. Baur[2], M.-P. Sanchez[1] & P. Croiseau[1]

1 GABI, INRA, AgroParisTech, Université Paris-Saclay, Jouy-en-Josas, France
2 ALLICE, Paris, France

**Keywords**
Genomic selection; haplotype; multibreed; regional breed.

**Correspondence**
D. Jónás, INRA-GABI; bât 211, Domaine de Vilvert, 78352 Jouy en Josas Cedex, France.
Tel: (33)1 34 65 29 65;
Fax: +33-1-3465-2478;
E-mail: david.jonas@jouy.inra.fr

Received: 29 July 2016;
accepted: 7 November 2016

## Summary

An important prerequisite for high prediction accuracy in genomic prediction is the availability of a large training population, which allows accurate marker effect estimation. This requirement is not fulfilled in case of regional breeds with a limited number of breeding animals. We assessed the efficiency of the current French routine genomic evaluation procedure in four regional breeds (Abondance, Tarentaise, French Simmental and Vosgienne) as well as the potential benefits when the training populations consisting of males and females of these breeds are merged to form a multibreed training population. Genomic evaluation was 5–11% more accurate than a pedigree-based BLUP in three of the four breeds, while the numerically smallest breed showed a < 1% increase in accuracy. Multibreed genomic evaluation was beneficial for two breeds (Abondance and French Simmental) with maximum gains of 5 and 8% in correlation coefficients between yield deviations and genomic estimated breeding values, when compared to the single-breed genomic evaluation results. Inflation of genomic evaluation of young candidates was also reduced. Our results indicate that genomic selection can be effective in regional breeds as well. Here, we provide empirical evidence proving that genetic distance between breeds is only one of the factors affecting the efficiency of multibreed genomic evaluation.

## Introduction

In order to obtain high accuracies, the current genomic selection methods require large training populations (i.e. animals with both phenotypic and genotypic records), typically consisting of several thousands of individuals (VanRaden *et al.* 2008). Genomic selection is currently implemented for the main dairy cattle breeds (e.g. for Holstein Friesian, in the USA: Wiggans *et al.* 2011; in France: Boichard *et al.* 2012b; Croiseau *et al.* 2015; in the Netherlands and in New Zealand: de Roos *et al.* 2009b; the Eurogenomics initiative: Lund *et al.* 2011). In regional breeds, the estimations of marker effects are less accurate as a result of small training populations, leading to lower selection efficiencies, when compared to large breeds. Indeed, as of today, genomic selection has not been implemented in regional dairy breeds. However, there is an increasing demand for it from breeders and breeding associations due to economical considerations as well as due to fear of a growing genetic gap between breeds with versus without genomic selection.

There are at least two different ways to increase the size of the training population for these breeds: the first one is the inclusion of females in the training population. However, in dairy cattle, much less information is available from the performance of individual females than on that of males due to a lower number of progeny per female, implying that many

  doi:10.1111/jbg.12249

more cows with records must be genotyped to improve the efficiency of genomic evaluation (Harris *et al.* 2013). The second approach is to merge the training populations of several breeds and estimate marker effects using the multibreed training populations. Although such a strategy can circumvent the problem of small training populations (especially if one or more large breeds are included as well), a multibreed genomic evaluation can be efficient only if (i) quantitative trait loci (QTL) affecting the traits of interest are shared across breeds, (ii) there is a conserved linkage disequilibrium (LD) between QTL and genetic markers among the breeds and (iii) the same QTL–marker phases are present in all of these breeds as well (de Roos *et al.* 2008). Indeed, Porto-Neto *et al.* (2015) have shown that consistent QTL–marker phases are essential for successful multibreed genomic evaluation. Given these requirements, markers from the single nucleotide polymorphism (SNP) chips can be split into two groups based on whether these conditions are met or not: if QTL are shared among the populations and the LD between the available markers and the shared QTL is conserved as well as the phases, then marker effects are expected to be more accurately estimated in a multibreed scenario. However, if at least one of these conditions is not met, the accuracy of marker effect estimation may decrease due to the additional noise introduced in the training population with the inclusion of breeds, in which either the QTL is not present or the linkage phases between the QTL and marker(s) are different. Consequently, to obtain the maximum gain possible, the optimal training population should be a population formed by individuals from breeds that are genetically as similar to each other as possible (de Roos *et al.* 2008).

In a classical validation study using a simulated multibreed experimental design derived from existing large training populations, Hozé *et al.* (2014) showed that multibreed training populations can improve prediction accuracy in breeds with small training populations. Hozé *et al.* (2014) also showed that breeds with small training populations benefit more from a multibreed training population than large breeds.

Multibreed genomic evaluations used in combination with haplotype markers can be expected to increase the prospect of conservation of LD between markers and QTL and therefore increase the accuracy of breeding value estimation. Haplotypes are combinations of N neighbouring SNP (Hayes *et al.* 2007; Villumsen *et al.* 2009; Garrick & Fernando 2014) and unlike SNP with two alleles, haplotypes can theoretically carry $2^N$ different alleles. Because of the increased number of alleles with haplotypes, there is a higher chance that at least one of these alleles will be linked to a QTL – when the latter is present – as compared to SNP markers. This assumption was confirmed by recent works (e.g. Croiseau *et al.* 2015; Jónás *et al.* 2016).

The main aim of this study was to assess the efficiency and the potential gains of genomic evaluations in four regional breeds. In addition to single-breed analyses, multibreed scenarios were studied in order to investigate the potential gains or losses in terms of accuracy due to the use of merged training populations and inclusion of females in the reference set.

## Materials and methods

### Data sets

Four regional French dairy cattle breeds were included in the analysis: Abondance, Tarentaise, Simmental and Vosgienne. Abondance and Simmental are the largest of these breeds with approximately 23 000 and 17 000 cows under performance recording in 2014, respectively, followed by the Tarentaise with ~7500 cows and finally the Vosgienne with ~1350 cows (Institut de l'Elevage, 2015). Performance records were daughter yield deviations (DYD) for males or yield deviations (YD) for females for the following five production traits: milk yield, fat content, fat yield, protein content and protein yield. (D)YD values were created by adjusting the observed performances for all fixed effects, which were estimated in the current genetic evaluation. When calculating the DYD values, genotyped female performances were excluded in order to avoid using the same phenotype data twice during the analysis. Genotype information from the Illumina Bovine SNP50 BeadChip® (manufactured by Illumina Inc., San Diego, CA, USA) was available; following a quality control filtering (minimum Hardy–Weinberg equilibrium p-value: $10^{-4}$, minor allele frequency: 5%, minimum call rate: 10%), 43 801 SNP were retained.

A classical validation study was performed, where the group of animals with both performance (as DYD and YD values for males and females, respectively) and genotype information was split into two populations based on birth date: a training population of the 80% oldest individuals and a validation population (20% youngest individuals). In a first step, allele effects were estimated using genotype and phenotype information from the training population. Once the estimated allele effects were available, they were used together with genotype information from the

validation population to estimate genomic estimated breeding values (GEBV) for the validation population. Finally, both the correlation coefficient and the regression slope of YD on GEBV of the validation population were calculated.

Table 1 shows the total number of genotyped animals from the four different breeds as well as the respective number of individuals in the reference and validation populations per breed. Although the training populations of the Abondance and Tarentaise breeds were relatively large, they mainly consisted of females. Proportion of females in the populations ranged from 36% (in Simmental) to 94% (in Vosgienne). It can be noted that in the case of Vosgienne, nearly all animals under performance recording have been genotyped. All individuals in the validation population of all breeds were females.

Because comparing the sizes of the training populations based on Table 1 is difficult due to the different amount of information represented by female and male records, the number of males that represent an equivalent amount of information as the females altogether within each breed was computed. For this purpose, the number of females with own performance corresponding to a single progeny-tested bull was obtained from Table 1 of Boichard *et al.* (2015). Due to a lower number of progenies per progeny-tested bull in the regional breeds, the reliability of these bulls was lower than that in the large dairy cattle breeds and was considered to be 60% here.

### Pedigree-based BLUP

Based on the same phenotypes, a pedigree-based BLUP analysis was also carried out to assess the benefits of the single-breed genomic selection scenarios. The BLUP model was as follows:

$$y_i = \mu_s + u_i + e_i \quad (1)$$

where $y_i$ is the performance value of individual $i$ (DYD for males and YD for females), $\mu_s$ is an overall mean effect calculated separately for males ($s = 1$) and females ($s = 2$), $u_i$ is the breeding value of animal $i$ ($u \sim \text{MVN}(0, \mathbf{A}\sigma_u^2)$, where MVN refers to a multivariate normal distribution, $\mathbf{A}$ is the additive relationship matrix and $\sigma_u^2$ is the genetic variance), and $e_i$ is the random error term of animal $i$ ($e \sim N(0, D\sigma_e^2)$, where D is a diagonal matrix with $\frac{1}{w}$ elements (where w is the equivalent daughter contribution for males and the number of record equivalent for females) and $\sigma_e^2$ is the residual error variance.

**Table 1** Population size and the number of genotyped males and females of the four analysed breeds

| | Number of animals | | | Number of animals in the ~ population | |
|---|---|---|---|---|---|
| Breed | Male | Female | Total | Training | Validation |
| Abondance | 344 | 1482 | 1826 | 1461 | 365 |
| Tarentaise | 297 | 1167 | 1464 | 1171 | 293 |
| Simmental | 324 | 183 | 507 | 406 | 101 |
| Vosgienne | 60 | 1008 | 1068 | 854 | 214 |

### Single-breed scenarios

In the single-breed scenarios, the routine French genomic evaluation procedure was applied to the four regional breeds. An outline of the applied method is given below.

Genomic evaluation in France is performed in a single-breed context in the four major dairy cattle breeds of the country: using phenotype and genotype information from bulls in the case of Holstein Friesian and Brown Swiss and from both bulls and cows in the case of the Normande and Montbéliarde breeds (Croiseau *et al.* 2015). For each trait of interest, a set of SNP linked to QTL were identified on the 50K SNP chip using a Bayesian approach (Bayes-C$\pi$) as implemented in the GS3 software (Legarra *et al.* 2013). The Bayes-C$\pi$ procedure was originally described by Habier *et al.* (2011), with two main originalities compared to Bayes-B: a single variance is used for all SNP effects and a proportion of markers without an effect on the trait (i.e. $\pi$) can be estimated in an iterative way. However, $\pi$ had to be fixed in the case of the regional breeds due to convergence problems (in other words, instead of a Bayes-C$\pi$ analysis, a Bayes-C was used for the regional breeds with $\pi$ fixed to 80%). The model used in this Bayes-C analysis was as follows:

$$y_i = \mu_s + p_i + \sum_{j=1}^{N} z_{ij} a_j \delta_j + e_i \quad (2)$$

where $p_i$ is the polygenic effect of animal $i$ ($p \sim \text{MVN}(0, \mathbf{A}\sigma_u^2)$; MVN, $\mathbf{A}$ and $\sigma_u^2$ are defined as for the pedigree-based BLUP model), $N$ is the total number of SNP in the model, $z_{ij}$ is an indicator variable representing the number of copies of one of the alleles at marker $j$ in animal $i$, and $a_j$ is the substitution effect for marker $j$, $\delta_j$ is a 0/1 variable indicating whether or not marker $j$ has an effect. All other terms are as defined previously. The model includes a residual polygenic effect in addition to the marker effects to account for the genetic variance not explained by the

markers. In practice, the genetic variance was split into two parts: a certain proportion (α) was attributed to the markers in the model and the remaining was assumed to be explained by the residual polygenic component. All α values between 10 and 90% (with 10% increases) were tested and the one resulting in the highest correlation coefficient between YD and GEBV measured in the validation population was selected for each trait separately. All variance components and the residual polygenic effect were estimated iteratively during the analysis as well as the effects and probabilities of inclusion of each marker in the model.

Following the Bayes-C analysis, markers with the highest probabilities of inclusion were selected (n = 250, 500 or 1000). Two consequences of this selection procedure are as follows:

1 Several selected markers might be linked to the same QTL, if the QTL has a large effect (e.g. the case of the diacylglycerol O-acyltransferase 1 (DGAT1) gene for fat content).
2 For each trait, the smaller sets were subsets of the larger set(s).

Once the SNP were selected, haplotypes of four SNP were constructed around these SNP using the Criterion-B haplotype selection procedure described by Jónás *et al.* (2016). This method constructs all possible haplotypes within a short genomic window of 10 SNP around the selected SNP. From these haplotypes, it selects the haplotype that combines the largest number of well-represented alleles and the lowest number of under-represented alleles. Such haplotype choice was proven to be better in genomic evaluation than the haplotypes built by merging the adjacent SNP into a haplotype (Jónás *et al.* 2016).

The selected haplotypes were then used as explanatory variables in the final step of the genomic evaluation process. Haplotype allele effects were estimated in a marker-assisted BLUP analysis and these estimated effects were used to estimate genomic breeding values for selection candidates (i.e. animals with only genotype information). Therefore, the model used in the MA-BLUP analysis is as follows:

$$y_i = \mu_s + \sum_{j=1}^{8218} z_{ij} a_j + \sum_{k=1}^{N_h} \left( \sum_{l=1}^{N_{ka}} \beta_{kl} \varepsilon_{ikl} \right) + e_i \qquad (3)$$

where $N_h$ is the number of haplotypes (i.e. 250, 500 and 1000), $N_{ka}$ is the number of segregating alleles at haplotype $k$, $\beta_{kl}$ is the estimated allele effect of allele $l$ at haplotype $k$, and $\varepsilon_{ikl}$ is an indicator variable indicating how many copies (0, 1 or 2) of allele $l$ at haplotype $k$ individual $i$ carries; all other terms were defined as in equations 1–2. In equation 3, the usual residual polygenic effect was replaced by the sum of the effects of the 8218 SNP from the BovineLD® BeadChip (Boichard *et al.* 2012a). This is equivalent to considering a genomic relationship matrix rather than a pedigree one to represent the covariance structure of the residual polygenic effect. The value of α (i.e. the proportion of the genetic variance allocated to the haplotype markers) was chosen with the same procedure as for the Bayes-C analysis. A more detailed description of the pipeline with initial results was given by Croiseau *et al.* (2015).

### Multibreed scenarios

In order to make multibreed evaluations possible, the performance values were standardized within each breed to have a genetic variance of 1 for each trait. After this scaling and assuming that the heritability did not differ significantly among breeds, the environmental variances were equal across the breeds as well.

The multibreed scenarios were conducted using the same pipeline as in the single-breed analyses. However, the training populations consisted of the merged sets of the training population of each breed. To test which breeds benefit from which other breed(s), 11 different training populations were constructed using the training populations of either two or three or four breeds (Table 2). The validation part of the pipeline was kept in a single-breed context. This allowed an unbiased comparison between the results of the single-breed and multibreed tests.

The multibreed genetic models were similar to those of the single-breed models, but the sex-specific overall mean effect was replaced by a breed- and sex-specific mean effect to account for all the differences in the genetic background of the breeds. The modified equations are shown below for both the Bayes-C

**Table 2** The 11 different training populations used in the multibreed tests

| | | | |
|---|---|---|---|
| Analyses with two breeds | | | |
| A + T | A + S | A + V | |
| T + S | T + V | S + V | |
| Analyses with three breeds | | | |
| A + T + S | A + T + V | A + S + V | T + S + V |
| Analyses with four breeds | | | |
| A + T + S + V | | | |

A, Abondance; T, Tarentaise; S, Simmental; V, Vosgienne.

(equation 4) and marker-assisted BLUP (equation 5) models:

$$y_{bi} = \mu_{bs} + p_i + \sum_{j=1}^{N} z_{ij} a_j \delta_j + e_i \quad (4)$$

$$y_{bi} = \mu_{bs} + \sum_{j=1}^{8218} z_{ij} a_j + \sum_{k=1}^{N_h} \left( \sum_{l=1}^{N_{ka}} \beta_{kl} \varepsilon_{ikl} \right) + e_i \quad (5)$$

where $y_{ib}$ is the performance value of animal $i$ from breed $b$ and $\mu_{bs}$ is the overall mean effect of breed $b$ and sex $s$. Other variables are defined as for equations 1–3.

## Results

Both correlation coefficients and regression slopes of DYD on GEBV were averaged over the five production traits, and only the average results are presented here. Furthermore, in all cases, the presented results are measured on the validation population. Differences between correlation coefficients were expressed in percentage point and in the case of the regression slopes, their average absolute deviations from 1 are shown instead of the slopes themselves, as the desirable value of the slope of regression is 1 and several of these values (particularly in case of the fat and protein content traits and the Vosgienne breed) exceeded 1.

Table 3 shows the number of male-equivalent individuals (i.e. the number of males plus the number of males representing the same amount of phenotypic information as the genotyped females) in the four populations studied in this study for two traits with different heritabilities. The number of progeny-tested bull-equivalent performances was the same for traits with the same heritability, that is for traits with a heritability of 0.3 (milk, fat and protein yield) and for traits with a heritability of 0.5 (fat and protein contents). However, due to the different heritabilities, the females represent a very different amount of phenotypic information for these groups of traits.

**Table 3** The number of males plus the number of male-equivalent females[a] in the analysed breeds

| | Milk yield | Fat content |
|---|---|---|
| Heritability | 0.3 | 0.5 |
| Abondance | 767 | 1332 |
| Tarentaise | 630 | 1075 |
| Simmental | 376 | 446 |
| Vosgienne | 348 | 732 |

[a]Calculated based on Boichard *et al.* (2015).

Based on both the total number of individuals (Table 1) and the number of male-equivalent individuals, Abondance and Tarentaise had the most phenotypic data available. However, the difference between the sizes of the two breeds was considerably smaller based on the number of male-equivalent individuals than based on the total number of individuals. Despite a relatively large number of females genotyped (Table 1), the number of male equivalents is the lowest in the Vosgienne breed (348) in the moderately heritable traits.

Linkage disequilibrium decay was compared between the single-breed and multibreed scenarios based on HD genotype data for more accurate estimates. LD patterns were remarkably similar between the single-breed and the 11 multibreed scenarios (see Figure S1).

### Single-breed scenarios

Figure 1 shows the part of genetic variance attributed to the haplotypes (i.e. $\alpha$) in the single-breed scenarios. Values are averaged across the five traits. As expected, this parameter increased with the increase in the number of haplotypes in the model; that is, when more QTL were included, a larger part of the genetic variance was explained by the markers. The increase in $\alpha$ was slower in the Simmental for reasons explained later. Results for the multibreed tests (data not shown) were very similar to the single-breed results presented in Figure 1.

Table 4 shows the correlation coefficients between GEBV and YD values for the four breeds in a single-breed context, as function of the number of haplotypes in the model. In addition, the results for the pedigree-based BLUP analysis are provided as well. The French routine genomic selection pipeline led to increased average correlations between YD and GEBV when compared to the correlations between YD and EBV from the pedigree-based BLUP analysis in nearly all traits and breeds. The gain [averaged across the five production traits and across the three different numbers of assumed QTL in the model (i.e. 250, 500 or 1000 haplotypes)] was 10.9, 5.7, 7.5 and 0.7% for the Abondance, Tarentaise, French Simmental and Vosgienne breeds, respectively. When compared to the pedigree-based BLUP analysis, the gain observed with the genomic evaluation was increasing with the number of haplotypes in all breeds except in the Simmental.

Apart from Simmental, there was a positive correlation between the number of animals in the training population (Tables 1 and 2) and the gain in terms of

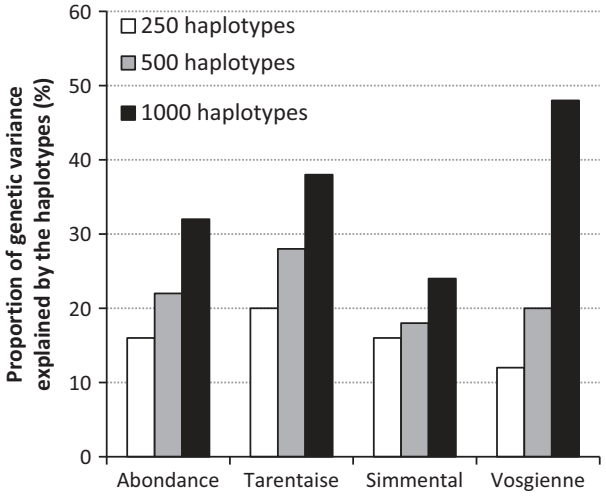



**Figure 1** Estimated proportion of genetic variance attributed to the haplotypes in the four single-breed scenarios. Average values over the five traits are plotted.

correlation coefficients with the genomic evaluation when compared to pedigree-based BLUP results. In spite of its smaller training population size, Simmental outperformed the Tarentaise in terms of extra gain in genomic selection when compared to the pedigree-based BLUP analysis.

In general, 500 and 1000 haplotypes in the model resulted in the highest correlations between YD and GEBV. However, Simmental was an exception again, with the highest observed correlation with only 250 haplotypes in the model. Differences in prediction accuracies with the different numbers of haplotypes in the model were relatively small, with a maximum of 1.1% in the Vosgienne.

Deviations from 1 of the regression slopes observed in the single-breed analyses are shown in Table 5. Once again, the applied genomic evaluation procedure outperformed the pedigree-based BLUP analysis. The deviation of the slopes from 1 was negatively correlated with the number of individuals with performance information. The average regression slope was closest to 1 in the Abondance and Tarentaise breeds, while it was the farthest within the Simmental. In general, the regression slopes were closest to 1 when 1000 haplotypes were included in the model. In addition, 500 haplotypes in the model resulted in slightly better slopes of regression than 250 haplotypes.

### Multibreed scenarios

The single-breed and multibreed tests were compared based on the average correlation coefficients and regression slopes observed across the three different numbers of haplotypes tested (250, 500 and 1000). The training populations of the multibreed scenarios always included the breed that was used in the validation step.

Figure 2 shows the correlation coefficients between YD and GEBV observed in the multibreed scenarios for the four different breeds. In the multibreed scenarios, an increased correlation coefficient between the GEBV and YD values was observed in the Abondance and Simmental breeds, while it decreased in the Tarentaise and Vosgienne breeds.

The Abondance breed benefited from all other breeds in the multibreed tests, when the basis of comparison was the correlation coefficient between the YD and GEBV measured on the validation population. When the training population of only one additional breed was added to the training population of the Abondance breed, an increase of 3.5 to 7.3% in correlation was observed. These values increased to 5.1 and 8.0%, when two additional training populations were merged with the training population of the Abondance breed and the gain in a multibreed test was 6.1%, when all the four breeds were used to estimate genomic breeding values in the Abondance breed.

Similarly, the Simmental benefited from the multibreed training populations, with an increase in correlation coefficient of 3.7% when the Abondance was included in the training population, and of 4.2% when the Vosgienne breed was added instead of the Abondance (Figure 2). When both breeds were included, the observed gain was lower (2.4%). In the case of the Simmental breed, the inclusion of the Tarentaise was detrimental, leading to an average 2.4%

**Table 4** Correlation coefficients between GEBV and YD values of the validation population in the single-breed scenarios. Results of the pedigree-based BLUP analysis are also provided. Average correlations over the five production traits for the four different breeds

| Method | Number of haplotypes | Abondance | Tarentaise | Simmental | Vosgienne |
|---|---|---|---|---|---|
| BLUP | – | 0.346 | 0.391 | 0.243 | 0.418 |
| Genomic selection | 250 | 0.454 | 0.446 | 0.323 | 0.420 |
| | 500 | 0.454 | 0.449 | 0.318 | 0.426 |
| | 1000 | 0.459 | 0.449 | 0.314 | 0.430 |

**Table 5** Regression slopes of DYD on GEBV in the single-breed scenarios. Presented values are averaged for the five production traits and measured as absolute deviations from 1

| Method | Number of haplotypes | Abondance | Tarentaise | Simmental | Vosgienne |
|---|---|---|---|---|---|
| BLUP | – | 0.111 | 0.121 | 0.394 | 0.155 |
| Genomic selection | 250 | 0.090 | 0.104 | 0.260 | 0.168 |
| | 500 | 0.092 | 0.099 | 0.257 | 0.150 |
| | 1000 | 0.092 | 0.079 | 0.244 | 0.114 |

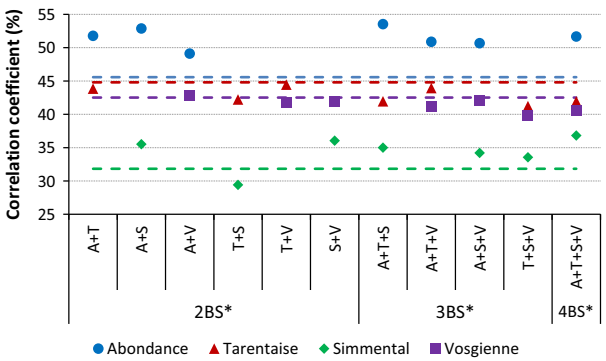



**Figure 2** Correlation coefficients observed in the validation population in the multibreed analyses in the four different breeds. Values are averages across the three tested haplotype sizes; the dashed lines correspond to the single-breed scenarios. Abbreviations on the *x*-axis labels: A, Abondance; T, Tarentaise; S, Simmental; V, Vosgienne. *2/3/4-breed scenarios

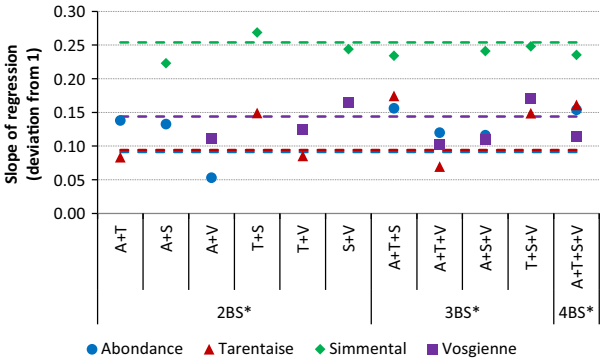



**Figure 3** Deviation of the slopes of regressions from 1 observed in the multibreed analyses in the four different breeds. Values are averages across the three tested haplotype sizes; the dashed lines correspond to the single-breed scenarios. Abbreviations on the *x*-axis labels: A, Abondance; T, Tarentaise; S, Simmental; V, Vosgienne. *2/3/4-breed scenarios

decrease in the correlations. When the Tarentaise was added together with the Abondance (or the Vosgienne) breed, the gain in terms of correlations was lower when only the Abondance (or Vosgienne) was included in the training population in addition to the Simmental. The highest correlation was observed, when the training population consisted of those from all four breeds (average gain: 5.0%).

The accuracy of genomic evaluation decreased in Tarentaise when multibreed training populations were used. A similar result was found in case of the Vosgienne breed, except with the Abondance+Vosgienne training population, for which the accuracy did not change compared to the single-breed scenario. The decrease ranged from 0.4 to 3.6% in Tarentaise and from 0.4 to 2.8% in Vosgienne.

Figure 3 shows the deviations of the regression slopes from 1. The results for all multibreed scenarios are plotted for all breeds. As for the correlation coefficients, deviations of the regression slopes were also averaged across the three tested numbers of haplotypes in the model and across the five traits. Similar to the single-breed results, the estimated regression slopes were better (i.e. closer to 1) in case of breeds with larger training populations (i.e. with Abondance and Tarentaise) than with the other ones. However, when the results are compared to the single-breed results, the conclusions are unclear: in general, the deviation of the regression slopes from 1 became smaller with the Simmental and Vosgienne breeds and increased with Abondance and Tarentaise.

### Statistical analysis of the observed gains

We investigated the significance of the obtained gains using Fisher's Z-transform (implemented in the 'cocor' R package by Diedenhofen & Musch 2015; based on Zou 2007). Our assumption was that the genomic evaluation results are superior compared to the BLUP results. Therefore, a one-tailed test with an $\alpha = 5\%$ was implemented. Gains were significant in case of two traits (fat content and protein content) in Abondance and Tarentaise (see Figures S1 and S2). In case of the multibreed scenarios, observed gains were mainly insignificant, when compared to the single-breed results (data not shown).

While only very high gains (>10%) would have been significant, a smaller gain was observed in most of the cases. To test whether a small gain can be consistently expected with genomic evaluation compared

to the pedigree-based BLUP results, a Wilcoxon signed-rank test was implemented. Genomic evaluation (with 1000 haplotypes) correlations were compared to those obtained with the pedigree-based BLUP. Once again, a one-tailed test was used with $\alpha = 5\%$ for the five pairs of correlations obtained in the five traits. The Wilcoxon signed-rank test was used because normality could not be assumed due to the small sample size (i.e. the number of traits) and because the correlations were paired by trait. Based on these tests, genomic selection can be expected to lead to an increased selection accuracy in Abondance ($W = 15$; $p \approx 0.03$) and in Simmental ($W = 15$; $p \approx 0.03$), but not in the other two breeds.

The same Wilcoxon signed-rank test was used to compare the highest multibreed correlations with those of the single-breed. In conclusion, in case of the Abondance and Simmental breeds, multibreed genomic evaluations led to systematically higher correlations ($p \approx 0.03$), when compared to the within-breed evaluation results.

## Discussion

In this study, we evaluated the performance of single-breed and multibreed genomic evaluations in four regional dairy cattle breeds in a classical validation study. The training populations consisted of both males and females, while the validation populations included only female individuals. The population sizes for these breeds ranged from 145 till 548 progeny-tested bulls after accounting for the differences between cows and bulls with respect to the represented amount of information. We showed that single-breed genomic evaluations were more accurate than a pedigree-based BLUP analysis even in regional breeds with a small training population. The obtained gains in terms of accuracy depended on the number of individuals in the training populations, and larger gains were observed with larger breeds (Tables 3 and 4). The Simmental breed had a particular population structure due to its large proportion of imported breeding animals and/or semen. Because the progeny of these animals had only a very limited amount of pedigree information available in France, overall performance of all breeding value estimation methods was inferior in Simmental when compared to the other breeds. This population structure of the Simmental explains why both the pedigree-based BLUP and the applied genomic evaluation procedures performed worse in Simmental than in the other breeds. In addition, this is also the reason why we observed a larger gain with genomic evaluations (compared to the pedigree-based BLUP) with Simmental (~7.54%) than with Tarentaise (~5.68%), in spite of the larger training population in the case of the latter breed (Table 3). The gain with genomic evaluation compared to pedigree-based BLUP was the smallest with the Vosgienne, which can be because of the higher average age of breeding animals within this breed, resulting in more accurate EBV from the pedigree-based BLUP tests. The deviations of the regression slopes from 1 also improved with the genomic evaluation, when compared to the pedigree-based BLUP results (Table 5).

Genomic evaluation has a positive impact on the quality of evaluation: all measured parameters showed some improvement with the genomic evaluation when compared to the pedigree-based BLUP results. As a consequence, routine genomic selection was implemented in the four regional breeds in France in early 2016. The most important expected benefits of genomic evaluation in the regional breeds are the possibility to have shorter generation intervals (if progeny testing is discontinued) and a larger number of evaluated animals, which has a positive influence on the within-breed genetic diversity as well.

Interpretation of the regression slopes is difficult in the multibreed tests, because they are not consistent for each trait within a breed. The unfavourable trends with the Abondance and Tarentaise are at least partly due to the positive correlation between the correlation coefficient and the slope of regression of linear regression models (i.e. given the $DYD = \beta_0 + \beta_1 * GEBV + e$ regression model, the regression slope can be written as $\beta_1 = r * \frac{\sigma_{DYD}}{\sigma_{GEBV}}$, where $r$ is the correlation coefficient between DYD and GEBV). In other words, the regression slope constantly increases with the increase in the correlation coefficient and this trend is either advantageous (when the slope of regression was lower than 1) or disadvantageous (when the slope of regression was higher than 1).

Hayes *et al.* (2009) demonstrated a large gain in the accuracy of the Jersey GEBV when analysing a Holstein–Jersey multibreed population using SNP information from the 50K chip. Using another combined Holstein–Jersey training population, Erbe *et al.* (2012) showed a 4% increase in prediction accuracy for the smaller breed (Jersey), when compared to the within-breed test, using the BovineHD BeadChip® (manufactured by Illumina Inc., San Diego, CA), but found a very limited gain when using 50K SNP chip data. Similar to Hayes *et al.* (2009), we also observed an improvement in terms of GEBV accuracies using the 50K SNP panel in several multibreed tests. While

Hayes *et al.* (2009) did not observe any gain in the Holstein Friesian (i.e. the large breed contributing to the multibreed population), we could demonstrate a large improvement of the accuracy even for the largest breed in our study. This is probably because of the shorter genetic distance between the breeds analysed in this current study (Gautier *et al.* 2010). Hozé *et al.* (2014) showed an improvement of 2.9% in selection accuracy compared to a single-breed scenario when analysing a Holstein–Normande–Montbéliarde multibreed population. In terms of correlation between YD and GEBV, we observed a maximum gain of 8 and 5% in the Abondance and Simmental breeds, respectively.

de Roos *et al.* (2009a) showed that genetic distance between the breeds participating in a multibreed genomic evaluation is an important factor with a significant effect on the efficiency of the evaluations. In our study, the Abondance breed benefited from the addition of the training population of all other breeds, while the Simmental benefited from the addition of the training populations of Abondance and Vosgienne. In contrast, neither the Tarentaise nor the Vosgienne benefited from any other breeds.

The level of accuracy of GEBV is partly due to a quite accurate estimation of the parent average and partly due to a relatively accurate estimation of QTL effects. The high accuracy of the BLUP breeding values in Vosgienne indicates that the training and validation populations were closely related. In addition, this breed had a small training population. Hence, in Vosgienne, the high accuracies of GEBV result mainly from an accurate estimation of the parent averages. Adding other breeds to the reference population led to more accurate QTL effect estimations (in the case of the shared QTL), but probably decreased the accuracy of the parent averages. Hence, the use of multibreed training population was detrimental in Vosgienne.

Linkage disequilibrium persistency is another factor that can explain the observed gains and losses in terms of accuracy. In order to measure the LD persistency, first we calculated the r values for the neighbouring SNP in each of the four breeds (Figure S1). Next, we calculated and plotted the correlations of the r values between the breeds for different marker distances (moving averages covering ~4Kb each are shown in Figure S4). This way of measuring the LD persistency is identical to that of de Roos *et al.* (2008). We did not observe the same decrease in correlation of r values with the increasing marker distance as de Roos *et al.* (2008) did. This is likely because of the much shorter range of marker distances covered by the neighbouring SNP in our analysis (20–60 Kb versus 0–1 Mb in de Roos *et al.* 2008). The correlations of r values ranged from 58% (between Abondance and Tarentaise) to 70% (between Simmental and Vosgienne). These correlations were generally lower with the Tarentaise breed (58–64%) and higher with the Simmental (64–70%). This can also partly explain our results, for example why the multibreed training population was detrimental for the Tarentaise breed and why was it beneficial for the Simmental.

These results suggest that in addition to the genetic distance between the breeds (Gautier *et al.* 2010), there are other relevant factors determining the efficiency of multibreed genomic selection (e.g. the frequency and relative importance of breed-specific QTL within each breed or the different QTL–marker allele frequencies in the different breeds). Indeed, if only the genetic distance would be relevant, genetically close breeds would benefit from each other in both ways.

Another essential condition for an efficient multibreed genomic evaluation is the consistency of phases between marker and QTL alleles among the different breeds. We found that the LD decay observed in the analysed breeds was remarkably similar. In addition, it was shown earlier that these breeds are very closely related (Gautier *et al.* 2010); therefore, it was reasonable to assume that these breeds would benefit from a multibreed genomic evaluation. In contrast, the use of a multibreed training population was detrimental for some breeds, suggesting the lack of conserved QTL–marker allele phases. A possible improvement would be to identify those markers (with significant effects) that influenced the traits in the same direction, as suggested by Porto-Neto *et al.* (2015).

## Conclusions

The French routine genomic evaluation method was applied to four regional breeds in both single-breed and multibreed contexts. We showed that genomic evaluation outperforms a pedigree-based BLUP analysis even though the available training population is of limited size. Both the Abondance and Simmental breeds benefited from at least two other breeds in multibreed genomic evaluations. In some cases, the introduction of multibreed training populations did not affect the estimated breeding values of the different breeds constituting to this multibreed training population in the same direction, suggesting that factors other than genetic distance between the breeds also influence the efficiency of multibreed genomic evaluations. Further research is required to better understand the background of multibreed genomic

evaluation. In particular, benefiting from known causative mutations identified in other dairy cattle breeds is especially promising when the aim is to develop an efficient genomic evaluation procedure for regional breeds.

## Acknowledgements

GEMBAL Project is funded by the Agence Nationale de la Recherche (ANR-10-GENM-0014), APISGENE, Races de France and INRA 'AIP Bioressources'. We also express our gratitude to the breeding organizations and AI companies that were involved in the genotyping of these breeds.

## References

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

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

Boichard D., Ducrocq V., Fritz S. (2015) Sustainable dairy cattle selection in the genomic era. *J. Anim. Breed. Genet.*, **132**, 135–143.

Croiseau P., Baur A., Jónás D., Hozé C., Promp J., Boichard D., Fritz S., Ducrocq V. (2015) Comparison of different marker-assisted BLUP models for a new French genomic evaluation. In: Book of abstracts of the 66th annual meeting of the European federation of animal science. Warsaw, (Poland). 31, August – 4 September 2015. Page 248. Copyright Wageningen Academic Publishers, Wageningen, Netherlands.

Diedenhofen B., Musch J. (2015) cocor: a comprehensive solution for the statistical comparison of correlations. *PLoS ONE*, **10**, e0121945.

Erbe M., Hayes B.J., Matukumalli L.K., Goswami S., Bowman P.J., Reich C.M., Mason B.A., Goddard M.E. (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.

Garrick D.J., Fernando R. (2014) Genomic prediction and genome-wide association studies in beef and dairy cattle. In: D.J. Garrick, A. Ruvinsky (eds), The Genetics of Cattle. (2nd edn). CABI, Wallingford, pp. 474–501.

Gautier M., Laloë D., Moazami-Goudarzi K. (2010) Insights into the genetic history of French cattle from dense SNP data on 47 worldwide breeds. *PLoS ONE*, **5**, e13038.

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

Harris B.L., Winkelman A.M., Johnson D.L. (2013) Impact of including a large number of female genotypes on genomic selection. *Interbull Bull.*, **47**, 23–27.

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

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

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.*, **97**, 3918–3929.

Institut de l'Elevage. (2015) Résultats de côntrole laitier – France 2014. (available at : http://idele.fr/no_cache/recherche/publication/idelesolr/recommends/resultats-de-controle-laitier-france-2014.html; last accessed 13 January 2016). .

Jónás D., Ducrocq V., Fouilloux M.-N., Croiseau P. (2016) Alternative haplotype construction methods for genomic evaluation. *J. Dairy Sci.*, **99**, 4537–4546.

Legarra A., Ricard A., Filangi O. (2013) GS3 software package and documentation. (available at: http://snp.toulouse.inra.fr/~alegarra; last accessed 1 January 2013).

Lund M.S., de Roos A.P.W., de Vries A.G., Druet T., Ducrocq V., Fritz S., Guillaume F., Guldbrandtsen B., Liu Z., Reents R., Schrooten C., Seefried F., Su G. (2011) A common reference population from four European Holstein populations increases reliability of genomic predictions. *Genet. Sel. Evol.*, **43**, 43.

Porto-Neto L.R., Barendse W., Henshall J.M., McWilliam S.M., Lehnert S.A., Reverter A. (2015) Genomic correlation: harnessing the benefit of combining two unrelated populations for genomic selection. *Genet. Sel. Evol.*, **47**, 84.

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

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

de Roos A.P.W., Schrooten C., Mullaart E., van der Beek S., de Jong G., Voskamp W. (2009b) Genomic selection at CRV. *Interbull Bull.*, **39**, 47–50.

VanRaden P.M., Van Tassell C.P., Wiggans G.R., Sonstegard T.S., Schnabel R.D., Taylor J.F., Schenkel F.S.

(2008) Invited review: reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci.*, **92**, 16–24.

Villumsen T.M., Janss L., Lund M.S. (2009) The importance of haplotype length and heritability using genomic selection in dairy cattle. *J. Anim. Breed. Genet.*, **126**, 3–13.

Wiggans G.R., VanRaden P.M., Cooper T.A. (2011) The genomic evaluation system in the United States: past, present future. *J. Dairy Sci.*, **94**, 3202–3211.

Zou G.Y. (2007) Toward using confidence intervals to compare correlations. *Psychol. Methods*, **12**, 399–413.

## Supporting Information

Additional Supporting Information may be found in the online version of this article:

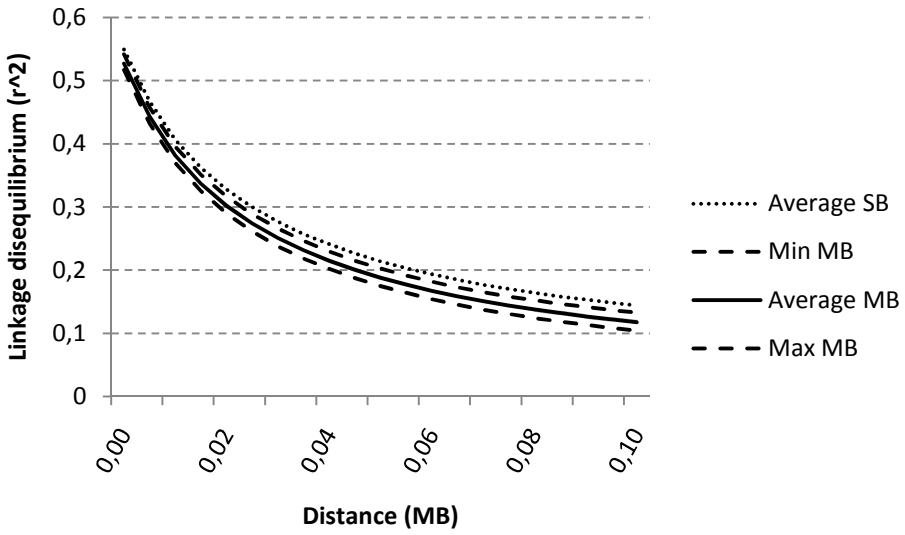
**Figure S1.** LD decay along the genome in both the single-breed (dotted line) and multibreed (solid line) scenarios. The slowest and fastest LD decays among the 11 different multibreed tests are also shown (dashed lines).

**Figure S2.** Results of the hypothesis testing indicating whether the observed gains in single-breed genomic evaluations are statistically significant from zero or not in the Abondance breed. Gains/losses in correlations observed with the single-breed genomic evaluation pipeline compared to the BLUP model are indicated (short horizontal lines). The lower confidence intervals for the gains/losses based on Fisher's Z-transform are also shown (black triangles). The following trait name abbreviations are used on the plot: MQ, milk quantity; FY, fat yield; PY, protein yield; FC, fat content; PC, protein content.

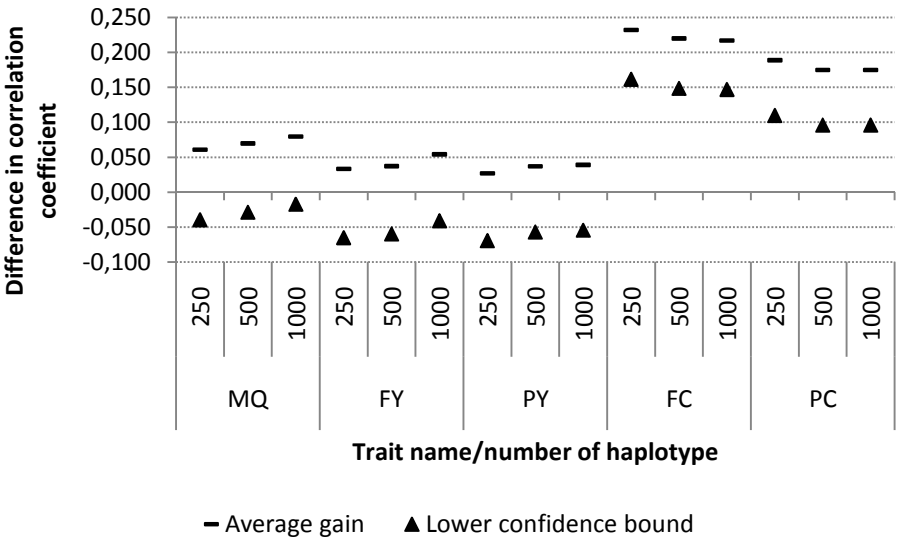
**Figure S3.** Results of the hypothesis testing indicating whether the observed gains in single-breed genomic evaluations are statistically significant from zero or not in the Tarentaise breed. Gains/losses in correlations observed with the single-breed genomic evaluation pipeline compared to the BLUP model are indicated (short horizontal lines). The lower confidence intervals for the gains/losses based on Fisher's Z-transform are also shown (black triangles). The following trait name abbreviations are used on the plot: MQ, milk quantity; FY, fat yield; PY, protein yield; FC, fat content; PC, protein content.

**Figure S4.** Between breeds correlation coefficients of r values calculated within breeds, as a function of markers distance. Different lines correspond to the different pairs of breeds (A, Abondance; T, Tarentaise; S, Simmental; V, Vosgienne).

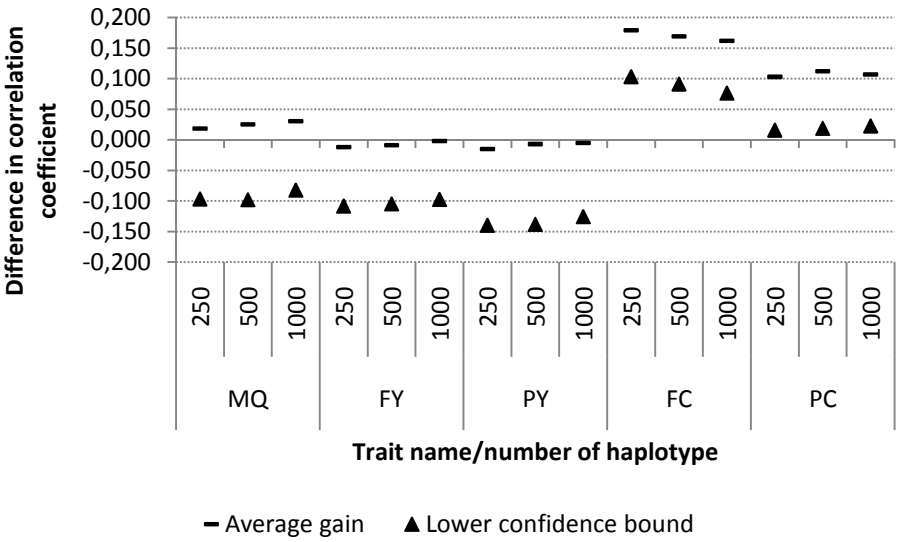
S. figure 1


0,6
0,5
0,4
0,3
0,2
0,1
0
Linkage disequilibrium (r^2)
0,00
0,02
0,04
0,06
0,08
0,10
Distance (MB)
Average SB
Min MB
Average MB
Max MB

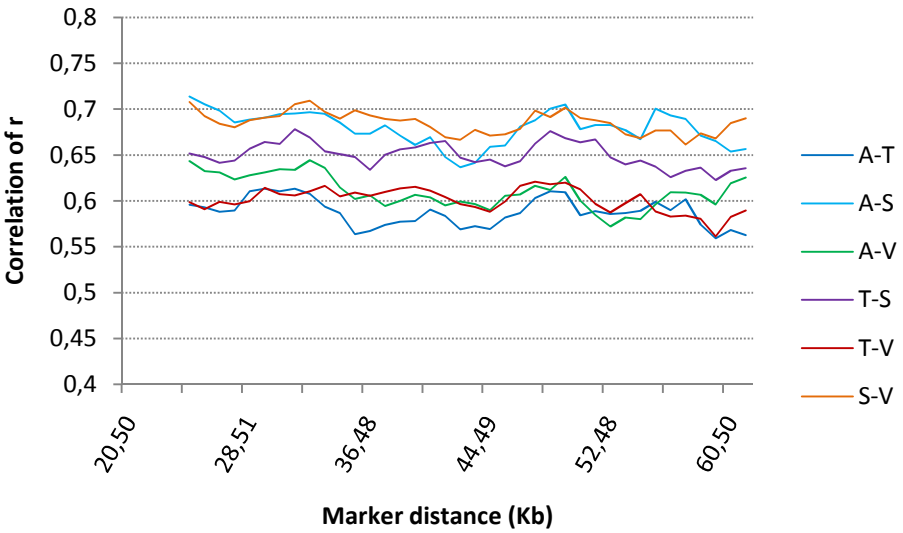
S. figure 2


Difference in correlation coefficient
0,250
0,200
0,150
0,100
0,050
0,000
-0,050
-0,100
250
500
1000
MQ
FY
PY
FC
PC
Trait name/number of haplotype
Average gain
Lower confidence bound

S. figure 3


0,200
0,150
0,100
0,050
0,000
-0,050
-0,100
-0,150
-0,200
Difference in correlation coefficient
250
500
1000
MQ
FY
PY
FC
PC
Trait name/number of haplotype
Average gain
Lower confidence bound

S. figure 4


0,8
0,75
0,7
0,65
0,6
0,55
0,5
0,45
0,4
Correlation of r
20,50
28,51
36,48
44,49
52,48
60,50
Marker distance (Kb)
A-T
A-S
A-V
T-S
T-V
S-V

### 4.3.3 BayesC results

The first step of the routine evaluation was a BayesC analysis, which was used as a QTL detection step. The proportion of SNP without an effect on the trait of interest ($\pi$) was fixed to 80% due to convergence issues with a variable $\pi$ with BayesC. BayesC was implemented in a validation study with the same training and validation set definitions as for the MA-BLUP analysis. The results of this BayesC analysis are presented here. The correlation coefficients and regression slopes of YD on GEBV (regression slopes expressed as a deviation from 1) averaged over the 5 production traits and measured in the validation population are shown in **Table 12** for the 4 regional breeds (for an easier comparison, the routine evaluation results with 1000 haplotypes in the model are also shown).

**Table 12**: Average correlation coefficients and regression slopes (expressed as deviations from 1) of the 5 traits measured on the validation set from a BayesC and from the routine genomic evaluation.

| Breed | BayesC | | Routine genomic evaluation | |
|---|---|---|---|---|
| | Correlation coefficient | Regression slope[1] | Correlation coefficient | Regression slope[1] |
| Abondance | 0.417 | 0.216 | 0.459 | 0.092 |
| Tarentaise | 0.446 | 0.109 | 0.449 | 0.079 |
| Simmental | 0.305 | 0.297 | 0.314 | 0.244 |
| Vosgienne | 0.431 | 0.081 | 0.430 | 0.114 |

1: Average absolute deviations from 1

In terms of correlations, the BayesC model outperformed the pedigree-based BLUP procedure, (also see **Table 4** and **5** from the paper), but not the French routine evaluation. In Abondance, Tarentaise and Simmental the correlation coefficients were higher with the routine evaluation than with the BayesC, while in Vosgienne the correlation coefficient in with BayesC is similar to the correlation obtained with the routine evaluation. Regression slopes with BayesC improved compared to BLUP in all breeds except Abondance (for regression slopes with BLUP, see **Table 5** from the

paper). The regression slopes were better (especially in the Abondance breed) with the routine evaluation except for Vosgienne.

The difference between the routine evaluation and the BayesC results, in terms of correlation coefficients and regression slopes were mainly in favor of the routine evaluation. A major advantage of the routine evaluation over the BayesC approach is that it uses the same markers over time and is much faster as well. However, the haplotype selection step of the routine might be repeated after a few generations of selection, as discussed in section 2.5 of Chapter 2.

### 4.3.4 Discussion

The French routine genomic evaluation was tested in four regional dairy cattle breeds. It was shown that the estimated GEBV reliabilities of the selection candidates were approximately the same compared to the reliabilities of progeny-tested bulls in these breeds (Sanchez et al., 2016) and therefore they are sufficiently high for official publications. Selection candidates in this context do not correspond to the validation population of the previous study (i.e. the 20% youngest – female – individuals) but to the population of young bulls without performance observations as of June, 2016.

Due to the lower costs of genotyping compared to progeny-testing, a much larger number of male candidates can be evaluated (between 55 and 226, depending on the breed) than under progeny testing (**Table 1**). This is expected to have a positive impact on the genetic diversity of the breeds, because artificial insemination (AI) cooperatives and breeders can now select from a wider range of young bulls with reasonable reliabilities. Furthermore, female reliabilities become as accurate as male reliabilities with genomic evaluation and breeding values also become available for fertility traits in females for the first time for these breeds. These are again important advantages compared to the previous breeding program.

As a consequence of these benefits, routine genomic evaluation was implemented in three of the four tested regional breeds in France (Abondance, Tarentaise and Vosgienne). The reference population for these breeds includes both males and females. Genomic evaluation was implemented in Simmental as well but using a

much larger, international reference population. Although this breed classifies as a regional breed in France, there is a substantial worldwide Simmental population (especially in Germany and Austria) and these countries assembled a large Simmental reference population in the previous years, which provides significantly more accurate GEBV for selection candidates and therefore promises larger annual genetic gains compared to the ones obtained in our study. Consequently, the French breeding association of the Simmental breed decided to participate in this international cooperation.

The use of a multi-breed training population in genomic evaluation was beneficial in two (Abondance and Simmental) of the four breeds. The three important requirements for an efficient multi-breed genomic evaluation are:

- QTL and SNP are shared across the breeds
- LD is conserved between the QTL and adjacent markers
- QTL-SNP linkage phases are shared

In the cases of the QTL where all of these 3 criteria are met, all of the 4 breeds benefit equally from the multi-breed training population. However, in the cases when at least one of these criteria is not fulfilled (e.g. the case of breed-specific QTL), the multi-breed training population introduces noise to the allele effect estimation. This latter phenomenon did not receive much attention until recently (e.g. Porto-Neto et al., 2015).

In the multi-breed tests, we observed that two out of the four analyzed breeds benefited from the multi-breed genomic evaluation while the other two did not. This indicates that the relative importance of breed-specific QTL differs among the breeds, which led to either a gain or a loss when a multi-breed genomic evaluation was performed. However, since the multi-breed genomic evaluation does not hold any promise to increase the estimation accuracy of allele effects for breed-specific QTL, it might be beneficial to identify these in a first step (for example by comparing QTL-detection analysis results from the different breeds) and estimate them separately in a within-breed context. This would efficiently avoid the noise introduced by the other breeds in which the QTL is not segregating. The same applies to the cases when

either the QTL-allele phases or the LD between the QTL and neighboring markers are not conserved. However, in these breeds the within-breed reference populations are likely to be too small to conduct such analyses with a high accuracy.

## 4.4 Genomic evaluation with high-density data

### 4.4.1 Introduction

In the previous chapter the benefits of genomic evaluation using 50K SNP-chip information in the regional breeds was presented. However, the HD chip was thought to improve the performance of multi-breed genomic evaluation due to the higher marker density, which leads to higher LD between markers and QTL. This could efficiently counterbalance the diminishing LD between markers when the training populations of multiple breeds are mixed.

The methodological developments presented in Chapter 3 allowed the combined use of HD data and haplotype markers in genomic evaluation, because the number of allele effects could be greatly reduced. If the windows of 144 SNP are used on the HD data in combination with haplotypes of 4 SNP, the number of haplotypes built from the 706,791 SNP of the HD chip could be reduced by 97% compared to the case when all consecutive haplotypes of 4 SNP are used.

Based on the results from the Montbéliarde breed (see section 3.4), the use of the HD data was detrimental to the selection accuracy and regression slopes. Accordingly, the HD data was not used in a within-breed context in the regional breeds because no gain can be expected from such an analysis. The results obtained with the HD data in a multi-breed context were compared directly to the results obtained with the 50K data (both single- and multi-breed).

### 4.4.2 Materials and methods

The same populations were used for this analysis than for the 50K tests. This population was presented in section 4.1 and 4.3 in detail. The multi-breed training population consisted of the training populations of the 4 breeds altogether. The implemented validation study was also identical, with the same animals in the training

and validation sets as the validation study with the 50K data. These allowed a direct comparison of the results obtained to the HD data with those obtained with the 50K SNP-chip data.

Three different analyses were implemented in a multi-breed context using HD data:

1. The first test was the routine evaluation (with all its steps). The window size was fixed to 144 SNP, which was the average number of SNP within the 10 SNP-wide windows from the 50K data in the Montbéliarde breed.
2. The first results were inferior compared to the single-breed tests with 50K data. This could be because the QTL-SNP (the SNP linked to QTL) could not be identified accurately in the QTL detection step of the routine evaluation. Therefore, in the second scenario, the QTL-SNP detected with the 50K data were used in the HD dataset (or the closest SNP from the HD panel, if the QTL-SNP was not available on this SNP-chip). Window size was again fixed to 144 SNP.
3. Although the results improved significantly compared to the first analysis, they were still inferior compared to the 50K results, which is against expectation. One explanation can be that it is detrimental to use the same window size for all regions, because most windows of 10 SNP from the 50K SNP-chip overlaps with either more or less SNP from the HD SNP-chip. **Figure 10** shows the distribution of the number of SNP from the HD SNP-chip under the windows of 10 SNP from the 50K chip in the Montbéliarde breed (the Montbéliarde is presented, because the average number of 144 SNP was also calculated from this breed). Although the average number of the windows is at 144 SNP, majority of the 10-SNP windows of the 50K overlap with either more or less than 144 SNP from the HD SNP-chip. Therefore, in the third analysis different window sizes were used for the different QTL-SNP, which covered the exact same genomic regions as the windows of the 50K.

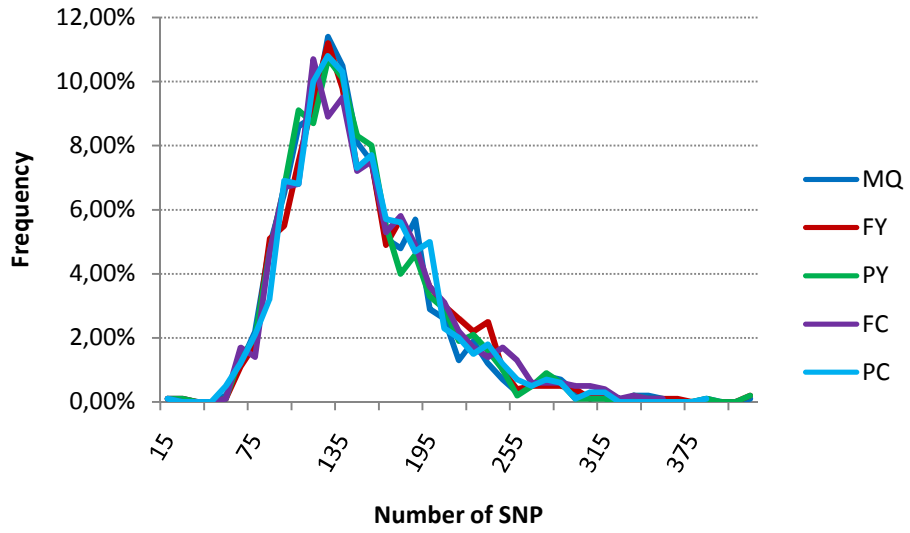



**Figure 10**: Frequency distribution of the number of SNP from the HD SNP-chip overlapping with the 10 SNP-wide windows from the 50K SNP-chip (Montbéliarde breed). Trait name abbreviations: MY – milk yield; FY – fat yield; PY – protein yield; FC – fat content; PC – protein content.

### 4.4.3 Results

The first scenario gave inferior results compared to the other two, while the second analysis pipeline gave on average over the 4 breeds worse results than the third scenario. Therefore, only the results of the third analysis are presented and discussed here.

The estimated correlation coefficients and regression slopes of YD on GEBV are shown in **Table 13** for the 4 regional breeds using a multi-breed training population. When these results are compared to the results of the single-breed analysis with 50K data, the correlation coefficients were higher, except for the Tarentaise breed, in which breed an average decrease of 1% was observed (also see **Table 4** from the previously inserted article). The average gains (over the 5 analyzed traits) in the other 3 breeds were between 0.2 (Vosgienne) and 4.5% (Simmental). Regression slopes improved in the same three breeds.

When the HD results are compared to the results of the multi-breed analysis with 50K data, the correlation coefficients presented in **Table 13** were inferior in Abondance and Simmental and an increase of 1.9% and 2.7% was observed in Tarentaise and Vosgienne, respectively. The decrease of the correlation coefficients in Abondance

and Simmental were relatively small (1.7% and 0.8%). Regression slopes improved considerably (i.e. were closer to 1) in Abondance and Vosgienne, but did not change in Simmental and declined in Tarentaise. Although in a multi-breed context slight improvements were observed in the correlation coefficients with the Tarentaise and Vosgienne breeds (HD vs. 50K), these gains were not large enough to surpass the correlations calculated for these breeds in a single-breed analysis with 50K data (**Table 13**).

**Table 13**: Correlations and regression slopes between the DYD and GEBV in the 4 regional breeds. Average single-breed (SB) and multi-breed (MB) results with 50K are also added.

| SNP-chip ID | Trait[1] | Correlation coefficient | | | | Regression slope | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | A | T | S | V | A | T | S | V |
| HD | MY | 0.42 | 0.29 | 0.35 | 0.38 | 0.98 | 0.64 | 0.80 | 1.00 |
| | FY | 0.45 | 0.36 | 0.37 | 0.29 | 1.11 | 0.79 | 0.81 | 1.01 |
| | PY | 0.36 | 0.25 | 0.41 | 0.33 | 0.96 | 0.59 | 0.92 | 0.97 |
| | FC | 0.65 | 0.69 | 0.30 | 0.55 | 1.01 | 1.04 | 0.48 | 1.02 |
| | PC | 0.61 | 0.61 | 0.37 | 0.60 | 1.00 | 0.99 | 0.80 | 1.22 |
| | Average[2] | 0.50 | 0.44 | 0.36 | 0.43 | 0.04 | 0.21 | 0.24 | 0.06 |
| 50K (SB) | Average[2] | 0.46 | 0.45 | 0.31 | 0.43 | 0.09 | 0.08 | 0.24 | 0.11 |
| 50K (MB) | Average[2] | 0.52 | 0.43 | 0.38 | 0.42 | 0.16 | 0.15 | 0.22 | 0.11 |

1: Trait name abbreviations: MY – milk yield; FY – fat yield; PY – protein yield; FC – fat content; PC – protein content

2: Average deviations from 1 are indicated for regression slopes

3: Breed name abbreviations: A – Abondance; T – Tarentaise; S – Simmental; V – Vosgienne

### 4.4.4 Conclusions

The multi-breed scenario with all 4 breeds contributing to the training population was performed as a pilot study. A consequence of the 3 analyses described earlier is that majority of the decrease in either the selection accuracy or in the bias with the HD SNP-chip was due to the poor performance of the QTL detection step with the HD chip (these results were not shown). When the SNP were identified using 50K SNP-chip data and HD was used only to build haplotypes, the performance of the genomic evaluation improved significantly. However, the analyses with the HD could not outperform those with the 50K.

Because the results of these first tests were not promising when compared to the 50K data results, this test was not continued with the other 10 multi-breed populations.

## 4.5 Genomic evaluation with causative mutations

### 4.5.1 Introduction

A possible way to improve the performance of genomic evaluation in regional breeds is the inclusion of candidate mutation information. These are specific SNP, which are likely to be either causative mutations underlying certain traits or in complete LD with such mutations; they were identified during the analysis of large dairy cattle breeds (Holstein, Normande and Montbéliarde in France). Since this information does not come from animals of regional breeds, there is uncertainty whether they can improve the performance of genomic evaluation in these breeds or not. This is because different QTL may be segregating in different breeds and QTL identified in one breed may not be present in other breeds (if the QTL is breed-specific) or it may not be segregating, if one of its alleles is fixed. Furthermore, when a QTL is present, its relative importance might be different in different breeds as well, depending on the genetic background (in particular, on the other QTL within the breed).

However, since the QTL detection power is much larger in the large breeds, QTL location could be narrowed down to a much smaller genomic region overlapping with a much lower number of putative mutations. Such fine resolution is currently not achievable in the regional breeds. In conclusion, no candidate mutations specific of regional breeds are currently available and any analysis using candidate mutation information in these breeds must rely on mutations indentified in larger breeds.

In this section, we aim to assess the possible gains with the inclusion of candidate mutation information in the regional breeds.

### 4.5.2 Materials and Methods

**Datasets**

Three of the four regional breeds were used for testing the impact of including potential mutations in their genomic evaluation. The Simmental breed was excluded, because only ~300 SNP from the LD SNP-chip could be imputed in this breed.

In case of the other breeds an enlarged reference population was used (i.e. the "February 2016" set from **Table 10**). These reference populations were ~40% larger than the ones used earlier. Most of the additional animals were females with own performance only. The number of additionally genotyped males ranged from 5 (in Vosgienne) to 44 (in Abondance).

The same SNP were used from the 50K data as used earlier in section 4.3, i.e. the 43.801 SNP that passed the quality control step. In addition, ~5,000 SNP unique to the LD SNP-chip were also available, from which approximately 3,000 were retained after removing the monomorphic SNP (**Table 14**). Most of the 5,000 SNP are candidate mutations linked to QTL affecting different recorded dairy cattle traits and they come from QTL detection studies conducted on the large dairy cattle breeds (Holstein, Montbéliarde and Normande). Because not all SNP are useful for all traits, it is important to identify – for each trait separately – which SNP should be used for prediction. In addition, some of the SNP from the LD-chip are linked to genetic disorders observed in some breeds and not to QTL affecting traits of interest. This data was also described in section 4.1.

**Table 14**: Number of imputed SNP and number of SNP retained from the LD SNP-chip after quality control.

| Breed | Number of SNP | |
|---|---|---|
| | **Imputed** | **Retained** |
| Abondance | 4,996 | 3,501 |
| Tarentaise | 4,764 | 2,976 |
| Vosgienne | 4,977 | 3,432 |

Phenotype data was used for the same 5 production traits as earlier: milk yield (MY), fat yield (FY), protein yield (PY), fat content (FC) and protein content (PC).

**Genomic evaluation methods**

The same implementation of the SNP-based BayesC approach was used as in sections 4.3 and 4.4. The value of π (the proportion of SNP without an effect on the analyzed trait) was fixed either to 80% or to 95%.

We also evaluated the BayesR procedure as implemented in the BESSiE software (Boerner and Tier, 2016) in addition to BayesC. Since the detection of large and medium sized QTL is the easiest (e.g. DGAT1, which gene has a major effect on fat content: Grisart et al., 2002), it is logical to assume that candidate mutations are either such QTL themselves or – more often – are linked such QTL. Hence, it seems advantageous to distinguish the different QTL based on their effect sizes, when including candidate mutation information. In contrast to BayesC, with BayesR the SNP can be divided into more than 2 groups, depending on their expected effect sizes (in practice, based on their associated variance). In our analyses, SNP were divided into 4 groups as indicated in **Table 15**. The proportions of the additive genetic variance explained by the SNP were identical to those used by Erbe et al. (2012), which values were regarded as standards. The proportions of SNP within each group were fixed, similarly to the value of π in the BayesC analysis. A total of 5% of the SNP was assumed to have an effect on the analyzed trait.

**Table 15**: Summary of the QTL groups used with BayesR.

| SNP group | Explained proportion of total $\sigma_a^2$ (%) | Proportion of SNP within the group (%) |
|---|---|---|
| No effect | 0 | 95 |
| Small effect | 0.01 | 4.49 |
| Medium effect | 0.1 | 0.485 |
| Large effect | 1 | 0.025 |

The underlying model used with both BayesC and BayesR is as follows:

$$y_i = \mu_{si} + p_i + \sum_{j=1}^{N} z_{ij} m_j \delta_j + e_i \quad (15)$$

where all parameters are as defined for equation 9.

In summary, the study compares different issues: 1) the effect of an increased reference population size (between 2015 and 2016) with the addition of mainly genotyped females with performances, 2) the effect of adding some putative mutations on the reliability of genomic evaluation, 3) the impact of using an *a priori* better method (BayesR) to account for the fact that putative mutations are expected to have a larger effect (i.e., to come from a distribution of effect with a larger variance).

### 4.5.3 Results and discussion

#### Correlation coefficient

**Table 16** presents the results obtained with the enlarged reference population and using only the 50K SNP-chip data while **Figure 11** show the observed gains with BayesC when candidate mutations were also included in the model (the same plot with BayesR are shown in **S. figure 6**).

Comparing the correlations in **Table 16** to the results obtained with the 2015 reference population (**Table 12**), we could observe an additional gain between 4.4% (Tarentaise) and 7.1% (Vosgienne). These gains were due to the genotyping of additional females and their inclusion in the reference population. Note that the value of π was also different: 95% here (**Table 16**) vs. 80% in 2015 (**Table 12**). However, more than 85% of the increase in correlations from 2015 to 2016 was observed with a π of 80% as well (data not shown).

BayesC outperformed BayesR in genomic evaluation, which was not expected as BayesR can differentiate QTL based on their effect sizes. This may be because no clear distinction could be done among the SNP with BayesR regarding their effect size: based on the output of the BESSiE software, every SNP had very similar

probabilities for being sorted in each of the 4 groups in which SNP were divided (i.e. the small, medium, large and “no effect” SNP groups). An alternative reason can be the improper choice of prior probabilities.

**Table 16**: Correlation coefficients obtained in the validation population with either BayesC or with BayesR (π=9%) using 50K SNP-chip information.

| Trait | BayesC[1] | | | BayesR[1] | | |
|---|---|---|---|---|---|---|
| | A | T | V | A | T | V |
| Milk yield | 0.344 | 0.432 | 0.401 | 0.301 | 0.436 | 0.386 |
| Fat yield | 0.339 | 0.446 | 0.352 | 0.284 | 0.432 | 0.321 |
| Protein yield | 0.257 | 0.439 | 0.451 | 0.196 | 0.414 | 0.438 |
| Fat content | 0.725 | 0.626 | 0.617 | 0.688 | 0.629 | 0.632 |
| Protein content | 0.654 | 0.508 | 0.689 | 0.602 | 0.447 | 0.698 |
| Average 2016 | 0.464 | 0.490 | 0.502 | 0.414 | 0.472 | 0.495 |
| Average 2015[2] | 0.417 | 0.446 | 0.431 | - | - | - |

1: Breed name abbreviations: A – Abondance; T – Tarentaise; V – Vosgienne

2: Results obtained with BayesC in 2015 (π=80%)

When comparing the effect of adding the candidate mutations to the genetic markers (**Figure 11** and **S. figure 6**), a small average gain (0.5% and 0.3% with BayesC and BayesR, respectively) was observed in the correlation coefficients. Larger gains were obtained for fat content (1-1.6% on average for the 3 breeds). Inclusion of the candidate mutations led to a moderate loss in selection accuracy only for protein yield and protein content with BayesR (maximum loss: -0.5% in Abondance). It is difficult to explain this loss of selection accuracy as the addition of a limited number of putative causative mutations is not expected to have a detrimental effect on the evaluation accuracy. Perhaps, the inclusion of many putative mutations not necessarily linked with the trait of interest led to an increased number of effects to be estimated (e.g. ~3,501 more SNP in Abondance), which may represent an extra noise responsible for the decrease of selection accuracy.

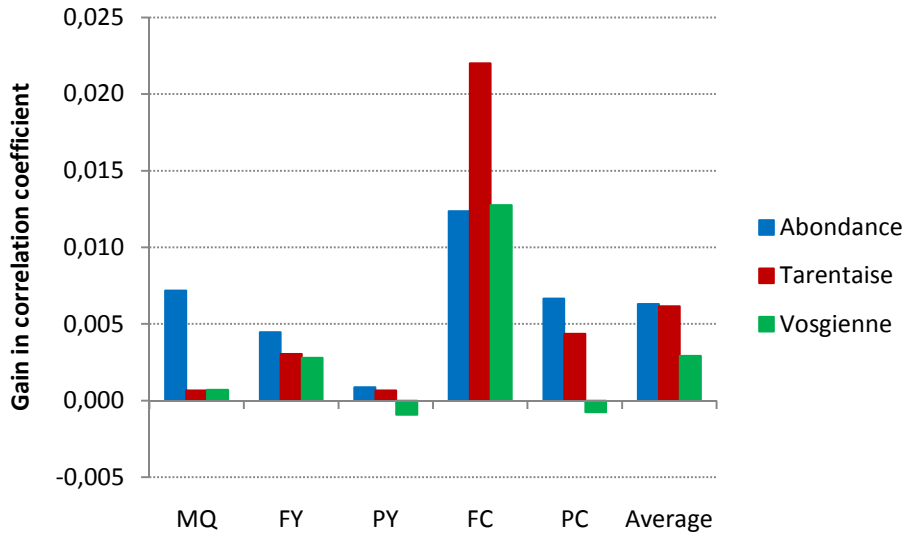



**Figure 11**: Effect of the inclusion of candidate mutations on the correlation between YD and GEBV measured on the validation population (BayesC).

**Regression slope**

The average absolute deviations of regression slopes from 1 are shown on **Figure 12**. Slopes were mainly below 1 in Abondance, but were always higher than 1 in Vosgienne. In Tarentaise, the regression slopes were slightly above 1 for the yield traits. Regression slopes exceeding 1 were usually higher with BayesR. The changes in the regression slopes with BayesC compared to those obtained in 2015 (**Table 12**) were slightly favorable in Abondance (average absolute deviation from 1: 0.168 in 2016 vs. 0.216 in 2016) and Tarentaise (average absolute deviation from 1: 0.085 in 2016 vs. 0.109 in 2015) but were disadvantageous in the Vosgienne breed with an average increase of 0.08 in the deviations of the regression slopes from 1.

Inclusion of the candidate mutations did not improve significantly the regression slopes with either BayesC or BayesR. An improvement was slightly more pronounced in the Vosgienne with BayesC and with Tarentaise with BayesR.

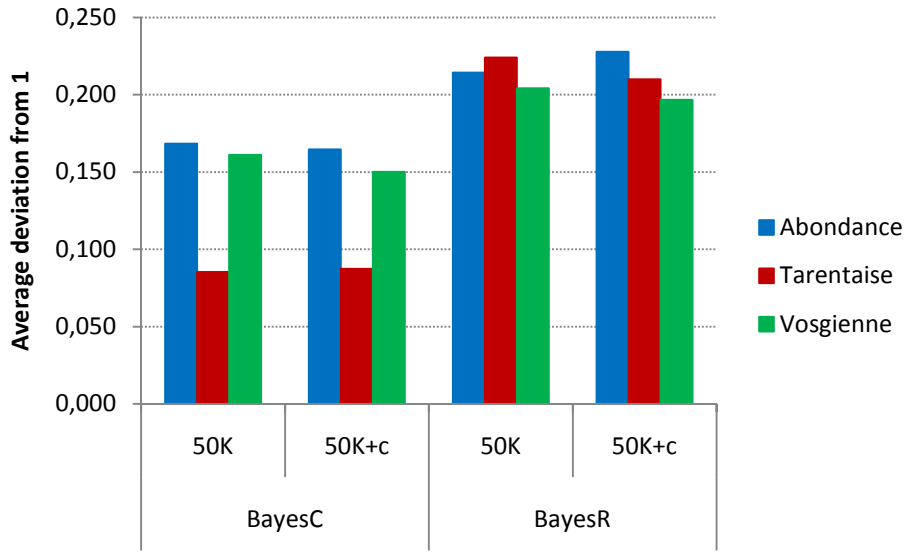



**Figure 12**: Average absolute deviation of regression slopes from 1 with either BayesC or BayesR and with the 50K and 50K+custom SNP-chip data.

### 4.5.4 Conclusions

Enlarging the reference population with additional females led to substantial (4-7%) increase in selection accuracy with BayesC while in two of the three breeds the regression slopes slightly improved as well. Therefore, extra genotyping of females can be expected to further improve the selection accuracy in the analyzed breeds.

Clear improvement of the selection accuracy by inclusion of candidate mutations was obtained only for fat content. With a BayesC procedure for the other traits, either only minor improvements were obtained (e.g. for fat yield) or no improvement at all (e.g. for protein yield). BayesR generally did not perform as good as BayesC, probably because the SNP effects could not be properly distributed into the different variance groups. This is frustrating, because only a proportion of the candidate mutations are expected to have a large effect, the others being likely without any effect as they were detected for other traits than the one being analyzed. Neither increasing the number of iterations by 10-fold nor allowing a variable π nor the combination of these two changes led to significantly different results from those presented here.

These observations are however not different from what was reported by Erbe et al. (2012) when they proposed the BayesR method: they analyzed the same 3 yield traits as in our study in a mixed Holstein-Jersey population. They compared the

performance of BayesR with BayesA and found only a very limited increase in selection accuracy and very similar regression slopes of DYD on GEBV.

Another version of the BayesR, called BayesRC was published recently (MacLeod et al., 2016). With this method, a set of SNP “enriched” in causative mutations can be created based on any prior information. Therefore, this method can be much more adequate to analyze the available candidate mutation information.

# Chapter 5

# General discussion

## 5.1 Introduction

Due to its economic advantages, genomic selection is more and more widespread in large dairy cattle breeds (e.g. García-Ruiz et al., 2016; Le Mézec et al., 2015). Genomic evaluation of animals assumes that information at DNA level is available on selection candidates as well as on a reference population (i.e. on genotyped animals with associated phenotype records). Since their development (2008 in bovine), SNP-chips are used to obtain DNA marker information. In several countries (e.g., France, Germany, Netherlands and USA), breeding organizations of different breeds genotyped a large number of progeny tested bulls with these SNP panels in order to obtain a large reference population. The larger the available reference population is, the better genomic selection performs. This puts regional breeds with limited total population size at a disadvantage compared to large (mainly international) breeds. In case of some breeds (e.g. the Brown Swiss) it is possible to create a large international reference population from the smaller national populations. However, this assumes that the breed is used in multiple countries, which is not the case in most of the regional breeds (e.g. Abondance, Tarentaise or Vosgienne).

Our main aim was to contribute to the development of a genomic evaluation procedure which can be efficient in regional dairy cattle breeds with a limited reference population. Moreover, a multi-breed reference population can be easily used to enlarge the reference populations of the regional breeds so we were also interested in the assessment of the performance of an evaluation based on such multi-breed reference population. This was appealing because the four regional breeds considered are closely related from an evolutionary perspective (Gautier et al., 2010; **Figure 3**), suggesting that multi-breed genomic evaluation might be beneficial for these breeds. Based on previous studies (Hozé et al., 2014; de Roos et al., 2008), it was hypothesized that the bovine high-density SNP panel would be required for multi-breed evaluations because the higher LD between the markers provided by this SNP-chip could capture the effects of the shared QTL.

## 5.2 **Biodiversity**

Biodiversity is essential in breeds and species of agricultural importance. About 50% of the total genetic variance within species used in agriculture can be found within breeds (Engels and Fassil, 2007). Therefore preserving the different breeds is important to maintain the genetic diversity in all species used in agriculture, including cattle. Moreover, the existence of genetic variability is a prerequisite for artificial selection: without genetic variance in the traits of interest, no breeding program can be efficient (e.g. see equation 11: if the genetic standard deviation ($\sigma_a$) is zero, the annual genetic gain is also zero). The preservation of across-breed genetic variation (which is the remaining 50% of the genetic variability) is equally important, especially to conserve the differences observed between the breeds, which is crucial for a sustainable agriculture. Therefore, the preservation of both within- and across-breed genetic variation is of great interest for the present and the future of agriculture.

Only a small number of bulls can be progeny tested within the regional breeds (**Table 1**), because increasing the number of bulls entering progeny testing would lead to an increased proportion of daughters coming from the progeny testing phase, i.e. from unproven bulls. However, the small number of proven bulls results only in a few number of selected proven bulls, which is detrimental for the genetic diversity of the breed. This is even more expressed if among the progeny tested bulls, only by

chance there is one with an extremely high estimated breeding value. As a consequence, such an excellent bull may have many more daughters but also sons, leading to a disproportionately large contribution to the next generation(s) and to an additional diminution in the genetic variability of the breed.

Genomic evaluation allows the simultaneous evaluation of many more selection candidates at a comparatively much lower cost. This can lead to a larger number of bulls selected for reproduction, while the annual genetic gain in the selected traits increases compared to the genetic gain observed with the breeding program including a progeny testing phase. The larger number of selected bulls will have a positive impact on the genetic diversity of the breed as well, contributing to an easier preservation of the breed.

Genomic evaluation has been implemented in the large dairy cattle breeds and the mentioned advantages have been observed. In addition to the economic advantages, the number of bull sires has increased in these breeds as well. This can have positive impact on the genetic gain: for example if an otherwise outstanding young bull has a strongly detrimental effect on one trait (e.g. fertility), breeders will not want to use it in breeding. However, with carefully planned matings, the bull might have a number of excellent male offspring, some without the detrimental characteristics. Such bulls can be then used by the farmers. The larger number of bull sires is a promising sign indicating that genetic diversity may decrease at a slower pace in these breeds as well (note that genetic diversity decreases in all populations when any form of selection is implemented). This is indirectly caused by the fact that not only sons of elite bulls are evaluated with genomic evaluation: bulls who previously would not have obtained a breeding value due to lack of sufficient progeny testing capacities can be evaluated and used in practice.

In addition to selection, the mating strategy also has an important role in management of genetic diversity. The larger number of selection candidates will give more room for population management decisions, for example to minimize the increase in inbreeding or perform assortative matings. This is a currently actively studied field of animal husbandry.

## 5.3 Effects of the slower genetic progress

The absence of genomic evaluation in the regional breeds would have led to indisputable economic disadvantages. The two most important drawbacks from an economical point of view are the high costs of progeny testing in any breed and the slower genetic progress in the regional breeds. These are disadvantageous both in the short and in the long term.

In the short term, either the presence or the absence of progeny testing is disadvantageous in the regional breeds compared to large breeds with genomic evaluation. If progeny testing is implemented in a breed, its high costs (compared to the costs of genomic selection) put the breeders in a difficult situation, because they have to remain competitive on a market they (partially) share with breeders of large breeds. Partly due to the smaller population size (especially the number of cows under performance recording) and partly due to the lower budget of breeding organizations devoted to regional breeds, progeny testing has also been limited by a lower number of progeny per bull. This has resulted in lower reliabilities compared to the reliabilities of either progeny tested or genomically evaluated bulls of large breeds.

In the long term, as soon as the difference between the genetic merit of regional and large breeds becomes too large, more and more breeders may want to switch from regional breeds to large (inter)national breeds, which could eventually lead to the disappearance of regional breeds. The French Bretonne Pie Noire breed is a good example of this negative trend: at the beginning of the 20$^{th}$ century, there were about 500,000 Bretonne Pie Noire cows in France, which decreased to about 15,000 by the middle of the 1970s (Colleau et al., 2002). In 1975 a conservation program was started to preserve the breed, which became the main focus of the population management by today. In parallel, although genetic improvement officially did not stop, the number of cows under performance recording continued to decrease to 125 by 2000 (Colleau et al., 2002), which prohibits any type of selection. Note that only a small proportion of the whole population is under performance recording. Although a slight improvement could be observed by the year 2014 (number of cows under performance recording: 199; Institut de l'Elevage, 2015b), it is still largely insufficient

for selection purposes. Furthermore, the number of farmers keeping animals of this breed was 270 in 2000 and presumably has further decreased since (Colleau et al., 2002).

Such trends are not only detrimental for the breeds, breeding organizations and regions themselves, but also for agriculture in a wider sense. Preservation of breeds is a crucial aspect of agro-ecology because neither future demands nor future production circumstances are known and therefore it is also unknown which breeds could produce efficiently in the future. In consequence, it is of great interest to maintain the biodiversity in agriculturally important animal species as well, in order to ensure that the indispensible genetic diversity will be preserved for the future.

Indeed, there are numerous initiatives to preserve and maintain biodiversity even in the agriculturally most important species and breeds. For instance, in 2005 in France, there were 132 different *in situ* conservation setups for livestock breeds, involving a huge variety of actors (Lauvie, 2011). Complementary to *in situ* programs, several genebanks conserve farm animal genetic resources (i.e. reproductive materials from both plants and animals) for the future. An example in case of plants is the European AEGIS initiative (http://www.ecpgr.cgiar.org/aegis/about-aegis/) and in case of animals, the EFABIS (EFABIS, 2016), both of which are organizations that coordinate multiple European genebanks (e.g. in France, the Cryobanque Nationale: http://www.cryobanque.org/index.php?lang=en; in Hungary, the Haszonállat Génmegőrzési Központ: http://genmegorzes.hu/). Moreover, there are European subsidies to farmers who keep breeds endangered to be lost for agriculture (e.g. in Hungary: Government of Hungary, 2015) as well as national and/or regional subsidies to organizations managing *in situ* conservation programs.

To support the preservation of regional breeds in dairy cattle breeding, the introduction of genomic selection in such breeds is seen as a great advantage.

## 5.4 Perspectives for the regional breeds

### Annual genetic gain

As it was discussed in section 2.6.1 of the General Introduction, genomic evaluation has a major impact on the annual genetic gain. A theoretical annual genetic gain can be calculated as shown in equation 11 and repeated below:

$$\Delta G = \frac{\left(i_{mf} * r_{IH,mf} + i_{mm} * r_{IH,mm} + i_{ff} * r_{IH,ff} + i_{fm} * r_{IH,fm}\right) * \sigma_a}{L_{mf} + L_{mm} + L_{fm} + L_{ff}} \qquad (11_c)$$

where ΔG is the annual genetic gain, $i_{..}$ is the selection intensity calculated for the four different paths, $r_{IH,..}$ is the selection accuracy calculated for the four paths, $\sigma_a$ is the standard deviation of the additive genetic effect of the trait (or composite breeding objective) under selection and $L_{..}$ is the generation intervals (expressed in years) again for the four paths. The distinction of the four paths is important, because the generation interval, selection intensity and accuracy change depending on whether males or females are selected and whether they are selected to create the next generation of bulls or cows.

The introduction of genomic evaluation should have similar impacts on the regional breeds as it had on the large breeds, although some of these are to a smaller extent. In case of males, the most important effect is the decrease in the generation intervals ($L_{mf}$ and $L_{mm}$), if progeny testing is discontinued. The accuracy of breeding values ($r_{IH,mf}$ and $r_{IH,mm}$) either do not change markedly (e.g., for lowly heritable traits, such as the fertility traits) or slightly increase (for moderately heritable traits, e.g. the production traits). Selection intensity ($i_{mf}$, $i_{mm}$) will also increase in males. In case of females, the accuracy of breeding values ($r_{IH,ff}$ and $r_{IH,fm}$) increases for lowly heritable traits, while the generation intervals of dams of cows ($L_{ff}$) is not expected to change markedly. Generation interval of dams of bulls ($L_{fm}$) can also decrease because genotyped heifers can be used now as bull dams while earlier, dams with 2 (or more) finished lactations were usually selected. Potentially, selection intensity of dams of cows ($i_{ff}$) can be expected to increase due to the combined effects of the availability of both more accurate breeding values on heifers and the use of sexed semen, consequently increasing the number of female selection candidates. Selection intensity of dams of bulls ($i_{fm}$) is likely to decrease slightly because more young bulls will be selected for breeding.

After the first year of availability of genomic evaluation in regional breeds, we can report the number of bull dams and the number of evaluated and selected male candidates planned by their breeding organizations (**Table 17**; S. Barbier, 2016, personal communication). The selection intensity of sires of bulls is expected to increase in all breeds. In Abondance ~30% of the selection candidates were retained for breeding with progeny testing, while with genomic evaluation this proportion decreases to ~20%. A larger decrease can be expected in Tarentaise (from 40% to ~15%) and in Vosgienne (from 40% to ~8%).

**Table 17**: Number of genotyped young candidates and selected bull sires and bull dams during the first year after the implementation of genomic selection in the regional breeds.

| Breed | Number of genotyped elite females[1] | Number of ~ male candidates | |
|---|---|---|---|
| | | Genotyped | Selected |
| Abondance | 200 | 150 | 20 |
| Tarentaise | 200 | 120 | 18 |
| Vosgienne | 160 | 50 | 4 |

1: Candidates to become dams of bulls

These changes in the regional breeds together with the estimated annual genetic gains are summarized in **Table 18** either with progeny testing or with genomic evaluation or with the mix of the two methods (i.e. when males to be progeny tested are retained based on their GEBV). For comparison purposes, the same parameters (with genomic evaluation) are also shown for a typical large breed. Note that all values presented in **Table 18** are rough estimates and serve only illustrative purposes. Furthermore, it is also assumed that sexed semen will be more widespread in all dairy breeds (including the regional ones), allowing an increase in the selection intensity on the “dams of cows” path. See Appendix E on page 218 for a detailed description of the calculations.

Based on the estimated values in **Table 18**, breeders can expect the annual genetic gain to increase by ~140% with the introduction of genomic selection, if they keep an organized progeny testing as well. Although this would lead to slightly higher genetic

gain *per generation* as a purely genomic evaluation selection scheme, the generation interval would be similar to that with progeny testing (apart from a small decrease on the "dams of bulls" and "sires of cows" path). If progeny testing is discontinued in the regional breeds, the genetic gain can further increase by ~28%, due to a large decrease in generation interval on the "sires of bulls" and "sires of cows" paths.

**Table 18**: Asymptotic annual genetic gain and different parameters affecting it in large breeds with genomic selection (GS) or in regional breeds with or without genomic selection (indicative values).

| Scenario | Path | SP (%) | i | $r_{IH}$ | L | $\Sigma i \cdot r_{IH}$ ($\sigma_a$) | $\Delta G$ ($\sigma_a$) |
|---|---|---|---|---|---|---|---|
| Large breeds with GS | Sires of bulls | 5 | 2.06 | 0.84 | 2.5 | 1.72 | |
| | Sires of cows | 5 | 2.06 | 0.84 | 2.5 | 1.72 | |
| | Dams of bulls | 5 | 2.06 | 0.84 | 2.5 | 1.72 | |
| | Dams of cows | 80 | 0.35 | 0.77 | 4 | 0.27 | |
| | Total | | | | 11.5 | 5.44 | 0.47 |
| Regional breeds with progeny testing | Sires of bulls | 40 | 0.97 | 0.71 | 7.5 | 0.69 | |
| | Sires of cows | 70 | 0.49 | 0.35 | 5.0 | 0.17 | |
| | Dams of bulls | 5 | 2.06 | 0.71 | 5.8 | 1.46 | |
| | Dams of cows | 100 | 0 | 0.59 | 5.2 | 0.00 | |
| | Total | | | | 23.5 | 2.31 | 0.10 |
| Regional breeds with GS (retaining progeny testing) | Sires of bulls | 4 | 2.15 | 0.84 | 7.5 | 1.80 | |
| | Sires of cows | 8 | 1.89 | 0.76 | 4.0 | 1.43 | |
| | Dams of bulls | 10 | 1.76 | 0.73 | 3.0 | 1.28 | |
| | Dams of cows | 90 | 0.2 | 0.73 | 5.2 | 0.15 | |
| | Total | | | | 19.7 | 4.65 | 0.24 |
| Regional breeds with GS | Sires of bulls | 10 | 1.76 | 0.73 | 2.5 | 1.28 | |
| | Sires of cows | 10 | 1.76 | 0.73 | 2.5 | 1.28 | |
| | Dams of bulls | 10 | 1.76 | 0.73 | 3 | 1.28 | |
| | Dams of cows | 90 | 0.2 | 0.73 | 5.2 | 0.15 | |
| | Total | | | | 13.2 | 3.99 | 0.30 |

Abbreviations: SP – selection proportion; i – selection intensity; $r_{IH}$ – selection accuracy; L – generation interval; $\Delta G$ – annual genetic gain

When compared to large dairy cattle breeds, the decreases in generation intervals can be expected to be similar in the regional breeds, because in both cases GEBV become available before maturity. By the time an animal can be used for breeding, it has a GEBV. Selection intensity (in case of males and females) is higher in large breeds, due to the larger number of candidates, while selection accuracy is also higher in the large breeds (in males and females alike) due to the availability of a larger reference population. Overall, the performance of genomic evaluation is expected to be significantly more efficient in the large breeds than in the regional ones.

However, the introduction of genomic selection is still a cardinal question in the regional breeds and must be implemented as quickly as possible. **Figure 13** illustrates the long-term effects of the existence or absence of genomic selection on the productivity of regional breeds compared to the productivity of large breeds. To create **Figure 13**, the estimated annual genetic gains from **Table 18** were used (for the large breeds without genomic evaluation, the $0.22\sigma_a$ estimate from Schaeffer (2006) was used). Before the genomic selection era, the annual genetic gain was already larger in the large breeds, partly due to the higher selection accuracy (achieved by a larger number of daughters under performance recording per bull) and partly due to a larger selection intensity (due to the larger number of bulls participating in progeny testing). As we entered the genomic evaluation era, the introduction of these modern evaluation methods in the large breeds doubled the annual genetic improvement of these breeds. As it was outlined earlier, genomic evaluation can have similar effects on the regional breeds, although to a lesser extent. In contrast, the absence of genomic selection in the regional breeds would result in a much more rapidly increasing gap between the genetic potential of large and regional breeds. This might have disastrous effects on both the regional breeds themselves and on agriculture in a broader sense, as discussed earlier.

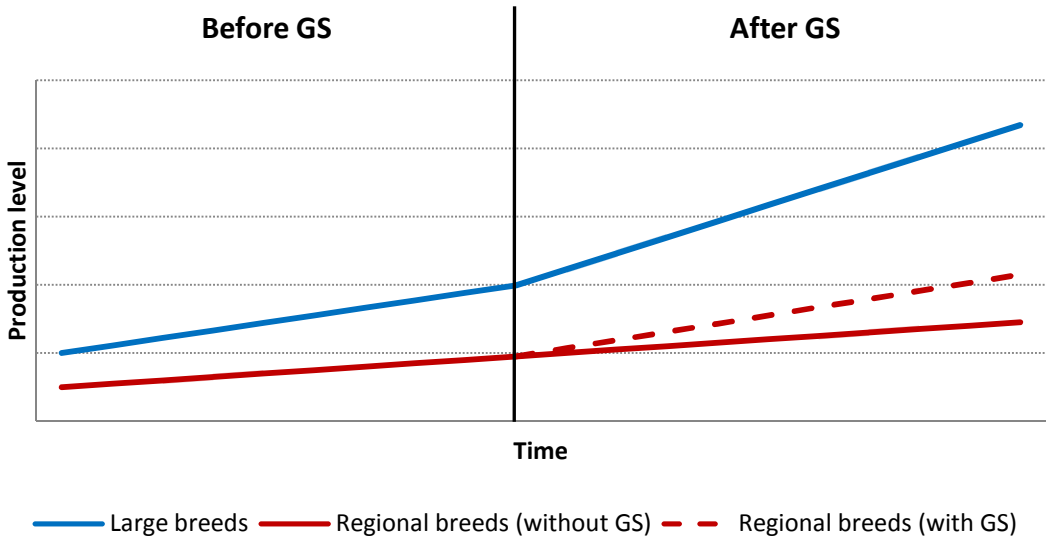



**Figure 13**: Illustration of the long-term effect of genomic selection (GS) on the production level of the regional and large breeds.

To prevent the difference between the genetic potential of these breeds from widening, efforts should be devoted to the improvement of the efficiency of genomic selection in the regional breeds. For this, there are several directions for future actions or research which are promising. These are:

- Increasing the reference population size: Breeding organizations should continue to genotype young heifers, which – by maximizing the available information for allele effect estimation – will contribute to a higher accuracy of the genomic evaluation in the regional breeds.
- Continue to study ways to implement an efficient multi-breed evaluation: The use of a multi-breed reference population was shown to be beneficial at least in Abondance.
- Find ways to benefit from larger breeds: Results of research work on large breeds, such as detection of putative candidate mutations can be transferred to the regional breeds to improve genomic evaluation.

## 5.5 Genomic evaluation in the regional breeds

The performance of different genomic evaluation methods were evaluated in four regional breeds. Our main aim was to assess the possible benefits and limits of

genomic evaluation in these breeds with particular interest in the haplotype-based methods: *haplotypic MA-BLUP* and *haplotypic GS3*.

A major challenge in the regional breeds is to create a reasonably large reference population that can be used for allele effect estimation because of the very limited number of progeny tested bulls in these breeds (**Table 1**). Furthermore, given the limited size of the population under performance recording in these breeds, progeny testing is also limited to approximately 25-30 female offspring per bull in order to obtain a reliability of (approximately) 50% for the production traits. This makes genomic evaluation in these breeds even more difficult. A possible way to enlarge the reference population is the genotyping of females. However, a female with her own performance only brings in less information than a progeny tested male (**Table 19**). Considering a 50% reliability level and the heritability of production traits (~0.3), the number of first lactation females representing the same amount of information as a single progeny tested bull is ~2.3. Multiple recordings on females improve their reliabilities based on performances and therefore older cows are more informative. This – at least in theory – might eventually lead to instances where females are more informative than males for selection purposes in the regional breeds.

**Table 19**: Number of females with one individual phenotype required to bring information equivalent to one male, according to heritability and male estimated breeding value (EBV) reliability based on progeny information only (Table 1 from Boichard et al., 2015).

| **Male EBV reliability** | **Heritability** | | | | |
|---|---|---|---|---|---|
| | **0.1** | **0.2** | **0.3** | **0.4** | **0.5** |
| 0.40 | 6.0 | 2.7 | 1.6 | 1.0 | 0.7 |
| 0.50 | 9.0 | 4.0 | 2.3 | 1.5 | 1.0 |
| 0.60 | 13.5 | 6.0 | 3.5 | 2.3 | 1.5 |
| 0.70 | 21.0 | 9.3 | 5.4 | 3.5 | 2.3 |
| 0.80 | 36.0 | 16.0 | 9.3 | 6.0 | 4.0 |
| 0.90 | 81.0 | 36.0 | 21.0 | 13.5 | 9.0 |

To maximize the reference population size and in turn the achievable gain from a genomic evaluation, breeding organizations (including those of the four regional breeds analyzed here) genotyped females with own performance(s) in addition to progeny tested bulls to form a mixed reference population. A mixed reference population is not a disadvantage in regional breeds, as it was demonstrated earlier that even in case of large breeds, the reference population will have to include females as the organized large-scale progeny testing of males has stopped or is likely going to stop (Boichard et al., 2015). The addition of a large number of cows to the reference population was shown to increase the prediction accuracies by ~4-5% while having no (or little) effect on the bias of the genomic breeding values (Kemper et al., 2015; S. Fritz, 2016, personal communication). We could also verify these results. However, in regional breeds the lack of a large number of animals with highly reliable performances (i.e. progeny tested bulls) is detrimental compared to the situation of the large breeds. Furthermore, the number of females that can be genotyped is also limited in these breeds. For example, in the Vosgienne breed, essentially all females under milk recording have been genotyped by 2016. Consequently, the two possible ways that remain to improve the efficiency of genomic evaluation in this breed are the improvements in genomic selection methods (e.g. exploiting genetic relationship information in a multi-trait analysis) and the opportunity of multi-breed genomic evaluations.

Although their genotypes were available, females were not included in the reference population in case of longevity due to the low heritability of the trait. For this trait the amount of information brought by all the genotyped females is only a fraction compared to the bulls (see **Table 19**). On the other hand, calculations (e.g. the number of record equivalents to be used for weighting them and to deregress them) become much more complicated with the inclusion of females and wrongly adjusted parameters could have detrimental effects on the final estimates.

The LD-decay pattern observed in the regional breeds were very similar to the LD-decay observed in the large dairy breeds in France. In early studies, an $r^2$ of 0.2 was often considered to be sufficient between adjacent markers for efficient genomic evaluations (de Roos et al., 2008: Calus et al., 2008). As pointed out by de Roos et

al. (2008), this was also the level of LD simulated by Meuwissen et al. (2001). This level of LD was observed in the regional breeds with an average marker distance of 52.5-72.5 Kb, depending on the breed (for Abondance and Taranteise it was slightly longer – 72.5 Kb – than for Simmental and Vosgienne – 52.5 Kb). As a comparison, the corresponding distance was 67.5 Kb for the Montbéliarde breed. In case of all of these breeds, the resolution of the 50K SNP-chip (average distance between adjacent markers: ~57,000 bp) can be predicted to be sufficient for an efficient genomic evaluation, given that there is a sufficiently large reference population available.

It is worth mentioning that the LD is measured between neighboring SNP and not between SNP and QTL. Indeed, QTL were assumed to be ungenotyped in all of the cited studies. As follows, QTL are expected to be located between the neighboring SNP and consequently, the distance between these QTL and the neighboring SNP can be predicted to be on average half of the average distance measured between the adjacent SNP. The LD corresponding to this distance (i.e. ~26.25-36.25 Kb) is approximately 30% in all breeds (including Montbéliarde). This is the LD that can be expected between SNP and (ungenotyped) QTL. This phenomenon could also (partially) explain why a D' threshold of 45% did perform better in our tests (as well as in Cuyabano et al., 2014) than a higher threshold when creating haploblocks.

**Single-breed evaluations**

In the following section, the performance of genomic evaluation methods applied to the regional breeds is discussed. It includes the application of the French routine evaluation on the 4 regional dairy cattle breeds. This evaluation incorporates part of the methodological improvements previously presented. Possible improvements, including the use of haploblock information, the use of HD SNP-chip and multi-breed tests are also reviewed. These studies can be divided into two parts, based on either the reference population (single-breed vs. multi-breed) or based on the SNP-chip density (50K vs. HD). Here, the division is based on the reference population, because the high-density SNP-chip was used only in the multi-breed context.

In the regional breeds, a BayesC model using 50K SNP-chip information resulted in higher selection accuracies (measured as the correlation between YD and (G)EBV in the validation population) compared to the performance of a pedigree-based BLUP model. Inflation of breeding values measured as the regression slope of YD on (G)EBV in the validation population was also closer to the optimal value of 1 with the BayesC analysis, except for the Abondance breed. The French routine genomic evaluation outperformed the BLUP tests in the regional breeds and showed a slight improvement compared to the BayesC model in most cases as well. Sanchez et al. (2016) also showed that the reliability of selection candidates were very close to the reliabilities of progeny tested bulls with a BLUP model. In some instances, the reliabilities of genomic evaluation (**Table 20**) even outperformed those of BLUP.

**Table 20**: Estimated reliabilities of selection candidates with the French routine evaluation (from Sanchez et al., 2016).

| Breed | Training population | | Trait group | | | |
|---|---|---|---|---|---|---|
| | Nr. of males | Nr. of females | Production | Somatic cell count | Fertility | Type traits |
| Abondance | 389 | 2769 | 54 | 51 | 40 | 51 |
| Tarentaise | 323 | 1569 | 52 | 48 | 34 | 49 |
| Vosgienne | 66 | 1171 | 54 | 45 | 33 | 49 |

As a consequence of the results obtained with the regional breeds, genomic evaluation was officially implemented in 2016 in Abondance, Tarentaise and Vosgienne. It is also implemented in Simmental, but in the framework of an international collaboration with Germany and Austria, a much larger reference population exists for this breed in Germany with a higher accuracy and lower bias than the ones obtained in France. As a result, the French Simmental breed association is currently relying on the German genomic evaluation. However, this is not optimal since French phenotypes are not included in the German evaluation. This situation may change in the future if a sufficiently large number of French cows are genotyped. Then the French Simmental breed may be officially added to the list of regional breeds with French genomic evaluation.

Genomic evaluation in the other three regional breeds is efficient and it also enables breeders to select for traits on which selection was not possible earlier. For example due to the low reliabilities of certain traits (e.g. the fertility traits) with BLUP, the breeding values of these traits were until now not published for females and in case of bulls, they were available with a sufficient accuracy only late in the bulls' life. This hindered selection on these traits. With genomic evaluation, the reliabilities of these traits slightly increased compared to progeny tested bulls and are equally high for both males (with or without progeny) and females, which now allows some selection on these traits.

Hayes et al. (2009) observed a positive correlation between the effective population size and the number of haplotypes: smaller effective population sizes lead to fewer and longer independent chromosome segments. We could observe the same trend: there were fewer haploblocks identified in Abondance (7,294), Tarentaise (6,485) and Vosgienne (8,296) than in Montbéliarde or in Simmental (8,393 and 9,918, respectively). This is partly due to the smaller effective population size of these regional breeds (51, 67 and 57 for Abondance, Tarentaise and Vosgienne, respectively according to Institut de l'Elevage (2015c). Simmental had more haploblocks than Montbéliarde which is also in accordance with the higher effective population size of this breed (73 vs. 141; Institut de l'Elevage, 2015c). A lower number of haploblocks also means that there are fewer effects that need to be estimated in a genomic evaluation study. However, in contrast with the Montbéliarde situation, the analysis using haploblock information in combination with haplotype selection did not improve the correlation coefficients nor the regression slopes of (D)YD on GEBV in the validation study for regional breeds (results not shown). This may be due to the much larger number of haplotype effects to estimate when haploblock information was used (~7,000-9,000, depending on the breed) compared to the number of haplotypes used in either the BayesC analysis or in the routine evaluation ($n_{hap} = 1{,}000$): when haploblock information was used, the number of haplotypes is not *a priori* determined and because of this, all haplotypes are used in the model, resulting in approximately 7-36 times more haplotype allele effects to estimate.

**Multi-breed evaluations**

Multi-breed genomic evaluations are expected to outperform their single-breed counterparts if the analyzed breeds are closely related, because in such a case they can be expected to share a larger proportion of QTL than when they diverged earlier during their evolution. The genetic distances between 47 cattle populations – including all the 5 breeds analyzed here and 2 Holstein and Jersey populations – was estimated by Gautier et al. (2010; also see **Figure 3**). The genetic distance between the three regional breeds and Montbéliarde was found to be much shorter than the distance between e.g., the Jersey and Holstein breeds, which are the most frequently studied breeds in a multi-breed context.

In our study, a multi-breed genomic evaluation was advantageous in Abondance and Simmental, but it was detrimental in Tarentaise and Vosgienne. The gains in accuracy in Abondance and Simmental were moderate (+5-8% at maximum), while the loss for the other two breeds were somewhat smaller (from <1% to 4%). There were no general trends regarding the traits (e.g. systematic decrease/increase with either the yield or the content traits, etc.) in either Tarentaise or Vosgienne. The loss in accuracy in these breeds was unexpected, again partly because these breeds were more closely related than Holstein and Jersey (for these breeds, other authors (Hayes et al., 2009b; Erbe et al., 2012) obtained a gain in accuracy) and partly because our reference populations were not smaller than those used in these other studies. In Abondance and Simmental, we obtained intermediate gains in accuracy compared to those published earlier (Hayes et al., 2009; Erbe et al., 2012; Zhou et al., 2014a). A contributing factor to the mainly higher gains observed with the Holstein-Jersey population can be the composition of the reference population, which included only progeny tested bulls (with a larger average number of daughters per bull) for the Holstein-Jersey tests, but consisted mainly of females in our case.

Following the analyses with the 50K SNP panel, we conducted a multi-breed analysis with the HD SNP-chip. This multi-breed analysis used a training population consisting of animals from all the four breeds. The use of the HD SNP-chip in a multi-breed context was of interest, because of its higher marker density. The HD SNP-chip was

beneficial only in the Abondance and Simmental breeds, leading to a 4-5% increase in selection accuracy compared to the 50K within-breed tests. However, compared to the 50K multi-breed tests, a 2% decrease was observed in accuracy in both of these breeds. Results were less biased in Abondance, but did not change in that respect in Simmental.

The genotyping of additional females was clearly beneficial, leading to an extra 4-7% increase in selection accuracy and in case of Abondance and Tarentaise, a significant decrease in bias. The interest of including candidate mutation information (identified in other breeds than the ones analyzed here) in the evaluation process was dubious: for certain traits it was beneficial, while for others it did not improve selection accuracy. Further research is required before this type of information can be exploited with the regional breeds.

## 5.6 Financial considerations

Until now the benefits of genomic evaluation in the regional breeds was discussed from a technical point of view. However, these benefits will occur only if genomic selection is used in practical animal breeding, which depends first and foremost on the breeders. Therefore it is essential to assure that breeders start using the results of genomic evaluation (i.e. the GEBV of young animals) and that they are encouraged to do so. According to S. Barbier (2016, personal communication for this whole section), to ensure that GEBV are used, the breeding organizations of the Abondance and Tarentaise breeds disseminate the semen of young bulls with GEBV values as if they were young selection candidates participating in progeny testing, that is without reporting detailed GEBV of the bulls. Breeders of these breeds receive these GEBV of the bulls only 20 months later (i.e., before the first mating of young heifers), allowing them to select the appropriate bulls for the heifers (planned matings). The objective of these breeding organizations is that 50% of the semen used for insemination in 2016 comes from young bulls with GEBV and to increase this proportion to 70% in the future. In the particular case of the Vosgienne breed, the breeding organization finances (with the help of regional subsidies) the genotyping of all heifers in performance recording herds, in order to make sure that all recorded animals will enter the reference population. Breeders are not required to contribute to

the costs. However, if they wish to receive GEBV, they will be asked to partly contribute to the costs of the establishment of the reference population. At this point, it has to be taken into account that breeders have their income from the (partially) open market, where they have to be competitive even on the short term. In the short term, breeders are mostly interested in maximizing their profits and therefore are most interested in either an increase in revenues or a decrease in costs (or optimally both at the same time). Although genomic evaluation leads to substantial savings in breeding schemes of breeds where it has been implemented, these savings mainly occur at the level of AI companies. These may not decrease their semen prices. In addition, in the large breeds, a substantial part of the realized savings was re-invested in further genotyping of males and of females. Therefore, in order to persuade breeders of regional breeds (especially the Vosgienne) to use the genomic breeding values in practice, the promise of long-term gains is insufficient and they may need to be convinced by certain economic advantages in the short term (e.g. under the form of subsidies or reduced prices).

Concerning the competitiveness of these breeds, their markets are protected to a certain degree, because part of it is very specific: for example, there are certain high level dairy products that *require* to be made from the milk of specific, regional breeds and the use of the milk of other breeds is strictly prohibited. This is the case for example of the Beaufort cheese, which can be made only from milk of Tarentaise or Abondance cows (http://www.fromage-beaufort.com/fr/index.aspx).

There might be also hesitation from breeders in the use of young bulls without progeny due to distrust towards new technological improvements or towards a slightly lower reliability of genomic evaluations (compared with evaluation based on actual daughter phenotypes). Today breeders trust the progeny testing system as it was implemented for several decades and resulted in reliable breeding value estimates for bulls. However, from the calculations in **Table 18**, it is clear that a maximum annual genetic gain requires to move as quickly as possible to a 100% use of semen from young bulls. Convincing breeders (or even the breeding organizations) to abandon progeny testing and instead use breeding animals with potentially less reliable breeding values can be very challenging. Indeed, this issue was experienced

previously in case of breeders of large breeds, with an intermediate period when many breeders hesitated in using young bulls because of their lower reliability. But their much higher average genetic merit finally convinced more and more of them. In case of farmers keeping regional breeds, this distrust can decrease as they will see the larger annual gains that are already manifesting in the large dairy breeds today and by the opportunity to select from a wider range of bulls. Nonetheless, proper trainings and dissemination of knowledge would be useful to tackle this issue.

## 5.7 Genomic evaluation with haplotypes

During the course of my PhD, we also proposed several methodological developments to improve the efficiency of genomic evaluation methods. These methods were then implemented and their performance assessed. Our primary focus was on haplotypes, their efficiency and the way it can be improved.

The combined use of haplotype markers and the HD chip is not straightforward, because their simultaneous use increases the number of allele effects to estimate to several millions, which is far beyond the capabilities of the available genomic evaluation procedures, given the limited reference population sizes. Therefore, the number of markers to be used from the HD chip has to be reduced prior to genomic evaluation. In our first methodological study, we addressed this issue. We also demonstrated the usefulness of haplotype markers in genomic evaluation.

We provided an empirical proof of the superiority of haplotypes over SNP in improving the performances of genomic evaluation. Both the correlation coefficient between the estimated breeding values and the observed performances (expressed as DYD) and the observed regression slopes of DYD on GEBV (which is expected to be close to 1 to avoid bias) in a validation population were improved with the use of haplotypes. We could also demonstrate that haplotype selection based on allele frequency information is beneficial. Such methods are relatively easy to implement and are computationally not too demanding. These properties make haplotype selection an attractive choice to improve the efficiency of genomic evaluation. Furthermore, a version of the proposed methods allows the implementation of haplotype selection *prior* to genomic evaluation at no additional costs. In this version,

information on the LD-pattern along the chromosomes is also taken into account in the haplotype selection process in addition to the minor allele frequency data. This helped minimizing the number of haplotypes to be used in genomic evaluation. When all of these haploblocks are used in the genomic evaluation process, all of them contribute to the final GEBV and all QTL are therefore necessarily "represented" by proxies. This is a major advantage compared to the model where only the largest QTL are included based on a prior QTL detection study. It also allows the use of the same haplotypes for all traits which makes practical implementation easier. Applying this method in the Montbéliarde breed led to improvements both in selection accuracy and in regression slopes similar in absolute values to those observed with the haplotype selection criteria.

The haplotype selection methods developed also allowed a large reduction in the number of allele effects to be estimated in the model. This was necessary for the combined use of the HD chip and haplotype markers. This reduction – when using fixed windows of 10 SNP (or 144 SNP in case of the HD) – was 60% with the 50K SNP-chip (97% with the HD chip) compared to a scenario in which all consecutive haplotypes of 4 SNP are built. When the developed criteria were used in combination with haploblock information, the reduction was somewhat smaller: on average ~26% with the 50K and ~90% with the HD data. Nevertheless, these reductions (especially in case of the HD data) were promising.

The accurate estimation of the numerous bi-allelic markers available from the HD SNP-chip is difficult for the current evaluation methods in most of the breeds. However, the rapid improvement of biotechnology has led to large scale whole-genome re-sequencing projects, which – combined with imputation – allows for the determination and prediction of tens of millions of SNP markers at a reasonable price for a large number of individuals (Daetwyler et al., 2014; Boussaha et al., 2016). Within the framework of the *1000 bull genomes project*, Boussaha et al. (2016) identified approximately 28 million SNP on the bovine genome. The most important advantage of such a dataset is that it *implicitly* includes *all* causative mutations (excluding – at least directly – those that are due to structural variations). However, its analysis is not feasible with the genomic evaluation methods available today. This

was most recently demonstrated by several presentations at the most recent (67th) *Annual Meeting of the European Federation of Animal Science* (e.g. Erbe et al., 2016). Indeed, no large improvements were obtained in terms of selection accuracy when using such a dataset and it was concluded that pre-selection of markers is essential when using whole-genome sequence (WGS) data in genomic evaluation studies (van den Berg et al., 2016). It is worth mentioning that the imputation of rare alleles is difficult when the number of re-sequenced bulls of the breed of interest is limited (Bouwman and Veerkamp, 2014), which further complicates the use of WGS data for genomic selection purposes.

The concept of haploblocks was first published by Knürr et al. (2013) and it led to slight improvements in reliabilities. Cuyabano et al. (2015) published more promising results: they showed that the use of LD-based haploblocks as predictors instead of individual SNP is beneficial when using HD SNP-chip data in dairy cattle. We demonstrated that the combined use of such haploblocks with haplotype selection methods based on allele frequency information can outperform individual SNP as genetic markers as well. These methods are therefore promising to decrease the number of effects to be estimated when analyzing either high-density or WGS data.

Use of HD SNP-chip was unsatisfactory in a single-breed context using the largest breed included in this study. This is not in accordance with the results of Cuyabano et al. (2015), who could show an improvement with the HD SNP-chip compared to the 50K SNP-chip when using haploblock information. However, this discrepancy may be related to the fact that Cuyabano et al. (2015) had ~30% fewer SNP for the analysis after editing (492,057 vs. 706,791 in our study) and more than twice as many progeny-tested bulls (5,214 vs. 2,235).

## 5.8 **Future perspectives**

The advent of whole-genome sequencing started only a few years ago. In the near future, more and more animals are expected to be imputed with better accuracy to tens of millions of SNP and the available genotype data may be of the same order of magnitude as the number of phenotype observations. In parallel, more effort will be devoted to the analysis of WGS data to successfully exploit it for genomic evaluation

purposes. Reduction of the number of SNP prior to any analysis will be unavoidable and any efficient method to do so will be of great relevance. The haplotype selection method developed here can be a good candidate for this as it relies on simple statistical assumptions and does not require additional information (i.e., other than the genotypes).

In addition, the 50K SNP-chip will continue to be used or new "custom" 50K chips can be assembled either from the high-density SNP-chip or from WGS results in order to further improve the performance of genomic evaluations. Haplotype selection/construction can play an important role in the exploitation of these new panels as well.

The French routine genomic evaluation applied to the regional breeds gave appealing results. However, most of the additional tests we implemented to improve the performance of the routine analysis in these breeds either improved it only slightly (e.g. use of causative mutations) or the improvement was breed-dependent (e.g. the use of HD data or multi-breed training populations). The following changes might improve the performance of the routine evaluations:

**Inclusion of causative mutations**

Probably the most promising improvement is the inclusion of information on candidate mutations in the evaluation. The main reason for this assumption is that these mutations were often identified as potential candidate mutations for the same traits, but in other breeds. Therefore there is strong prior information that these SNP might be causative mutations in the regional breeds as well, when they segregate in such breeds. In our analyses, we could not completely exploit this information and therefore further research should address this question. BayesRC is a promising method, because this approach can incorporate the strong prior information that some SNP are present in a functional part of the genome and therefore are more likely to be causative mutations (MacLeod et al., 2016).

**Subsets of high-density data**

Instead of using all SNP from the HD chip, using only a subset of them can reduce the number of effects to be estimated in genomic evaluation. A subset can be created by, for example, excluding the SNP that are in very high LD with neighboring SNP or the SNP, which are far away from genes or regulatory regions. The average distance between neighboring SNP from the 50K is ~3,500 bp, while $r^2$ was on average 64% for SNP with <5,000 bp between them (**Figure 8**; ~61% in a multi-breed case: **S. figure 5**). This suggests that there is room to decrease the number of SNP without risking a diminishing selection performance, since de Roos et al. (2008) recommended 20% or Cuyabano et al. (2014) used 45%. For example the creation of a "transcriptome set" (i.e. the set of SNP located either on genes or +/- 1Kb from genes) was shown to improve the efficiency of multi-breed genomic evaluation (Erbe et al., 2012). In their study the "transcriptome" panel included ~58,500 SNP and it increased the selection accuracy compared to the 50K and measured in the smaller breed (Jersey) for milk yield (+12%) and protein yield (+10%). However, the selection accuracy diminished for fat yield (-5%).

**Exploit LD-phase information**

Inclusion of LD-phase information (e.g. as it was done by Porto-Neto et al., 2015) can be a step towards distinguishing common and breed-specific QTL. Porto-Neto et al. (2015) identified the SNP that had similar effects in two different breeds based on a within-breed analysis and considered them as SNP linked to common QTL. The published results are promising. If the breed-specific and common SNP can be accurately distinguished, the breed-specific QTL effects could be estimated independently from the other breeds. This can significantly contribute to an accurate allele effect estimation of breed-specific QTL in an otherwise multi-breed analysis. However, the accurate distinction of breed-specific QTL from the shared QTL is difficult in large breeds and even more difficult in regional ones.

**Combine haploblock information and HD data**

The combined use of haploblocks and the HD SNP-chip data in a multi-breed context was unfortunately not possible due to the too large number of SNP within

haploblocks (up to 542). This was beyond the capacity of the available software, but it is still a promising direction for future research. However, there were ~26,000 haploblocks created in total (with ~275,000 allele effects) when the 4 regional breeds were included together in the dataset, which might be disproportionately large compared to the number of available phenotypes.

In conclusion, there are still promising opportunities to improve the performance of genomic evaluation methods in the regional breeds. Eventually, these improvements might further decrease the differences between the genetic potential of large and regional breeds.

# Chapter 6

# Concluding remarks

Genetic improvement of livestock helps to increase the production level of breeds, the adaptation of the breeds to farming systems as well as to the ever changing production environments. It is an important component to improve the cow productivity in all aspects which significantly contributes to the competitiveness of the farmers on an open market.

A revolutionary change has occurred in the past decade, which culminated with the introduction of genomic evaluation in the largest dairy cattle breeds in multiple countries. The lack of genomic evaluation in the remaining (mainly regional) breeds put these breeds into a difficult situation with weaknesses that cannot be avoided using traditional selection. During this PhD work, we addressed the increasing demand of breeding organizations of such breeds for a genomic evaluation method that is efficient in small breeds.

We chose to use haplotype-based genomic evaluation methods to address this question, because the linkage disequilibrium between the (usually unknown) causative mutations and the haplotype markers is expected to be higher than the linkage disequilibrium with individual SNP. We could provide empirical proof to support this claim. In several independent analyses, we found that a haplotype size

of 4 SNP was performing best. The first observations made on a large breed led us to the conclusion that using all haplotypes in a regional breed with a very limited reference population would be most likely inefficient due to over-parameterization. Therefore, we had to decrease the number of haplotype markers used in the models. We showed that this can be efficiently done in large breeds by either selecting markers based on a prior QTL detection analysis or by exploiting information on the linkage disequilibrium pattern along the genome.

We developed a methodology for haplotype selection relying on haplotype allele frequency information which outperformed the haplotypes built from flanking markers in genomic evaluation. Using this approach, we could also confirm that statistical parameters, such as the haplotype allele frequencies or the linkage disequilibrium can be used to pre-select haplotypes for genomic evaluation purposes and that the selected haplotypes can improve the efficiency of genomic evaluation. The number of haplotypes could be greatly reduced along the genome as well. However, the combined use of both selection criteria (i.e. allele frequency and linkage disequilibrium) required a large reference population. Given these results, the haplotype selection method based on haplotype allele frequency information was incorporated into the French routine genomic evaluation in April, 2015.

We evaluated the performance of the routine French genomic evaluation in four regional breeds and found that genomic evaluation was efficient in these breeds. As a consequence, genomic evaluation was officially implemented in three of the four breeds (namely, Abondance, Tarentaise and Vosgienne) in 2016. Genomic evaluation can be predicted to have a large and positive impact on the realized annual genetic gain (increasing it by 3-fold, compared to the annual genetic gain obtained with progeny testing in these breeds) and on their genetic variability as well. However, none of the benefits will be existent if farmers do not use young bulls with genomic breeding values in practice. Therefore it is fundamental that farmers are encouraged to use young bulls based on their GEBV. Furthermore, trainings should be also organized for farmers to ensure that their information about genomic evaluation is up to date and to create a forum where their questions can be addressed.

Genotyping of cows and young heifers will likely continue in all regional breeds in which genomic evaluation was implemented. Breeding organizations of these breeds receive funding mainly from the regional governments as incentives to preserve and improve them, not only to maintain biodiversity in livestock species but also because they are important parts of the economy of their regions of origin. Furthermore, research programs aiming at further improving the performance of genomic evaluations in breeds with a reference population of limited size should continue. These works could include among others, multi-breed genomic evaluation studies or the use of candidate mutations to enhance the performance of genomic evaluations. They can contribute to an increased efficiency of genomic evaluation in regional breeds in the future.

Through this work, we demonstrated that genomic evaluation is efficient in four French regional breeds and that there are opportunities for further development of genomic selection in these breeds. Maintenance of regional breeds is essential both for agriculture and for the society and in this context, the introduction of genomic evaluation will play a significant role. The apparently fast practical implementation of genomic selection since the first genomic evaluation is a good sign for the future. In the longer term, the continuation of an efficient genomic selection will continue to require the collaboration of farmers, breeding organizations, scientists and representatives of the (regional) governments.

# References

Alexandratos, N. and Bruinsma, J. 2012. World agriculture towards 2030/2050: the 2012 revision. ESA Working paper No. 12-03. Rome, FAO.

Baird, S. J. E. 2015. Exploring linkage disequilibrium. Mol. Ecol. Resourc. 15(5): 1017-1019.

Beissinger, T. M., Rosa, G. J. M., Kaeppler, S. M., Gianola, D. and de Leon, N. 2015. Defining window-boundaries for genomic analyses using smoothing spline techniques. Genet. Sel. Evol. 47: 30.

Bennewitz, J., Reinsch, N., Szyda, J., Reinhardt, F., Kühn, C., Schwerin, M., Erhardt, G., Weimann, C. and Kalm, E. 2003. Marker assisted selection in German Holstein dairy cattle breeding: Outline of the program and marker assisted breeding value estimation. Page 5 in Book of abstracts of the 54th annual meeting of the European Federation of Animal Science. Rome, Italy.

Bentley, D. R. 2006. Whole-genome re-sequencing. Curr. Opin. Genet. Dev. 16(6): 545-552.

Boerner, V. and Tier, B. 2016. BESSiE: A program for best linear unbiased prediction and Bayesian analysis of linear mixed models. BESSiE version 1.0 software documentation. Downloaded 1st of March, 2016.

Boichard, D., Ducrocq, V. and Fritz, S. 2015. Sustainable dairy cattle selection in the genomic era. J. Anim. Breed. Genet. 132(2): 135-143.

Boichard, D., Fritz, S., Rossignol, M. N., Boscher, M. Y., Malafosse, A. and Colleau, J. J. 2002. Implementation of marker-assisted selection in French dairy cattle. Communication no. 22-03. In: Proc. of the 7th WCGALP. Montpellier, France.

Boichard, D., Chung, H., Dassonneville, R., David, X., Eggen, A., Fritz, S., Gietzen, K. J., Hayes, B. J., Lawley, C. T., Sonstegard, T. S., Van Tassell, C. P., VanRaden, P. M., Viaud-Martinez, K. A., Wiggans, G. R. and for the Bovine LD Consortium.

2012a. Design of a Bovine Low-Density SNP Array Optimized for Imputation. PLoS ONE, 7(3), e34130.

Boichard, D., Guillaume, F., Baur, A., Croiseau, P., Rossignol, M. N., Boscher, M. Y., Druet, T., Genestout, L., Colleau, J. J., Journaux, L., Ducrocq, V. and Fritz, S. 2012b. Genomic selection in French dairy cattle. Anim. Prod. Sci. 52: 115-120. http://dx.doi.org/10.1071/AN11119.

Boussaha, M., Michot, P., Letaief, R., Hozé, C., Fritz, S., Grohs, C., Esquerré, D., Duchesne, A., Philippe, R., Blanquet, V., Phocas, F., Floriot, S., Rocha, D., Klopp, C., Capitan, A. and Boichard, D. 2016. Construction of a large collection of small genome variations in French dairy and beef breeds using whole genome sequences. Genet. Sel Evol. 48(1): 87.

Bouwman, A. C. and Veerkamp, R. F. 2014. Consequences of splitting whole-genome sequencing effort over multiple breeds on imputation accuracy. BMC Genet. 15: 105.

Brøndum, R. F., Rius-Vilarrasa, E., Strandén, I., Su, G., Guldbrandtsen, B., Fikse, W. F. and Kund, M. S. 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. and Browning, B. L. 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(5): 1084-1097.

Calus, M. P. L., Meuwissen, T. H. E., de Roos, A. P. W. and Veerkamp, R. F. 2008. Accuracy of genomic selection using different methods to define haplotypes. Genetics. 178(1): 553-561.

Chen, L., Li, C., Miller, S. and Schenkel, F. 2014. Multi-population genomic prediction using a multi-task Bayesian learning model. BMC. Genet. 15: 53.

Colleau, J. J., Quéméré, P., Larroque, H. Sergent, J. and Wagner, C. Gestion génétique de la race bovine Bretonne Pie-Noire : bilan et perspectives. INRA Prod. Anim. 15(3): 221-230.

Coop, G., Wen, X., Ober, C., Pritchard, J. K. and Przeworski, M. 2008. High-resolution mapping of crossoversreveals extensive variation in fine-scale recombination patterns among humans. Science. 319(5868): 1395-1398.

Croiseau, P., Baur, A., Jónás, D., Hozé, C., Promp, J., Boichard, D., Fritz, S. and Ducrocq V. 2015a. A new Marker-Assisted BLUP genomic evaluation for French dairy breeds. Interbull Interbull meeting, Orlando, FL, USA.

Croiseau, P., Baur, A., Jónás, D., Hozé, C., Promp, J., Boichard, D., Fritz, S. and Ducrocq, V. 2015b. Comparison of different marker-assisted BLUP models for a new French genomic evaluation. Page 248 in Book of abstracts of the 66th annual meeting of the European Federation of Animal Science. Warsaw University of Life Sciences, Poland.

Croiseau, P., Fouilloux, M-N., Jónás, D., Fritz, S., Baur, A., Ducrocq, V., Phocas, F., and Boichard, D. 2014. Extension to haplotypes of genomic evaluation algorithms. AB#708 in Proc. 10th World Congress of Genetics Applied to Livestock Production. Vancouver, Canada.

Cuyabano, B. C. D., Su, G. and Lund M. S. 2014. Genomic prediction of genetic merit using LD-based haplotypes in the Nordic Holstein population. BMC Genomics 15: 1171.

Cuyabano, B. C., Su, G. and Lund,M. 2015. Selection of haplotype variables from a high-density marker map for genomic evaluation. Genet. Sel. Evol. 47:61.

Daetwyler, H. D., Capitan, A., Pausch, H., Stothard, P., van Binsbergen, R., Brøndum, R. F., Liao, X., Djari, A., Rodriguez, S. C., Grohs, C., Esquerré, D., Bouchez, O., Rossignol, M-N., Klopp, C., Rocha, D., Fritz, S., Eggen, A., Bowman, P. J., Coote, D., Chamberlain, A. J., Anderson, C., VanTassell, C. P., Hulsegge, I., Goddard, M. E., Guldbrandtsen, B., Lund, M. S., Veerkamp, R. F., Boichard, D. A., Fries, R. and Hayes, B. J. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat. Genet. 46(8): 858-865.

Devlin, B. and Risch, N. 1995. A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 29: 311-322.

de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K. and Cotes, J. M. 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182(1):375-385.

de Roos, A. P. W., Hayes, B. J., Spelman, R. J. and Goddard, M. E. 2008. Linkage disequilibrium and persistence of phase in holstein-friesian, jersey and angus cattle. Genetics 179(3): 1503-1512.

de Roos, A. P. W., Hayes, B. J. and M. E. Goddard. 2009a. Reliability of genomic predictions across multiple populations. Genetics. 183(4): 1545-1553.

de Roos, A. P. W., Schrooten, C., Mullaart, E., van der Beek, S., de Jong, G. and Voskamp, W. 2009b. Genomic selection at CRV. Interbull Bull. 39: 47-50.

Ducrocq, V., Fritz, S., Guillaume, F. and Boichard, D. 2009. French report on the use of genomic evaluation. Page 17 in Proc. of the Interbull technical workshop. Uppsala, Sweden.

Druet, T. and Georges, M. 2009. A hidden markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait loci fine mapping. Genetics. 184(3): 789-798.

Engels, J. M. M. and Fassil, H. 2007. Plant and animal genebanks. Pages 144-175 in: The role of food, agriculture forestry, and fisheries in human nutrition. Edited by V. R. Squires in Encyclopedia of Life Support Systems (EOLSS), Developed under the Auspices of the UNESCO, Eolss Publishers, Paris, France, [http://www.eolss.net].

ENSEMBL. 2016. http://www.ensembl.org/Bos_taurus/Info/Annotation. Accessed on the 7th of September, 2016.

Erbe, M., Hayes, B. J., Matukumalli, L. K., Goswami, S., Bowman, P. J., Reich, C. M., Mason, B. A. and Goddard, M. E. 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J. Dairy Sci. 95(7): 4114-4129.

Erbe, M., Ni, G., Pausch, H., Emmerling, R., Meuwissen, T. H. E., Cavero, D., Götz, K.-U. and Simianer, H. 2016. Experiences from genomic prediction with imputed sequence data in different species. Page 103 in Book of abstracts of the 67th annual

meeting of the European Federation of Animal Science. The waterfront conference and exhibition centre, Belfast, Northern Ireland.

Fallin, D. and Schork, N. J. 2000. Accuracy of Haplotype Frequency Estimation for Biallelic Loci, via the Expectation-Maximization Algorithm for Unphased Diploid Genotype Data. Am. J. Hum. Genet. 67(4): 947-959.

FAO. 2006. Livestock's long shadow – environmental issues and options. Rome, FAO.

FAO. 2015. World population prospects – 2015 revision. New York, FAO.

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

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. S., Daly, M. J. and Altshuler, D. 2002. The structure of haplotype blocks in the human genome. Science. 296(5576): 2225-2229.

Garrick, D. J. and Fernando, R. 2014. Genomic prediction and genome-wide association studies in beef and dairy cattle. Pages 474-501 in: The genetics of cattle. D. J. Garrick and A. Ruvinsky, ed. CABI (2nd edition), Wallingford, UK.

Gibson, G. and Muse, S. V. 2009. A primer of genome science. (3rd edition). ISBN 9780878932368. Sinauer Associates, Sunderland, MA, USA.

García-Ruiz, A., Cole, J. B., VanRaden, P. M., Wiggans, G. R., Ruiz López, F. J. and Van Tassel, C. P. 2016. Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. Proc. Natl. Acad. Sci. U. S. A. 113(28): E9995-4004.

Gautier, M., Laloë, D., and Moazami-Goudarzi. K. 2010. Insights into the genetic history of French cattle from dense SNP data on 47 worldwide breeds. PLoS ONE, 5(9), e13038.

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

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

Goddard, K. A. B., Hopkins, P. J., Hall, J. M. and Witte, J. S. 2000. Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. Am. J. Hum. Genet. 66(1): 216-234.

Government of Hungary, 2015. Több támogatás az őshonos állatok tartóinak Accessed 2nd of October, 2016. http://www.kormany.hu/hu/foldmuvelesugyi-miniszterium/hirek/tobb-tamogatas-az-oshonos-allatok-tartoinak.

Guillaume, F., Fritz, S., Boichard, D. and Druet, T. 2008a. Correlations of marker-assisted breeding values with progeny-test breeding values for eight hundred ninety-nine French Holstein bulls. J. Dairy Sci. 91(6):2520-2522.

Guillaume, F., Fritz, S., Boichard, D. and Druet, T. 2008b. Estimation by simulation of the efficiency of the French marker-assisted selection program in dairy cattle. Genet. Sel. Evol. 40(1):91-102.

Grisart, B., Coppieters, W., Farnir, F., Karim, L., Ford, C., Berzi, P., Cambisano, N., Mni, M., Reid, S., Simon, P., Spelman, R., Georges, M. and Snell, R. 2002. Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Res. 12(2): 222-231.

Habier, D., Fernando, R. L., Kizilkaya, K. and Garrick, D. J. 2011. Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics. 12: 186. http://dx.doi.org/10.1186/1471-2105-12-186.

Harris, B. L. and Johnson, D. L. 2010a. Genomic prediction for New Zealand dairy bulls and integration with national genetic evaluation. J. Dairy. Sci. 93: 1243-1252.

Harris, B. L. and Johnson, D. L. 2010b. The impact of high density SNP chips on genomic evaluation in dairy cattle. Interbull Bull. 42: 40-43.

Hayes, B. J., Bowman, P. J., Chamberlain, A. J. and Goddard, M. E. 2009a. Invited review: genomic selection in dairy cattle: progress and challenges. J. Dairy Sci. 92(2): 433-443.

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

Hayes, B. J., Chamberlain, A. J., McPartlan, H., Macleod, I., Sethuraman, L. and Goddard, M. E. 2007. Accuracy of marker-assisted selection with single markers and marker haplotypes in cattle. Genet. Res. 89: 215-220. http://dx.doi.org/10.1017/S0016672307008865.

Hayes, B. and Goddard, M. E. 2001. The distribution of the effects of genes affecting quantitative traits in livestock. Genet. Sel. Evol. 33:209-229.

Henderson, C. R. 1984. Applications of linear models in animal breeding. University of Guelph, Ontario, Canada. ISBN 0-88955-030-1.

Holland, P. M., Abramson, R. D., Watson, R. and Gelfand, D. H. 1991. Detection of specific polymerase chain reaction product by utilizing the 5' → 3' exonuclease activity of Thermus aquaticus DNA polymerase. Proc. Natl. Acad. Sci. U. S. A. 88(16): 7276-7280.

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

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

Institut de l'Elevage. 2014. Bilan de l'indexation des races bovines laitières. Accessed 9th of September, 2015. http://idele.fr/no_cache/recherche/publication/idelesolr/recommends/bilan-dindexation-des-races-bovines-laitieres-2014-bil.html.

Institut de l'Elevage. 2015a. Chiffres clés Bovins 2015. Accessed 28th of December, 2015. http://idele.fr/no_cache/recherche/publication/idelesolr/recommends/chiffres-cles-bovins-2015.html.

Institut de l'Elevage. 2015b. Résultats de Contrôle Laitier. Accessed 24th of May, 2016. http://idele.fr/no_cache/recherche/publication/idelesolr/recommends/resultats-de-controle-laitier-france-2015.html.

Institut de l'Elevage. 2015c. Indicateurs de variabilité génétique – races bovines – Edition 2015. Accessed 2nd of October, 2016. http://idele.fr/no_cache/recherche/publication/idelesolr/recommends/indicateurs-de-variabilite-genetique-races-bovines-edition-2015.html.

Institut de l'Elevage. 2016. Le point sur l'insémination en semence sexée en 2015. Accessed 11th of August, 2016. http://idele.fr/no_cache/recherche/publication/idelesolr/recommends/le-point-sur-linsemination-en-semence-sexee-en-2015.html.

Jeffreys, A. J., Neumann, R., Panayi, M., Myers, S. and Donnelly, P. 2005. Human recombination hot spots hidden in regions of strong marker association. Nat. Genet. 37(6):601-606.

Khansefid, M., Pryce, J. E., Bolormaa, S., Miller, S. P., Wang, Z., Li, C. and Goddard, M. E. 2014. Estimation of genomic breeding values for residual feed intake in a multibreed cattle population. J. Anim. Sci. 92(8): 3270-3283.

Lauvie, A., Audiot, A., Couix, N., Casabianca, F., Brives, H. and Verrier, E. 2011. Diversity of rare breed management programs: Between conservation and development. Livest. Sci. 140: 161-170.

Legarra, A., Ricard, A. and Filangi, O. 2013. GS3 software package and documentation. http://snp.toulouse.inra.fr/~alegarra. Accessed 1st of January, 2013.

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

Le Mézec, P., Benoit, M., Moureaux, S. and Patry, C. 2015. Genomics, sexed semen: changes in reproduction choices in French dairy herds. Page 249 in Book of abstracts of the 66th annual meeting of the European Federation of Animal Science. Warsaw University of Life Sciences, Poland.

Li, Y., Willer, C., Sanna, S. and Abecasis, G. 2009. Genotype imputation. Annu. Rev. Genomics Hum. Genet. 10: 387-406.

Liu, Z., Reinhardt, F., Bünger, A. and Reents, R. 2004. Derivation and calculation of approximate reliabilities and daughter yield-deviations of a random regression test-day model for genetic evaluation of dairy cattle. J. Dairy. Sci. 87:1896-1907.

Liu, Y., Qin, X., Song, X-Z. H., Jiang, H., Shen, Y., Durbin, K. J., Lien, S., Kent, M. P., Sodeland, M., Ren, Y., Zhang, L., Sodergren, E., Havlak, P., Worley, K. C., Weinstock, G. M. and Gibbs, R. A. 2009. Bos Taurus genome assembly. BMC Genomics 10: 180.

Lourenco, D. A., Tsuruta, S., Fragomeni, B. O., Chen, C. Y., Herring, W. O. and Misztal, I. 2016. Crossbreed evaluations in single-step genomic best linear unbiased predictor using adjusted realized relationship matrices. J. Anim. Sci. 94(3): 909-919.

Lukić, B., Pong-Wong, R., Rowe, S. J., de Koning, D. J., Velander, I., Haley, C. S., Archibald, A. L. and Woolliams, J. A. 2015. Efficiency of genomic prediction for boar taint reduction in Danish landrace pigs. Anim. Genet. 46: 607-616.

Lund, M. S., de Roos, A. P. W., de Vries, A. G., Druet, T., Ducrocq, V., Fritz, S., Guillaume, F., Guldbrandtsen, B., Liu, Z., Reents, R., Schrooten, C., Seefried, F. and Su, G. 2011. A common reference population from four European Holstein populations increases reliability of genomic predictions. Genet. Sel. Evol., 43, 43.

Lynch, M. and Walsh, B. 1998. Genetics and analysis of quantitative traits. (1st edition). ISBN 978-0878934812. Sinauer Associates, Sunderland, MA, USA.

Knürr, T., Strandén, I., Koivula, M., Aamand, G. P. and Mäntysaari, E. A. 2013. Haplotype-assisted genomic evaluations in Nordic red dairy cattle. Page 454 in Book of Abstracts of the 64th Annual Meeting of the European Federation of Animal Science, Nantes, France.

Ma, L., O'Connell, J. R., VanRaden, P. M., Shen, B., Padhi, A., Sun, C., Bickhart, D. M., Cole, J. B., Null, D. J., Liu, G. E., Da, Y. and Wiggans, G. R. 2015. Cattle sex-specific recombination and genetic control from a large pedigree analysis. PLOS Genet. 11(11):e1005387.

MacLeod, I. M., Bowman, P. J., Vander Jagt, C. J., Haile-Mariam, M., Kemper, K. E., Chamberlain, A. J., Schrooten, C., Hayes, B. J. and Goddard, M. E. 2016. Exploiting biological priors and sequence variants enhance QTL discovery and genomic prediction of complex traits. BMC Genomics 17:144.

Matukumalli, L. K., Lawley, C. T., Schnabel, R. D., Taylor, J. F., Allan, M. F., Heaton, M. P., O'Connell, J., Moore, S. S., Smith, T. P. L., Sonstegard, T. S. and Van Tassel, C. P. 2009. Development and characterization of a high density SNP genotyping assay for cattle. PLoS One 4(4): e5350.

Matukumalli, L. K., Schroeder, S., DeNise, S. K., Sonstegard, T., Lawley, C. T., Georges, M., Coppieters, W., Gietzen, K., Medrano, J. F., Rincon, G., Lince, D., Eggen, A., Glaser, L., Cam, G. and Van Tassel, C. 2011. Analyzing LD blocks and CNV segments in cattle: Novel genomic features identified using the BovineHD BeadChip. Pub. No. 370-2011-002, Illumina Inc., San Diego, CA.

McRae, A. F., McEwan, J. C., Dodds, K. G., Wilson, T., Crawford, A. M. and Slate, J. 2002. Linkage disequilibrium in domestic sheep. Genetics. 160: 1113-1122.

Meuwissen, T. H. E. and Goddard, M. E. 1996. The use of marker haplotypes in animal breeding schemes. Genet. Sel. Evol. 28: 161-176.

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

Misztal, I. 1999. Complex models, more data: simpler programming. Proc. Inter. Workshop Comput. Cattle Breed. 1999 March 18-20, Tuusala, Finland. Interbull Bul. 20:33-42.

Misztal, I. BLUPF90 software package and documentation. http://nce.ads.uga.edu/wiki/doku.php?id=documentation. Accessed 5th of September, 2016.

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

Park, T. and Casella, G. 2008. The Bayesian Lasso. J. Am. Stat. Assoc. 103:681-686.

Porto-Neto, L. R., Barendse, W., Henshall, J. M., McWilliam, S. M., Lehnert, S. A. and Reverter, A. 2015. Genomic correlation: harnessing the benefit of combining two unrelated populations for genomic selection. Genet. Sel. Evol., 47: 84.

Pryce, J. E., Gredler, B., Bolormaa, S., Bowman, P. J., Egger-Danner, C., Fuerst, C., Emmerling, R., Sölkner, J., Goddard, M. E. and Hayes, B. J. 2011. Short communication: Genomic selection using a multi-breed, across-country reference population. J. Dairy. Sci. 94:2625-2630.

Rauw, W. M. and Gomez-Raya, L. 2015. Genotype by environment interaction and breeding for robustness in livestock. Front. Genet. 6: 310.

Rincon, G., Weber, K. L., Van Eenennaam, A. L., Golden, B. L. and Medrano, J. F. 2011. Hot topic: Performance of bovine high-density genotyping platforms in Holsteins and Jersyes. J. Dairy Sci. 94: 6116-6121.

Rendel, J. M. and Robertson, A. 1950. Estimation of genetic gain in milk yield by selection in a closed herd of dairy cattle. J. Genet. 50(1): 1-8.

Saintilan, R., Capitan, A., Benoit, M., Barbier, S. and Fritz, S. 2015. Implementing a genomic preselection tool in the three main French beef cattle breeds. 22nd International 3R Congress. Paris, France.

Sanchez, M-P., Jónás, D., Baur, A., Ducrocq, V., Hozé, C., Saintilan, R., Phocas, F., Fritz, S., Boichard, D. and Crouseau, P. 2016. Implementation of genomic selection in three French regional dairy cattle breeds. Page 601 in Book of abstracts of the 67th annual meeting of the European Federation of Animal Science. Belfast, Northern Ireland.

Sargolzaei, M., Chesnais, J. P. and Schenkel, F. S. 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genomics. 15:478.

Schaeffer, L. R. 2006. Strategy for applying genome-wide selection in dairy cattle. J. Anim. Breed. Genet. 123(4): 218-223.

Shen, R., Fan, J-B., Campbell, D., Chang, W., Chen, J., Doucet, D., Yeakley, J., Bibikova, M., Garcia, E. W., McBride, C., Steemers, F., Garcia, F., Kermani, B. G., Gunderson, K. and Oliphant, A. 2005. High-throughput SNP genotyping on universal bead arrays. Mutat. Res. 573(1-2): 70-82.

Stephens, J. C., Schneider, J. A., Tanguay, D. A., Choi, J., Acharya, T., Stanley, S. E., Jiang, R., Messer, J. C., Chew, A., Han, J-H., Duan, J., Carr, J. L., Lee, M. S.,

Koshy, B., Kumar, A. M., Zhang, G., Newell, W. R., Windemuth, A., Xu, C., Kalbfleisch, T. S., Shaner, S. L., Arnold, K., Schulz, V., Drysdale, C. M., Nandabalan, K., Judson, R. S., Ruaño, G. and Vovis, G. F. 2001. Haplotype variation and linkage disequilibrium in 313 human genes. Am. J. Hum. Genet. 63(2): 595-612.

Støier, S., Larsen, H. D., Aaslyng, M. D. and Lykke, L. 2016. Improved animal welfare, the right technology and increased business. Meat. Sci. 120: 71-77.

Szyda, J., Ptak, E., Komisarek, J. and Zarnecki, A. 2008. Practical application of daughter yield deviations in dairy cattle breeding. J. Appl. Genet. 49: 183-191.

Thaxton, Y. V., Christensen, K. D., Mench, J. A., Rumley, E. R., Daugherty, C., Feinberg, B., Parker, M., Siegel, P. and Scanes, C. G. 2016. Symposium: Animal welfare challenges for today and tomorrow. Poult. Sci. 95(9): 2198-2207.

The International HapMap Consortium. 2005. A haplotype map of the human genome. Nature 437: 1299-1320.

Tobler, A. R., Short, S., Andersen, M. R., Paner, T. M., Briggs, J. C., Lambert, S. M., Wu, P. P., Wang, Y., Spoonde, A. Y., Koehler, R. T., Peyret, N., Chen, C., Broomer, A. J., Ridzon, D. A., Zhou, H., Hoo, B. S., Hayashibara, K. C., Leong, L. N., Ma, C. N., Rosenblum, B. B., Day, J. P., Ziegle, J. S., De La Vega, F. M., Rhodes, M. D., Hennessy, K. M. and Wenz, H. M. 2005. The SNPlex Genotyping System: A Flexible and Scalable Platform for SNP Genotyping. J. Biomol. Tech. 16(4): 398-406.

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

Verbyla, K. L., Hayes, B. J., Bowman, P. J. and Goddard, M.E. 2009. Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle. Genet. Res. 91: 307-311.

Villumsen, T. M., Janss, L. and Lund, M. S. 2009. The importance of haplotype length and heritability using genomic selection in dairy cattle. J. Anim. Breed. Genet. 126: 3-13. http://dx.doi.org/10.1111/j.1439-0388.2008.00747.x.

Wang, T., Chen, Y.-P. P., Goddard, M. E., Meuwissen, T. H. E., Kemper, K. E. and Hayes, B. J. 2015. A computationally efficient algorithm for genomic prediction using a Bayesian model. Genet. Sel. Evol. 47: 34.

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

Weng, Z-Q., Saatchi, M., Schnabel, R. D., Taylor, J. F. and Garrick, D. J. 2014. Recombination locations and rates in beef cattle assessed from parent-offspring pairs. Genet. Sel. Evol. 46:34.

Wientjes, Y. C. J., Calus, M. P. L., Goddard, M. E. and Hayes, B. J. 2015. Impact of QTL properties in the accuracy of multi-breed genomic prediction. Genet. Sel. Evol. 47: 42.

Wiggans, G. R., VanRaden, P. M. and Cooper, T. A. 2011. The genomic evaluation system in the United States: Past, present future. J. Dairy Sci. 94, 3202-3211.

Zhang, Z., Liu, J., Ding, X., Bijma, P., de Koning, D-J. and Zhang, Q. 2010. Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. PLoS One. 5(9): pii: e12648.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P.A., Heath, A. C., Martin, N.G., Montgomery, G. W., Goddard, M. E. and Visscher, P. M. 2010. Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42(7): 565-569.

# Appendix A

**S. table 1**: GBLUP results of 5 production traits in the Montbéliarde breed. Calculated correlation coefficients and regression slopes are shown in the table below. These results were better than those of the pedigree-based BLUP (results not shown), but slightly inferior compared to those obtained with *haplotypic GS3* (**Table 4**).

**S. table 1**: Correlation coefficients and regression slopes of DYD on GEBV values obtained with the GBLUP analysis (Montbéliarde breed).

| Trait name | Correlation coefficient | Regression slope |
|---|---|---|
| MY | 0.490 | 0.810 |
| FY | 0.551 | 0.850 |
| PY | 0.478 | 0.738 |
| FC | 0.570 | 0.785 |
| PC | 0.584 | 0.987 |
| Average[2] | 0.535 | 0.166 |

1: Trait name abbreviations: MY – milk yield; FY – fat yield; PY – protein yield; FC – fat content; PC – protein content

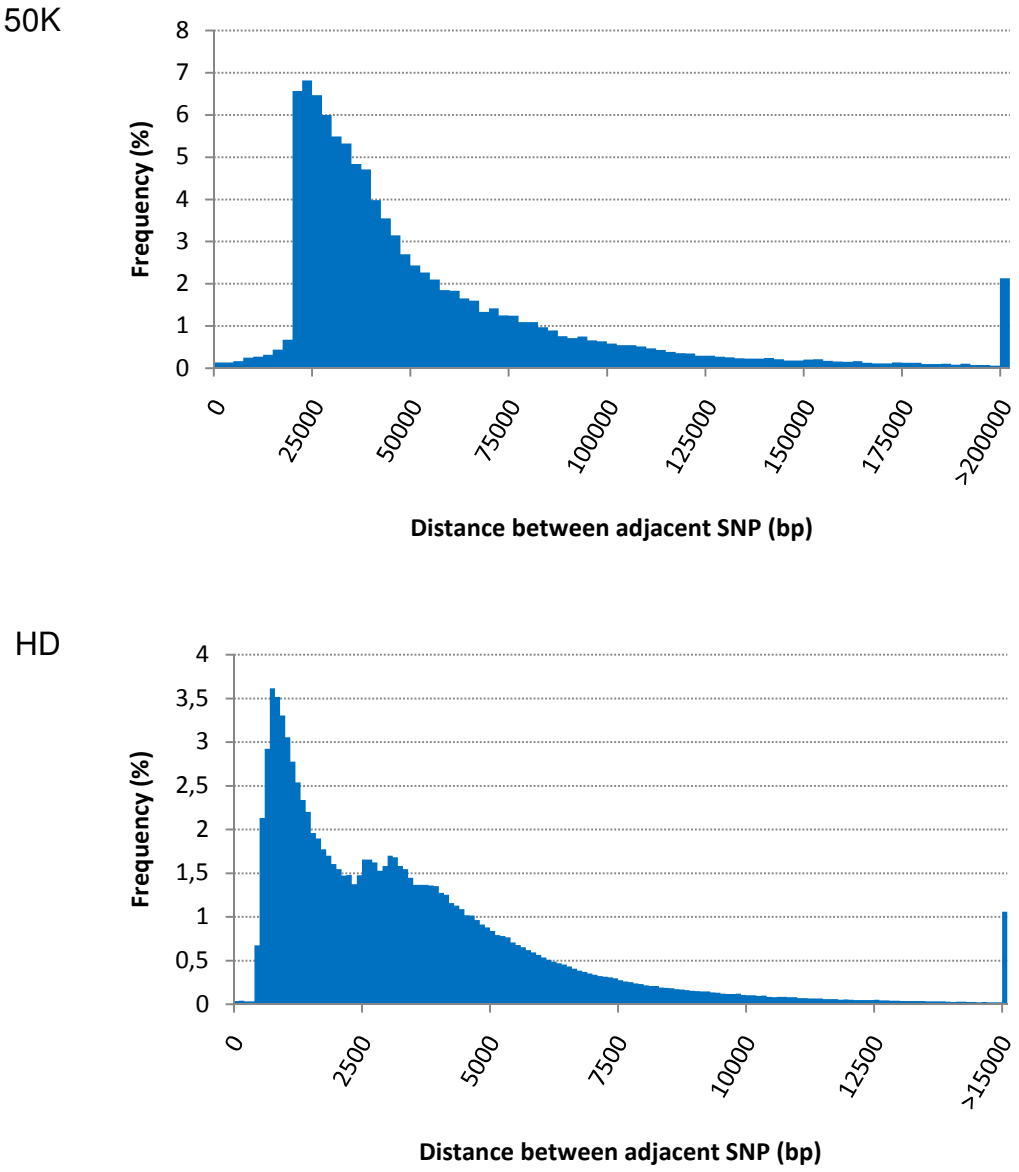
2: Average deviations from 1 are indicated for the regression slope

# Appendix B

Additional figures and tables related to the discussion of the new haplotype selection procedure presented in section 3.4. A short explanation is added to each table/figure.

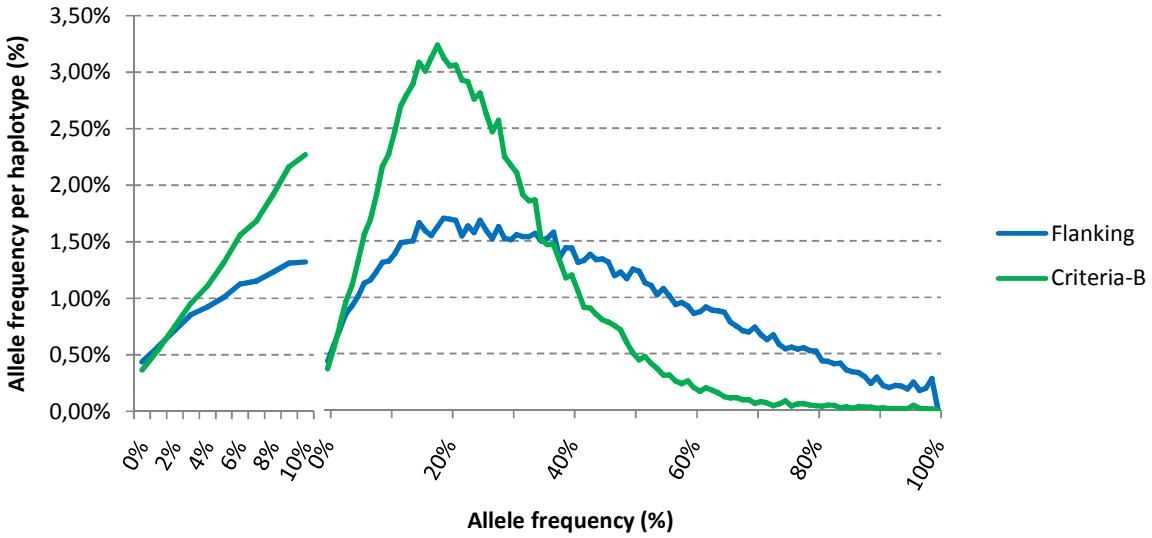
**S. figure 1**: Frequency distribution of the distances between adjacent SNP from either the 50K or the HD chip. Note the 1 order difference in magnitude between x-axis values of the 2 figures.

(See next page for the plots.)




**S. figure 1**: Frequency distribution of the distances between neighboring SNP from the (A) 50K and (B) HD SNP panels. Frequencies are calculated for every bins of 100 bp and 2500 bp for the HD and 50K SNP panels, respectively.

**S. figure 2**: Distribution of haplotype allele frequencies with 2 haplotype construction methods (Criterion-B and flanking haplotypes) using HD chip data and haplotypes of 3 SNP with both methods. Window size was 80 SNP in case of Criterion-B. Criterion-B resulted in better distribution of allele frequencies with less over-represented alleles and more alleles with an intermediate (10-40%) frequency.




**S. figure 2**: Overall distribution of haplotype allele frequencies according to the haplotype construction approach (haplotype size: 3 SNP; 6,000 QTL-SNP). The 0-10% region is also depicted with a more detailed scale on the x-axis.

**S. table 2** and **S. table 3**: Correlation coefficients (**S. table 2**) and regression slopes (**S. table 3**) of DYD on GEBV with different SNP-based and haplotype-based genomic evaluation methods using HD data with the Montbéliarde breed.a) QTL-SNP test: analysis using SNP identified in a prior QTL detection step; b) flanking haplotypes: using haplotypes built from the QTL-SNP and the neighboring SNP; c) flanking SNP: using the same markers as with the flanking haplotypes but as independent, single-SNP markers; d) Criterion-B haplotypes: haplotypes selected by Criterion-B from a 10 SNP-wide window surrounding the QTL-SNP; e) Criterion-B SNP: using the same markers as with the Criterion-B haplotypes but as independent, single-SNP markers.

Flanking haplotypes outperformed the analyses using only the QTL-SNP as genetic markers, while Criterion-B outperformed the flanking haplotypes. These are true for both the correlation coefficients and regression slopes and for both cases when the markers were used as haplotypes or as individual SNP.

**S. table 2**: Correlations between genomic estimated breeding values and DYD in the validation population for the scenario with an optimal number of QTL are presented. Window size: 80 SNP; Montbéliarde breed.

| Haplotype selection method | Marker type | Haplotype size | Milk quantity | Fat yield | Protein yield | Fat content | Protein content | Average |
|---|---|---|---|---|---|---|---|---|
| QTL-SNP | SNP | 1 | 0.467 | 0.478 | 0.412 | 0.560 | 0.574 | 0.498 |
| Flanking markers | SNP[1] | 3 | 0.456 | 0.491 | 0.415 | 0.563 | 0.591 | 0.503 |
| | | 4 | 0.455 | 0.490 | 0.418 | 0.560 | 0.591 | 0.503 |
| | haplotype | 3 | 0.481 | 0.530 | 0.433 | 0.565 | 0.604 | 0.523 |
| | | 4 | 0.483 | 0.536 | 0.440 | 0.570 | 0.618 | 0.529 |
| Criterion-B | SNP[1] | 3 | 0.462 | 0.503 | 0.434 | 0.588 | 0.614 | 0.520 |
| | | 4 | 0.477 | 0.511 | 0.445 | 0.588 | 0.610 | 0.526 |
| | haplotype | 3 | 0.476 | 0.539 | 0.452 | 0.585 | 0.614 | 0.533 |
| | | 4 | 0.494 | 0.543 | 0.461 | 0.575 | 0.614 | 0.537 |

1:All the SNP used for haplotypes are included in the BayesC analysis but they are used separately, as independent explanatory variables.

**S. table 3**: Regression slopes of DYD on GEBV in the validation population for the scenario with an optimal number of QTL are presented. Window size: 80 SNP; Montbéliarde breed.

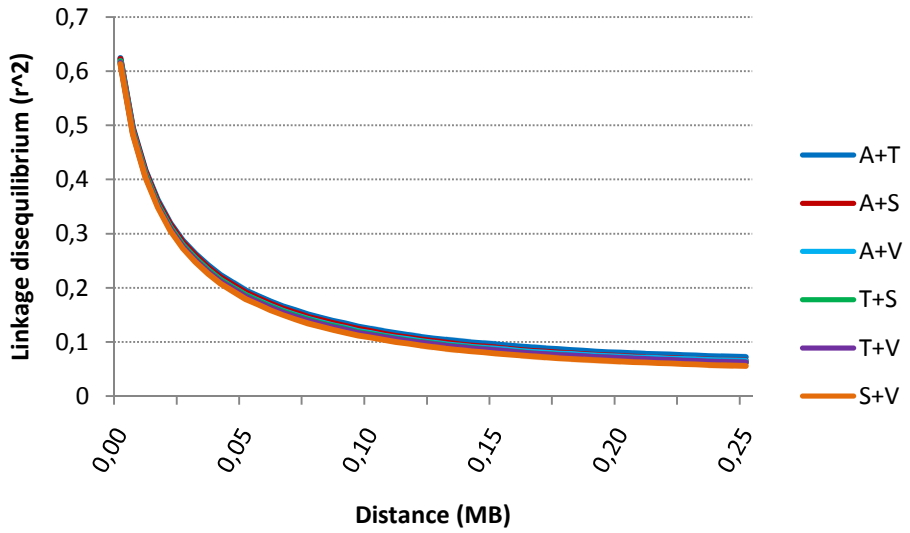
| Haplotype selection method | Marker type | Haplotype size | Milk quantity | Fat yield | Protein yield | Fat content | Protein content | Average |
|---|---|---|---|---|---|---|---|---|
| QTL-SNP | SNP | 1 | 0.631 | 0.594 | 0.519 | 0.758 | 0.780 | 0.656 |
| Flanking markers | SNP[1] | 3 | 0.632 | 0.649 | 0.545 | 0.791 | 0.808 | 0.685 |
| | | 4 | 0.635 | 0.652 | 0.550 | 0.788 | 0.809 | 0.687 |
| | haplotype | 3 | 0.705 | 0.739 | 0.594 | 0.804 | 0.868 | 0.742 |
| | | 4 | 0.722 | 0.778 | 0.622 | 0.833 | 0.884 | 0.768 |
| Criterion-B | SNP[1] | 3 | 0.677 | 0.675 | 0.598 | 0.828 | 0.895 | 0.735 |
| | | 4 | 0.721 | 0.702 | 0.621 | 0.819 | 0.894 | 0.751 |
| | haplotype | 3 | 0.747 | 0.781 | 0.676 | 0.835 | 0.945 | 0.796 |
| | | 4 | 0.804 | 0.824 | 0.691 | 0.830 | 0.974 | 0.825 |

1:All the SNP included in the haplotypes are included in the BayesC analysis but they are used separately, as independent explanatory variables.

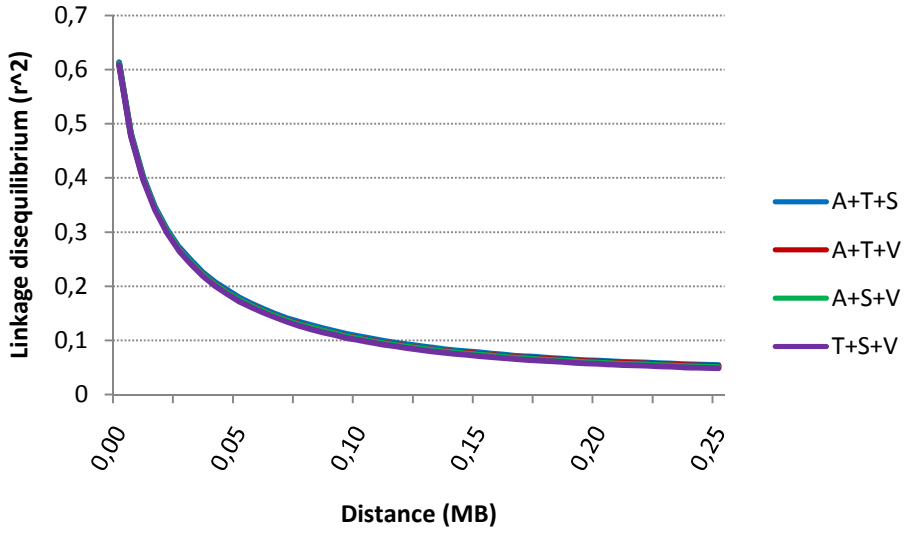
# Appendix C

LD-decay pattern in the multi-breed scenarios are shown either with 2 breeds (**S. figure 3**) or with 3 breeds (**S. figure 4**) or with all the 4 regional breeds contributing to the multi-breed population (**S. figure 5**).

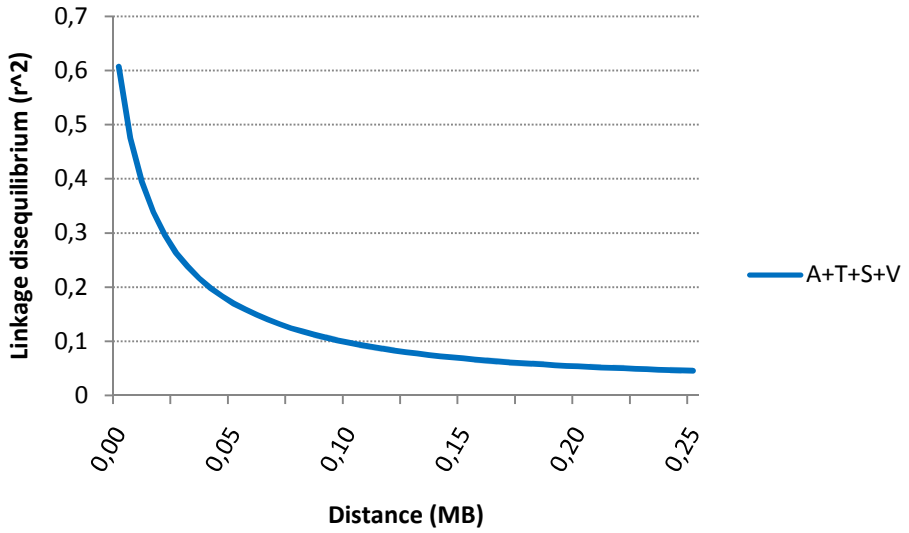
The LD-decay is faster with more breeds contributing to the evaluated population, which is in accordance with the expectations: due to the between-breed genetic diversity, if more breeds are included in the analysis, a faster LD-decay is expected. However, the difference between the curves is minor, which is due to the short evolutionary distance between these breeds (e.g. see **Figure 3**). This also explains why the LD-decay in the multi-breed scenarios is also remarkably similar to that in the single-breed scenarios (**Figure 9**).




**S. figure 3**: Linkage disequilibrium decay in the multi-breed (2-breed) scenarios. Breed name abbreviations: A – Abondance; T – Tarentaise; S – Simmental ; V – Vosgienne.




**S. figure 4**: Linkage disequilibrium decay in the multi-breed (3-breed) scenarios. Breed name abbreviations: A – Abondance; T – Tarentaise; S – Simmental ; V – Vosgienne.

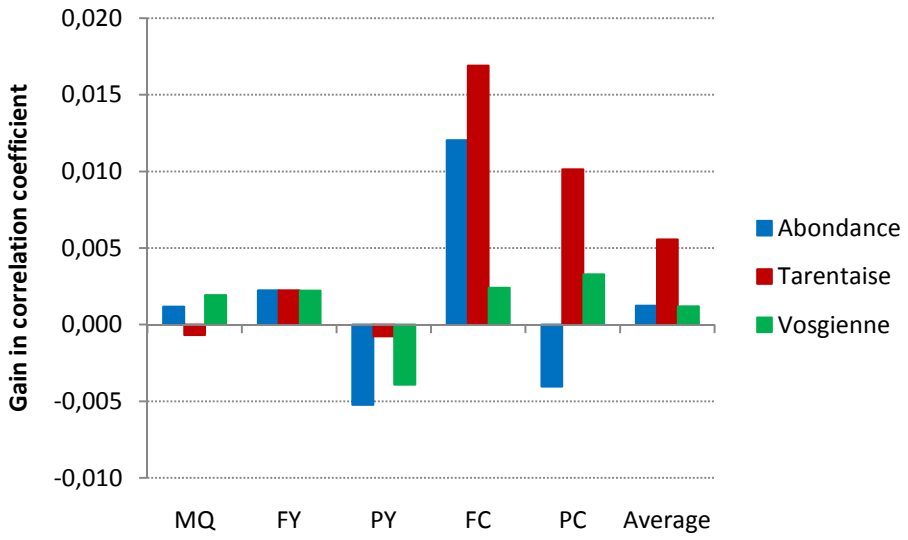



**S. figure 5**: Linkage disequilibrium decay in the multi-breed (4-breed) scenario. Breed name abbreviations: A – Abondance; T – Tarentaise; S – Simmental ; V – Vosgienne.

# Appendix D

**S. figure 6**: Effect of including the candidate mutations in the analysis with the BayesR method, compared to the scenario with only the 50K data used. Gain/loss in correlations between YD and GEBV of the animals in the validation population are shown.

These results were almost exclusively inferior compared to the same values obtained with BayesC (**Figure 11**).




**S. figure 6**: Effect of the inclusion of candidate mutations on the correlation between YD and GEBV measured on the validation population (BayesR).

# Appendix E

Most of the values presented in **Table 18** are based on real-life information from 2 breeds (Abondance and Tarentaise; Vosgienne is not considered due to its very particular situation), in which genomic evaluation was implemented in France. The way we obtained these estimates is presented in detail here. To create **Table 18**, the following four selection schemes were compared:

- Large breeds with genomic evaluation
- Regional breeds with progeny testing
- Regional breeds with genomic evaluation, but retaining progeny testing (i.e., with a proportion of AI done using “unorganized progeny tested” bulls, that is evaluated on the basis of their first crop daughter records)
- Regional breeds purely with genomic evaluation, i.e., all offspring born (daughters and bulls) are from young bulls

These four scenarios are discussed in detail below.

**Large breeds with genomic evaluation implemented**

Real-life (rough) estimates were available for the large breed scenario. The final estimate for ΔG (0.47) is the same as in Schaeffer (2006).

**Regional breeds with progeny testing**

Before 2016, a breeding program based on progeny testing was implemented in the regional breeds, without any genomic information. The three parameters (selection intensity, selection accuracy and generation interval) are discussed here for each of the 4 paths (see section 2.6.1 of Chapter 2).

**Selection intensity**

- Sires of bulls

**Table 1** shows the number of progeny tested bulls every year in the 4 regional breeds. The average number of bulls selected each year based on their progeny test results is 5 and 4 for Abondance and Tarentaise, respectively (D. Boichard and S. Barbier, 2016, personal communication). This means that ~40% of the tested males are selected (i.e. selection intensity: 0.97) on the "sires of bulls" path.

- Sires of cows

Approximately 50% of the cows are used for progeny testing, therefore these cows are inseminated with semen from unproven bulls; these bulls are assumed to represent the mean of the population of progeny of elite bulls and cows (whose selection intensity is taken into account in the sires of bulls and dams of bulls paths) and therefore the selection intensity is ~0 for them. The other 50% of the cows are inseminated with proven bulls (selection intensity ~0.97). A weighted selection intensity is calculated for the "sires of cows" path and it gives ~0.49 (or ~70% selection proportion).

- Dams of bulls

No data was available to estimate this parameter. Schaeffer (2006) used 2% for large breeds, but it is likely to be higher for the regional breeds and here it was assumed to be 5%.

- Dams of cows

Dams of cows are largely unselected as (nearly) all of the females are required to maintain the constant population size (selection proportion: 100%; selection intensity: 0,0).

**Selection accuracy**

- Sires of bulls

Bulls were progeny tested with ~25-30 individuals to obtain a reliability of ~0.50 for these animals (D. Boichard, personal communication). The corresponding accuracy is ~0.71.

- Sires of cows

Accuracy of progeny tested bulls: ~0.71; accuracy of unproven bulls: 0. Similarly to the selection intensities, these accuracies are weighted with 0.5 and 0.5, respectively, because 50% of the cows are “used” for progeny testing and 50% of them are inseminated with semen of proven bulls.

- Dams of bulls

Dams of bulls are required to have at least 2 finished lactations. Therefore their accuracy is larger than the accuracy on the “dams of cows” path, but lower than that of the progeny tested bulls. It was assumed to be ~0.70.

- Dams of cows

Dams of cows have own performance records only; the accuracy of these animals was assumed to be ~0.60. Note that since selection intensity in cows is 0, this value is of no importance when calculating the annual genetic gain (that is because the genetic gain on this path is supposed to be zero irrespective of the selection accuracy).

**Generation interval**

All generation interval values were inspired by real data (Institut de l’Elevage, 2015c).

**Regional breeds with genomic evaluation**

In this scenario, only a genomic evaluation is assumed with no progeny testing. Again, the selection intensity, selection accuracy and generation interval are discussed separately:

**Selection intensity**

- Sires of bulls

The ~10% figure was calculated from **Table 17**, which was created based on the information provided by the breeding organizations (S. Barbier, 2016, personal communication). The calculated intensities were averaged over the 2 breeds.

- Sires of cows

Sires of cows were assumed to be the same as the sires of bulls.

- Dams of bulls

The ~10% figure was again calculated from **Table 17**. The calculated intensities were averaged over the 2 breeds.

- Dams of cows

Assuming an increase in the use of sexed semen (as observed in large breeds), it is expected that the number of female selection candidates will increase. Furthermore, genomic evaluation gives equally accurate GEBV for females as for males. As a consequence of these, selection intensity is expected to increase in females. However, still a large proportion will be needed to maintain the population size. The 90% proportion (selection intensity equal to 0.2) in females is a rough estimate to express these expectations.

**Selection accuracy**

Selection accuracy is equally high for all paths. The 0.73 (a reliability of ~0.53) was chosen based on our estimates (Sanchez et al., 2016).

**Generation interval**

Generation intervals for the “sires of cows” “sires of bulls” are expected to be similar to those of the corresponding large breeds generation intervals. That is because GEBV are available before maturity. The generation interval in the “dams of sires” path is expected to increase slightly.

Generation interval in the “Dams of cows” path is not expected to be affected by the introduction of genomic selection in the regional breeds, because cows were used for breeding at the age of maturity even in the previous selection program. Therefore, no decrease is expected on this path (unless using sexed semen is essentially on heifers).

**Regional breeds with genomic evaluation, but retaining progeny testing**

In this scenario, ~10% of the bulls with GEBV are retained for progeny testing (18-20 animals, depending on the breed) and 30% of them (5-6) are kept after progeny testing.

Compared with the previous scenario, the “dams of bulls” and “dams of cows” paths are unaffected by the fact that progeny testing is retained.

**Selection intensity**

- Sires of bulls

Sires of bulls come from the progeny tested bulls. Therefore, the 5-6 bulls passing progeny testing will become sires of bulls from a total of 120-150, which is ~4% of all the candidates. The corresponding intensity is 2.15.

- Sires of cows

The long-term aim of the breeding organizations is to inseminate 70% of the cows by the selection candidates with GEBV only (S. Barbier, 2016, personal communication), while the remaining 30% of the cows are going to be inseminated with semen from progeny tested bulls. The intensities with and without progeny testing are 1.76 and 2.15, respectively. Weighting these gives a combined intensity for sires of cows of 1.88 (~8% of the population selected).

**Selection accuracy**

- Sires of bulls

Compared to the genomic evaluation scheme, the selection accuracy will increase due to progeny testing. Here we assume that the reliability will be ~70%. The corresponding accuracy is ~84%.

- Sires of cows

Similarly to selection intensity, weighting the selection accuracy of progeny tested bulls (0.84) with 30% and the accuracy of bulls with GEBV only (0.73) will result in an overall accuracy of 0.76 in this path.

**Generation interval**

- Sires of bulls

Generation interval in this path is the same as with progeny testing.

- Sires of cows

Generation interval in this path is the weighted average of the generation interval with progeny testing (7.5 years with a weight of 30%) and with genomic evaluation only (2.5 years with a weight of 70%). Combined together, it results in a generation interval of 4 years for this path.

# Publications and trainings

## Scientific publications

Jónás, D., Ducrocq, V. and Croiseau, P. In press. The combined use of LD-based haploblock and allele frequency-based haplotype selection method enhances genomic evaluation in dairy cattle. *J. Dairy Sci.*

Jónás, D., Ducrocq, V., Fritz, S., Baur, A., S., M-P., Croiseau, P. 2016. Genomic evaluation of regional dairy cattle breeds in single-breed and multi-breed contexts. *J. Anim. Breed. Genet.* 134(1): 3-13.

Jónás, D., Ducrocq, V., Fouilloux, M-N., Croiseau, P. 2016. Alternative haplotype construction methods for genomic evaluation. *J. Dairy Sci.* 99(6): 4537-4546.

## Scientific conferences

Sanchez, M. P., Jónás, D., Baur, A., Ducrocq, V., Hozé, C., Saintilan, R., Phocas, F., Fritz, S., Boichard, D. and Croiseau, P. 2015. Implementation of genomic selection in three French regional dairy cattle breeds. Oral presentation at the 67$^{th}$ Annual meeting of the European Federation of Animal Science. Belfast (Northern Ireland). 29 August - 2 September, 2016.

Jónás, D., Ducrocq, V., Fouilloux, M-N., Croiseau, P. 2015. Haplotype construction methods to enhance genomic evaluation. Oral presentation at the 66$^{th}$ Annual meeting of the European Federation of Animal Science. Warsaw (Poland). 31 August - 4 September, 2015.

Jónás, D., Hozé, C., Boichard, D., Croiseau, P. 2014. Application of a three-haplotype LDLA model to the French Holstein population. Oral presentation at the 10$^{th}$ World congress on genetics applied to livestock production. Vancouver (Canada). 17-22 August, 2014.

## Seminars and workshops

ABIES Seminar: Paris (France); 25-26 February, 2014.

17th Séminaire des thésards du Département de Génétique Animale: Jouy en Josas (France); 23-24 April, 2014.
Poster presentation: Jónás, D., Croiseau, P. and Ducrocq, V. 2014. Evaluation of haplotype-based genomic selection methods in a multi-breed context.

Doc'J PhD seminar: Jouy en Josas (France); 3 December, 2014.

ABIES Seminar: Paris (France); 14-15 April, 2015.
Poster presentation: Jónás, D., Ducrocq, V., Fouilloux, M-N., Croiseau, P. 2015. Alternative haplotype construction methods for genomic evaluation.

Doc'J seminar: Jouy en Josas (France); 16 April 2015.

18th Séminaire des thésards du Département de Génétique Animale: La Rochelle (France). 21-22 May, 2015.
Oral presentation: Jónás, D., Croiseau, P. and Ducrocq, V. 2015. Evaluation of haplotype-based genomic selection methods in multi-breed context.

EAAP workshop: Poznań (Poland); 7-11 September, 2015.

SelGen seminar: Paris (France); 26 May, 2016.
Oral presentation: Jónás, D., Croiseau, P. and Ducrocq, V. 2015. Evaluation of haplotype-based genomic selection methods in multi-breed context.

19th Séminaire des thésards du Département de Génétique Animale: Toulouse (France); 16-17 March, 2016.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Agriculture, alimentation, biologie, environnement, santé (ABIES)

**Titre :** Evaluation des performances des méthodes de sélection génomique basées sur des haplotypes et intérêt de ces approches dans un contexte multiracial

**Mots clés :** bovins laitiers, sélection génomique, multiraciale, haplotype, haploblock

**Résumé :** En sélection génomique, des marqueurs de l'ADN sont utilisés pour l'évaluation des grandes races laitières. La plupart des méthodes d'évaluation génomique actuelles utilisent des SNP, bien que l'utilisation d'haplotypes de SNP apporte un plus grand polymorphisme. Il n'y avait pas d'évaluation génomique en place en 2014 pour les races régionales (Abondance, Tarentaise, Vosgienne), plaçant ces races en position de faiblesse.
Notre objectif principal a été de mesurer l'intérêt de l'utilisation d'haplotypes en évaluation génomique, y compris à partir d'une population d'apprentissage multiraciale. Nous avons montré que les haplotypes conduisent à de meilleurs résultats que les SNP et que la fréquence des allèles et l'étendu du déséquilibre de liaison sont importants pour une construction optimale des haplotypes. Nous avons développé deux critères incorporant ces informations qui améliorent la précision des évaluations tout en réduisant le nombre de marqueurs utilisés.
Depuis 2015, un de ces critères a été inclus dans les évaluations génomiques officielles en France. Notre approche a donné dans les races régionales une précision similaire à celle obtenue après testage sur descendance. Une évaluation génomique de routine est en place pour 3 races régionales en France depuis Juin 2016. L'utilisation d'une puce Haute Densité n'a pas amélioré sa précision, alors qu'une population d'apprentissage multiraciale a été bénéfique uniquement pour certaines races. Le génotypage des nouvelle femelles a augmenté la précision de la sélection mais l'inclusion de mutations candidates détectées dans les grandes races laitières n'a conduit qu'à une légère amélioration chez les races régionales.

**Title:** Evaluation of haplotype-based genomic selection methods with focus on their performances in a multi-breed context in dairy cattle

**Keywords:** dairy cattle, genomic evaluation, multi-breed, haplotype, haploblock

**Abstract:** In genomic selection, DNA marker information is exploited for evaluation purposes in large dairy cattle breeds. Most of the current genomic evaluation methods rely today on SNP information, although haplotypes are expected to perform better due to their higher polymorphism. In 2014, genomic evaluation had not yet been implemented in regional breeds (Abondance, Tarentaise, Vosgienne), resulting in economic weaknesses for these breeds.
Our aim was to assess the use of haplotypes in genomic evaluation with focus on their performance in combination with multi-breed reference populations. We found that haplotypes outperformed individual SNP markers for genomic evaluation. We also showed that information on haplotype allele frequency and on linkage pattern are relevant to select haplotypes for evaluation purposes. Our haplotype selection criteria also allowed a significant reduction of the number of markers used for genomic prediction.
One of these criteria was incorporated into the French routine genomic evaluation in 2015. The performance of such an evaluation was then assessed in four regional breeds, leading to similar or higher accuracies than current progeny testing. Consequently, routine genomic evaluation was implemented in these breeds in 2016. The use of high density genotypes did not improve the performance of genomic evaluation in these breeds, while multi-breed training populations were beneficial only in some of them. Additional genotyped females led to notable increases in selection accuracies. Inclusion of candidate mutations identified in large breeds led to only minor improvements in regional breeds.

**Université Paris-Saclay**
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France