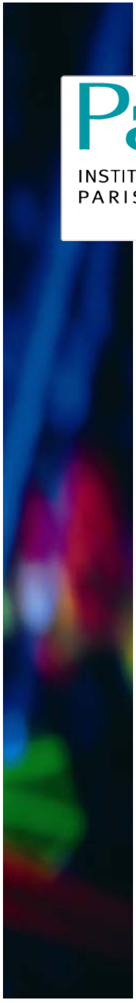
Genomic selection of dairy cows

Romain Dassonneville

To cite this version:

Romain Dassonneville. Genomic selection of dairy cows. Animal genetics. AgroParisTech, 2012. English. NNT : . pastel-00954581



Doctorat ParisTech

THÈSE

pour obtenir le grade de docteur délivré par

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

(AgroParisTech)

Spécialité : Génétique animale

présentée et soutenue publiquement par

Romain DASSONNEVILLE

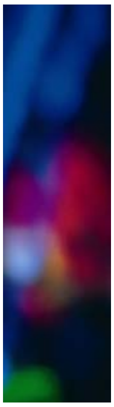
le 28 septembre 2012

Genomic selection of dairy cows

Sélection génomique des vaches laitières

Directeur de thèse : **Vincent DUCROCQ**
Co-encadrement de la thèse : **Didier BOICHARD**

Représentant entreprise, Institut de l'Élevage : **Sophie MATTALIA**



Jury

M.Georg THALLER, Professeur, CAU - Kiel - Allemagne	Rapporteur
M. Johan SÖLKNER, Professeur, BOKU – Vienne - Autriche	Rapporteur
M. Sander DE ROOS, Docteur, responsable schéma de sélection CRV – Arnhem – Pays-Bas	Examinateur
M. Pierre-Louis GASTINEL, chef du département génétique, Institut de l'Élevage, Paris	Examinateur
M. Thomas HEAMS, Docteur, Maître de conférence AgroParisTech	Examinateur
M. Didier BOICHARD, Docteur, Directeur de recherche, INRA Jouy-en-Josas	Examinateur

AgroParisTech
INRA GABI
Domaine de Vilvert, 78350 Jouy-en-Josas

Key-words

Genomic Selection – Imputation – Low density panel – Bias – Preferential treatment

Abstract

Genomic selection has revolutionized breeding in dairy cattle, at least on the male pathway. This thesis focuses on the female side. First, the genotyping tool most adapted to females was defined. The first study conducted within the Eurogenomics consortium assessed the value of using the commercially available Illumina® 3K SNP chip. The allelic imputation error rate was 4% with the national reference population, and the loss in reliability of GEBV when using imputed genotypes instead of real genotypes was 0.05 (2% and 0.02 respectively with the combined Eurogenomics reference population). In a second study, alternative in silico low density chips were described. Their imputation accuracy was 1 to 2.5% higher than the initial commercial 3K. The imputation accuracy not only depends on the number of markers, but also on MAF and spacing. A novel imputation strategy, fast and accurate, based on existing software, was described. Then, the construction of the new Bovine LD panel, adapted to many breeds and specifically dedicated to imputation, was detailed. This tool is well adapted for the genotyping of females in dairy cattle at a reasonable cost.

A second main aspect of this thesis was to study how performances of genotyped cows fit within the current genomic prediction model. An experimental design was set up to assess the effect of potential biases such as preferential treatment on genomic predictions. Two genomic evaluations were performed, one including only daughters performances of proven bulls, and another one including phenotypes for both males and females. Two traits were studied: milk yield, which is prone to preferential treatment and somatic cell count. Two groups were considered: elite females genotyped by breeding companies and randomly selected cows genotyped in a side project. For several measures potentially related to bias, the elite group presented for milk yield a different pattern than for the other trait/group combinations. The study demonstrated that including own milk performances of elite females induced over-estimated genomic evaluations. Such a bias has two major consequences: it may affect genomic predictions equations, and it may induce overestimated breeding values for the cow and her close relatives. Different possible solutions to properly include such performances in genomic predictions were described and their potential impacts were compared.

Finally, the benefits of genotyping heifers either by breeding companies or by farmers were discussed. A review of several simulation studies was conducted. Selecting bulls dams based on their genotypes appears to be crucial within a breeding scheme. Indeed, it is as important as using young bulls for artificial insemination. Using genotyping tools to select heifers to replace culled cows is more controversial. The return on investment for the famers depends on the cost of genotyping, the replacement rate as well as the economic value of the expected genetic improvement. Several herd management decisions could be facilitated when using genomic breeding values. A positive interaction exists between genomic selection within herd and several reproduction practices such as embryo transfer or use of sexed semen. Their combination may help in solving the issue that dairy cattle faces today related to the decrease of performances for health traits such as fertility.

This research was conducted from October 2009 to September 2012 with the financing from **Agence Nationale de la Recherche** (ANR,project AMASGEN) and **APIS-GENE**, at:

Institut National de la Recherche Agronomique (INRA)

Génétique Animale et Biologie Intégrative (GABI)
Equipe G2B, génétique et génomique des bovins laitiers et allaitants
Domaine de Vilvert
78350 Jouy-en-Josas
France

With the support of :

Institut de l'Elevage (« CIFRE », **Association Nationale de la Recherche de la Technologie** fundings)

Institut de l'Elevage

Département Génétique
Maison Nationale des Eleveurs
149, rue de Bercy
75595 Paris Cedex 12
France

A ma famille,

“I have hope
because I saved one seed
that I will plant and grow again.”

Palestinian poem

.

Acknowledgments / Remerciements

Tout d'abord je tiens à remercier mes directeurs de thèse. Vincent, merci de m'avoir incité à réaliser cette thèse, elle m'aura permis de « m'éclater » sur ce sujet. Tu as été à mes côtés tout au long de la thèse, et je sais que je pouvais compter sur toi pour tout aspect théorique relatif aux évaluations.

Didier, merci pour le temps consacré à ma thèse malgré l'emploi du temps chargé.

Sophie, c'est toi qui m'as fait tomber dans la marmite de la sélection génétique, quand je n'étais qu'un étudiant en 2ème année d'agro. Merci. C'était un plaisir que d'intégrer ton équipe. J'ai beaucoup apprécié que tu sois toujours prête à discuter des impacts de nos travaux sur la filière.

I would like to thank the members of the jury. Georg, Hans and Sander, thank you for reading this manuscript and for coming to the defense.

Merci à Pierre-Louis, également membre du jury. Je souhaite aussi en profiter pour remercier l'Institut de l'Elevage pour avoir participé au financement de ce projet.

Merci à Thomas, mon professeur référent, qui a été un interlocuteur privilégié de l'école doctorale.

Je tiens ensuite à remercier les membres de l'équipe AMASGEN.

Sébastien d'abord, le capitaine, tant sur le terrain de foot qu'au bureau. Tu incarnes à toi seul l'esprit d'équipe. Tu as toujours été disponible pour m'aider dans ma thèse (mise à disposition de fichiers, …) ou pour partager des discussions « terrain » sur la sélection.

François, ensuite, mon « professeur geek ». Tu m'as enseigné la quasi-totalité des outils informatiques que j'utilise, des logiciels d'imputation aux scripts shell ou awk. Sans parler des bières échangées, à Versailles, lors de congrès, ou dans les contrées nordiques.

Clotilde, la plus courageuse, puisque tu m'as supporté en tant que coloc' ! Nos échanges pour refaire le monde de l'élevage furent savoureux.

Pascal, le musicien de l'équipe, toujours jovial, c'est un plaisir que de te compter parmi mes collègues.

Aurelia, la spécialiste de la chaine d'évaluation, tu m'as été d'une grande aide à plusieurs reprises lors de cette thèse.

Merci également aux membres toulousains de l'équipe, Christèle, Andrés et Carine.

Merci à Bérénice, qui a partagé mon bureau tout au long de la thèse, également « jumelle CIFRE » au sein de l'équipe de l'Institut, pour ton dynamisme communicatif. Nous étions complémentaires, entre les stats et la connaissance du terrain. Nous nous sommes mutuellement motivés, que ce soit pour rejoindre un groupe bayésien ou pour suivre une formation management.
Merci aussi à Sandrine, qui a partagé notre bureau à mi-temps, et nous a ramené l'air frais breton un lundi sur 2.

Merci à tous les membres de l'équipe G²B. Une pensée particulière pour Aurélien, avec qui c'est toujours un plaisir de dialoguer.

Merci à tous les membres du services races laitières de l'Institut de l'Elevage, en particulier Hélène, dont la porte a toujours été ouverte, mais aussi Mickaël, Stéphanie, Julie, Amandine et Sophie (Mo) malgré ces croche-pieds et pieds de nez incessants.

Merci à tous les membres de l'ex-SGQA, avec qui j'ai partagé d'agréables moments en salle café cochon et tout particulièrement Romain, Thierry et Thierry.

Merci à Jean-Pierre, directeur d'unité mais aussi co-équipier en défense centrale.

I would like to thank Rasmus for our fruitful collaboration, and all the members of the Eurogenomics consortium.

I would also like to thank the Interbull team, in particular Jette, Hossein and Freddy that taught me a lot about genetics.

Je voudrais remercier tous les collègues de l'INRA, de l'Institut de l'Elevage ou de l'UNCEIA que j'ai côtoyé.

Merci aux castors et en particulier Fred, tant pour nos parties de fous rires que de ballon rond.

Enfin, merci à mes proches, enfin, à Perrine d'abord pour son soutien constant, à mon père qui m'a transmis sa passion de l'élevage et la génétique, à ma mère, à ma sœur et à toute ma famille. Sans oublier mes amis de toujours, Jérémy, Matthieu, Greg et Rémi qui m'ont accompagné au bout du monde ou au sommet de l'aiguille.

Abbreviations and Definitions

AI: Artificial Insemination

BLUP: Best Linear Unbiased Predictor is the model used for conventional genetic evaluations.

DAG: Directed Acyclic Graph

DNA: DeoxyriboNucleic Acid

DYD: stand for Daughter Yield Deviations (DYD) and were defined by VanRaden and Wiggans (1991). They correspond to the average daughters performances corrected for all fixed effects (such as the herd, year, season effects among others), the permanent environment effect, and also for the genetic contribution of the bull's mate (i.e., half the additive genetic value of the cow's dam).

DGV refer to Direct Genomic Values. They are obtained after genomic evaluation (when model includes a polygenic effect, DGV refer to the non-polygenic genetic effect).

EBV: refers to Estimated Breeding Values. They are obtained after conventional genetic evaluation.

EDC: Equivalent Daughter Contribution

EN: Elastic Net

GBLUP: Genomic BLUP, for which the relationship matrix A is replaced by a genomic relationship matrix G, based on markers information.

GEBV: refers to Genomic EBV (or genomically enhanced EBV). They are obtained after genomic evaluation, and also account for pedigree information. They are either derived from a genomic model which includes a polygenic effect, or from a blending of DGV and conventional EBV.

GG: GoldenGate

LD: Linkage Disequilibrium: non-random association of alleles between loci.

LPI: Lifetime Profit Index

MAF: Minor Allele Frequency

MAS: Marker Assisted Selection

MS: Mendelian Sampling term

NM: Net Merit

PTA: Predicted Transmitting Ability (equals EBV/2)

QTL: Quantitative Trait Locus

Reference population: this term refers to all the genotyped animals with phenotypic data (progeny tested bulls for instance). This reference population is usually split into 2 groups in most studies (training and validation).

SCC: Somatic Cell Count

Single step: procedure for which both conventional and genomic evaluations are performed at once.

SNP: means Single Nucleotide Polymorphism. It corresponds to a DNA sequence for which one single nucleotide present two possible forms.

TMI: Total Merit Index

Training population: refers to individuals for which phenotypic data are used in the model.

US/USA: United States (of America)

Validation population: refers to individuals for which phenotypic data are removed from the model. The model is then used to predict these phenotypic data. Estimates can be compared to “true” phenotypic data.

YD: Yield Deviation is the equivalent phenotypic measure for females and correspond to the performance of the cow herself (not her progeny), corrected as well for all the effect but the genetic effect.

CONTENTS

General Introduction

- **Background**

During the XXth century, genetic improvement in livestock relied on performance recording and pedigree registration without information on the genome. In dairy cattle, sophisticated progeny testing schemes were implemented. Male candidates were randomly mated in order to have a given number (often 100) of daughters. These daughters obtained performances on production, type, and functional traits, when bulls were about 5-year old. Daughters' performances were used in specific statistical analyses in order to estimate breeding values of their sires. Such breeding schemes were efficient. For instance, in France the annual genetic gain for milk yield was around 100kg or 0.2 genetic standard deviation over the last two decades. However, the generation interval (time elapsed between two successive generations) was quite long, generating important drawbacks such as high costs of breeding programs and a long delay between selection decisions and the observation of their effect on performances. This approach was based on the polygenic model proposed by Fisher, assuming an infinite number of genes involved in the genetic determinism of the trait. This model is biologically wrong (limited amount of genetically inherited material) but very effective in practice. However, its efficiency to predict the Mendelian sampling effect (i.e. the part of the breeding value which is not predictable from the parental value) is low when the individuals has no own performance nor progeny, or when the heritability is low.

For selection, animals do not have phenotypes (e.g. dairy bulls) or have phenotypes relatively late (dairy cows for instance). To achieve efficient selection, it would be desirable to directly have access to the information hidden in the genetic material (DNA) early in the life of the animal. A first solution is to know the genotype at some known major genes. For example in beef cattle, the myostatin gene, a gene involved in muscular hypertrophia was discovered (Grobet et al. 1997). It is possible to use such information for selection (Lande and Thompson, 1990). Sometimes, a specific location on a chromosome is known to have an effect on a trait of interest but the genes involved remain unknown. This location is called QTL (Quantitative Trait Locus, Georges et al., 1995). It is possible to capture the information at a given QTL based on the information at adjacent markers given the fact that long chromosomes segments are inherited from parents to progeny. Shrimpton and Robertson (1998) demonstrated that only a few genes have large effects whereas many genes have small effects. Capturing all this information requires to know the genotype at markers all over the genome. This technique is called genomic selection. Its implementation became possible with the reduction of genotyping costs.

Three main scientific questions were related to the implementation of genomic selection. First, which genotyping tool should be used? The Illumina company developed the Bovine SNP50® (Matukillami et al., 2009) and this tool was found to be well adapted to the need, with a good adequacy between its marker density and the effective population size of most breeds. Second, some methodological questions were addressed such as: which statistical prediction model should be used to estimate markers effects? Meuwissen et al. (2001) described the original Bayesian models (BayesA and BayesB) and other approaches were proposed subsequently, especially GBLUP (VanRaden, 2008). Last, how breeding schemes should evolve after inclusion of genomic selection? Schaeffer (2006) demonstrated that progeny-testing was no more economically efficient when genomic selection is implemented. Emphasis was set on the estimation of genomic breeding values of young male candidates and their early use through artificial insemination in breeding scheme.

- **Aim of the thesis**

This thesis aims at addressing the same questions as above but focusing on the female population. First, which genotyping tool is best adapted to females? Second, do individual performances of genotyped cows fit within the current prediction model? Last, what are the benefits of genotyping females, both for the farmer and at the breeding scheme level?

- **the AMASGEN research project**

In 2008, a genome-wide markers assisted selection was implemented in France and became official in 2009. Some improvements of the method were required. The research project called AMASGEN was launched for 3 years in 2009 in France by INRA (the French research institute for agriculture) with the collaboration of Institut de l'Elevage (French technical institute for livestock productions) and UNCEIA (Union Nationale des Coopératives d’Elevage et d’Insémination animale, umbrella association of cooperatives for artificial insemination and breeding). AMASGEN stands in French for Methodological Approaches and Application for GENomic Selection in dairy cattle. The main aim of this project was to develop a method to combine genomic information from genotyped animals with the information from phenotypes and pedigree for a fast and large implementation of genomic selection in the French dairy cattle breeding schemes. The fourth work package of this project was dedicated to the aim of this study.

- **Outlines of the thesis**

In Chapter 1, after addressing the reasons for using low density panels, theoretical aspects of imputation are introduced and several dedicated software are compared. The various ways of measuring imputation accuracy are presented. Considering the possible use of a commercial 3K chip, a first study measured the impact of using imputed genotypes on reliability of GEBV. Results are presented in a first article: **Dassonneville R., R.F. Brøndum, T. Druet, S. Fritz, F. Guillaume, B. Guldbrandtsen, M.S. Lund, V. Ducrocq, G. Su. 2011. Impact of imputing markers from a low density chip on the reliability of genomic breeding values in Holstein populations. J Dairy Sci 94 :3679–3686** (article I).

Chapter 2 aims at defining the most adapted low density panel. Considering some deficiencies of the initial commercial product, two alternative in silico chips optimized for markers allelic frequencies and spacing, are proposed. Their imputation accuracy is compared to commercial chip. Results for several breeds are presented in a second article: **Dassonneville R., S. Fritz, F., V. Ducrocq, D. Boichard. 2012. Short Communication: Imputation performances of three low density marker panels in beef and dairy cattle. J. Dairy Sci. 95:4136–4140** (article II). Taking into account the room for improvement for the low density panel, a new chip was designed and later commercialized. Its design is described in the third article: **The Bovine LD Consortium. Boichard D., H. Chung, R. Dassonneville, X. David, A. Eggen, S. Fritz, K. J. Gietzen, B. J. Hayes, C. T. Lawley, T. S. Sonstegard, C. P. Van Tassell, P. M. VanRaden, K. A. Viaud-Martinez, G. R. Wiggans. 2012. Design of a Bovine Low-Density SNP Array Optimized for Imputation. PLoS ONE 7(3): e34130** (article III).

In Chapter 3, after introducing historical aspects related to the bias induced by preferential treatment and its possible impact over genomic predictions, we describe an experimental design which was set up in order to properly measure the bias induced by preferential treatment on performances of genotyped cows. Results are presented in a fourth article: **Dassonneville R., A. Baur, S. Fritz, D. Boichard, V. Ducrocq. 2011. Inclusion of cows performances in genomic evaluations and its impact on bias due to preferential treatment. Submitted** (article IV). After the existence of a bias was demonstrated, some possible solutions to this incoming problem were described and compared.

A discussion on the expected benefits of genotyping females follows in Chapter 4. First the theoretical background of the measure of genetic gain is detailed, and then several simulations studies are reviewed. Major conclusions regarding the genotyping of bull dams are drawn. A survey considering the potential returns on investment for a farmer is outlined. Possible interactions with reproduction practices and herd management decisions are considered with special focus on the genetic improvement of health traits.

CHAPTER 1 – Imputation and low density SNP chip

1.1. Presentation of low density panels and imputation

1-1.1. THE NEED OF A CHEAP LOW DENSITY SNP PANEL

The use of Bovine SNP50 ® chip has been a huge success in dairy cattle and hundreds of thousands of genotypes were performed across the world with this tool. This success is not only explained by the possibility to double the genetic gain offered by genomic selection, but also the fact that the cost of the chip is considerably low compared with the cost of progeny-tested bulls. In Europe, it is usually considered that one progeny-tested bulls costs around €40,000. Schaeffer (2006) reported a value of $50,000 per bull in Canada. The Bovine SNP50 chip cost $208 in 2009 (official price) and this price has been divided by 2 since then. The complete total price for a genotyping (including chip price and lab costs) was around $500 and keeps decreasing. Genotyping dozens of male candidates and selecting the best of them as new AI bull clearly appears as a nice opportunity for breeding companies to reduce costs, compared to progeny testing.

However, this chip may not be the optimal tool to genotype females. Potential bull dams may be genotyped by AI centers, and their high genetic merit may justify a more expensive genotyping. But the large amount of females in commercial herds may not benefit from this chip because of its price. For this reason, and considering the potential market, the Illumina Company developed a cheaper genotyping tool. This chip contained fewer markers and was more affordable. The first low density chip (Golden Gate GG3K) was launched in 2009 and contained 2900 SNP. It was studied in article I. Another SNP panel arose in 2011 and replaced the GG3K. This new Bovine LD chip contains around 6900 SNP (see article III). Both low density chips were released at a much lower price ($33) compared to the standard Bovine SNP50 (see Figure 1).

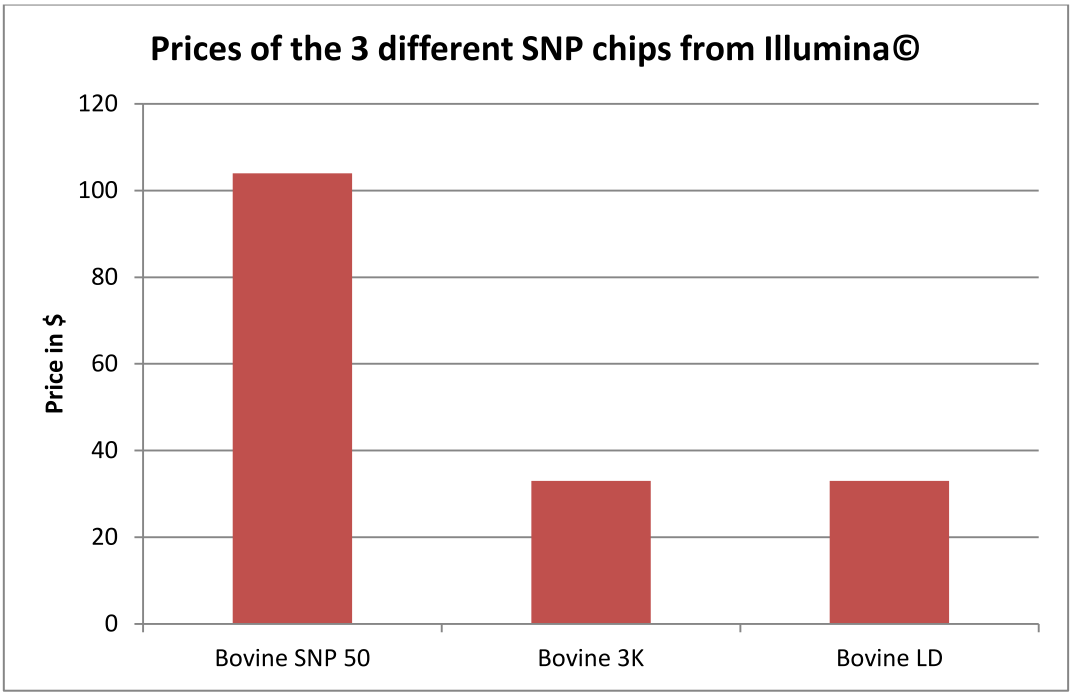


Figure 1: Official 2012 prices of the 3 different Illumina SNP chips developed for cattle

1-1.2. CRITERIA TO SELECT MARKERS TO INCLUDE IN THE LOW DENSITY PANEL

From the standard 50K SNP panel, there are 2 obvious ways to select a subset of markers to create a new low density chip. The first one consists in considering the best markers (SNP with the largest effect) for a given trait. The trait of interest could be the total merit index. The second approach makes use of linkage disequilibrium across chromosome segments and considers evenly spaced markers.

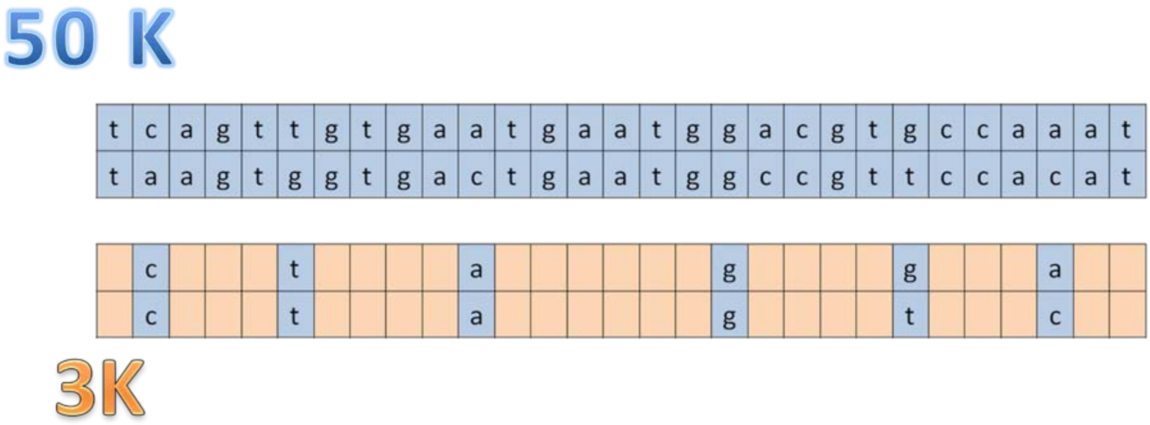
The main drawback of the first approach is that it would lead to different SNP panels for different breeds. SNP effects for a trait are usually not consistent across breeds. It would also lead to different choices of markers subsets for different traits (different selection objectives). This would imply several SNP chips which is not compatible with the cost reduction initially sought: to decrease the cost of the SNP chip, it is required to produce and sell a large amount of the same product.

Weigel et al. (2009) or Moser et al. (2010) compared these 2 approaches. Results from the best markers were (as expected) slightly better for the trait of interest, however differences were small and imputation of missing markers (between evenly spaced markers) was not performed. Indeed, the genomic evaluation carried out in their studies only considered a limited number of SNP, and did not take full advantage of linkage disequilibrium between missing markers and genotyped markers of the low density panel. Indeed, some specific statistical methods can be used to fully benefit from low density panels by reconstructing genotypes at the standard density.

1-1.3. Defining imputation

Imputation consists in pred_cting mis_ing l_tt_rs w_thin wo_ds or se_t_nces. It us_al_y reli_s on s_mple r_les. In statistics, imputation is the substitution of missing data by the most likely value. In genetics terms, imputation can be defined as the estimation of unmeasured genotypes.

Imputation requires 2 distinct data sets, including genotypes of different individuals and corresponding to 2 different marker panels (Figure 2). Some markers are usually included in both panels in order to create a link between the 2 data sets (actually, the more markers in common, the more accurate the imputation). Pedigree information (relationships between individuals) and the genetic marker map (position of markers on the genome) bring additional information.



t	c	a	g	t	t	g	t	g	a	a	t	g	a	a	t	g	g	a	c	g	t	g	c	c	a	a	a	t
t	a	a	g	t	g	g	t	g	a	c	t	g	a	a	t	g	g	c	c	g	t	t	c	c	a	c	a	t

	c				t					a							g					g				a		
	c				t					a							g					t				c		

Figure 2 Diagram describing imputation. The rows correspond to sequences of bases (a,c,g,t) on the paternal and maternal haplotyes. Imputation consists in filling in the gaps (in orange here) of the low density using information extracted from high density

1-1.4. Statistical basis of imputation

Imputation usually relies on hidden Markov model. A Markov model is a stochastic process that assumes that the conditional probability distribution of future states depends only upon the present state.

Let us consider a series of markers present on a given chromosome segment and distributed according to their position on the genetic map. The Markov property implies that only the information at a marker n is required to impute the genotype at marker n+1.

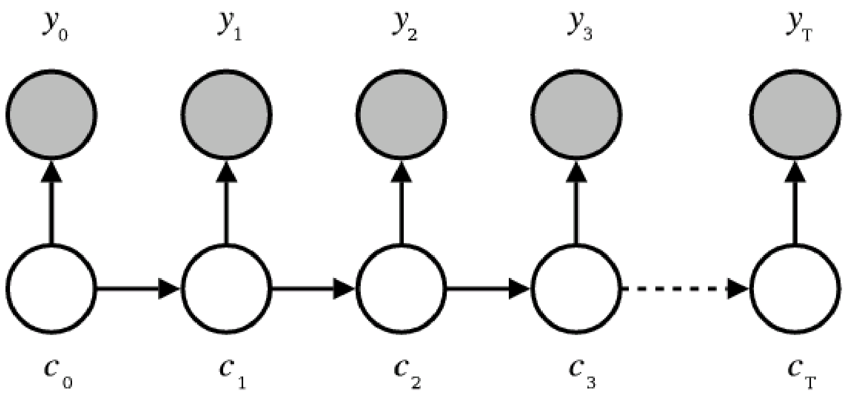


Figure 3 A hidden Markov model. c0, c1, ... cT follow the Markow property : to determine c3, only the information at c2 is necessary. In the case of hidden Markov model, c0 to cT are not observable. y0, y1 ... yT represent the observation. Observation y 3 only relies on the hidden state of c3.

Instead of directly considering the measurable variable (here the genotype), we consider a hidden variable which could correspond to ancestors' haplotypes following the Markov property. When we deal with SNP, the information at each single marker is binary, so it is very limited. Allowing a larger number of different states at every position gives more flexibility to the statistical tool to sum up the information corresponding to the previous positions. For a given position at a locus n, a hidden state is assigned based on the information gathered in the hidden state at the position n-1. This hidden state (the haplotype for instance) at position n corresponds to one state of the observed variable (here the genotype) at position n.

During the process of imputation, hidden Markov model usually creates a mosaic structure (see Figure 4).

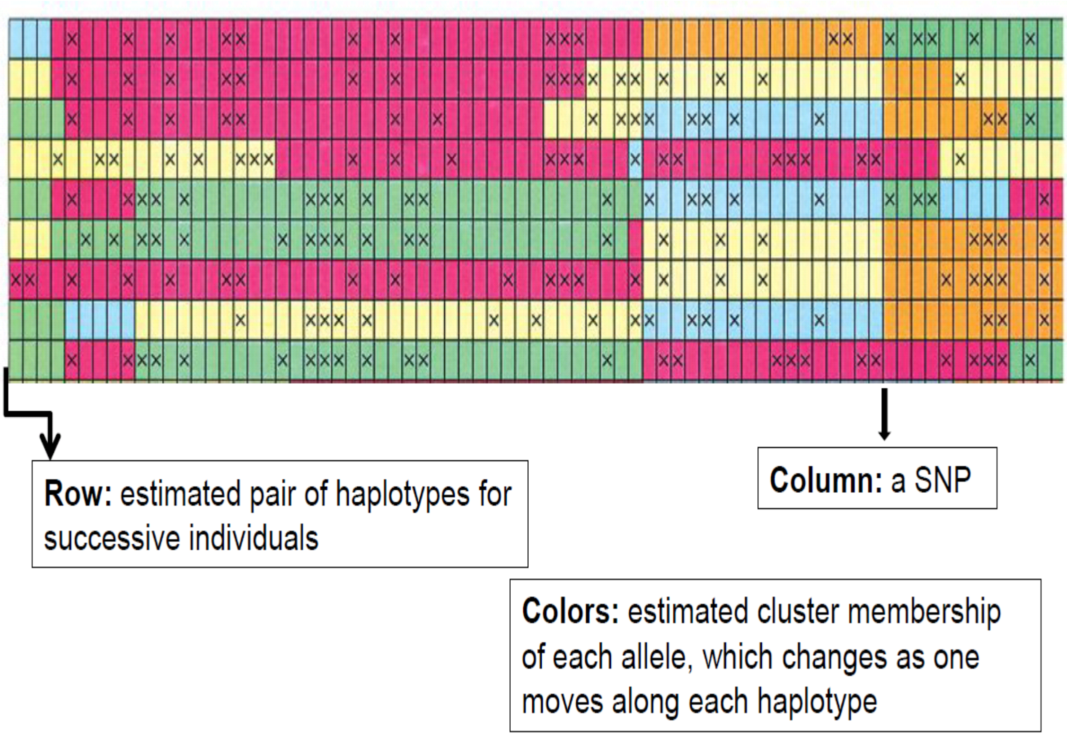

Row: estimated pair of haplotypes for successive individuals
Column: a SNP
Colors: estimated cluster membership of each allele, which changes as one moves along each haplotype

Figure 4 Mosaic structure obtained with an imputation software based on a hidden Markov model (from Scheet and Stephens, 2006). For each individual correspond 2 rows (diploid organisms). And every marker is represented by a column. The color of each spot refers to the hidden state. X correspond to one allele of the SNP when blank correspond to the other allele, defining the observed variable (i.e. the genotype). Colors can be seen as ancestors' haplotypes.

1.2. The different imputation software

1-2.1. THE FIRST EXPERIENCE OF IMPUTATION IN HUMAN GENETICS WAS BASED ON POPULATION LINKAGE DISEQUILIBRIUM

As often in genetics, human genetics are a few steps ahead. As the first genome to be sequenced, the human genome was studied using different SNP chips. In order to aggregate data coming from different studies, the need for imputation arose. The main focus while using genomic data in human genetics is association studies. Most of the time, the aim is to find a causal mutation in a gene involved in a specific disease. The data includes individuals from small families (compared to large half-sibs families encountered in dairy cattle) and may come from sub-populations usually not related to each other. For these reasons, imputation is based on the linkage disequilibrium observed at the population level.

Two software specifically dedicated to imputation of human data are briefly presented here. There exist many more but these two became standards and are heavily used worldwide.

The first one is fastPHASE, developed by Scheet and Stephens (2006), derived from PHASE (Stephens et al. 2001) allowing the analysis of larger data sets. It is based on a hidden Markov model. The idea is that over short regions, haplotypes tend to cluster into groups. The model specifies a given number K of unobserved states (cluster of haplotypes). The mosaic structure seen in Figure 4 is produced by fastPHASE. Colours can be seen as founders' haplotypes segregating in the population. The software can be used for both haplotyping and imputation. For imputation, the best guess is sampled from the conditional distribution of the observed genotype given the hidden state. Haplotyping consists in allocating the paternal or maternal origin of a given chromosome segment.

The second one is Beagle, developed by Browning and Browning (2007) and heavily used in the field. It is also based on an hidden Markov model. It has some similarities with fastPHASE, the main difference being that the haplotype-cluster model is localized. While the number K of unobserved states is required as an input in fastPHASE, and remains the same all along the chromosome, this value can differ for every marker position in Beagle. A Directed Acyclic Graph (DAG) is produced and summarizes the LD pattern. It gives the different emission probabilities from one hidden state at one marker position to the possible hidden state at the next position (see Figure 5). The DAG (which can be seen as a special kind of tree where branches can merge) is simplified using the Viterbi algorithm. In Beagle theory, recombination is modeled as merging edges.

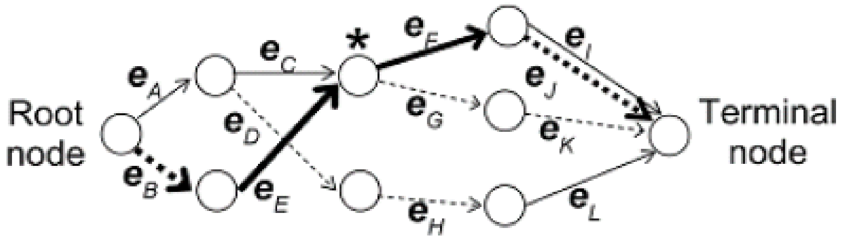


Figure 5 Example of a DAG representing localized-cluster model for 4 markers. For each marker allele 1 is represented by a solid line and allele 2 by a dashed line. From Browning and Browning (2007)

Note: The Viterbi algorithm, initially developed to remove noise in communication, is also used in bio-informatics nowadays. It allows to simplify a tree -such as a hidden Markov model, or here, the DAG- while constructing it, matching similar edges.

There exist many other software, such as IMPUTE (Marchini et al. 2007), and several are well described in a review from Marchini and Howie (2010).

1-2.2. IMPUTATION SOFTWARE SPECIFICALLY DEDICATED TO ANIMAL POPULATIONS, BASED ON LINKAGE AND MENDELIAN SEGREGATION RULES

Human population and livestock populations are much different (in terms of LD for instance) but the data sets built for genetics studies are even more different. For both species they are based on SNP data. But human SNP chips include many more markers (usually around one million of SNP) while livestock population are studied with medium density SNP chip (dozens of thousands SNP). The main difference consists in the “depth” of the pedigree. Human genetic data sets are usually built on small families or even unrelated individuals. On the other hand, genetics data sets of livestock populations are usually formed of related individuals across several generations. For example, in dairy cattle, very large half-sibs families are studied across several generations.

Considering these differences, some geneticists specialized in animal populations developed specifically dedicated software that take advantage of the very specific family structure of livestock populations. Indeed, when one or both parents are genotyped, some specific rules can be applied in order to partially determine with certainty the genotype of the offspring.

Those rules are based on Mendelian segregation. The simplest example is when the sire is homozygous at one locus. One knows for sure the paternal allele of the offspring.

Druet et al. (2008) described such a strategy. The software, which was not publicly released at first (it is now included in a package) is called linkPHASE. It follows 4 different steps. First, homozygous alleles are assigned to both (paternal and maternal) phases. Then, markers that can be unambiguously determined (based on homozygous markers of parents for instance) also are assigned. Then anchoring markers (heterozygous markers for parents for which phases offspring phases are already determined) are defined. They are used as informative flanking markers. Emission probabilities are then computed based on genetic distances. Finally, the most probable haplotypes are assigned if they reach a given probability threshold (95%). This method is very fast, and takes full advantage of the family structure of livestock data sets.

	linkPHASE	*fastPHASE* or *Beagle*
developed for	livestock populations	human population
data usually used	large half-sibs families	unrelated individuals
account for	linkage	LD (linkage disequilibrium)
based on	Mendelian segregation rules and informative flanking markers	hidden Markov model
markers it was initially developed for	microsatellites (multi-allelic)	SNP (bi-allelic)
accuracy	medium to high	excellent at high density
speed	very fast	slow (especially when data includes thousands of individuals)

Table 1 Main differences between linkPHASE and fastPHASE/Beagle

1-2.3. Combining the 2 sources of information: population based linkage disequilibrium and linkage using family information

Both approaches (hidden Markov model based on population LD in human genetics and mendelian segregation rules and linkage based on family structure in livestock populations) are not exclusive. Traditional human genetics approaches can be used to impute livestock SNP data. However they are slow. Moreover, it is a pity not to take into account simple Mendelian segregation rules to determine unambiguously an important fraction of the genotypes to be imputed. Accounting for family information can also avoid some errors when the parent carry rare haplotypes in the population. Druet and Georges (2010) proposed a method to combine these 2 sources of information. It is included in a package called PHASEBOOK. In the imputation process they proposed, Mendelian segregation rules first are applied. Depending on the value of the probability threshold, linkage information can be used to assign the most probable haplotype for some regions (given informative flanking markers). This is done with the linkPHASE software and partially reconstructed genotypes are obtained. Then either the fastPHASE probability model or the Beagle probability model are used (with some modifications) to exploit linkage disequilibrium. This is performed using programs called dualPHASE or DAGphase. For instance, DAGphase exploits the DAG created by Beagle on a base population. To sum up and simplify, this method is very similar to a regular hidden Markov model (as used in human genetics). The main difference is that the input file includes genotypes partially reconstructed using family information. It is a way to speed up the process with no loss in accuracy since only genotypes assigned with certainty are added.

There exist other kinds of methods to perform imputation. One can think of long range phasing. Initially proposed by Kong et al. (2008), it considers that long identity by descent blocks may be inherited from a hypothetical common ancestor, even for unrelated individuals. In practice, “libraries” of long haplotype blocks are created based on genotypes of the training population. Hickey et al. (2011) proposed such an approach adapted to dairy cattle populations with a software called Alphaimpute. One possible drawback is that some of these software leave some missing markers uncalled. Depending on the genomic evaluation method used, this downside may unable the use of such an imputation method. Such software were not studied in this document. Johnston et al. (2011) compared several different software on dairy cattle data. Long range phasing methods are more recent and one may consider that they are really promising. One could find a way to combine this approach with other sources of information in order to speed up the process and to increase even more accuracy.

Findhap is a software developed by VanRaden et al. (2011) that combines pedigree reconstruction of genotypes and population haplotyping. It uses libraries where long block of haplotypes are sorted according to their frequency in the training population, and checks whether the most frequent haplotypes fit with the low density genotypes and nearby markers in order to impute missing SNP.

Imputation clearly appears as a very interesting approach: it fully benefits from the links (either at the population or the family level) that exist between individuals genotyped at the various densities to predict complete genotypes from lower density genotypes at a reduced cost.

1.3. Measures of imputation accuracy

There are two kinds of methods to assess imputation accuracy. The first one is to double genotype the same individuals. The animals are genotyped on both the low density and high density chips and imputed SNP are compared with real ones. The main advantage of such a method is that it is the most realistic, and it accounts for both genotyping and imputation issues. However this technique has two main drawbacks: first it is more expensive (2 genotypes per animal). Second, it requires that the studied chip already exist technically, whereas sometimes, one just want to test a given set of markers to determine what the imputation accuracy would be if such a chip were developed.

For these reasons, another way to measure imputation accuracy is to simulate in silico low density genotypes. Only high density genotypes are used. Then, a low density genotype is created by erasing markers that are not present on the low density chip (but present on the initial high density chip). It is possible to test as many different chips as required. Differences observed between imputed and real genotypes only result from imputation mistakes and are not due to genotyping errors of common markers of the 2 chips.

The second option was chosen in our studies. It is also the most used in the literature.

- **Direct measures obtained only comparing true and imputed genotypes.**

The advantage of simulation study is to know the "true" genotype for a given marker, present on the high density chip, and missing on the low density chip. The genotypes obtained after imputation can be compared with this "true" genotype.

1-3.1. Error rate or Concordance rate

The easiest measure of imputation accuracy is to count the number of markers for which true and imputed genotypes differ. This number, divided by the total number of missing markers, gives a ratio, usually expressed in %, called the imputation error rate.

This measure was used by e.g. Zhang and Druet (2010), Dassonneville et al. (2011).

Simply derived from the error rate, the concordance rate relates to the proportion of missing markers that were correctly imputed, meaning the number of missing markers for which true and imputed genotypes are the same divided by the total number of missing markers.

The concordance rate appears to be a more optimistic way of presenting the same result: 95% of markers correctly imputed sounds better than 5% of markers incorrectly imputed.

This measure was used by e.g. Weigel et al. (2010), Dassonneville et al. (2012).

One can easily get the concordance rate from the error rate figures (or the other way around) as:

concordance rate = 100 – error rate (or 1 - error rate when not expressed in%).

1-3.2. Counting per genotype or per allele

There are 2 ways of considering imputation results (or genotyping results) for a given marker:

- Considering the genotype

During the genotyping procedure, fluorescence is used to assign individuals' genotypes for a given SNP to 3 clusters, corresponding to the 2 different homozygous possibilities, and one for the heterozygous. One possible way of measuring imputation is to check whether the genotype was properly assigned after imputation to the correct group. Easily, and as stated in Table 2, if the imputed and true genotypes are different, then the imputed genotype is considered as wrong.

- Considering the 2 alleles

One may also consider that the genotype for one locus actually hides 2 separate sources of information corresponding to the 2 different alleles (inherited from the 2 parents). One imputed allele (of one haplotype, either the paternal or the maternal one) may be right while the other one may be wrong. Considering either the 2 right or the 2 wrong as one error may appear simplistic then.

case	true genotype	imputed genotype	# of genotype errors	# of allele errors	ratio of genotype errors	ratio of allele errors
1	homozygous	(same) homozygous	0	0	0/1	0/2
2	homozygous	heterozygous	1	1	1/1	1/2
3	homozygous	opposite homozygous	1	2	1/1	2/2
4	heterozygous	heterozygous	0	0	0/1	0/2
5	heterozygous	homozygous	1	1	1/1	1/2

Table 2 genotype and allele error rates for all the different cases which can be observed when comparing true and imputed genotype.

Obviously, when imputation is correct, i.e. when true and imputed genotypes are identical (cases 1 and 4 in the table), the error rate is 0 and both genotype or allele measures are identical. We expect this situation to be the most frequent.

When the true genotype is homozygous and the imputed genotype is heterozygous, or vice-versa (cases 2 and 5), the error rate is 1 for the genotype measure, and ½ for the allele measure. It is the main difference between the 2 measures. This situation is expected to be the main source of errors.

When true and imputed genotypes are opposite homozygous (case 3), the error rate is 1 for both measures. However, if the error rate p is small, the probability of such an error is even smaller, as it is proportional to p^2.

Considering the last situation as rare, one can approximate the genotype error rate as twice the allele error rate. This approximation can be used to “translate” results from one study to be compared with another one.

Some authors (e.g. Weigel et al., 2010) prefer to use genotype error rate. They argue that allele error rate is a way to present better results, as the error rate appears lower (or concordance rate appears higher). As Druet et al. (2010), we chose to consider the allele error rate. Imputed genotypes will be used in genetic evaluations, which are based on an additive model. For this reason, if the imputed genotyped for one marker is half correct (cases 2 and 5), we can expect the resulting genomic evaluation as half right, and not completely wrong. This is the reason why we prefer to use allele error rate.

1-3.3. Correlation between true and imputed genotypes

Most of the authors report concordance or error rates, based on alleles or genotypes. These measures are simple to calculate, and easily derived from the other one. There exist some alternative measures that present some better properties. Hickey suggested to use correlation between true and imputed genotypes in order to account for MAF.

As we stated in part 2.1. , “Using concordance rate as imputation efficiency criterion may be misleading as it depends on MAF. The lower the MAF, the higher the concordance rate, for the same efficiency and this should be accounted for in the interpretation. For example, with an average MAF of 0.2, 80% of the results would be correct after random sampling of the missing alleles. Getting a 95% concordance rate corresponds only to a 75% (= (95-80)/(100-80)) imputation efficiency. Correlations are an alternative criterion less dependent on MAF.“

One hidden message is that, a marker with low MAF presents less variation across individuals, bringing less information to the genomic model, therefore good concordance rate for such marker is misleading when one wants to predict the information brought from imputed genotypes to the evaluation model. In the study of Hickey et al. (2012), performed on maize data, they plotted both concordance rate and correlation as a function of MAF (Figure

6). One may thereby observe differences between the 2 measures, low MAF markers present high concordance rate but low correlation.

We chose to report the 2 kind of measures.

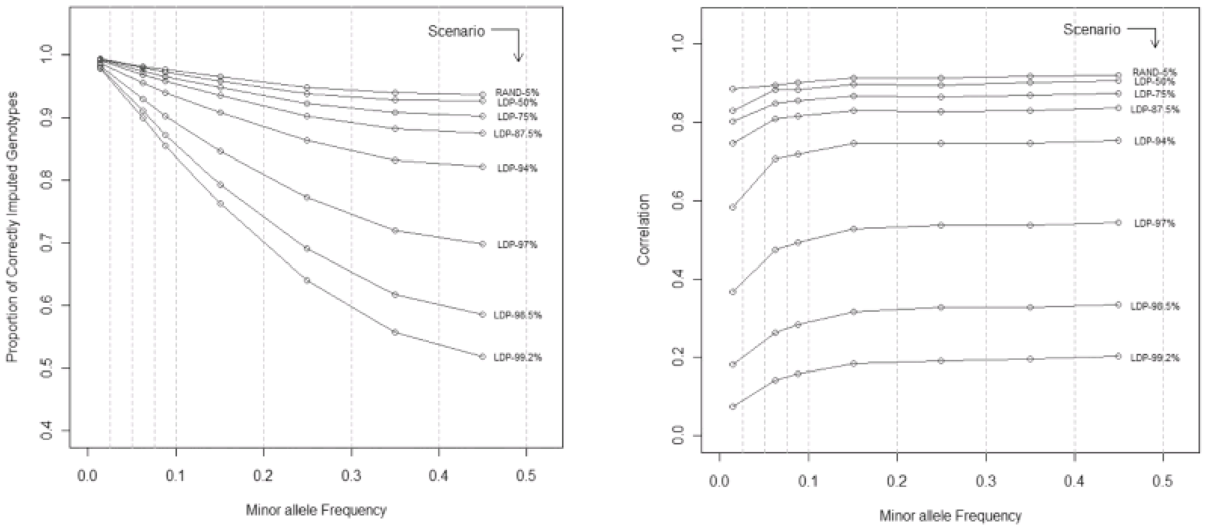


Figure 6 (from Hickey et al. 2012) 2 measures of imputation accuracy depending on MAF of markers. On the graph on the left is represented the concordance rate whereas on the right is presented the correlation.

1-3.4. Comparing phases or genotypes ?

The French genomic evaluation model relies on haplotypes (Boichard et al., 2012). These haplotypes are derived from the phased genotypes obtained after imputation. One may want to measure imputation accuracy looking at the phases produced. However it is very difficult to compare phases outputs. For the same genotypes, it is possible to count the number of switches (Druet, personal communication), i.e. the number of apparent recombinations on the imputed phase but not present on the “true” phase. An additional drawback is that this measure requires to “know” the “true” phases. But phases can only be obtained from genotype data after running a phasing software (and this software may also induce some errors).

1-3.5. Measuring consequences of imputation errors on genomic evaluations

- Graphical representation of G matrices

- The G matrix

Many countries (ref) have chosen the GBLUP approach to implement genomic evaluation. One possibility is to use a G matrix as in VanRaden (2009). Every line/column corresponds to an individual, and based on marker information, a “genomic relationship” is calculated between every 2 individuals. Then, this G matrix can be used within the mixed model equations as if it were the A matrix (which is based on pedigree relationships).

- The G matrix as a measure of imputation accuracy

When one wants to compare genomic evaluations based on true or imputed genotypes using the GBLUP method, the G matrix has first to be derived. Zukowski (personal comunication) propose to directly compare G matrices (either the “true” one or the one based on imputed genotypes). Figure 7 proposes a graphical representations of the G matrices obtained with several imputation methods. Graphical representations are not as easy to compare as numerical values. However, they give a nice and quick overview. It is also to measure distances between matrices.

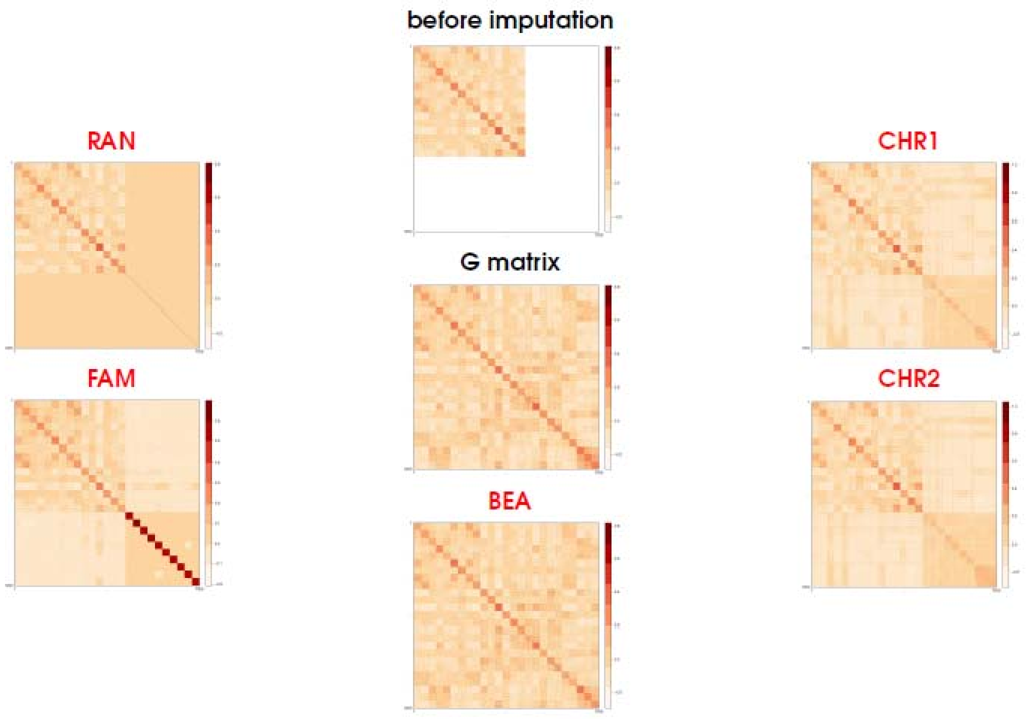


Figure 7 graphical representations of the G matrices obtained with several imputation methods (from Zukowski) The top matrix only represent relationships between individual of the training population (before imputation). The matrix in the middle is the "true" G matrix. The top left matrix has been computed after random imputation (RAN) of the missing markers for the validation population. The matrix down on the left is obtained after imputation based on family structure (FAM, Wimmer et al., 2012). The 2 matrices on the right are obtained after imputation of missing markers using ChromoPhase.f90 (CHR1 and CHR2, Daetwyler et al., 2012). The matrix on the bottom (BEA) is based on imputed genotypes performed using Beagle software.

With no surprises, random imputation gave poor results but can be considered as a "control". Chromophase and Beagle gave the best imputation accuracy figures (results not shown). However, comparing graphical representations clearly shows that the matrix most "similar" to the "true" G matrix is the one obtained after Beagle imputation.

While here we want to predict relationships between individuals, Beagle which is only based on population linkage disequilibrium (and does not take into account complex pedigree relationships) clearly better predicts "genomic relationships" between individuals. This illustrates how well Beagle works, and is consistent with the better imputation of genotypes of individuals closely related to the reference population.

1-3.6. Comparing genomic evaluations based on imputed genotypes

In animal breeding, the main use of imputed genotypes is to include them in genomic evaluation. Obviously, the best way to measure the consequences of using imputed genotypes on genomic selection is to run genomic evaluations on both imputed and true genotypes and compare the DGV or GEBV obtained. Phenotypes and complete genotypes of the training population are used, as well as imputed genotypes for the validation population. Genomic breeding values based on imputed and true genotypes can be compared. Moreover, they can be compared to phenotypes (DYD or deregressed proofs of validation animals). One can then estimate the loss in reliability when using imputed genotypes instead of true genotypes.

1.4. Article I Impact of imputing markers from a low density chip on the reliability of genomic breeding values in Holstein populations

1-4.1. Background

In 2010, Illumina developed a new genotyping tool : the Bovine 3K Beadchip ® based on the GoldenGate technology (GG 3K). The cost of this SNP chip was reduced by 68% compared to the standard Bovine SNP50®. This did not mean that the total genotyping cost was substantially reduced since lab costs remained more or less the same. The main target of this product was the large population of females. The use of a cheap tool was thought to be crucial to launch female genotyping as a new service.

Also in 2010, a European consortium, Eurogenomics, was formed. It covered 4 large Holstein populations (Dutch-Flemish, French, German and Nordic - Denmark, Finland, Sweden -). Its members decided to join their reference population (15,966 progeny-tested genotyped bulls) in order to achieve more reliable genomic predictions. Their scientific partners also decided to cooperate in order to conduct common studies. This article is one of such studies.

1-4.2. Objectives

The purpose of the study is to measure imputation accuracy and to quantify the loss in reliability when using imputed genotypes instead of real genotypes. The aim is to determine whether such a low density tool can be used in practice.

The novelty of the paper (among others studies on imputation) lies on the fact that genomic evaluations based on imputed genotypes were calculated and their quality was compared to the situation without imputed genotypes. That was done using 2 different methods for genomic prediction (GBLUP in Nordic countries, GMAS in France) considering 4 different traits. It was also desired to compare the loss in reliability due to the use of low density panels to the gain achieved when increasing the reference population size.

Article I

Dassonneville R., R.F. Brøndum, T. Druet, S. Fritz, F. Guillaume, B. Guldbrandtsen, M.S. Lund, V. Ducrocq, G. Su. 2011

Impact of imputing markers from a low density chip on the reliability of genomic breeding values in Holstein populations

J Dairy Sci 94 :3679–3686.

ABSTRACT

The purpose of this study was to investigate the imputation error and loss of reliability of direct genomic values (**DGV**) or genomically enhanced breeding values (**GEBV**) when using genotypes imputed from a 3K single nucleotide polymorphism (**SNP**) panel to a 50K SNP panel. Data consisted of genotypes of 15,966 European Holstein bulls from the combined EuroGenomics reference population. Genotypes with the low density chip were created by erasing markers from 50K data. The studies were performed in the Nordic countries (Denmark, Finland, and Sweden) using a BLUP model for prediction of DGV and in France using a genomic marker assisted selection approach for prediction of GEBV. Imputation in both studies was done using a combination of the DAGPHASE 1.1 and Beagle 2.1.3 software. Traits considered were protein yield, fertility, somatic cell count and udder depth. Imputation of missing markers as well as prediction of breeding values were performed using two different reference populations in each country; either a national reference population or a combined EuroGenomics reference population. Validation for accuracy of imputation and genomic prediction was done based on national test data. Mean imputation error rates when using national reference animals was 5.5% and 3.9% in the Nordic countries and France, respectively, whereas imputation based on the EuroGenomics reference dataset gave mean error rates of 4.0% and 2.1%, respectively. Prediction of GEBV based on genotypes imputed with a national reference dataset gave an absolute loss of 0.05 in mean reliability of GEBV in the French study, whereas a loss of 0.03 was obtained for reliability of DGV in the Nordic study. When genotypes were imputed using the EuroGenomics reference a loss of 0.02 in mean reliability of GEBV was detected in the French study, and a loss of 0.06 was observed for the mean reliability of DGV in the Nordic study. Consequently, the reliability of DGV using the imputed SNP data was 0.38 based on national reference data, and 0.48 based on EuroGenomic reference data in the Nordic validation, and the reliability of GEBV using the imputed SNP data was 0.41 based on national reference data, and 0.44 based on EuroGenomic reference data in the French validation.

Key words: Genomic selection, imputation, reliability, reference population.

INTRODUCTION

Genomic selection (Meuwissen et al., 2001) is becoming a routine tool for genetic evaluation in dairy cattle breeding. Currently, an SNP panel with 54,000 markers is widely used. A new low density panel with only 3,000 markers at a lower price potentially reducing genotype costs is now also available (Illumina, San Diego). Using the low density panel instead of the current one may allow cattle breeders to genotype more bulls and cows.

Several options for selecting a low density panel have been suggested. One option is to select a number of markers with large effects for a given trait, another is to use markers evenly spaced across the genome. Previous studies showed that the difference in reliability of the genomic breeding values, when using 3,000 markers with large effect or 3,000 markers evenly spread across the genome, is small (Moser et al., 2010). The option of evenly spaced markers removes the need for trait and breed specific low density SNP panels. The efficiency of a trait specific marker panel also depends on the linkage disequilibrium (**LD**) between the markers with large effect and the actual QTL. This LD might decline through generations. The other advantage of evenly spread markers is the possibility to use statistical methods to impute the missing markers, thus extending the 3,000 markers to 50,000 markers albeit with some uncertainty. This is also possible with unevenly spread markers, but then the accuracy of imputation is expected to be lower.

It has been reported that a lower marker density leads to lower reliability of genomic prediction (Moser et al., 2010). A feasible strategy is to extend the low density markers to the current 50K markers by imputation. Several methods for imputation of SNP markers, relying on either linkage based on family information (Daetwyler et al., 2010) or LD based on population information (Browning and Browning, 2007; Scheet and Stephens, 2006), have been proposed. It is also possible to combine both types of information (Druet and Georges, 2010). In a study using this combined approach to impute from 3,000 to 50,000 markers, where the 3,000 markers were specially selected for high minor allele frequency, Zhang and Druet (2010) found an allele error rate, i.e. the proportion of incorrectly predicted alleles, of approximately 3%. A study by Weigel et al. (2010) on American Jersey cattle has shown that using 3,000 SNPs for candidates imputed to a 50K SNP panel can provide approximately 95% of the predictive ability achieved using the real 50K SNP panel.

The accuracy of imputation can be increased by increasing the size of the reference population. EuroGenomics is a collaboration between four European AI companies and scientific partners: DHV-VIT (Germany), UNCEIA-INRA (France), CRV (Netherlands, Flanders) and Viking Genetics-Aarhus University (Denmark-Finland-Sweden). The collaboration includes the sharing of reference populations for genomic selection, where each country initially contributed 4,000 genotyped Holstein bulls with progeny tested breeding values. A previous study showed a significant increase in reliability of genomic breeding values using this combined reference population (Lund et al., 2010). We expect that the accuracy of imputation based on EuroGenomics reference data will be higher than that based on national reference data.

The objective of this study is to investigate the imputation error, when imputing from a 3K SNP panel to a 50K SNP panel using a group of reference animals with 50K information. The 3,000 markers were the same as the Illumina 3K SNP panel. The imputed SNP markers were used for genomic prediction to assess how the imputation error rate affects the reliability of genomic breeding values and the ranking of the animals. This assessment was carried out in the Nordic countries and France. For both analyses, a validation population consisting of national test animals with 3K genotype was imputed to 50K genotype using a reference population made of either national or EuroGenomics data.

MATERIAL AND METHODS

DATA

The combined EuroGenomics reference population contains 15,966 progeny tested bulls with genotypes from the Illumina Bovine 50K SNP panel (Matukumalli et al., 2009). 4,000 Dutch bulls were genotyped using a customized CRV 60K chip, but by double genotyping 972 influential bulls with the Illumina 50K chip, it was possible to impute markers from the Illumina chip for all Dutch bulls with an imputation error of less than 1% (Druet et al., 2010).
Measurement of imputation error rate and reliability of genomic predictions for Nordic and French bulls were carried out separately, using either national or EuroGenomics reference data. Deregressed proofs (**DRP**) on the scale of the target population calculated from Interbull 2010-01 MACE proofs were used for predicting and validating DGV and GEBV, if the equivalent daughter contribution (**EDC**) was at least 20 (Lund et al., 2010). In the French study, daughter yield deviations (**DYD**) from the October 2009 national evaluation were used as phenotypes for the French bulls. The reference and validation populations were divided

according to the bulls' birth date. The cut-off dates were October 1, 2001 and June 13, 2002 in the Nordic and French case, respectively. Thus, about 25% of national genotyped bulls were taken as a validation set.

The traits studied were protein yield, somatic cell count (**SCC**), fertility (defined as non return rate (**NRR**) in the Nordic countries and conception rate (**CR**) in France), and udder depth. Heritabilities and number of animals available for the specific traits are shown in Table 1.

Table 1: Heritabilities (h^2) and number of animals used for protein yield, somatic cell count (SCC), fertility and udder depth (UD) in the Nordic and French study.

	Nordic				**French**			
Trait	**h^2**	**Nordic reference**	**Euro reference**	**Nordic validation**	**h^2**	**French reference**	**Euro reference**	**French validation**
Protein	0.39	3,038	10,701	899	0.3	3,071	12,078	966
SCC	0.15	3,077	10,800	899	0.15	3,071	12,078	966
Fertility	0.02	3,069	10,712	895	0.02	3,071	12,078	966
UD	0.37	2,958	10,755	900	0.36	3,071	12,078	966

Marker data were edited according to procedures used in Nordic countries and in France.

Nordic marker editing:

The genotypic data was edited both per animal and per locus. At the animal level, the requirements were a call rate above 95% except for some old animals which were accepted with call rates of at least 85%. Marker loci were accepted, if they had a call rate of at least 95% in a large reference sample. Loci with a minor allele frequency less than 5% were excluded. Loci without a known map position in the Btau 4.0 assembly or mapped on the X chromosome were discarded. Animals with an average Gen Call score (Illumina, 2005) of less than 0.65 were excluded. Individual marker typings with a Gen Call score of less than 0.6 were also discarded.

French marker editing:

The French genotypic data was first edited per locus. Markers without a known map position in the Btau 4.0 assembly, or mapped to the X-chromosome were removed. Markers were then filtered for Hardy Weinberg equilibrium (q value < 0.01). Markers with call rates below 0.85 were removed. Markers with MAF strictly equal to 0 were removed. Genotype data were finally checked for Mendelian inconsistencies between parents and offspring. Inconsistent genotypes were set to missing. Marker editing procedures differed slightly between France and the Nordic countries (including Gen Call score for example).

While checking for inconsistencies between parents and offspring, Mendelian segregation rules were also applied in order to determine marker types of ungenotyped ancestors. Inferred marker data was not complete. However, it is important for ancestors with large numbers of progeny. Thus, the French national training population included 3,071 animals with real observed marker types (Table 2) and a total of 3,505 when ancestors with imputed genotypes are included. The corresponding figures for the EuroGenomics population are 12,078 and 13,947 animals, respectively. This might help for further imputation, especially through linkage information.

Table 2: Number of animals and number of markers used.

	National		**EuroGenomics**		**No. of Markers**	
	Reference	**Validation**	**Reference**	**Validation**	**Reference**	**Validation**
Nordic	3,058	1086	10,880	1,086	38,545	2,285
France	3,071/ 3,505*	966	12,078/ 13,947 *	966	43,582	2,635

*Including bulls with partially reconstructed genotypes.

Simulating Illumina Bovine 3K Bead Chip Data

The 2900 SNPs in the Illumina Bovine 3K Bead chip are all included in the Bovine 50K chip (except for 14 markers located on the Y-chromosome). To mimic the low density chip, marker types of test animals, i.e. animals born after the cut-off date, were obtained by erasing markers from the 50K marker type (i.e. in silico chip). As 3K genotypes are simulated from 50K data, they do not account for a possibly higher genotyping error rate with the 3K chip. After marker editing as outlined above, 2,285 and 2,635 markers were kept for the Nordic and French data, see Table 2.

Imputation of missing SNP markers

Imputation of markers was done using the PHASEBOOK package (Druet and Georges, 2010) in combination with Beagle 2.1.3 (Browning and Browning, 2007). The method was applied as a stepwise procedure using both linkage and LD information. The same procedure as in Zhang and Druet (2010) was applied. First, all markers that can be determined unambiguously using Mendelian segregation rules were phased using the LinkPHASE software. In the first step, both training and test animals were included. An iterative procedure was then applied, where a directed acyclic graph (DAG) describing the haplotype structure of the genome was fitted to the partially phased data from the previous step. This was, however, only done for the reference animals. This was done for 10 iterations and then, the final DAG, the genotype file and the output from LinkPHASE (partially phased data) were used to reconstruct haplotypes and impute missing markers for both test and training animals using the Viterbi algorithm. With Beagle and PHASEBOOK, all markers are imputed, and the method does not leave any missing markers. More details on the imputation procedure can be found in Druet et al. (2010) and Zhang and Druet (2010).

Allele imputation error rate calculation

The number of errors was counted as 0 when the imputed and observed marker types were identical, 1 if the real marker type was homozygous and the imputed genotype was heterozygous (or vice versa), and 2 if real and imputed marker types were opposite homozygous. Error counting only considered markers/animals where observed marker types were not missing in the original non-imputed dataset. The error rate was calculated as the total number of errors divided by twice the number of imputed loci. This gives the number of falsely predicted alleles, which is an appropriate measure when using an additive prediction model, as in this study. For other purposes, the genotype error rate could be easily found as approximately twice the allele error rate (Zhang and Druet, 2010).

Prediction of direct genomic values in Nordic countries:

Prediction of DGV was performed using a BLUP model at SNP level (VanRaden, 2008). Specifically, the model is given by

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

Where **y** is the vector of phenotypic observations, μ is the mean, **u** is a vector of SNP effects, **e** is the random error vector and **Z**=**M**-**P** is a design matrix for the random effects. The marker matrix **M** is an m by n matrix, where m is the number of animals and n is the number of markers. Entries in the i'th row of **M** are the genotypes for the i'th animal and is given by -1 if the animal is homozygous aa, 0 if the animal is heterozygous and 1 if the animal is homozygous AA. The matrix **P** has n columns where the elements in column j are $\mathbf{P}_j=2p_j -1$, where p_j is the frequency of allele A at locus j. Subtraction of the allele frequencies standardizes the allele effects to a population mean of zero. Thus **a**=**Zu** gives the direct genomic values.

DRP were used as phenotypic values in the model. The weighting factor $r^2_{DRP}/(1-r^2_{DRP})$ was used to scale the inversed residual variance of an observation,

DGV Reliability was calculated as the weighted squared correlation between DRP and DGV, divided by the mean reliability of DRP. The weights were given by $r^2_{DRP}/(1-r^2_{DRP})$ standardized to a mean weight of 1.

Prediction of genomic breeding values in France

The French genomic prediction is an extension of the marker-assisted evaluation method by Fernando and Grossman (1989). The model is the following:

$$y = 1\mu + Zu + \sum_{i=1}^{nQTL} (h_{i1} + h_{i2}) + e$$

Where **y** is the vector of phenotypic observations, µ is the overall mean, **u** is a vector of random pedigree-based residual polygenic effects, h_{ij} is the random effect of haplotype j for QTL i, and **e** is a vector of residuals, with heterogeneous residual variances inversely proportional to EDC.

The selection of QTL included in the model was the result of a combination of 2 approaches (Boichard et al., 2010). First, dozens of QTL per trait were detected after QTL fine mapping as described below. Then, hundreds of haplotypes were chosen using the Elastic Net algorithm (**EN**).

For QTL mapping, a linkage disequilibrium linkage analysis (LDLA) combining both within-family linkage information and population-based LD was used on the EuroGenomics training population (12,078 animals) following the approach described by Druet et al. (2008). Identity by descent probabilities were calculated as in Meuwissen and Goddard (2001). The likelihood ratio test threshold to retain a QTL was arbitrarily set to 6. This resulted in the selection of 80 to 100 QTL depending on the trait. Fine-mapped QTL were traced by haplotypes of 5 flanking markers.

Then, an EN procedure was run on the French training population (3,071 animals) following the approach described by Croiseau et al. (2010). The selected SNPs were grouped into haplotypes of 3 to 5 SNPs. The two sets of haplotypes were included in the model. For computational reasons, the number of markers detected by the EN procedure included in the model was limited so that the total number of QTL was at maximum 700.

The genetic variance attached to each QTL detected through LDLA mapping was proportional to the variance estimated in the single QTL analysis. The variance explained by each haplotype selected by EN was assumed to be constant and their sum over all EN haplotypes was set to 30% of genetic variance. 442 to 693 QTL were included in the model, explaining 51 to 57% of genetic variance (Table 3).

Table 3: Number (n) of quantitative trait loci (QTL) selected for the French prediction model using either linkage disequilibrium linkage analysis (LDLA) or an elastic net (EN) procedure and percentage of allocated genetic variance (%var) for protein yield, somatic cell count (SCC), fertility and udder depth (UD).

	LDLA		EN		Total	
	n	**% var**	**n**	**% var**	**n**	**% var**
Protein	100	24	593	30	693	54
SCC	80	27	362	30	442	57
Fertility	80	21	392	30	472	51
UD	80	27	482	30	562	57

RESULTS AND DISCUSSION

Accuracy of imputation

Imputation in the Nordic population showed a mean error rate of 5.5% when using only Nordic animals as the reference set (Table 4). The extension of this reference set with the EuroGenomics animals gave an error rate of 4.0%. The same pattern was found in the French population where a French reference set gave a mean imputation error rate of 3.9 % whereas increasing the reference set with EuroGenomics animals reduced it to 2.1%. The lower error rate in the French study is likely due to the inclusion of more markers for the study on the 3K chip (2635 vs. 2285) due to different marker editing rules (such as selection on MAF), giving a denser genome coverage and a higher homozygosity. A previous study by Zhang and Druet (2010) showed that both, the number of reference animals and the number of markers in the low-density panel affect the imputation error rate. This error rate is also affected by the relationship between validation and reference animals.

Table 4: Imputation allele error rates (%) for Nordic and French test animals using a national reference population or the EuroGenomics reference population.

Test population	National		Eurogenomics	
Nordic	N	Error rate	N	Error Rate
All	1,086	5.5	1,086	4.0
Sire in ref.	795	4.5	1,039	3.8
Sire not in ref.	291	8.3	47	7.0
Sire and Maternal grandsire in ref.	650	4.3	953	3.8
French				
Sire in ref.	966	3.9	966	2.1

ref : reference population

In the Nordic study, it was observed that the mean error rate depended on whether or not the animals had their sire in the reference data, confirming that a closer relationship to the reference population reduces the imputation error rate in the test population. All of the animals in the French validation population had their sire in the reference population, and an additional step based on Mendelian segregation rules was carried out to partially reconstruct genotypes of ungenotyped ancestors. The results indicate that if low-density genotyping and imputation are widely used in the future, the imputation accuracy might decrease unless all breeding bulls are genotyped with the 50K panel.

The results in the present study are consistent with the error rates obtained by Zhang and Druet (2010) using the same method for imputation, i.e. between 2.1 and 4%. Their reference population was smaller (500 to 2000 animals) but their 3K panel was optimized according to MAF for their population, thus all markers on the 3K panel were available, whereas some were excluded during the quality control in the present study. Comparing imputation error rates based on different studies is however difficult because the relationship between training and validation populations differs, and because the number of reference individuals and the number of markers vary.

Reliability of genomic prediction

Prediction of DGV based on either true or imputed genotypes in the Nordic data (Table 5) showed that using the Nordic reference population the observed marker types had a mean reliability of DGV over the four traits of 0.41, whereas the imputed marker types led to a mean reliability of 0.38. Using the EuroGenomics data as the reference population for prediction of DGV resulted in a mean reliability of 0.54 with the observed marker types while using imputed genotypes resulted in a mean reliability of 0.48.

Table 5: Reliabilities of direct genomic values for Nordic candidates with full (50K) or imputed (3K imp) marker data for protein yield, somatic cell count (SCC), non return rate (NRR) and udder depth (UD) using either Nordic reference population (Nor-ref) or Eurogenomics reference population (EU-ref).

Trait	N	Nor-ref 50 K	Nor-ref 3K imp	EU-ref 50 K	EU-ref 3K imp
Protein	899	0.41	0.32	0.56	0.51
SCC	899	0.41	0.39	0.55	0.49
NRR	895	0.44	0.42	0.49	0.45
UD	900	0.40	0.36	0.55	0.49
Average		**0.41**	**0.38**	**0.54**	**0.48**

For the prediction of GEBV based on either observed or imputed marker types for the French validation data (Table 6), with a French national reference population and observed marker types, a mean reliability across four traits of 0.46 was obtained. The corresponding value for imputed marker types was 0.40. Using the EuroGenomics data as a training population, the mean reliability of GEBV of young animals was 0.48 and 0.46 for observed and imputed marker types.

Table 6: Reliabilities of genomically enhanced breeding values for French candidates with full or imputed marker data for protein yield, somatic cell count (SCC), conception rate (CR) and udder depth (UD) using either French reference population (FR-ref) or Eurogenomics reference population (EU-ref).

.

Trait	**N**	**FR-ref 50 K**	**FR-ref 3K imp**	**EU-ref 50K**	**EU-ref 3K imp**
Protein	966	0.40	0.32	0.37	0.36
SCC	966	0.55	0.52	0.58	0.57
CR	966	0.44	0.41	0.47	0.44
UD	966	0.45	0.40	0.51	0.48
Average		**0.46**	**0.41**	**0.48**	**0.46**

Lund et al. (2010) reported that reliabilities of genomic prediction using the EuroGenomics reference data were considerably higher than those using national data, because of the increased size of the reference data. The French validation in this study however, showed a small difference between reliabilities of GEBV predicted from the national and the EuroGenomics reference data. The small difference can be explained by the way the QTL were chosen for the prediction model. For both, the prediction model based on national reference data and the prediction model based on EuroGenomics reference data, the QTL selected using the LDLA procedure were based on EuroGenomics data, and the QTL selected using the EN procedure were based on national data. On one hand, the genomic prediction based on French data gained from LDLA based on the whole EuroGenomics population. On the other hand, genomic predictions based on EuroGenomics data were probably suboptimal since the EN procedure used only French data. The only way to properly measure the impact of increasing the reference population on genomic reliability based on real genotypes would have been to do 2 LDLA QTL mappings (as in Lund et al., 2010), but the main focus of this study was on imputation. The haplotype effects were however estimated either on EuroGenomics data or national data leading to a gain in reliability when increasing the reference population.

The patterns of differences between reliabilities of genomic predictions using observed 50K marker types and the imputed marker types were not consistent between the Nordic and French validations. The difference was smaller when using national reference data than when using EuroGenomics reference data in the Nordic evaluation, while an opposite pattern was observed in the French validation. The reasons for the inconsistent pattern were not clear. A possible reason was that the markers with high imputation error rate might give different contribution to genomic prediction when using different reference datasets. For example, MAF for the loci with high imputation error rate might be smaller (less informative) in one set of reference data, while larger (more informative) in another set of reference data.

Correlations between DGV/GEBV based on imputed or observed marker types were high (Table 7). The correlation ranged from 0.92 to 0.95 using national reference data and from 0.93 to 0.96 using EuroGenomics data in Nordic validation. Similarly, the correlations ranged from 0.91 to 0.94 using national reference data and from 0.94 to 0.97 using EuroGenomics data in the French validation. These results indicate no serious re-ranking of animals when using imputed data.

Table 7: Correlations between direct genomic values or genomically enhanced breeding values predicted using observed or imputed marker data for Nordic and French candidates for protein yield, somatic cell count (SCC), fertility and udder depth (UD) using either Nordic reference population (Nor-ref), French reference population (FR-ref) or Eurogenomics reference population (EU-ref).

	Nordic		**France**	
Trait	**EU ref**	**NOR ref**	**EU ref**	**FR ref**
Protein	0.94	0.92	0.97	0.94
SCC	0.93	0.92	0.95	0.94
Fertility	0.96	0.95	0.96	0.94
UD	0.93	0.93	0.94	0.91

CONCLUSION

Imputation of the commercially available low-density bovine 3K chip to the bovine 50K chip gave allele error rates between 2.1 and 5.5 %. The accuracies of imputation were higher when using the EuroGenomics reference datasets than when using national reference datasets. Imputation was more accurate when the sire of the candidate was genotyped on the 50K panel. Using the imputed markers for candidates, the mean reliability of DGV was 0.38 based on based on national reference data, and 0.48 based on EuroGenomics reference data in the Nordic validation, and the reliability of GEBV using the imputed SNP data was 0.41 based on national reference data, and 0.44 based on EuroGenomics reference data in French validation. Therefore, a 3K SNP chip imputed to 50K could be a feasible alternative for pre-selection of young animals. One may also consider 3K genotyping as an attractive tool for a large pre-screening of the female population.

ACKNOWLEDGEMENTS

Data for this study were provided by the Eurogenomics consortium. This study was partially supported by Apis-Gene and the French ANR project AMASGEN. Tom Druet is research associate from the F.R.S.–FNRS (Brussels, Belgium).

REFERENCE LIST

Boichard, D., F. Guillaume, A. Baur, P. Croiseau, M. N. Rossignol, M. Y. Boscher, T. Druet, L. Genestout, A. Eggen, L. Journaux, V. Ducrocq, and S. Fritz. 2010. Genomic Selection in French Dairy Cattle. Manuscript 716 in WCGALP 2010, Leipzig. Germany.

Browning, S. R., and B. L. Browning. 2007. Rapid and accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. The American Journal of Human Genetics 81:1084-1097.

Croiseau, P., C. Colombani, A. Legarra, F. Guillaume, S. Fritz, A. Baur, R. Dassonneville, R. Patry, C. Robert-Granié, and V. Ducrocq. 2010. Improving genomic evaluation strategies in dairy cattle through SNP pre-selection. Manuscript 360 in WCGALP 2010, Leipzig. Germany.

Daetwyler, H. D., G. R. Wiggans, B. J. Hayes, J. A. Wooliams, and M. E. Goddard. 2010. Imputation of Missing Genotypes From Sparse to High Density Using Long-Range Phasing. Manuscript 539 in WCGALP 2010 Leipzig. Germany.

Druet, T., S. Fritz, M. Boussaha, S. Ben-Jemaa, F. Guillaume, D. Derbala, D. Zelenika, D. Lechner, C. Charon, D. Boichard, I. G. Gut, A. Eggen, and M. Gautier. 2008. Fine mapping of quantitative trait loci affecting female fertility in dairy cattle on BTA03 using a dense single-nucleotide polymorphism map. Genetics 178(4):2227-2235.

Druet, T., and M. Georges. 2010. A Hidden Markov Model Combining Linkage and Linkage Disequilibrium Information for Haplotype Reconstruction and Quantitative Trait Locus Fine Mapping. Genetics 184:789-798.

Druet, T., C. Schrooten, and A. P. W. de Roos. 2010. Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. J. Dairy Sci. 93(11):5443-5454.

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

Illumina, Inc. 2005. Illumina GenCall Data Analysis Software - GenCall software algorithms for clustering, calling, and scoring genotypes. Illumina. Pub. No. 370-2004-009.

Lund, M. S., A. P. W. de Roos, A. G. de Vries, T. Druet, V. Ducrocq, S. Fritz, F. Guillaume, B. Guldbrandtsen, Z. Liu, R. Rents, C. Schrooten, M. Seefried, and G. Su. 2010. Improving Genomic Prediction by EuroGenomics Collaboration. Manuscript number 880 in WCGALP 2010, Leipzig. Germany.

Matukumalli, L. K., C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan, M. P. Heaton, J. O'Connell, S. S. Moore, T. P. L. Smith, T. S. Sonstegard, and C. P. Van Tassell. 2009. Development and Characterization of a High Density SNP Genotyping Assay for Cattle. Plos One 4(4):e5350

Meuwissen, T. H. E., and M. E. Goddard. 2001. Prediction of identity by descent probabilities from marker-haplotypes. Genet. Sel. Evol. 33(6):605-634.

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.

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

Scheet, P., and M. Stephens. 2006. A Fast and Flexible Statistical Model for large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. The American Journal of Human Genetics 78:629-644.

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

Weigel, K. A., G. de los Campos, A. I. Vazquez, G. J. M. Rosa, D. Gianola, and C. P. Van Tassell. 2010. Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle. J. Dairy Sci. 93(11):5423-5435.

Zhang, Z., and T. Druet. 2010. Marker imputation with low-density marker panels in Dutch Holstein cattle. J. Dairy Sci. 93(11):5487-5494.

1-4.3. Main results

Results show a moderate allelic imputation error rate (2-5%).This result is quite satisfactory, especially when the sire is genotyped with the 50k, but there is some room for improvement. The loss in reliability of genomic proofs was also moderate (0.02 to 0.05): GEBV based on imputed genotypes and on the Eurogenomics reference population presented higher reliability than GEBV based on real genotypes but with a national reference population size.

Comparison between Nordic and French results revealed some potential factors affecting imputation accuracy such as the number of effective markers. The importance of having genotyped sires to properly impute their offspring was also shown.

The accuracy observed after imputation of low density genotypes may not be high enough to select among candidates for artificial insemination, but it could be an efficient tool for screening the female population.

CHAPTER 2 – Defining the most adapted low density panel

2.1. – Article II Short Communication: Imputation performances of three low density marker panels in beef and dairy cattle

2-1.1. BACKGROUND

In article 1, we have seen that a low density panel is a potential tool which could be used for pre-screening candidates or for genotyping females. However, one can wonder whether the commercial 3K chip proposed is the best one and how it can be enhanced. Article 1 presents a European joint study and only the Holstein breed is considered. After some preliminary work, the 3K SNP minor allele frequencies (MAF) for other important French dairy breeds, such as the Montbéliarde and Normande were calculated and they appeared to be rather low. One could expect that other SNP panels jointly adapted to several breeds would give better results.

Initially, the Golden Gate technology was chosen for the low density chip because of its low price. However, it was observed that it induces lower call rates, and results in more difficult lab conditions for genotyping. Furthermore, out of the 2900 SNP included in the chip, only 2635 are kept after quality edits. This loss in effective markers (about 10%) is a drawback in terms of imputation efficiency.

2-1.2. OBJECTIVES

The objectives of this study were to develop some alternative in silico SNP chips and to compare them to the commercially available product in terms of imputation efficiency. The markers included in these custom chips were chosen in order to optimize MAF and spacing across various breeds. Our hypothesis was that imputation accuracy depends on these 2 characteristics of the panel.

Article II

Dassonneville R., S. Fritz, F., V. Ducrocq, D. Boichard. 2012

Short Communication: Imputation performances of three low density marker panels in beef and dairy cattle

J. Dairy Sci. 95 :4136–4140

ABSTRACT

Low density chips are appealing alternative tools contributing to the reduction of genotyping costs. Imputation enables to predict missing genotypes in order to recreate the denser coverage of the standard 50K genotype. Two alternative in silico chips were defined. They included markers selected to optimize Minor Allele Frequency and spacing. The objective of this study was to compare imputation accuracy of these custom low density chips with the commercially available 3K chip. Data consisted of genotypes of 4,037 Holstein bulls, 1,219 Montbéliarde bulls and 991 Blonde d'Aquitaine bulls. Criteria to select markers to include in low density marker panels are described. In order to mimic a low density genotype, all markers except the markers present on the low density panel were masked in the validation population. Imputation was performed using the Beagle software. Combining the Directed Acyclic Graph obtained by Beagle with the PHASEBOOK package provides fast and accurate imputation which is suitable for routine genomic evaluations based on imputed genotypes. Ninety five to ninety nine percent of alleles were correctly imputed depending on the breed and the low density chip. The alternative low density chips gave better results than the commercially available Golden Gate 3K chip. A low density chip with 6,000 markers is a valuable genotyping tool suitable for both dairy and beef breeds. Such a tool can be used for pre-selection of young animals or large-scale screening of the female population.

Key words: low density chip, imputation, genomic selection, SNP chip

INTRODUCTION

Genomic selection (Meuwissen et al., 2001) is now widely used in dairy cattle breeding to select bulls at an early stage; it requires estimation of effects for Single Nucleotide Polymorphisms (SNP) covering the whole genome at a sufficient density. Thousands of bulls have been genotyped on a SNP panel with 54,000 markers, commercially available from Illumina and developed by Matukumalli et al. (2009). One of the next challenges in order to take full advantage of genomic selection is to genotype a large fraction of the female population. This requires to substantially reduce genotyping costs. For this purpose, low density chips can be considered as an alternative tool. The Golden Gate Bovine ® 3K chip (GG 3K) was developed in 2009 by Illumina with a cheap technology (Golden Gate). The drawback of having genotypes for a smaller number of markers can be overcome by applying imputation. Imputation is a statistical method which predicts unobserved genotypes offering the possibility to infer a dense (e.g. 50K) genotype based on low density chip data. Reconstructing denser coverage such as the standard BovineSNP50 panel from low density chips is now recognized as the best way of using these cheaper tools (Weigel et al, 2010). Therefore, low density chips should be developed in such a way that imputation accuracy is maximized. Genomic breeding values can then be computed from evaluations based on imputed genotypes (Weigel et al, 2010, Berry and Kearney, 2011, Dassonneville et al, 2011, VanRaden et al.2011). In addition to genomic evaluation, low density chips can be used for sexing and parentage assignation or verification. Minor Allele Frequency (MAF), quality edits and spacing between markers are assumed to influence imputation performance. To check this hypothesis, two custom in silico low density marker panels were developed optimizing both MAF and spacing. The objective of this study is to compare imputation accuracy of different chips: the commercially available chip and two custom low density chips of various marker density (3K and 6K).

MATERIAL AND METHODS

Data

The reference population includes individuals that were genotyped on the Bovine 50K chip. The validation population is a subset of the reference population, including the youngest animals that were considered as selection candidates in this study. For this reason, low density genotypes were mimicked for these animals. The training population consisted of the remaining individuals from the reference population.

Three breeds were chosen for this study and studied separately: the Holstein breed, the French Montbéliarde dairy breed, and the Blonde d’Aquitaine beef breed. The reference population of the Holstein and Montbéliarde dairy breeds included 4,037 and 1,219 progeny-tested bulls, respectively, distributed across several generations. The validation population for these two breeds was defined through a cut-off date such that approximately 25% of the bulls of the reference population forming the validation population were born after that date. All of the selection candidates had their sire in the training population and most of the male ancestors were genotyped and included in the training population, whereas no female was genotyped. The reference population of the Blonde d’Aquitaine beef breed included 961 young bulls and their 30 sires. 237 young bulls were randomly selected to form the validation population. Therefore, for the Blonde d’Aquitaine, the sire was the only densely genotyped ancestor of the validation animals. Table 1 summarizes the number of animals included in training and validation population for the three different breeds.

Table 1. Number of animals for the three different breeds

	Reference	Validation
Montbéliarde	997	222
Holstein	3,071	966
Blonde d'Aquitaine	754	237

Only 50k genotype data were used. The genotypic data was first edited per locus. The study focused on autosomes and markers mapped on the X chromosome were deleted. Markers with an unknown map position in the Btau 4.0 assembly, departing from Hardy Weinberg equilibrium, with a call rate below 0.85, or with a MAF strictly equal to 0 were removed. Genotype data were finally checked for Mendelian inconsistencies between parents and offspring. Inconsistent genotypes were set to missing values.

Three different low density chips were defined. The GG 3K chip was similar to the commercially available Golden Gate Bovine 3k of Illumina. But no 3k genotyping was performed and 3k genotypes were simply obtained from the 50k by selecting the corresponding markers. This approach is somewhat optimistic because it does not account for a lower call rate due to the different chemistry used. The two other chips were created in silico. They can be considered as based on the Infinium technology since markers were chosen among those of the Bovine 50K SNP, and with the same call rates.

To be included in the 2 custom in silico panels, the selection criteria were the following:

- Markers had to be present on versions 1 and 2 of the Bovine SNP50 (Matukumalli et al., 2009) and on the Bovine HD chip from Illumina.
- Markers had to have a known position on Btau 4.0 (Elsik et al, 2009) and UMD3 (Zimin et al, 2009) assemblies,
- The marker position had to be consistent between the 2 assemblies, i.e. less than 10 Mb apart.
- Call rates needed to be above 0.98 with no technical problem observed in the sample of genotyped animals at INRA.

- Markers were checked for Hardy-Weinberg equilibrium (q value > 0.01).

From the set of markers meeting these criteria, SNP were chosen in order to maximize the Minor Allele Frequency (MAF) within chromosome segments. MAF were available for 8 French dairy and beef breeds (Blonde d'Aquitaine, Brown Swiss, Charolaise, Holstein, Limousine, Normande, Montbéliarde and Maine-Anjou), with 110 to 16,055 samples per breed. For the custom 3K chip, the genome was divided into one megabase segments and within each segment, the SNP with the highest average MAF over the 3 main French dairy breeds (Holstein, Normande, and Montbéliarde) was kept. For the custom 6K chip, in a first step, the SNP set of the custom 3K was retained. In a second step, the SNP with the highest average MAF over the 8 breeds was added within each Mb. Finally, a few more SNP were added to cover every half Mb and to ensure a better coverage of chromosome extremities (4 markers per Mb instead of 2). For these 2 custom low density chips, MAF was optimized and each Mb (custom 3k chip) or each half Mb (custom 6k chip) was covered. These rules were quite simple and did not account for linkage disequilibrium between markers.

The GG 3K chip includes 2900 SNP (Table 2) among which 2635 were kept after edits and quality control, as described above. The custom chips included 2,929 and 6,052 SNP covering the 29 autosomes. Markers used for parentage testing were included. It must be mentioned that additional SNP from sexual chromosomes should be integrated for gender determination.

Table 2. Number of markers included in the 3 low density chips

Low density chip	Number of markers	after quality control
GoldenGate Bovine3K	2900	2635
French custom 3K	2929	2929
French custom 6K	6144	6052

Imputation method

The SNPs included in the low density chips are all included on the Bovine 50K chip. To mimic the low density chip, marker genotypes of validation animals were obtained by erasing markers from the 50K and not present in the low density chip.

Imputation of markers was performed using Beagle 3.2 (Browning and Browning, 2007) with the –unphased option. Consequently, pedigree information was not used for imputation. It should be noticed that with Beagle, all markers are imputed, and the method does not leave any missing markers.

Efficiency was first measured by the number of correctly imputed alleles (as in Zhang and Druet, 2010). Therefore the number of errors was counted as 0 when the imputed and observed marker genotypes were identical, 1 if the real genotype was homozygous and the imputed one was heterozygous (or vice versa), and 2 if real and imputed marker genotypes were opposite homozygous. The error rate was calculated as the total number of errors divided by twice the number of imputed (masked) loci.

Several authors (Weigel et al. 2010, Druet et al. 2010, VanRaden et al. 2011) report the same kind of measures (error rate or concordance rate). Hickey et al. (2012) suggest to use correlation between true and imputed genotypes in order to account for MAF. Indeed, a high concordance rate is expected for low MAF and may overestimate imputation performances. In this study, both measures were reported.

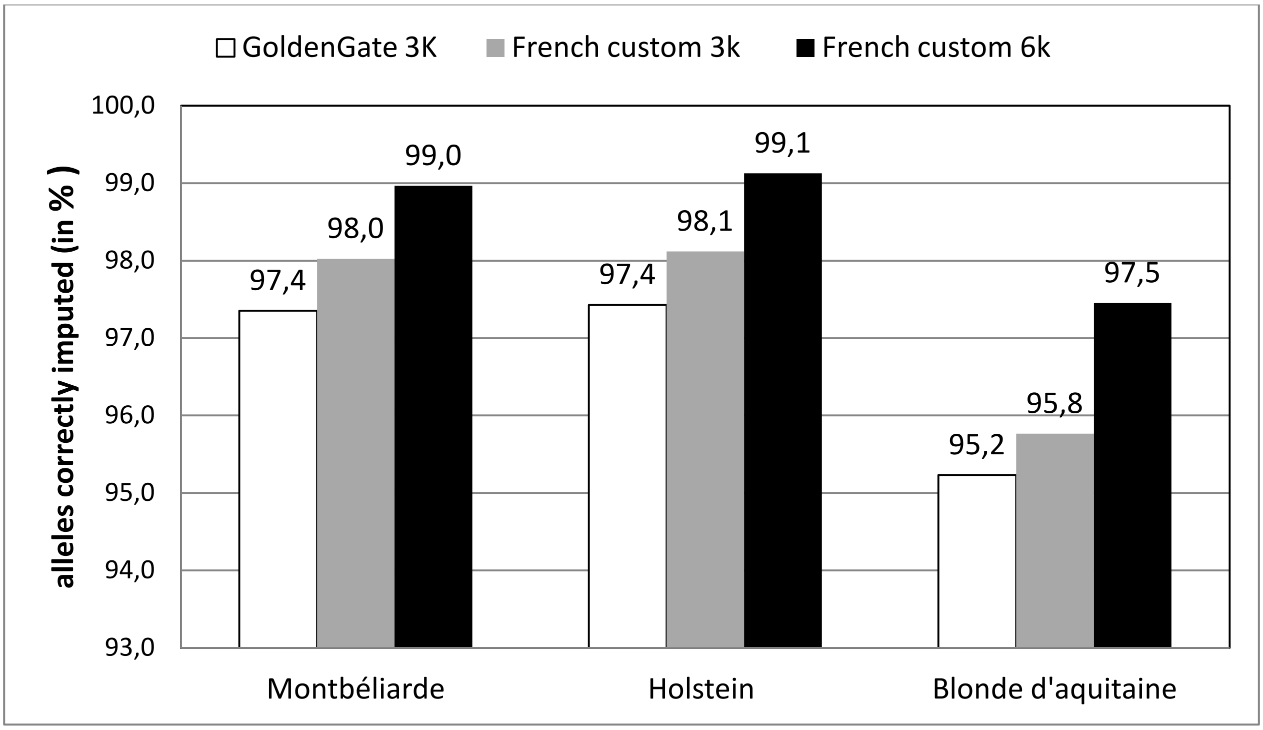
RESULTS AND DISCUSSION

The percentage of alleles correctly imputed is presented in Figure 1 for the 3 different breeds and the 3 different low density chips. In the Montbéliarde breed, 97.4% of the masked alleles were correctly imputed from the GG 3K chip. This result seems to be high enough to implement genomic selection based on low density chips; indeed, Weigel et al. (2010) found lower imputation accuracy and were still able to run appropriate genomic evaluation for the Jersey breed. However, imputation accuracy can be improved with other marker panels. The custom 3K chip included more effective markers and was optimised for French dairy breeds including Montbéliarde. For this reason, this optimized 3k chip gave better imputation accuracy (98.0%) than the GG3K (97.4%). With the 6K chip, both marker density and MAF optimisation were improved, resulting in higher imputation accuracy (99%).

Holstein was one of the 3 breeds involved in the choice of markers for the GG 3K (Illumina). For this reason, one could have expected similar imputation results between the two 3K chips. But the custom 3K chip gave better results (imputation accuracy of 98.1% instead of 97.4%). One possible explanation is the lower number of effective markers on the GG 3K chip. Another explanation is the Golden Gate chemistry constraints in the choice of markers, limiting the possible optimization on MAF and spacing. As expected, results were better with the 6K chip (more than 99% of alleles correctly imputed).

In the French Blonde d'Aquitaine beef breed, imputation accuracy was lower compared to the two other breeds, probably because of a smaller and different reference population with few ancestors genotyped and also to a larger effective population size. Imputation accuracy was slightly better with the custom 3K chip (95.8%) than with the GG 3K (95.2%) although MAF of that breed were not taken into account when constructing these 2 chips. This may be related to a better optimisation of MAF and spacing or an increased number of efficient markers as reported above. The biggest gain for the 6K chip compared to the 3K chips was obtained with the beef breed (97.5% vs. 95-96%). On the one hand, Blonde d'Aquitaine MAF was accounted for in the design of the 6k chip. On the other hand, the advantage of the 6k chip was larger because the performances of the 3k chips were lower than in dairy breeds, leaving a larger margin for improvement.

Figure 1. Dassonneville et al.



Proportion of masked alleles that are correctly imputed in the validation population, for three different breeds, and three different low density chips.

Using concordance rate as imputation efficiency criterion may be misleading as it depends on MAF. The lower the MAF, the higher the concordance rate, for the same efficiency and this should be accounted for in the interpretation. For example, with an average MAF of 0.2, 80% of the results would be correct after random sampling of the missing alleles. Getting a 95% concordance rate corresponds only to a 75% (= (95-80)/(100-80)) imputation efficiency. Correlations are an alternative criterion less dependent on MAF. Table 3 presents the correlations between true and imputed genotypes. They ranged from 0.88 to 0.97. Comparison across breeds, and moreover, across marker panels leads to the same conclusions as studying the fraction of alleles correctly imputed. The ranking of the chips was the same and Blonde d'Aquitaine breed results were lower than in dairy breeds.

Table 3. Correlation between true and imputed genotypes

	Montbéliarde	**Holstein**	**Blonde d'Aquitaine**
GoldenGate Bovine3K	0.94	0.93	0.88
French custom 3K	0.95	0.94	0.89
French custom 6K	0.97	0.96	0.92

A previous study by Zhang and Druet (2010) showed that both the number of reference animals and the number of markers in the low-density panel affect the imputation error rate. This error rate is also affected by the relationship between validation and reference animals. Comparing imputation error rates based on different populations is therefore difficult because the relationship between training and validation populations differs, and because the number of reference individuals and the linkage disequilibrium between markers vary.

One major concern regarding the use of Beagle software is the computational time. In a large population from the Eurogenomics reference population (Lund et al., 2011) with 12,068 animals in the training population and 3987 animals in the validation population, computing time with Beagle alone was from 40 to 80 hours per chromosome. Alternative methods using long range phasing and pedigree information (VanRaden et al., 2011; Hickey et al., 2011) are known to be much faster and have been compared by Johnston et al. (2011). However, we can propose an alternative which benefits from the accuracy of Beagle and the fast algorithm of the PHASEBOOK package. This package was developed by Druet and Georges

(2010) and allows to take advantage of pedigree relationships and to use this family information in addition to population-based linkage disequilibrium through the DAGphase software. Indeed, it is possible to run properly Beagle from scratch, in order to obtain and store the directed acyclic graph (DAG). This demanding analysis needs to be run only once. Then the DAG obtained from Beagle could be used within DAGphase. This solution is as accurate as Beagle (in terms of error rate, results not shown), and much less time consuming (1 hour per chromosome to be compared with the 40 to 80 hours of Beagle on the same dataset including 16,055 individuals). This 2-step approach is quite efficient. The second step, corresponding to only one iteration of DAGphase, is fast and can be run as often as necessary (i.e., monthly or weekly for genomic evaluations). An additional advantage of this combined approach over other methods (Hickey et al., 2011, Sargolzaei et al., 2011) is that all masked markers are called.

The DAGphase software was used by Zhang and Druet (2010) to impute from 3,000 to 50,000 markers, where the 3,000 markers were especially selected for high minor allele frequency. They found an allele error rate, i.e. the proportion of incorrectly predicted alleles, around 3% with a reference population of 4,734 Holstein bulls.

Our hypothesis was that a better optimization of MAF and spacing for the choice of markers to include in low density chips may lead to a gain in imputation accuracy. Comparison of imputation results of the GG3K and the 3K custom in silico chip (same marker density), for both Holstein and Montbéliarde breeds confirmed this hypothesis. The relative lower performances of the GG3K chip may be due to the constraints related to Golden Gate chemistry and the lower effective number of markers kept after quality control (2635 markers instead of 2900 leading to a drop of 10% in marker density).

CONCLUSIONS

Imputation using Beagle software was efficient to reconstruct a dense - 50K - genotype from low density chip data. Accuracy, measured by the allelic concordance rate, ranged between 95 and 99%. The highest values were obtained with the highest density 6k chip and the Holstein population characterized by a large reference population. Using the DAG obtained from Beagle into the PHASEBOOK algorithm allows to take advantage of family information and speed up the imputation process with no loss in imputation accuracy.

Low density chips are appealing alternative tools which reduce genotyping costs. This could allow to genotype more animals. They can be used for pre-selection of young animals. It is most interesting for large scale genomic selection of females.

The existing Golden Gate Bovine 3K chip presents satisfactory results. However, other choices of markers are possible for low density chips in order to optimize MAF and spacing for various breeds so that imputation is more accurate. The 6K chip appears to be the method of choice and provides a high imputation efficiency, even for a beef breed with a small reference population such as the French Blonde d'Aquitaine breed. Consequently, a low density chip with around 6,000 markers is an appealing genotyping tool that is suitable for dairy and beef breeds. This option was chosen by Illumina to produce the new LD chip in collaboration with an international consortium (Boichard et al., 2012).

ACKNOWLEDGEMENTS

Agence Nationale de la Recherche and APISGENE are gratefully acknowledged for funding the AMASGEN project. Blonde d'Aquitaine data were obtained from the Qualvigene project.

REFERENCES

Berry D. P. and Kearney J. F.. 2011. Imputation of genotypes from low- to high-density genotyping platforms and implications for genomic selection. Animal, 5 , pp 1162-1169 doi:10.1017/S1751731111000309

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 K., Wiggans G.R., 2012. Design of a Bovine Low-Density SNP Array Optimized for Imputation. Plos One, accepté.

Browning, S. R., and B. L. Browning. 2007. Rapid and accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. The American Journal of Human Genetics 81:1084-1097.

Druet, T., and M. Georges. 2010. A Hidden Markov Model Combining Linkage and Linkage Disequilibrium Information for Haplotype Reconstruction and Quantitative Trait Locus Fine Mapping. Genetics 184:789-798.

Druet T., C. Schrooten, and A.P. de Roos. 2010. Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. J. Dairy Sci. 93(11):5443-5454.

Dassonneville R., R.F. Brøndum, T. Druet, S. Fritz, F. Guillaume, B. Guldbrandtsen, M.S. Lund, V. Ducrocq, and G.Su. 2011. Effect of imputing markers from a low density chip on the reliability of genomic breeding values in Holstein populations. J. Dairy Sci. 94(7), 3679-3686.

Elsik C.G., R.L. Tellam, K.C. Worley, R.A. Gibbs, D.M. Muzny, G.M. Weinstock, D.L. Adelson, E.E. Eichler, L. Elnitski, R. Guigo, 2009. The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science, 324(5926):522-528.

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

Hickey J.M., J. Crossa, R. Babu, and G. de los Campos. Factors Affecting the Accuracy of Genotype Imputation in Populations from Several Maize Breeding Programs. Crop Science. doi: 10.2135/cropsci2011.07.0358; Published online 8 Dec. 2011.

Johnston J., Kistemaker G., and Sullivan P.G.. 2011. Comparison of different imputation methods. Preliminary proceedings of 2011 Interbull meeting, August 27–29, Stavanger, Norway, 7 pages.

Lund M.S., A.P.W. de Roos, A.G. de Vries, T. Druet, V. Ducrocq, S. Fritz, F. Guillaume, B. Guldbrandtsen, Z. Liu, R. Reents, C. Schrooten, M. Seefried, and G. Su. 2011. A common reference of four European Holstein populations increases reliability of Genomic Predictions. Genet Sel Evol. 43:43

Matukumalli, L. K., C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan, M. P. Heaton, J. O'Connell, S. S. Moore, T. P. L. Smith, T. S. Sonstegard, and C. P. Van Tassell. 2009. Development and Characterization of a High Density SNP Genotyping Assay for Cattle. Plos One 4(4).

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.

Sargolzaei, M., J. P. Chesnais, and F. S. Schenkel. FImpute. 2011. An efficient imputation algorithm for dairy cattle populations. J. Anim. Sci. 89(E-Suppl. 1)/J. Dairy Sci. 94(E-Suppl. 1): 421 (abstr. 333).

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

Weigel, K. A., G. de los Campos, A. I. Vazquez, G. J. M. Rosa, D. Gianola, and C. P. Van Tassell. 2010. Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle. J. Dairy Sci. 93(11):5423-5435.

Zhang, Z., and T. Druet. 2010. Marker imputation with low-density marker panels in Dutch Holstein cattle. J. Dairy Sci. 93(11):5487-5494.

Zimin A.V., A.L. Delcher, L. Florea, D.R. Kelley, M.C. Schatz, D. Puiu, F. Hanrahan, G. Pertea, C.P. Van Tassell, T.S. Sonstegard, G. Marcais, M. Roberts, P. Subramanian, J.A. Yorke, S.L. Salzberg. 2009. A whole-genome assembly of the domestic cow, Bos taurus. Genome Biol 10(4):R42.

2-1.3. Main results

Several French beef and dairy cattle breeds were taken into account when selecting the markers to be included in the SNP panel. Criteria to choose among the possible SNP were described. Results in terms of imputation efficiency were shown for 4 different beef and dairy breeds. Two different measures of imputation accuracy were presented - including the correlation between true and imputed genotypes - in order to avoid bias due to the dependency between MAF and error rate. Imputation was performed with the Beagle software and accuracy ranged between 95% and 99% depending on the breed and the SNP panel considered. For all the breeds, the custom chips gave better results than the commercial 3K chip, reflecting a higher MAF and an even spacing along the genome. As expected, the 6K chip gave the best results, reflecting the effect of the number of markers and, therefore, their higher density.

An innovative imputation procedure is also described in the article. It relies on existing software but aims at increasing speed and accuracy. While the method described in Druet and Georges (2010) using the PHASEBOOK package produces the required DAG through successive iterations of both Beagle and DAGphase on a reduced set of individuals, here the DAG is produced by Beagle alone and then stored. Then it can be re-used with a single iteration of DAGphase. This drastically reduces computation time with no loss in accuracy, and appears to be suitable for routine imputation of genotypes for genomic evaluations.

Other research teams in different countries made the same assertion that other low density panels could improve imputation accuracy and enhance the interest for such a tool. The scope of these studies was purely scientific and just looked at simulation of in silico chips. The industry now has the opportunity to catch up the idea and develop a new low density chip.

– Article III Design of a bovine low-density SNP array optimized for imputation

2-1.4. Background and objectives

In 2010, a low density SNP chip was developed in dairy cattle : the Bovine 3K Beadchip ® from Illumina. Some studies (including the previous one) pointed out that such a tool can be enhanced in order to benefit to more breeds and to be more accurate in terms of imputation. For this reason, Illumina gathered a new consortium (the Bovine LD consortium) in order to develop a new low density tool for bovine population adapted to imputation. The following article describes how it was designed.

Some of the drawbacks of the previous 3K chip were specifically considered. For instance, the Golden Gate 3K chip (GG3K) was adapted to three North American dairy cattle breeds: Holstein, Jersey and Brown Swiss. The consortium designing the new Bovine LD SNP chip intended to at adapt the new chip to as many various breeds as possible.

Note: The commercial name given to the new tool, Bovine LD, may be misleading since here, LD means low density, whereas it usually refers to Linkage Disequilibrium in the genetics field.

Article III

The Bovine LD Consortium. 2012

Boichard D., H. Chung, R. Dassonneville, X. David, A. Eggen, S. Fritz, K. J. Gietzen, B. J. Hayes, C. T. Lawley, T. S. Sonstegard, C. P. Van Tassell, P. M. VanRaden, K. A. Viaud-Martinez, G. R. Wiggans

Design of a Bovine Low-Density SNP Array Optimized for Imputation

PLoS ONE 7(3): e34130

ABSTRACT

The Illumina BovineLD BeadChip was designed to support imputation to higher density genotypes in dairy and beef breeds by including single-nucleotide polymorphisms (SNPs) that had a high minor allele frequency as well as uniform spacing across the genome except at the ends of the chromosome where densities were increased. The chip also includes SNPs on the Y chromosome and mitochondrial DNA loci that are useful for determining subspecies classification and certain paternal and maternal breed lineages. The total number of SNPs was 6,909. Accuracy of imputation to Illumina BovineSNP50 genotypes using the BovineLD chip was over 97% for most dairy and beef populations. The BovineLD imputations were about 3 percentage points more accurate than those from the Illumina GoldenGate Bovine3K BeadChip across multiple populations. The improvement was greatest when neither parent was genotyped. The minor allele frequencies were similar across taurine beef and dairy breeds as was the proportion of SNPs that were polymorphic. The new BovineLD chip should facilitate low-cost genomic selection in taurine beef and dairy cattle.

Funding: The National Research Agency (ANR) and Apisgene funded the French BovineSNP50 data. The Dairy Futures Cooperative Research Centre and the Beef Genetic Technologies Cooperative Research Centre funded the dairy and beef Australian genotyping respectively. SMH: Were there other funders or grants? The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

INTRODUCTION

Genetic improvement of several key agricultural species is accelerating with the adoption of genomic selection [1,2,3]. With this method, animals or plants can be selected for breeding on the basis of their genetic merit predicted by markers spanning the entire genome. Particularly in dairy cattle, this method has been shown to be more efficient than conventional progeny testing of bulls (up to double the rate of genetic gain) as well as substantially less expensive [4]. Moreover, genomic selection opens new opportunities for sustainable management of populations by more efficiently selecting for traits that have low heritability, e.g. fitness traits, or traits that are difficult to measure. This method is also useful for managing the accumulation of inbreeding within breeds with a small effective population size. In dairy cattle, genomic selection has been deployed at a rapid pace, and most countries with major dairy breeding programs now rely heavily on this new technology [5].

A major challenge in implementing genomic selection in most species is the cost of genotyping. The expected value of the information gained by genotyping must exceed the cost of obtaining the genotypes. During the early stages of genomic selection in the dairy industry, the cost of high-density genotyping could be justified. The primary application was to evaluate bulls that were potential candidates for production of commercial semen. Using SNP information for those evaluations resulted in more accurate selection of bulls to acquire and extensively market. Once increased accuracies of genome-enhanced breeding values had been demonstrated, breeders and buyers quickly adopted this technology to improve accuracy of selection [6]. This example of a genomic-selection application has extreme value compared with other animal food production paradigms. In contrast, profit from genomic selection is likely to be much lower for beef bulls and dairy females [5,7]. An appealing approach in situations with much lower returns from genotyping is to use a more economical, reduced-density SNP chip with markers optimized for imputation.

Imputation is the process of predicting unknown genotypes for animals from observed genotypes and often uses information from a reference population with dense genotypes to predict missing genotypes for animals with lower density genotypes. It is also applied to merge genotypes of similar densities but different SNPs. Most imputation algorithms use information from relatives and population linkage disequilibrium. A number of software programs for imputation have been developed based originally on human genetics [8,9] and more recently on animal genetics [10,11,12,13]. The limited effective population sizes and population structures in livestock allow the possibility of imputation of high-density genotypes from quite low-density genotypes [14,15,11,16].

In 2010, a low-density bovine SNP chip, the Illumina GoldenGate Bovine3K Genotyping Beadchip, was developed and made commercially available. That product offered a significant advance toward low-cost genomic selection in cattle; however, imputation accuracy was highly dependent on the relationship of the individual genotyped with the Bovine3K chip to the reference population genotyped at a higher density [17]. In addition, some samples failed to provide genotypes of adequate quality for use in genomic predictions. The SNP call rate performance of the Bovine3K chip was slightly reduced compared with the

BovineSNP50 chip because GoldenGate chemistry relies on two hybridization events for proper SNP detection as opposed to a single event for Infinium chemistry.

In this study, the Illumina Infinium BovineLD Genotyping Beadchip was developed to provide high imputation accuracy for higher density SNP genotypes in taurine dairy and beef populations. The main objective was to provide a tool that would enable genomic estimated breeding values to be calculated from accurately imputed genotype data from an Infinium-based SNP array with very low rates of failed samples. The main features of the new BovineLD chip are presented along with its imputation performance in a range of breeds and reference populations.

MATERIAL AND METHODS

SNP selection

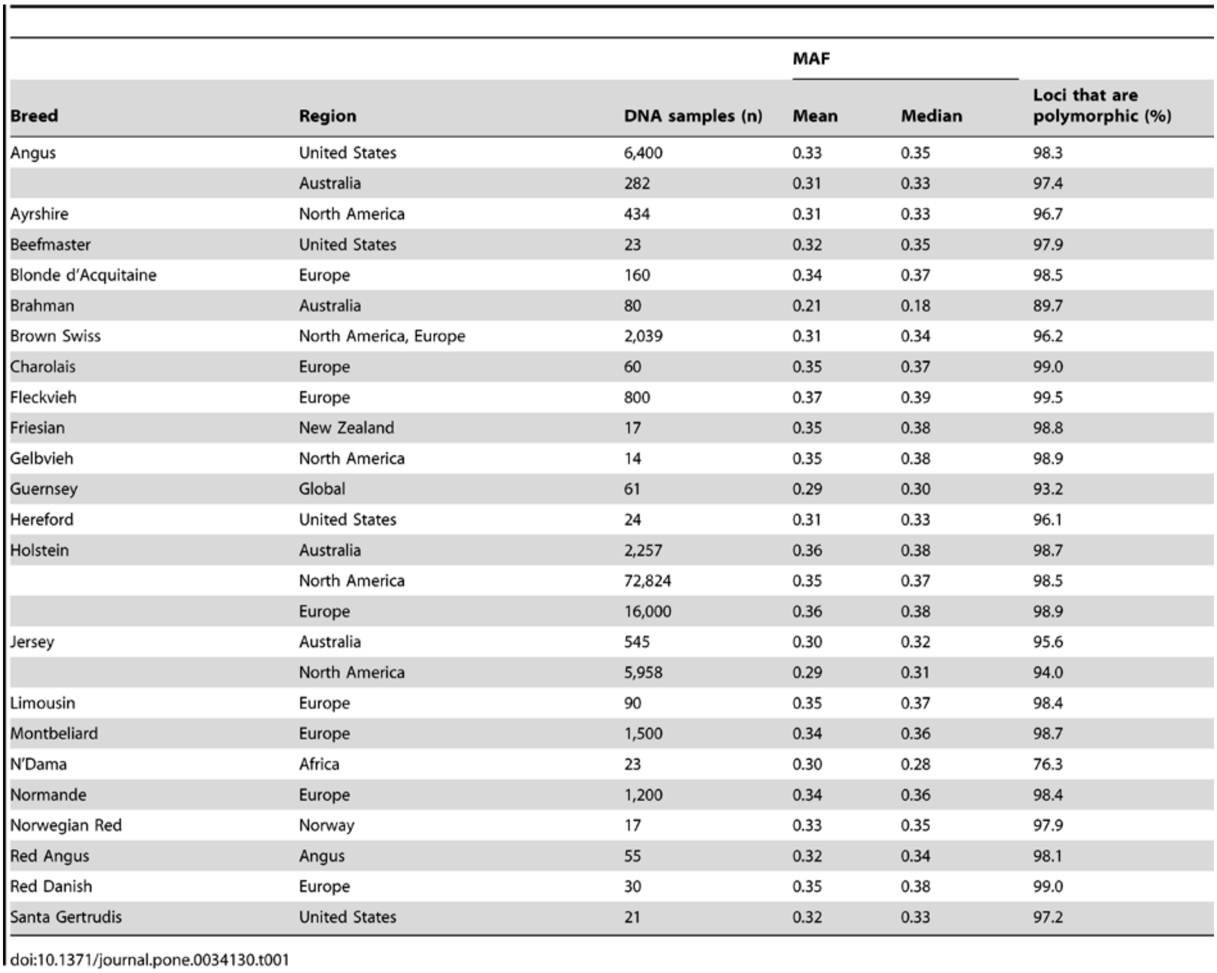
To provide highly accurate imputation to BovineSNP50 genotypes in global taurine breeds, SNPs were selected from validated assays from existing higher density chips and similar SNP detection technology, i.e. the Illumina BovineSNP50 and BovineHD SNP arrays, with priority given to BovineSNP50 content. From the known and validated SNPs, selection priority was 1) high minor allele frequencies (MAFs) in targeted breeds, 2) uniform spacing at a minimum of 2 SNPs per Mbp, with increased SNP density within 500 kilobase pair (kbp) of chromosomal ends, 3) inclusion of SNPs for determination of sex, parentage, Y haplotypes, and subspecies and maternal lineages, 4) SNP quality and fidelity criteria for robust reproducibility (>98% call rate and <0.01% Mendelian inconsistency), and 5) a target overlap of 2,000 SNPs with the Bovine3K chip to ensure backward compatibility. The anticipated SNP spacing, with 2 SNP per Mb, obviated the need to check for highly correlated SNPs.

The SNPs were selected to be highly informative with a high MAF over a large range of breeds from around the world (Table 1). The reference MAF estimates were from breeds in 10 countries from North America, Europe, and Oceania. Content selection was optimized using taurine allele frequencies. To achieve regular spacing, the UMD3 bovine genome assembly was used to define 500-kbp segments over the 29 autosomes. A lack of flanking information at the end of each chromosome had resulted in lower imputation efficiency in preliminary tests. To correct that problem, the SNP density was doubled in the first and last segments of each chromosome. Reflecting the diverse membership of the Bovine LD Consortium, initial SNP selection was made by one member and updated by the others. The initial SNP selection was based on two independent criteria. First, SNPs with the highest mean MAF in each 500-kbp segment were selected over a broad range of European breeds including European Holstein, Montbéliarde, Normande, Jersey, Brown Swiss, Norwegian Red, Swedish Red and White, Finnish Ayrshire, Charolais, Limousine, Blonde d'Aquitaine, and Maine Anjou, with Holstein receiving double weight; the top two SNPs were selected in the segment at each end of the chromosome. Second, SNPs with the highest mean minimum MAF for six major European dairy breeds (European Holstein, Montbéliarde, Normande, Jersey, Brown Swiss, and Norwegian Red) were selected for each 500-kbp segment, with again 2 SNPs selected at each end of the chromosome. Selecting those SNPs with the highest mean of the two selection criteria within each 500-kbp segment (with doubling at the chromosome ends) resulted in 8,000 SNPs. Those 8,000 SNPs were subjected to a similar selection process using MAFs from North America and Oceania along with the European populations. For Holstein and Jersey breeds, the MAF used was the mean across the 3 populations; for Brown Swiss, only North America and Europe were included. The mean MAF was computed from Holstein, Jersey, Brown Swiss, Angus, and Brahman. The minimum MAF was from Jersey, Brown Swiss, and Angus. Again, the SNPs with the highest mean of the two selection criteria were selected with doubling at the chromosome ends.

Next, some of the selected SNPs were replaced by Bovine3K SNPs that were in nearby locations to ensure backward compatibility. In addition, SNPs used for breed determination and parentage testing that had not already been selected were included, and some SNPs were added to fill gaps generated by map inconsistencies.

For the X chromosome, Bovine3K SNPs with high MAFs were selected and supplemented with BovineSNP50 SNPs, with consideration given to spacing, MAF, and fidelity. Because large gaps remained after that initial selection, additional X- chromosome SNPs were chosen from the BovineHD assay.

For the Y chromosome and mitochondrial DNA (mtDNA), 9 Y-specific and 13 mtDNA SNP markers were identified from the BovineHD chip based on assay fidelity and performance across 27 breeds, MAF across those breeds, and ability of a SNP to discern subspecies and geographic locations of breed origins.



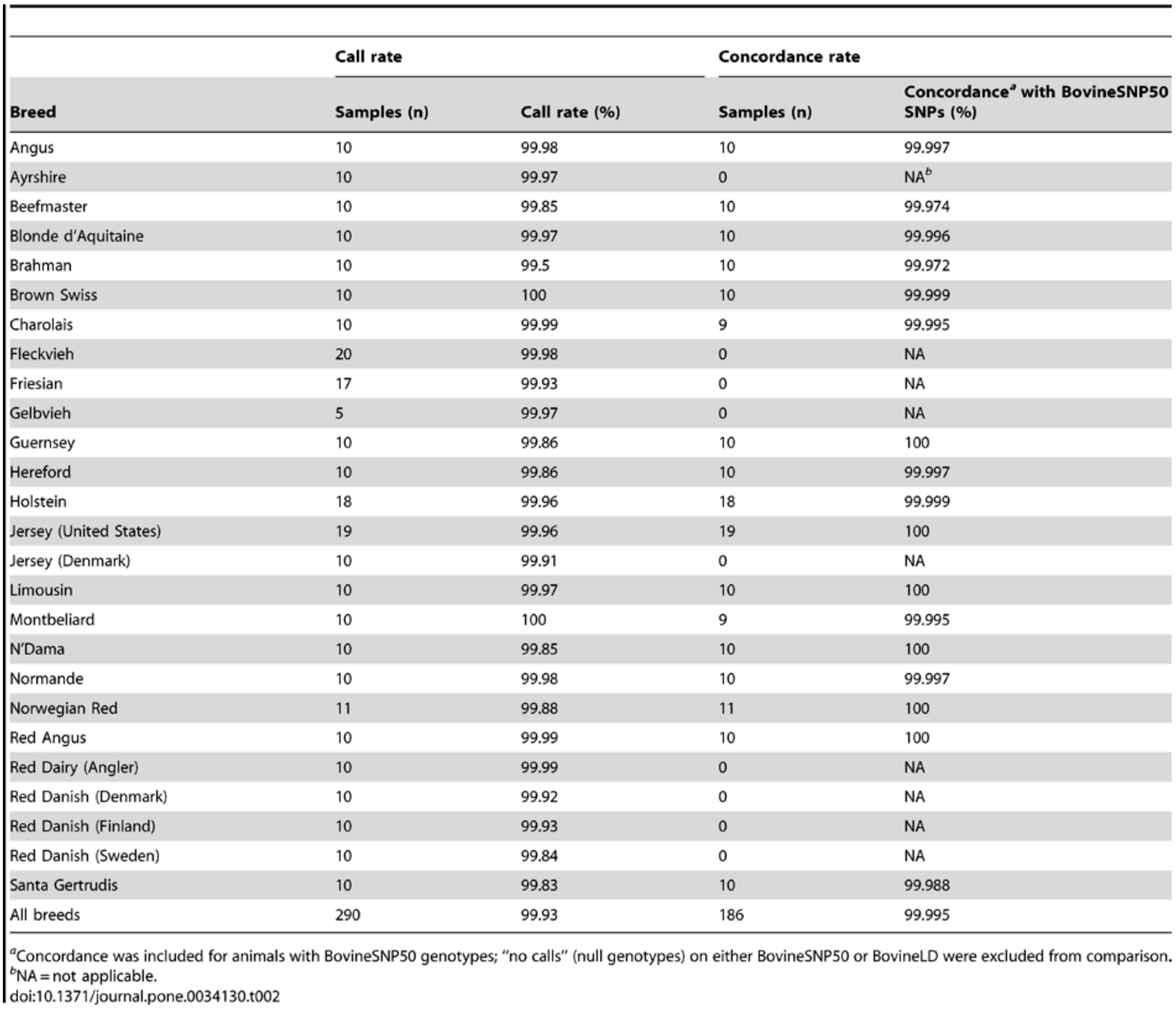
			MAF		
Breed	**Region**	**DNA samples (n)**	**Mean**	**Median**	**Loci that are polymorphic (%)**
Angus	United States	6,400	0.33	0.35	98.3
	Australia	282	0.31	0.33	97.4
Ayrshire	North America	434	0.31	0.33	96.7
Beefmaster	United States	23	0.32	0.35	97.9
Blonde d'Acquitaine	Europe	160	0.34	0.37	98.5
Brahman	Australia	80	0.21	0.18	89.7
Brown Swiss	North America, Europe	2,039	0.31	0.34	96.2
Charolais	Europe	60	0.35	0.37	99.0
Fleckvieh	Europe	800	0.37	0.39	99.5
Friesian	New Zealand	17	0.35	0.38	98.8
Gelbvieh	North America	14	0.35	0.38	98.9
Guernsey	Global	61	0.29	0.30	93.2
Hereford	United States	24	0.31	0.33	96.1
Holstein	Australia	2,257	0.36	0.38	98.7
	North America	72,824	0.35	0.37	98.5
	Europe	16,000	0.36	0.38	98.9
Jersey	Australia	545	0.30	0.32	95.6
	North America	5,958	0.29	0.31	94.0
Limousin	Europe	90	0.35	0.37	98.4
Montbeliard	Europe	1,500	0.34	0.36	98.7
N'Dama	Africa	23	0.30	0.28	76.3
Normande	Europe	1,200	0.34	0.36	98.4
Norwegian Red	Norway	17	0.33	0.35	97.9
Red Angus	Angus	55	0.32	0.34	98.1
Red Danish	Europe	30	0.35	0.38	99.0
Santa Gertrudis	United States	21	0.32	0.33	97.2

doi:10.1371/journal.pone.0034130.t001

Table 1. Number of DNA samples, minor allele frequencies (MAFs), and estimated frequency of loci that were polymorphic by breed and region.

Imputation

Imputation efficiency was assessed in 10 populations (North American, French, and Australian Holsteins; North American and Australian Jerseys; North American Brown Swiss; Australian Angus; French Montbéliarde; French Normande; and French Blonde d'Aquitaine). Beagle software [9] was used for the Australian and French populations and findhap.f90 [13] for the North American populations. Using existing genotypes from the BovineSNP50 chip, imputation efficiency was determined by comparing imputed and true genotypes. Part of the population was retained as a "reference," while target individuals for imputation had their genotypes reduced in silico to either BovineLD or Bovine3K genotypes. Results were assessed as the proportion of genotypes that were correct in the target population. For example, if the imputed genotype was a heterozygote and the BovineSNP50 genotype was a homozygote, that genotype was counted as incorrectly imputed. The count of correct genotypes included both observed and imputed genotypes to measure the overall success of a lower density genotype in approximating a BovineSNP50 genotype.



	Call rate		Concordance rate	
Breed	**Samples (n)**	**Call rate (%)**	**Samples (n)**	**Concordance[a] with BovineSNP50 SNPs (%)**
Angus	10	99.98	10	99.997
Ayrshire	10	99.97	0	NA[b]
Beefmaster	10	99.85	10	99.974
Blonde d'Aquitaine	10	99.97	10	99.996
Brahman	10	99.5	10	99.972
Brown Swiss	10	100	10	99.999
Charolais	10	99.99	9	99.995
Fleckvieh	20	99.98	0	NA
Friesian	17	99.93	0	NA
Gelbvieh	5	99.97	0	NA
Guernsey	10	99.86	10	100
Hereford	10	99.86	10	99.997
Holstein	18	99.96	18	99.999
Jersey (United States)	19	99.96	19	100
Jersey (Denmark)	10	99.91	0	NA
Limousin	10	99.97	10	100
Montbeliard	10	100	9	99.995
N'Dama	10	99.85	10	100
Normande	10	99.98	10	99.997
Norwegian Red	11	99.88	11	100
Red Angus	10	99.99	10	100
Red Dairy (Angler)	10	99.99	0	NA
Red Danish (Denmark)	10	99.92	0	NA
Red Danish (Finland)	10	99.93	0	NA
Red Danish (Sweden)	10	99.84	0	NA
Santa Gertrudis	10	99.83	10	99.988
All breeds	290	99.93	186	99.995

[a] Concordance was included for animals with BovineSNP50 genotypes; "no calls" (null genotypes) on either BovineSNP50 or BovineLD were excluded from comparison.
[b] NA = not applicable.
doi:10.1371/journal.pone.0034130.t002

Table 2. Numbers of samples, call rates, and BovineSNP50 concordance for validation of BovineLD single-nucleotide polymorphisms (SNPs) by breed.

Content validation

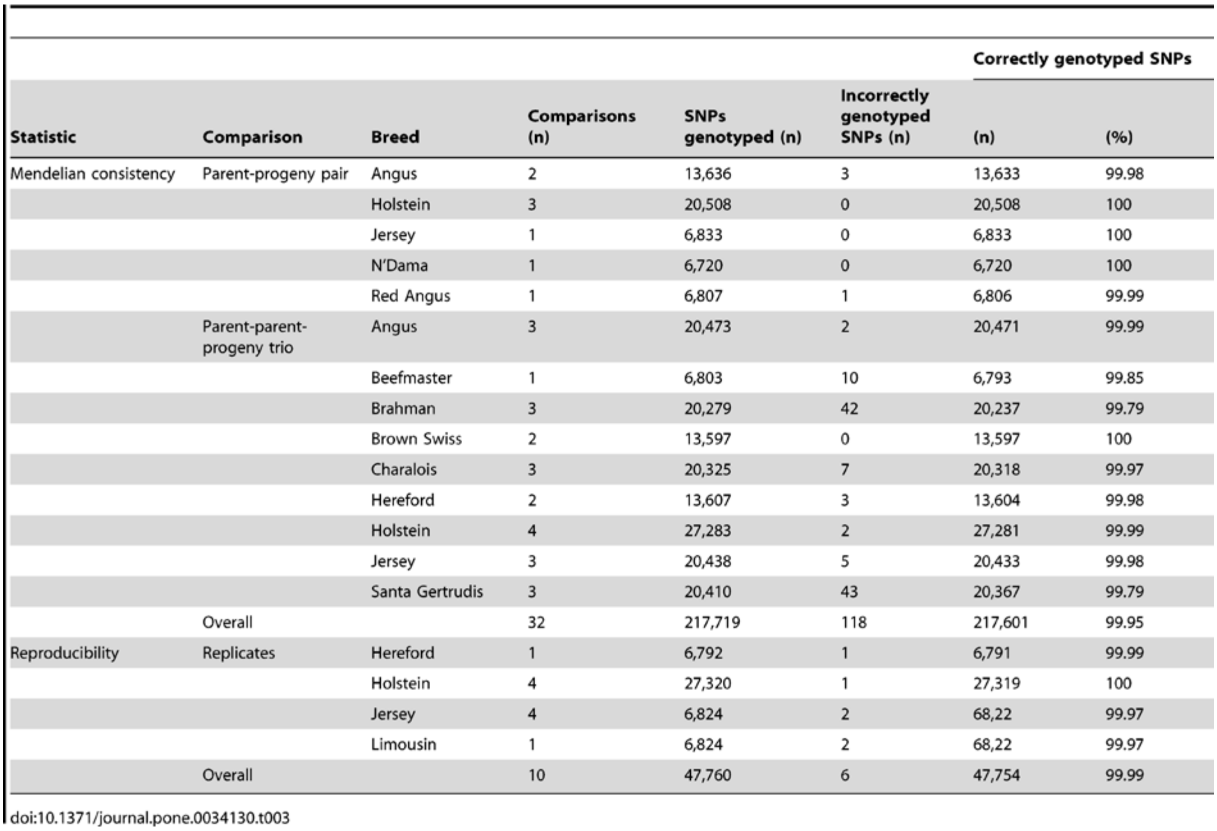
The SNP assays for 6,914 loci were validated using data from 290 samples that represented 26 global dairy and beef breeds (Table 2) and included Bovine Hapmap samples [18]. The 290 samples (234 males, 56 females) included 286 unrelated samples, 2 trios, and 2 replicates. All markers were assessed for clustering of the genotypes using Illumina GenomeStudio genotyping software (version 2010.3). A total of 6,909 clearly identifiable and scorable clusters were retained for robust utility of the panel. The cluster positions were defined with priority given first to data from dairy breeds and second to beef breeds. The purpose of the resulting cluster position file is to apply known robust cluster positions to future genotyping data for high throughput genotype calling. For phylogenetic analysis based on Y and mtDNA SNPs, individual sequences for each breed were clustered to construct consensus sequences using SNPs from 9 Y-chromosome loci and 13 mtDNA loci with the DNASTAR SeqMan program (version 6.1).

RESULTS

SNP call rates and accuracy

The BovineLD chip, consisting of 6,909 final loci, was validated for 290 individuals from 26 major dairy and beef breeds (Table 2). The mean call rate was 99.94% among dairy breeds, 99.90% among beef breeds, and 99.93% among all samples. Mendelian consistency was examined using two Holstein trios, which showed a single error on BTB-01149046 out of 13,797 total possible comparisons. Reproducibility was 100% across two Holstein replicated samples. Mendelian consistency and reproducibility were also examined for the overlapping 6,844 SNPs of the BovineHD and BovineLD chips. Those data included 8 parent-progeny, 24 parent-parent-progeny, and 10 replicate comparisons that represented 11 taurine, 2 indicine, and 1 hybrid breeds (Table 3). Mendelian consistency was 99.95%, and reproducibility was 99.99%.

Concordance between SNP calls from the BovineLD and other assays was evaluated by comparing BovineLD genotyping data used for validation against a subset of genotyping data collected for the BovineSNP50 assay. For taurine breeds, discordant calls represented <0.01% of all genotyping calls (Table 2). The concordance rate for 2,088 SNPs in common between BovineLD and Bovine3K assays was 98.78% for 281 females genotyped with both chips. The most likely cause of the differential performance between the BovineLD and Bovine3K chips is the chemistry difference between the Infinium and GoldenGate assays.



						Correctly genotyped SNPs	
Statistic	Comparison	Breed	Comparisons (n)	SNPs genotyped (n)	Incorrectly genotyped SNPs (n)	(n)	(%)
Mendelian consistency	Parent-progeny pair	Angus	2	13,636	3	13,633	99.98
		Holstein	3	20,508	0	20,508	100
		Jersey	1	6,833	0	6,833	100
		N'Dama	1	6,720	0	6,720	100
		Red Angus	1	6,807	1	6,806	99.99
	Parent-parent-progeny trio	Angus	3	20,473	2	20,471	99.99
		Beefmaster	1	6,803	10	6,793	99.85
		Brahman	3	20,279	42	20,237	99.79
		Brown Swiss	2	13,597	0	13,597	100
		Charalois	3	20,325	7	20,318	99.97
		Hereford	2	13,607	3	13,604	99.98
		Holstein	4	27,283	2	27,281	99.99
		Jersey	3	20,438	5	20,433	99.98
		Santa Gertrudis	3	20,410	43	20,367	99.79
	Overall		32	217,719	118	217,601	99.95
Reproducibility	Replicates	Hereford	1	6,792	1	6,791	99.99
		Holstein	4	27,320	1	27,319	100
		Jersey	4	6,824	2	68,22	99.97
		Limousin	1	6,824	2	68,22	99.97
	Overall		10	47,760	6	47,754	99.99

doi:10.1371/journal.pone.0034130.t003

Table 3. Mendelian consistency and reproducibility comparisons for a set of 6,844 SNPs in common for the BovineHD and BovineLD BeadChips.

Performance for MAF, mean spacing, and paternal and maternal lineages

Data for calculating mean MAF (Table 1) were primarily BovineLD markers extracted from BovineSNP50 data. However, if BovineSNP50 data were not available, BovineLD markers from the validation data were used. That method allowed MAFs to be calculated more accurately. Mean MAF for the 6,909 SNPs was > 0.29 for all taurine breeds (Table 1). For Brahman (a Bos primigenius indicus breed), mean MAF was lower (0.18). Overall, >89% of the SNPs were polymorphic in Brahman, which suggested that the BovineLD chip may be useful for imputation in this breed.

For the 6,909 SNPs selected for the BovineLD chip, median spacing was 0.348 Mbp, with only 82 (1.1%) of intervals greater than 1 Mbp (Fig. 1). The strategy of increasing SNP density at chromosome ends substantially improved imputation accuracy for those regions compared with the Bovine3K array (Fig. 2).

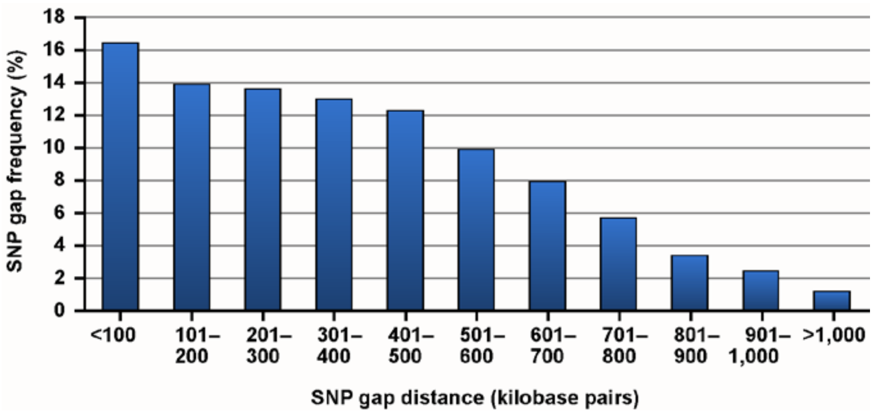


Figure 1. BovineLD single-nucleotide polymorphism (SNP) gap distribution.

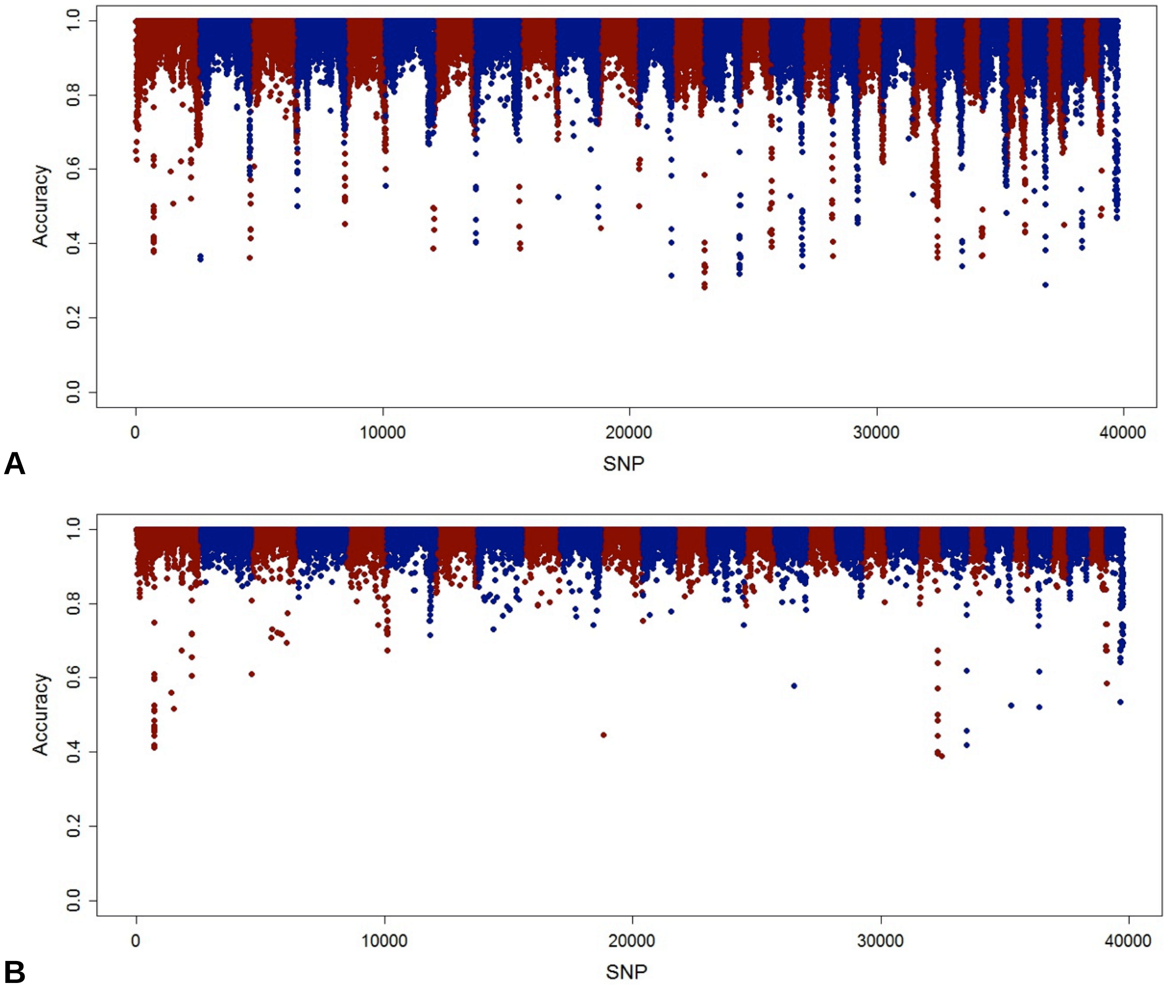
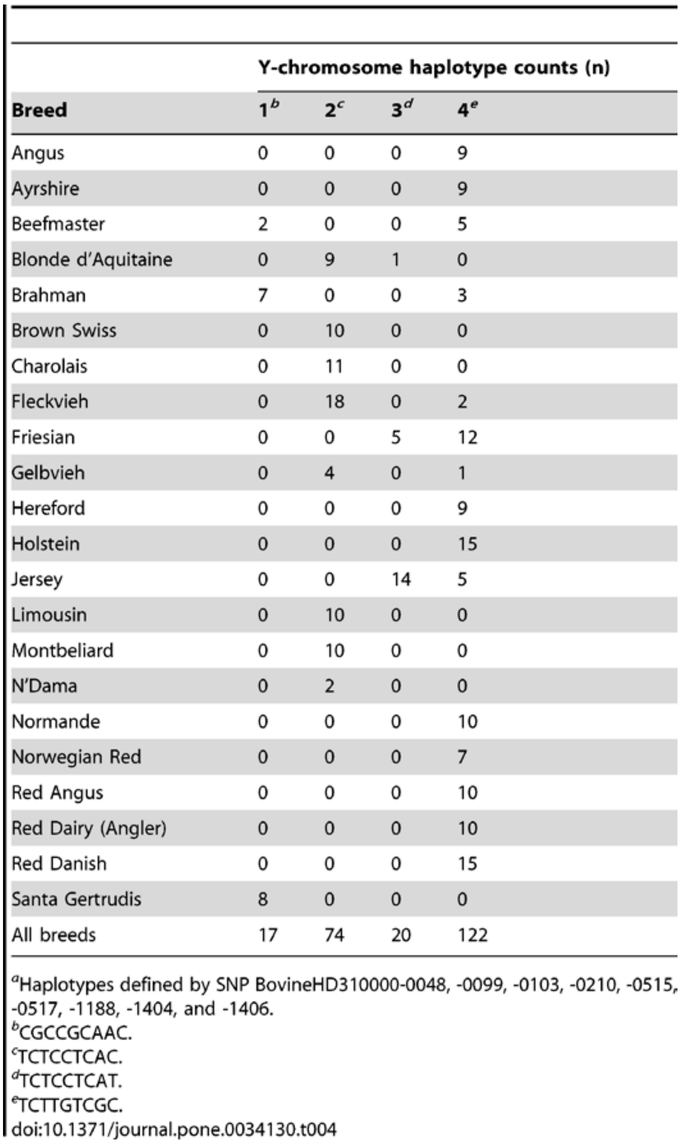


Figure 2. Imputation accuracy for Bovine3K and BovineLD genotypes.

Imputation was performed for A) Bovine3K and B) BovineLD genotypes using Beagle software; imputation accuracy is reported by single-nucleotide polymorphism (SNP).

The sex-specific and lineage SNPs also performed well. The nine Y-chromosome SNPs had a 100% call rate across 230 males of different breeds and no genotype calls for the 55 females. For the five animals of unknown sex, these markers indicated that four of the animals were male and one was female. Four unique Y-chromosome haplotypes were identified (Table 4): haplotype 1, (CGCCGCAAC), indicine paternal lineage; haplotype 2 (TCTCCTCAC), central European lineage; haplotype 3 (TCTCCTCAT), 1 base different from haplotype 2 and probably animals that came to the island of Jersey from France or Spain; and haplotype 4 (TCTTGTCGC), northern European lineage, including islands. Only a few breeds had more than one haplotype, for example Santa Gertrudis and Beefmaster, both of which are taurine – indicine hybrids. Common haplotypes across breeds reflect common origin. Phylogenetic analysis separated the 26 breeds into four distinctive clades, which agrees with a previous report on the dual origins of dairy cattle breeds in Europe [19].

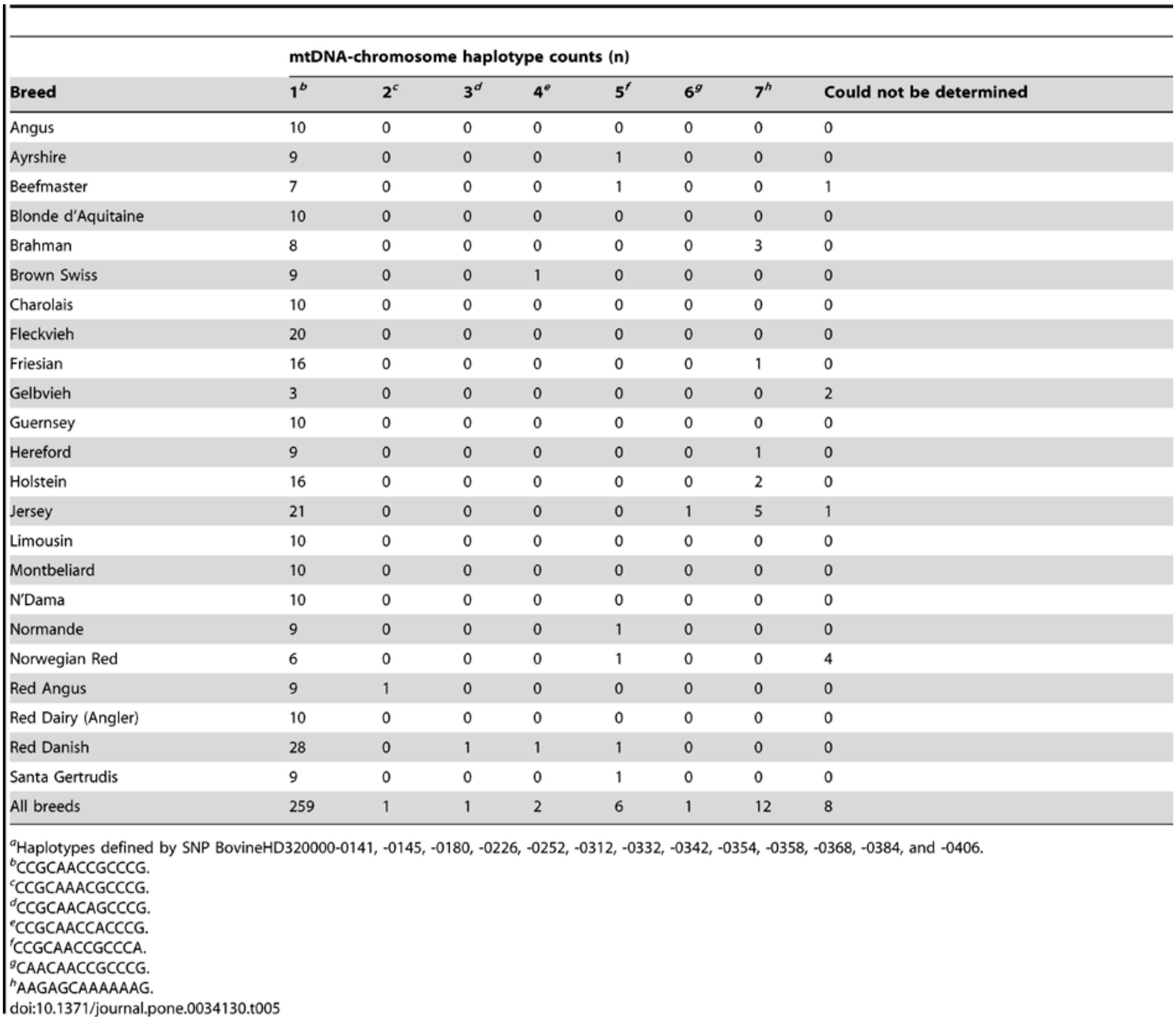


	Y-chromosome haplotype counts (n)			
Breed	1[b]	2[c]	3[d]	4[e]
Angus	0	0	0	9
Ayrshire	0	0	0	9
Beefmaster	2	0	0	5
Blonde d'Aquitaine	0	9	1	0
Brahman	7	0	0	3
Brown Swiss	0	10	0	0
Charolais	0	11	0	0
Fleckvieh	0	18	0	2
Friesian	0	0	5	12
Gelbvieh	0	4	0	1
Hereford	0	0	0	9
Holstein	0	0	0	15
Jersey	0	0	14	5
Limousin	0	10	0	0
Montbeliard	0	10	0	0
N'Dama	0	2	0	0
Normande	0	0	0	10
Norwegian Red	0	0	0	7
Red Angus	0	0	0	10
Red Dairy (Angler)	0	0	0	10
Red Danish	0	0	0	15
Santa Gertrudis	8	0	0	0
All breeds	17	74	20	122

[a] Haplotypes defined by SNP BovineHD310000-0048, -0099, -0103, -0210, -0515, -0517, -1188, -1404, and -1406.
[b] CGCCGCAAC.
[c] TCTCCTCAC.
[d] TCTCCTCAT.
[e] TCTTGTCGC.
doi:10.1371/journal.pone.0034130.t004

Table 4. Animal counts for Y-chromosome haplotypes [a] by breed.

For mtDNA SNPs (Table 5), 259 of the animals sampled had the same mitochondrial haplotype, and seven mitochondrial haplotypes were found. For the mtDNA haplotypes, only haplotype 7 (AAGAGCAAAAAAG) is associated with indicine cattle. Some indicine influence was evident for animals primarily from Australia, New Zealand, and Texas: 5 Jerseys, 3 Brahmans, 2 Holsteins, 1 Friesian, and 1 Hereford. Most taurine indicine cattle were derived from taurine cows. Therefore, the lack of haplotype 7 for taurine breeds in most regions is not unexpected. The BovineLD markers should be useful in determining lineage origin between taurine and indicine breeds or identifying potential admixture within a population of locally adapted animals.



	mtDNA-chromosome haplotype counts (n)							
Breed	**1[b]**	**2[c]**	**3[d]**	**4[e]**	**5[f]**	**6[g]**	**7[h]**	**Could not be determined**
Angus	10	0	0	0	0	0	0	0
Ayrshire	9	0	0	0	1	0	0	0
Beefmaster	7	0	0	0	1	0	0	1
Blonde d'Aquitaine	10	0	0	0	0	0	0	0
Brahman	8	0	0	0	0	0	3	0
Brown Swiss	9	0	0	1	0	0	0	0
Charolais	10	0	0	0	0	0	0	0
Fleckvieh	20	0	0	0	0	0	0	0
Friesian	16	0	0	0	0	0	1	0
Gelbvieh	3	0	0	0	0	0	0	2
Guernsey	10	0	0	0	0	0	0	0
Hereford	9	0	0	0	0	0	1	0
Holstein	16	0	0	0	0	0	2	0
Jersey	21	0	0	0	0	1	5	1
Limousin	10	0	0	0	0	0	0	0
Montbeliard	10	0	0	0	0	0	0	0
N'Dama	10	0	0	0	0	0	0	0
Normande	9	0	0	0	1	0	0	0
Norwegian Red	6	0	0	0	1	0	0	4
Red Angus	9	1	0	0	0	0	0	0
Red Dairy (Angler)	10	0	0	0	0	0	0	0
Red Danish	28	0	1	1	1	0	0	0
Santa Gertrudis	9	0	0	0	1	0	0	0
All breeds	259	1	1	2	6	1	12	8

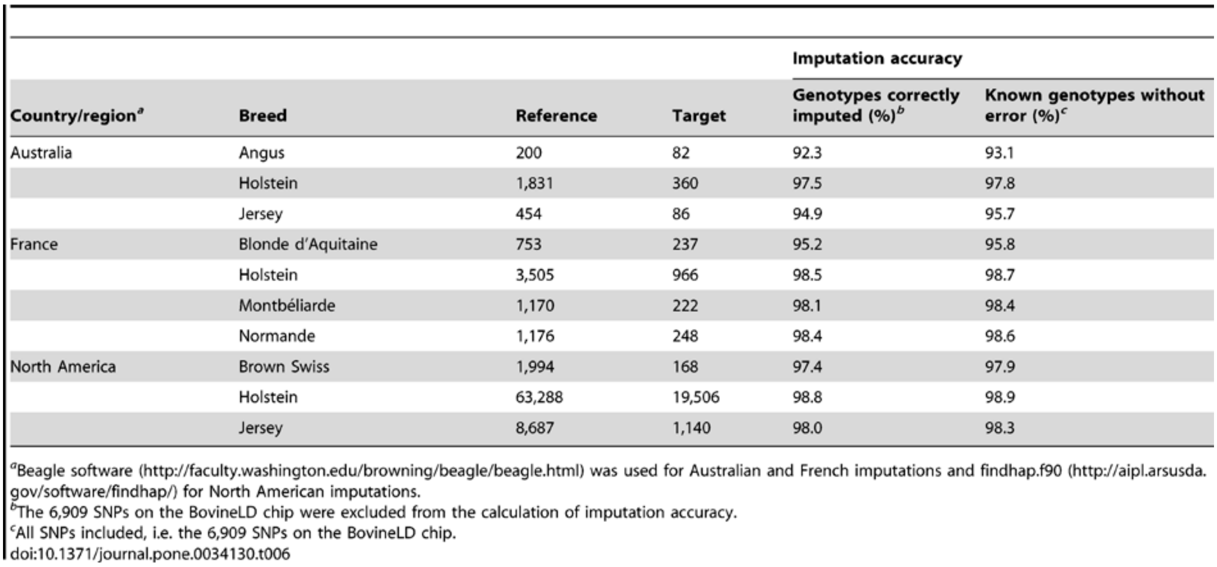
[a] Haplotypes defined by SNP BovineHD320000-0141, -0145, -0180, -0226, -0252, -0312, -0332, -0342, -0354, -0358, -0368, -0384, and -0406.
[b] CCGCAACCGCCCG.
[c] CCGCAAACGCCCG.
[d] CCGCAACAGCCCG.
[e] CCGCAACCACCCG.
[f] CCGCAACCGCCCA.
[g] CAACAACCGCCCG.
[h] AAGAGCAAAAAAG.
doi:10.1371/journal.pone.0034130.t005

Table 5. Animal counts for mtDNA-chromosome haplotypes [a] by breed.

Accuracy of imputation

Imputation accuracy was assessed in Australian, French, and North American cattle populations. In all cases, the accuracy of imputation to BovineSNP50 genotypes was □95% (Table 6). Most imputation results were > 97%, particularly for dairy breeds. The results were lower for some breeds, likely because of the limited reference population size used. For example, the considerably larger size of the North American reference set of Holsteins compared with the Australian set could explain why the North American imputation accuracy was 1.1 percentage points higher than for Australia. The effect of a smaller reference set of genotypes on imputation accuracy was further demonstrated by imputation from BovineLD genotypes for Australian Angus – this breed had the smallest reference size in the data set. For French populations, imputation efficiency also varied, with the highest accuracy for

Holsteins and the lowest for Blondes d'Aquitaine (Table 6); imputation accuracy for Normandes and Montbéliardes was slightly lower than for Holsteins. Again, much of the variation is likely explained by reference population size.

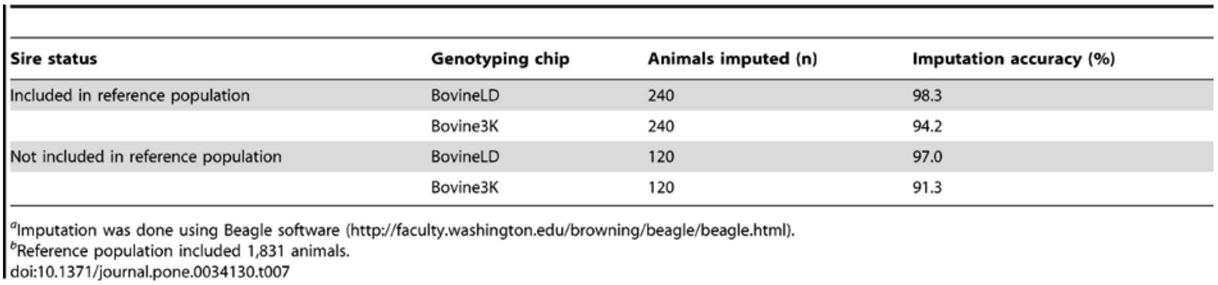


				Imputation accuracy	
Country/region[a]	Breed	Reference	Target	Genotypes correctly imputed (%)[b]	Known genotypes without error (%)[c]
Australia	Angus	200	82	92.3	93.1
	Holstein	1,831	360	97.5	97.8
	Jersey	454	86	94.9	95.7
France	Blonde d'Aquitaine	753	237	95.2	95.8
	Holstein	3,505	966	98.5	98.7
	Montbéliarde	1,170	222	98.1	98.4
	Normande	1,176	248	98.4	98.6
North America	Brown Swiss	1,994	168	97.4	97.9
	Holstein	63,288	19,506	98.8	98.9
	Jersey	8,687	1,140	98.0	98.3

[a] Beagle software (http://faculty.washington.edu/browning/beagle/beagle.html) was used for Australian and French imputations and findhap.f90 (http://aipl.arsusda.gov/software/findhap/) for North American imputations.
[b] The 6,909 SNPs on the BovineLD chip were excluded from the calculation of imputation accuracy.
[c] All SNPs included, i.e. the 6,909 SNPs on the BovineLD chip.
doi:10.1371/journal.pone.0034130.t006

Table 6. Accuracy of imputation from BovineLD genotypes to BovineSNP50 genotypes for Australian, French, and North American breeds.

For Australian and North American Holsteins, accuracy of imputation to BovineSNP50 genotypes was better for BovineLD genotypes than for Bovine3K genotypes. For Australian Holsteins, imputation accuracies were up to almost 6 percentage points higher with the BovineLD chip than with the Bovine3K chip using the same data (Table 7). Mean imputation accuracy was 92.8% for Australian Holstein Bovine3K genotypes compared with 97.6% for BovineLD genotypes. For North American Holsteins, accuracies of imputation to BovineSNP50 genotypes from Bovine3K genotypes ranged from 93.0 to 96.7% (depending on number of parents genotyped) for 2,456 animals genotyped with both Bovine3K and BovineSNP50 chips [17]. Corresponding values for BovineLD genotypes (Table 8) are 96.6 to 99.3%.

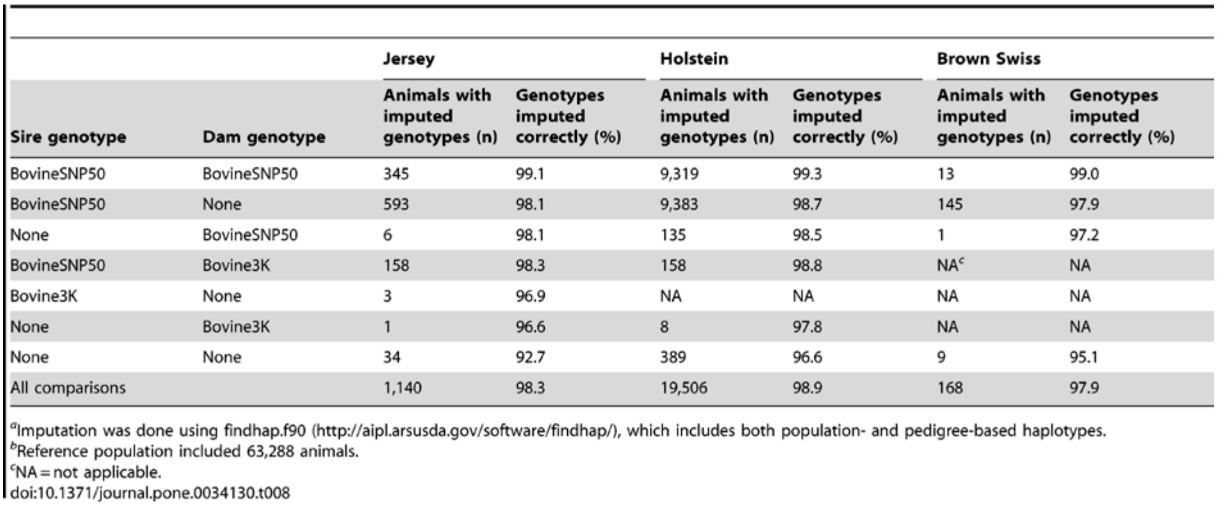


Sire status	Genotyping chip	Animals imputed (n)	Imputation accuracy (%)
Included in reference population	BovineLD	240	98.3
	Bovine3K	240	94.2
Not included in reference population	BovineLD	120	97.0
	Bovine3K	120	91.3

[a]Imputation was done using Beagle software (http://faculty.washington.edu/browning/beagle/beagle.html).
[b]Reference population included 1,831 animals.
doi:10.1371/journal.pone.0034130.t007

Table 7. Accuracy of imputation[a] from BovineLD or Bovine3K genotypes to BovineSNP50 genotypes for Australian Holsteins with and without a sire in the reference population[b].

The greatest improvement in imputation for BovineLD genotypes compared with Bovine3K genotypes was for individuals with no genotyped parents. For Australian Holsteins, difference in mean imputation accuracy with and without a sire in the reference population was 2.9 percentage points for Bovine3K genotypes but only 1.3 percentage points for BovineLD genotypes. The improvement was smaller for North American Holsteins: a difference of 2.7 percentage points between both parents genotyped and no genotyped parents for Bovine LD genotypes (Table 6) compared with 3.7% for Bovine3K genotypes [17]. Compared with North American Holsteins, BovineLD imputation accuracy for animals without a parent in the reference population was slightly poorer for North American Jersey and Brown Swiss populations (Table 8). However, the more than doubling of markers and the different SNP selection criteria [20] compared with the Bovine3K chip allowed high imputation accuracies across a wider range of dairy breeds as well as some beef breeds.



		Jersey		Holstein		Brown Swiss	
Sire genotype	Dam genotype	Animals with imputed genotypes (n)	Genotypes imputed correctly (%)	Animals with imputed genotypes (n)	Genotypes imputed correctly (%)	Animals with imputed genotypes (n)	Genotypes imputed correctly (%)
BovineSNP50	BovineSNP50	345	99.1	9,319	99.3	13	99.0
BovineSNP50	None	593	98.1	9,383	98.7	145	97.9
None	BovineSNP50	6	98.1	135	98.5	1	97.2
BovineSNP50	Bovine3K	158	98.3	158	98.8	NA[c]	NA
Bovine3K	None	3	96.9	NA	NA	NA	NA
None	Bovine3K	1	96.6	8	97.8	NA	NA
None	None	34	92.7	389	96.6	9	95.1
All comparisons		1,140	98.3	19,506	98.9	168	97.9

[a]Imputation was done using findhap.f90 (http://aipl.arsusda.gov/software/findhap/), which includes both population- and pedigree-based haplotypes.
[b]Reference population included 63,288 animals.
[c]NA = not applicable.
doi:10.1371/journal.pone.0034130.t008

Table 8. Accuracy of imputation [a] from BovineLD genotypes to BovineSNP50 genotypes for North American Brown Swiss, Holsteins, and Jerseys with and without parents in the reference population[b].

DISCUSSION

The Illumina BovineLD BeadChip includes 6,909 SNPs selected to provide optimized imputation to BovineSNP50 genotypes in dairy breeds. The SNPs have MAFs of >0.3 in most breeds, and nearly uniform spacing across the genome except at the ends of the chromosome where densities were increased. The chip also includes SNPs on the Y chromosome and mtDNA loci that are useful for determining subspecies classification and certain paternal and maternal breed lineages. Accuracy of imputation to BovineSNP50 genotypes using the BovineLD chip was >99% when both parents were genotyped in the North American BovineSNP50 reference population. That high accuracy suggests that the design criteria for the BovineLD chip would be useful to consider in other species for which an “imputation chip” could dramatically lower the cost of implementing genomic selection. BovineLD imputation was about 3 percentage points more accurate across multiple populations compared with Bovine3K imputation. The improvement was greatest when neither parent had been genotyped. The gain in imputation accuracy is attributed primarily to the increased overall density of the BovineLD chip compared with the Bovine3K chip and also to the even further increased density at the ends of chromosomes. The high MAFs also contribute to the improved imputation accuracy. The MAFs were similar across taurine beef and dairy breed as was the proportion of SNPs that were polymorphic. The similar SNP characteristics suggest that the BovineLD chip will perform well in imputation of taurine beef cattle, but that will be dependent on the size of the population genotyped with a higher density SNP assay. Overall, the new BovineLD BeadChip should facilitate low cost genomic selection in Bos primigenius taurus beef and dairy cattle.

ACKNOWLEDGMENTS

The authors would like to acknowledge the Beef Cooperative Research Centre for providing data. Cécile Grohs (National Institute for Agricultural Research, Centre de Recherches de Jouy en Josas), Christian Bendixen and Rikke Vingborg (GenoSkan), Reiner Emmerling (Institute of Animal Breeding, Bavarian State Research Center for Agriculture), Ekkehard Schütz (Institute of Veterinary Medicine, Georg August Universität), John Flynn (Weatherby’s DNA Laboratory, Irish Equine Centre), Jürg Moll (Qualitas), and Marie-Noëlle Rossignol (Labogena) are acknowledged for providing Gentrain samples. Pfizer Animal Genetics contributed both BovineLD and Bovine3K genotypes for 221 females. The authors thank Suzanne Hubbard (Animal Improvement Programs Laboratory) for her invaluable assistance in manuscript review.

Disclaimers: Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. The U.S. Department of Agriculture is an equal opportunity provider and employer. TSS, CPV, PMV, and GRW were U.S. Government employees at the time they contributed to this work. The work of those individuals was prepared as part of official government duties. Title 17 U.S.C. §105 provides that “Copyright protection under this title is not available for any work of the United States Government.” Title 17 U.S.C. §101 defines a U.S. Government work as a work prepared by a military service member or employee of the U.S. Government as part of that person's official duties. The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the U.S. Department of Agriculture or the U.S. Government.

Author Contributions

Designed the experiment: DB, BJH, TSS, CPV, SF, XD. Collected DNA samples: TSS, CPV, AE, CTL, XD. Conducted Y-chromosome and mtDNA phylogenetic analyses: HC, TSS. Conducted genotype imputation studies: RD, BJH, PMV, GRW. Selected SNPs: SF, GRW (supplemental SNPs). Sample validation: KJG, KV, CTL. Wrote the paper: AE, DB, RD, BJH, TSS, CPV, PMV, GRW, CTL.

REFERENCES

1. Meuwissen THE, Hayes BJ, Goddard ME (2001). Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819–1829.
2. Heffner EL, Jannink JL, Sorrells ME (2011) Genomic selection accuracy using multifamily prediction models in a wheat breeding program. The Plant Genome 4: 65–75.
3. Wiggans GR, VanRaden PM, Cooper TA (2011) The genomic evaluation system in the United States: past, present, future. J Dairy Sci 94: 3202–3211.
4. Schaeffer LR (2006) Strategy for applying genome-wide selection in dairy cattle. J Anim Breed Genet 123: 218–223.
5. Pryce JE, Goddard ME, Raadsma HW, Hayes BJ (2010) Deterministic models of breeding scheme designs that incorporate genomic selection. J Dairy Sci 93: 5455–5466.
6. Pryce, JE, Daetwyler H (in press) Designing dairy cattle breeding schemes under genomic selection—a review of international research. Anim Prod Sci.
7. Van Eenennaam AL, van der Werf JHJ, Goddard ME (2011) The value of using DNA markers for beef bull selection in the seedstock sector. J Anim Sci 89: 307–320.
8. Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78: 629–644.
9. Browning SR, Browning BL (2011) Haplotype phasing: existing methods and new developments. Nat Rev Genet 16: 703–714.
10. Druet T, Georges M (2010) A hidden Markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. Genetics 184: 789–798.

11. Daetwyler HD, Wiggans GR, Hayes BJ, Woolliams JA, Goddard ME (2011) Imputation of missing genotypes from sparse to high density using long-range phasing. Genetics 189: 317–327.

12. Hickey JM, Kinghorn BP, Tier B, Wilson JF, Dunstan N, et al. (2011) A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. Genet Sel Evol 43: 12.

13. VanRaden PM, O'Connell JR, Wiggans GR, Weigel KA (2011) Genomic evaluations with many more genotypes. Genet Sel Evol 43: 10.

14. Druet T, Schrooten C, de Roos AP (2010) Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. J Dairy Sci 93: 5443–5454.

15. Weigel KA, Van Tassell CP, O'Connell JR, VanRaden PM, Wiggans GR (2010) Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels and population-based imputation algorithms. J Dairy Sci 93: 2229–2238.

16. Dassonneville R, Brøndum RF, Druet T, Fritz S, Guillaume F, et al. (2011) Effect of imputing markers from a low-density chip on the reliability of genomic breeding values in Holstein populations. J Dairy Sci 94: 3679–3686.

17. Wiggans GR, Cooper TA, VanRaden PM, Olson KM, Tooker ME (in press) Use of the Illumina Bovine3K BeadChip in dairy genomic evaluation. J Dairy Sci.

18. The Bovine Hap Map Consortium (2009) Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. Science 324: 528–532.

19. Edwards CJ, Ginja C, Kantanen J, Pérez-Pardal L, Tresset A, et al. (2011) Dual origins of dairy cattle farming—evidence from a comprehensive survey of European Y-chromosomal variation. PLoS ONE 6: e15922.

20. Dassonneville R, Fritz S, Boichard D, Ducrocq V (2011) Imputation efficiency with different low density chips in French dairy and beef breeds. Preliminary proceedings of 2011 Interbull meeting, August 27–29, Stavanger, Norway, 4 pages.

2-1.5. MAIN RESULTS

Designing a new Bovine LD SNP chip adapted to as many various breeds as possible was done setting up well described criteria to select markers. Imputation accuracy was measured and compared to existing chips using large data sets of Bovine 50K genotypes. The imputation accuracy results were also compared to those of the former Bovine 3K chip.

The new SNP panel is not a new in silico chip just developed for the purpose of a scientific study, but it is a new tool to be adopted by the breeding industry. For this reason, a real validation procedure was run, testing the new SNP chip in lab conditions.

After stating the need for a reduced-cost tool in order to embrace the huge market of genotyping both dairy cattle females and beef cattle animals, the paper precisely describes the different criteria used for SNP selection. Results of the validation study are shown. Call rates appear to be very high, due to the high technical quality of the selected markers. Concordance rate between 50k and LD genotypes is also very high, showing that these markers display very few genotyping errors, a critical point for further imputation. Concordance rates between true and imputed genotypes are presented for the different breeds. They are high, in fact 3 points higher than with the former 3K chip. With such a high imputation accuracy, the gap between using this new Bovine LD chip rather than the standard Bovine SNP50® is divided by at least 2 in terms of reliability, compared with the former Bovine 3K panel. This study demonstrated that three criteria were essential, technical quality, spacing, and MAF in a variety of populations. Because of the quite large spacing between markers, LD was not a critical parameter.

Many Bos taurus breeds have been considered in the design, and one may expect rather high MAF for all Bos taurus breeds (including those which were not explicitly covered). It would have been even better to also consider Bos indicus breeds. However, the 2 populations are so different that it was not possible to find markers that would fit both. Considering their economic weight, priority was given to Bos taurus.

2.2. – A brief description of the routine imputation procedure implemented in France

First, genotypes and pedigree are checked for parent-offspring inconsistencies. Pedigree relationships or genotypes are erased from the data set when needed. This is important since in the next step, pedigree and genotypes are considered as certain and any mistake may induce wrong imputation or phasing. Simultaneously, Mendelian segregation rules are applied marker by marker taking into account pedigree relationships in order to impute missing markers. This is particularly useful for ungenotyped parents with many progeny.

As a second step, Mendelian segregation rules and pedigree relationships are used in order to impute and phase the different alleles or haplotypes that can be filled in with certainty. This is done using the Linkphase software (described in Druet et al., 2008). The threshold parameter of this software is set to 1.00, meaning that only alleles that can be derived with absolute certainty are accepted. The outputs of this step are partially reconstructed genotypes (and phases).

Note: Initially, this threshold parameter of Linkphase was set to 0.95. However, since imputation accuracy is often above this value, it was preferred to set it up to 1.00. A threshold of 1.00 means that the linkage is poorly used. Indeed, haplotypes segregate during meiosis. But double recombinations, although rare and unrealistic over short distances, do not have a probability strictly equal to 0. Maybe other values for the threshold, such as 0.9999 would be better.

The next step consists in running Beagle (Browning and Browning, 2007) on a sufficiently large data set, in order to build the DAG. It is performed once for good. The main output of such a run of Beagle is not the phases and imputed genotypes, but simply the DAG which summarizes the LD information along the chromosome. This step is very long: on large datasets, it takes days for a single chromosome. But it doesn’t need to be run again at each routine imputation. Indeed the DAG can be stored, and re-used for successive routine imputations. This step needs to be run again only when the genetic map changes, or if DAGs need to be updated, e.g. when the initial reference population was not large enough and is heavily increased.

Finally, one iteration of DAGphase (Druet and Georges, 2010), considering as input files the partially reconstructed genotypes and the DAG, is run in order to fill in the gaps. The software uses LD information from the DAG. The final output is fully imputed genotypes (and phases).

The imputation accuracy is at least as high as with Beagle alone. The whole procedure is quite fast (few hours) because each step is quick (except the Beagle run, but again, it is run just once, and then stored for successive imputations).

This procedure is now implemented in French routine evaluation, both for phasing and imputing low density genotypes.

CHAPTER 3 - Preferential treatment and bias in genomic evaluations

In this chapter, we want to check if performances of genotyped cows can fit within the genomic prediction model. First, potential bias that may affect these phenotypes are described.

3.1. The bias induced by preferential treatment

3-1.1. AN OLD ISSUE IN GENETIC EVALUATIONS

- **Definition of preferential treatment**

Preferential treatment can be defined as management practices which modify production. These practices are selective, they are applied to some cows, but they are not applied to all their herd mates (Kuhn et al., 1994). They may be related to housing, feeding or reproduction practices. Some preferential treatment may occur inadvertently (e.g. feeding cows according to their production). Intentional preferential treatment also occurs and is usually used to enhance the likelihood that a cow will be chosen as bull-dam.

- **Consequences of preferential treatment on genetic evaluations**

The bias induced by preferential treatment was first observed through the inconsistencies between the parent average of a bull (mostly influenced by the breeding value of his dam) and performances of his daughters (included to calculate his breeding value after progeny testing). Van Vleck (1987) demonstrated that EBV of bull dam did not predict performances of daughters of her son as accurately as theory predicts. Colleau (1989) also showed that the use of bovine somatotropine applied to some cows may induce bias in genetic evaluations.

Kuhn et al. (1994) used a simulation study to assess the magnitude of the bias induced by preferential treatment. Bias ranged between 15 and 450 kg (representing 6 to 39 % of the overestimation of the production initially inserted). Bias highly depends on the extent of preferential treatment, the number of records involved, and mostly, whether relatives are also preferentially treated. Powel and Norman (1986) also demonstrated that bias becomes even more important when preferential treatment is also applied to relatives.

- **Progeny testing protects (somehow) from such consequences**

For the last decades, progeny testing has been the standard in terms of evaluation of young bulls before their extensive use for artificial insemination. As the bulls were mated to cows randomly selected among the commercial population, one can consider that their daughters were not subject to preferential treatment. Only their dam may have been affected by such a potential bias. However when dozens of daughters bring information to properly estimate the bull's breeding value, the contribution of his dam's performances to his EBV is reduced.

Therefore the potentially overestimated performances of the dam have a very limited impact on the ranking of the bull. This is why one can consider that progeny testing "protects" from preferential treatment bias.

When a farmer had to choose within a batch of progeny-tested bulls, his choice was not biased because of preferential treatment. This is different for breeding organizations. They needed to choose young males to enter progeny-testing and this choice was based on parent average which may be heavily biased because of overestimation of the breeding value of the bull dam.

- **Accounting for heterogeneity of variance**

Environmental factors as well as management practices induce differences among herds in terms of variability of performances. Thus, genetic and residual variances may differ among herds. If this difference is not accounted for, Vinson (1987) showed that it leads to overvaluation and selection of more individuals from the variable herds. This results in a reduced response to selection. Wiggans and Van Raden (1991) reported that, when the genetic evaluation model is adjusted for heterogeneous variances, a slight gain is observed for the correlation between parent and offspring information, and, moreover, the genetic trend (for milk yield) is increased by about 5 kg/year. Van der Werf et al. (1994) showed that when such heterogeneity of variance is accounted for in genetic evaluation model, 20 % of the bias (between parent average and breeding value based on progeny records) can be removed.

3-1.2. Why this old issue is highlighted by genomic selection

Genomic selection allows the systematic use of young bulls for AI. In some countries, these young bulls are evaluated with direct genomic values (DGV) which do not include any polygenic effect, but most of the evaluation centers chose to evaluate young candidate with GEBV which include information from performances of relatives. For a young bull, the residual polygenic contribution is the parent average, and is not very reliable. The impact of the breeding value of the dam is very important, and contributes as much as the sire. For this reason, bias due to preferential treatment is likely to occur, since an increased (overestimated) index for the dam will induce an increased (overestimated) GEBV for the young candidate.

Farmers using semen from young bulls are not as “protected” from preferential treatment bias as they used to with progeny-tested bulls. On the other hand, breeding organizations used to select young bulls candidates to progeny testing only based on parent average. They now have a much more efficient tool: indeed, GEBV are more reliable than parent average and less biased by preferential treatment. The same situation occurs with selection of bull dams.

- **Effect of biased performances on genomic prediction equations**

Wiggans et al. (2011) used an elegant approach to demonstrate what non-adjusted performances of the female reference population may induce on prediction equations. The method to adjust performances will be described hereafter.

The female reference population consisted of cows, genotyped on the standard 50K chip, with own performances. They most likely are potential bull dams, and are the most concerned by preferential treatment.

As reported below, United States have chosen to include genotyped cows in the reference population. Wiggans et al. (2011) proposed a method to adjust cows performances (see paragraph 3-5.3). A very convincing result is the “Manhattan plot” they compare. They reported marker effects across the genome (usually called a “Manhattan plot”) in 2 situations, one where cows performances are adjusted, one where performances are not adjusted. There was no sex effect included in the model, and region of the X chromosome is also shown.

What we can see from the Figure 8 is that with unadjusted performances, SNP effects are homogeneous across chromosomes 1 to 29, but a high peak is present on the X chromosome. This is no longer the case with adjusted performances.

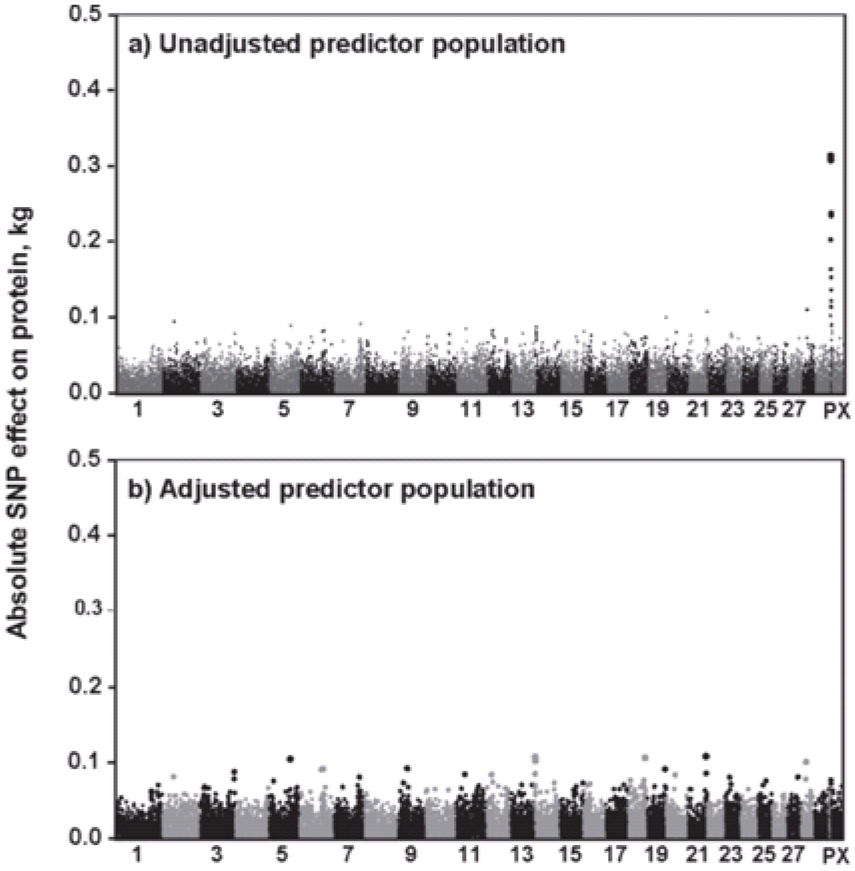


Figure 8 (from Wiggans et al. 2011) Manhattan plots with either adjusted or unadjusted cows performances for the reference population

When looking at the Manhattan plot with unadjusted predictor population, one may conclude that there exists a big QTL on this region of the X chromosome. There is not such a big QTL known to be involved in milk production in that region, especially because it does not appear anymore when performances are adjusted (or when only males are in the reference population). Markers of the X chromosome are present in 2 copies in females’ genotypes, but only one copy in males genotypes. This region accounts for the difference between cows EBV and bulls EBV due to bias in cows EBV. So, we have a clear picture of the consequences of biased performances over prediction equations. One can easily consider that using such prediction equations may be misleading in terms of proper estimation of breeding values.

3-1.3. Preliminary study

- **Introduction**

Several national genomic evaluation centers wondered in early 2010 whether females should be included in the reference population, and, if yes, how to properly include their performances.

As stated in section 3.3. , Canada, or the different countries of Eurogenomics chose not to include females in the reference population. On the other hand, the US have included cows in the reference since the beginning of the genomic era. In 2010, considering that performances of dams were also bringing bias in evaluations, the USDA center responsible for evaluations in the US decided to adjust cow performances (Wiggans et al., 2011).

The French genomic evaluation model involves a polygenic effect. Until 2010, the whole pedigree information was used, including dams of genotyped individuals. Their performances were integrated in the QTL BLUP and were used to compute GEBV.

The Eurogenomics consortium suggested to remove potentially biased cow performances from the reference population in order to get more reliable GEBV. Before removing such performances from the equations, it was decided to conduct a study to assess the impact of including direct individual cows performances in genomic evaluations.

Required definitions

When using a reference population consisting of AI bulls, the phenotypic information of daughters performances is summarized in DYD. Daughter Yield Deviations (DYD) were defined by Van Raden and Wiggans (1991). They correspond to average daughter performances corrected for fixed effects such as herd or season, the permanent environment effect, as well as the additive genetic contribution of the dam). Afterwards, these DYD are used as if they were the performances of the bulls themselves.

The equivalent phenotypic measure for females is the yield deviation, YD. It corresponds to the performance of the cow herself (not of her progeny), also corrected for the other effects than the genetic effect.

- **Material and methods**

Data

Data consisted of 3966 progeny-tested Holstein bulls. This reference population was split into 2 populations, a training population including the older bulls for which both phenotypes and genotypes were included in the model and a validation population including bulls for which phenotypes (even when available) were removed from the equations. The genomic evaluation aimed at properly predicting their DYD.

In the validation procedure estimated breeding values of validation bulls were compared to their phenotypic value (i.e. their DYD).

phenotypic performances

Two different datasets were considered : one only including the DYD of the 3,505 male animals present in the pedigree file, and one including both DYD of male and YD of female individuals. 3,830 cows with own performances were added in the second data set. When YD were included, the performances of genotyped cows were not used in the calculations of DYD. Among the individuals with phenotypic records (DYD or YD), some may have been genotyped, but not all of them. These phenotypic records included in the model were used to estimate jointly polygenic effects and QTL effects.

Traits studied

Two traits were specifically studied, as they are known to be opposite as far as preferential treatment is concerned. The first one was milk production, expressed in kg. It is the trait for which management practices have the largest influence, either unintentionally (amount of concentrate as feed intake according to production) or on purpose (distinct milking procedures, distinct feeding based on the status -bull dam/regular cow). It is the trait usually considered in preferential treatment studies.

The second trait is somatic cell count (SCC), expressed as a log transformed function of the number of somatic cells counted per ml of milk. It is considered as one of the traits the least prone to preferential treatment. Indeed, management practices (housing, hygiene during milking) are very likely to affect all of the contemporaries in a given herd and it is very difficult for a breeder to change on purpose the performances of some specific cows. Even if separate housing may have some effect on performances, it is likely to have a very limited impact for somatic cell count.

Not only the type of trait is different (production vs health trait) but also the heritability differs between milk production and somatic cell count. Heritability for milk production is around 0.3 whereas heritability of SCC is around 0.15. This difference in heritability has an impact on the amount of information that comes from one single own performance for a given cow. This performance will have a bigger impact on EBV (or GEBV) of the animal for milk just because of this higher heritability. Note that this difference in heritability may be misleading in the interpretation of the results: what we would consider as an effect of preferential treatment may be related to the genetic parameter of the traits.

Excepted results under the assumption of bias due to preferential treatment

Our hypothesis is that the evolution of correlation between GEBV and DYD in the validation population when YD are removed from the model differ for these 2 traits: it is likely to increase for milk, since potentially biased performances are removed and this should increase accuracy. On the other hand, this correlation is likely to remain the same for SCC (no biased performances initially) or even decrease (some interesting information is being removed).

- **Results**

	DYD	GEBV (DYD)
GEBV (DYD)	0,609	
GEBV (DYD+YD)	0,608	0,967

MILK

	DYD	GEBV (DYD)
GEBV (DYD)	0,698	
GEBV (DYD+YD)	0,707	0,991

SCC

Table 3 Correlation among males in the validation population between phenotypes (expressed as DYD) and GEBV, calculated with or without including female own performances (expressed as YD).

From Table 3 Correlation among males in the validation population between phenotypes (expressed as DYD) and GEBV, calculated with or without including female own performances (expressed as YD)., the correlations between DYD and GEBV appear to be more or less the same whether YD are included or not in the calculations (differences only appear at the third digit level). This is true for both (milk and SCC) traits. It does not fit with our initial hypothesis. One may argue that removing YD induced a slight decrease of the correlation between DYD and GEBV for SCC.

Note : Correlations for SCC were higher than for milk. One may consider at first that genomic prediction better predict somatic cells count rather than milk production. This is not the true reason. Heritability of SCC is lower, so the performance of an individual brings less information to estimate the breeding value, and the DYD of candidates (validation population) are not as accurate (and close to the true genetic value) for SCC than for milk. In the same way, correlation between GEBV (with or without YD) is closer to 1 for SCC. One first explanation is that GEBV is less affected by biased performances and the ranking remains the same. Another possible explanation is the lower heritability of the trait.

When analyzing these results, it is like looking at a glass as half full or half empty. Indeed, some information has been removed (YD) with no loss in accuracy (no decrease in correlation) which means that this additional information was (at least partly) biased, bringing as much noise as fruitful information. However we were not able to clearly demonstrate the difference between the two traits whereas one is susceptible to preferential treatment and the other one should not.

One possible explanation to such a low impact of removing YD from equations is that we were here looking at the impact on male individuals. One can expect a higher impact when looking at cows’ genomic predictions.

Additionally, we were looking at one single population, with no clue on whether some individuals may or may not have been more affected by preferential treatment.

In relation with this preliminary study, the decision was taken to remove own cows’ performances (YD : yield deviations) from the genomic evaluations in France for two reasons. The first is that other members of the Eurogenomics consortium chose not to include females’ performances in the genomic evaluations in order to get “a priori” more reliable genomic breeding values. Second, our study confirmed that YD brought as much noise as they brought additional useful information since the correlation remained the same whether YD were included or not. However, additional studies are required in order to find a way to properly include YD in the model.

From a scientific point of view, the difference between our initial hypothesis and the conclusions we drew from this preliminary study convinced us to look deeper in this issue. The objective would be to look at female GEBV and to distinguish 2 cow populations, one more prone to preferential treatment than the other.

3.2. Article IV Inclusion of cows performances in genomic evaluations and its impact on bias due to preferential treatment

Article IV

Dassonneville R., A. Baur, S. Fritz, D. Boichard, V. Ducrocq. 2011

Inclusion of cows performances in genomic evaluations and its impact on bias due to preferential treatment

Submitted

ABSTRACT

BACKGROUND

Genomic evaluations now are an essential feature of dairy cattle breeding. While young bulls were the initial target of genomic selection, a rapidly increasing number of females (both heifers and cows) are now being genotyped. A rising issue is whether and how own performances of genotyped cows should be included in genomic evaluations. The purpose of this study was to assess the impact of including yield deviations (YD) – i.e., the own performances of cows – in genomic evaluations.

METHOD

Two different genomic evaluations were performed, one including only reliable DYD (daughter yield deviations) of proven bulls, and another one including both YD for females and DYD for males based on their non genotyped daughters. Two traits were studied: milk yield (kg), which is the trait the most prone to preferential treatment and somatic cell count (SCC) for which such a bias is very unlikely. Data consisted of two different groups of animals from the three main dairy breeds in France: 11,884 elite females genotyped by breeding companies and 7,032 cows genotyped in a side research project (and considered as randomly selected among the commercial population).

RESULTS

For several measures potentially related to preferential treatment bias, the elite group presented a different pattern than the other trait/group combinations for milk yield: for instance, the average difference between breeding values including YD or not was significantly different from 0. The correlations between breeding values for milk yield from evaluations with or without YD were lower for elite females compared to randomly selected cows. For SCC, they were very similar.

CONCLUSIONS

The study demonstrated that explicitly including own milk performances of elite females induced biased (over-estimated) genomic evaluations. There is a need for a special treatment of milk production performances of elite cows in genomic evaluations.

Key words: genomic selection, dairy cattle, reference population, preferential treatment, bias

BACKGROUND

Preferential treatment and the bias it induces are an old issue in genetic evaluations. Preferential treatment can be defined as management practices which modify production. These practices are selective: they are applied to some cows, but they are not applied to most of their herd mates (Kuhn et al., 1994). They may be related to housing, feeding or reproduction practices. The bias induced by preferential treatment was observed through the inconsistencies between the parent average of a bull (influenced by the breeding value of his dam) and performances of his daughters (included to calculate his breeding value after progeny testing) (Van Vleck, 1987).

The issue related to bias due to preferential treatment got highlighted again with the development of genomic selection. In genomic selection, the reference population consists of individuals with both genotypes and performances which are used to estimate marker effects. The larger the reference population size, the more reliable the genomic evaluations (Goddard and Hayes, 2009). At an early stage, the reference population was only composed of progeny-tested bulls and genomic evaluations were only based on reliable averaged performances of the bulls' daughters. Considering the rapidly increasing number of genotyped cows with own performances, it is very appealing to include these genotyped cows in the reference population and it will be necessary in the future to upgrade the reference population if the number of bulls with a progeny evaluation is much lower than in the past. Within the female population, potential bull dams were the first target for genotyping. The use of potentially biased performances of these genotyped elite females in genomic evaluations may have two major impacts, the first one on GEBV of these cows and of their relatives, the second on prediction equations.

Ways to deal with this issue vary across countries. For example, the U.S. chose very early to include genotyped cows in the reference population (Wiggans et al. 2011). Fearing potential bias, Canada but also the Eurogenomics consortium (Lund et al., 2011) decided not to include cows in the reference population. Later, Wiggans et al. (2011) proposed a method to adjust female performances before inclusion in genomic predictions. Genomic evaluations in the U.S. have been corrected since then. There is a need to more precisely assess the impact on the reliability of genomic predictions of the inclusion of genotyped cows in the reference populations on the reliability of genomic predictions.

When using a reference population consisting of AI bulls, the phenotypic information of daughters' performances is summarized in DYD. Daughter Yield Deviations (DYD) were defined by VanRaden and Wiggans (1991). They correspond to the average daughters performances corrected for all fixed effects (such as the herd, year, season effects among others), the permanent environment effect, and also for the genetic contribution of the bull's mate (i.e., half the additive genetic value of the cow's dam). Subsequently, they are used as if they were the performance of the bulls themselves.

The equivalent phenotypic measure for females is the yield deviation, YD. It correspond to the performance of the cow herself (not her progeny), corrected as well for all the effect but the genetic effect.

In a preliminary study where female own performances were either included or excluded (results not shown), the correlations between phenotypic (DYD) values and GEBV for bulls of the validation population were measured for several production traits and for somatic cell count. This value can be regarded as the square root of realized reliability. These correlations did not decrease when the own performances of the genotyped cows which are ancestors of males candidates were removed from the training set. It can be seen from two opposite angles. On one hand, removing female information did not result in a loss in accuracy which means that this additional information was (at least partly) biased, bringing as much noise as fruitful information. However on the other hand, it should be noted that one could have expected a gain in correlation if the removed information had been heavily biased.

The objective of this study was to compare predictions obtained after two distinct genomic evaluations. For the first one, only bulls were included in the reference population while in the second one, genotyped cows were also added to the reference population. Two traits (milk yield and somatic cell count (SCC)) were considered. They differ by their particularities concerning preferential treatment. In contrast with other studies including our preliminary work, two distinct cow populations were considered, one including only elite dams and another one including cows (nearly) randomly selected among the commercial population. The latter is supposed to be less affected by preferential treatment. Under the assumption of the existence of a bias induced by preferential treatment, different characteristics of genomically enhanced breeding values (GEBV) would be expected for the elite group for milk yield compared to the other combinations of trait and group.

METHODS

Data

The study focused on genotyped dairy cows. Three French dairy breeds were investigated separately for this study: the Holstein, Montbéliarde, and Normande breeds. Two distinct populations were looked at: elite females and randomly selected cows. Females (both heifers and cows) genotyped by breeding companies were considered as elite females. It was assumed that if a breeding organization was interested in genotyping a particular female candidate seen as a potential bull dam, and ready to pay to obtain a genomic evaluation of this female, then such an animal could be defined as "elite". As a reference, non preferentially treated group, a representative subset of the commercial population was needed. On a side research project studying genetic and environmental parameters of milk fatty acids composition, about 8,000 dairy cows were genotyped. They were specifically chosen to be representative of the commercial population. Cows to genotype were determined based on constraints on a set of sires (20 for Normande and Montbéliarde, 30 for Holstein) from which a limited number of daughters per sire were randomly selected within a given set of partner herds. We considered these cows as "randomly selected". Elite cows are more likely to be preferentially treated with over-estimated performances while the other group of cows is less prone to such a bias.

Obviously, the reference populations also included progeny-tested bulls, distributed across several generations. Table 1 summarizes the number of animals included in the training population for the three breeds, the number of genotyped elite females and the number of genotyped "randomly selected" cows. In total, 1,798, 2,157 and 19,485 genotyped progeny-tested bulls were included in the reference population for Normande, Montbéliarde and Holstein breeds respectively. The Normande and Montbéliarde bulls reference populations only included individuals genotyped in France whereas the Holstein reference population also comprised bulls genotyped by European partner breeding organizations which genotypes were exchanged within the Eurogenomics consortium (Lund et al., 2011). All the individuals were genotyped with the Bovine 50K chip (Illumina inc., San Diego).

Table 1. Number of genotyped individuals for the three different breeds for each group

Breed	AI bulls	Elite cows	Randomly selected cows
Montbéliarde	1,798	2,190	1,826
Normande	2,157	2,129	2,374
Holstein	19,485	7,565	2,832

Performances included for the reference population

For each breed, two kinds of genomic evaluations were computed. In the first one, only males were included in the training population, and only performances related to males were used to estimate markers effects. A second genomic evaluation was performed on the similar datasets, in which genotyped cows with own performances were also added to the training population. For this second evaluation, performances from both males and some females were used to estimate markers effects.

Both YD and DYD are by-products of the official polygenic evaluations. Data used for this study were obtained after the official evaluation of November 2011. They were used as inputs for the genomic evaluations. Note that when YD were used for genotyped cows, their contribution was removed from their sires' DYD in order to avoid double counting their performances.

Both lactation milk yied and SCC were evaluated with an animal model (polygenic, pedigree-based) evaluation. For milk yield, heterogeneity of variances were accounted for in the model as described by Robert et al. (1999). YD and DYD were then corrected for heterogeneity of variance and expressed as in a standardized (reference) environment.

Genomic evaluation model

The French genomic prediction is an extension of the marker-assisted evaluation approach of Fernando and Grossman (1989), making use of haplotypes of 5 SNP. The QTL-BLUP model can be written as:

$$y = 1\mu + Zu + \sum_{i=1}^{nQTL} (h_{i1} + h_{i2}) + e$$

where **y** is the vector of phenotypic observations, µ is the overall mean, **u** is a vector of random (pedigree-based) residual polygenic effects, h_{ij} is the random effect of haplotype j for QTL i, and **e** is a vector of residuals, with heterogeneous residual variances inversely proportional to EDC (equivalent daughter contributions).
The selection of QTL included in the model was the result of a combination of 2 approaches (Boichard et al., 2012a). First, dozens of QTL per trait were detected after QTL fine mapping using a linkage disequilibrium linkage analysis (LDLA) as defined by Druet et al. (2008). Then, hundreds of haplotypes were chosen using the Elastic Net algorithm (**EN**) (Croiseau et al., 2011).

Individuals were included alltogether to estimate genomic breeding values. Elite and “randomly selected” subsets were analyzed separately later, but the GEBV of individuals present in these two populations came from the same evaluation.

RESULTS

Mean and standard deviations of the EBV for the 2 cow populations (elite and randomly selected) are reported in table 2. This table underlines the existing difference in the genetic merit of the two populations for the traits of interest. EBV come from the official polygenic evaluation of November 2011. The difference between the average genetic merit of the two populations ranged from 359 to 737 kg for milk yield (corresponding to 0.5 to 0.95 genetic standard deviation) and from 0.22 to 0.45 (expressed in genetic standard deviation units) for somatic cell count. For both traits, the elite population presented a genetic superiority over the randomly selected population. The difference was more important for milk yield than for SCC.

Table 2. Statistics (mean and standard deviation) of official national EBV for the two cow populations (elite and randomly selected) and for two traits (milk yield and Somatic Cell Count (SCC) expressed in kg of milk and genetic standard deviation (for SCC) for the three breeds

		elite		random	
	breed	mean	s.d	mean	s.d
Milk yield	Montbéliarde	663	348	304	320
	Normande	717	318	203	333
	Holstein	1055	462	318	386
SCC	Montbéliarde	0.26	0.69	0.04	0.59
	Normande	0.29	0.64	-0.16	0.73
	Holstein	0.4	0.63	-0.05	0.69

Correlations between GEBV obtained with the two reference populations are shown in table 3. GEBV were obtained when including either both bulls' DYD and cows' YD (GEBV$_{(DYD+YD)}$), or only DYD (GEBV$_{(DYD)}$). Under the assumption of the existence of a bias induced by preferential treatment affecting only milk production of elite cows, these correlations should be similar when looking at elite or randomly selected group for SCC but lower for the elite group and milk yield.

Table 3. Correlation between GEBV calculated including both DYD and YD and GEBV calculated only including DYD for the two cow populations (elite and randomly selected) and for two traits (milk yield and Somatic Cell Count (SCC)

breed	trait	elite	random
Montbéliarde	Milk yield	0.740	0.770
	SCC	0.915	0.893
Normande	milk yield	0.768	0.820
	SCC	0.900	0.909
Holstein	Milk yield	0.931	0.938
	SCC	0.966	0.965

For the Normande breed, the correlations for SCC were essentially the same whether elite or randomly selected cows populations were included (difference lower than 0.01). On the other hand, the correlation was substantial higher for the randomly selected population compared to the elite group for milk yield (0.82 instead of 0.77). For the Montbéliarde, the correlation was also higher for the randomly selected group for milk trait (0.77 instead of 0.74) but an opposite pattern was observed for SCC. Indeed, the correlation was lower for the randomly selected group (difference of 0.02). For the Holstein, all observed differences in correlation were small: there was almost no difference (0.001) for the two populations for SCC and only a slight difference (0.007) for milk yield.

Boxplots of the differences between GEBV$_{(DYD+YD)}$ and GEBV$_{(DYD)}$ for the three breeds are presented in figures 1 to 3. For each breed, 4 boxplots are displayed, one per trait x population combination. Note that the values for milk yield is expressed in kg (one genetic standard deviation equals 591, 661 and 759 kg for Normande, Montbéliarde and Holstein breeds respectively) whereas they are expressed in genetic standard deviation for SCC. In absence of bias in the data, these boxplots should be centered around 0, and symmetrically distributed around 0.

Figure 1 Dassonneville et al.

Boxplot representing the difference between GEBV calculated including both DYD and YD and GEBV calculated only including DYD for the two cow populations (elite and randomly selected) and for two traits (milk yield and Somatic Cell Count (SCC) expressed in kg of milk and genetic standard deviation (for SCC) for the Montbeliarde breed

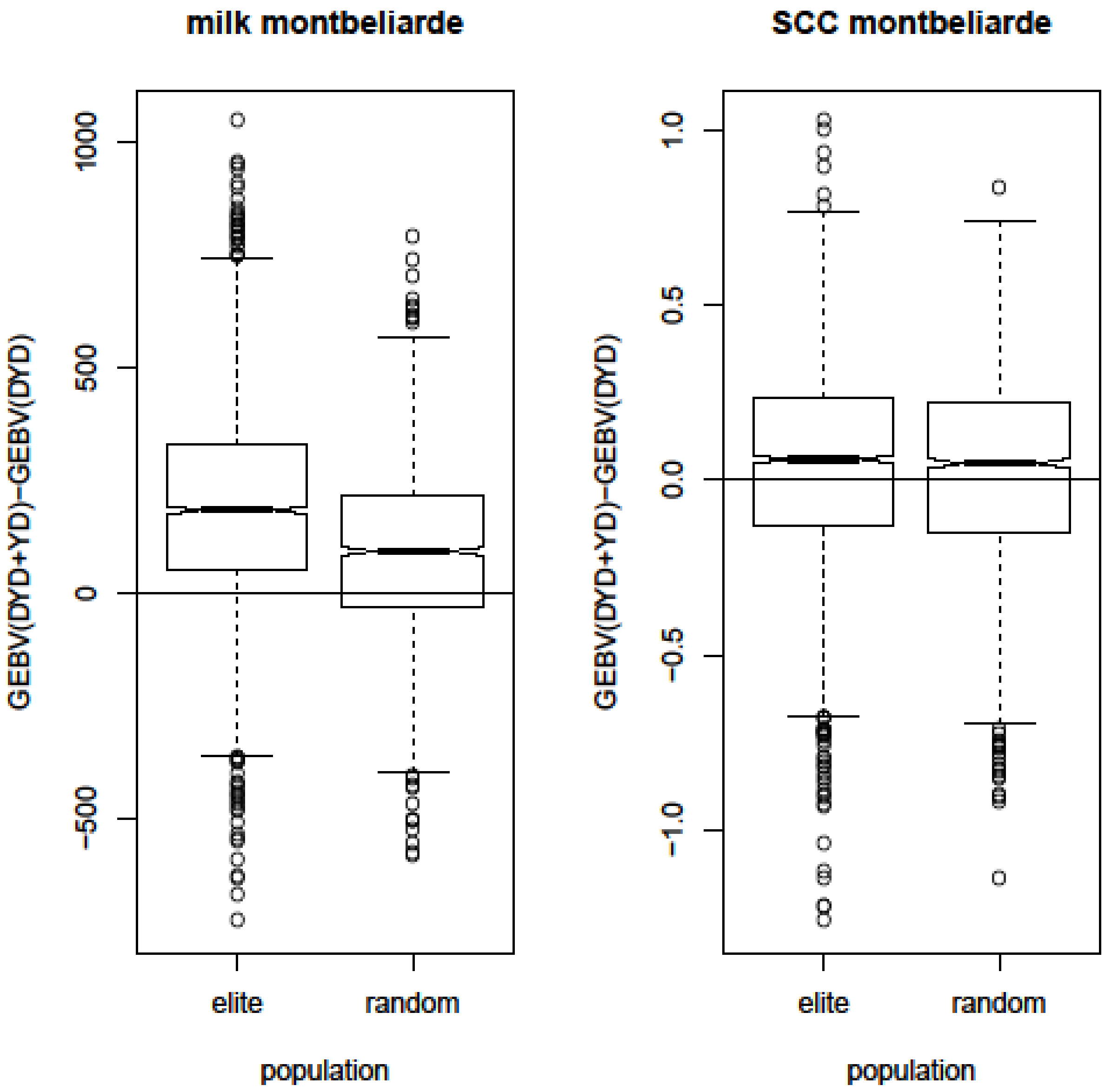


Figure 2 Dassonneville et al.

Boxplot representing the difference between GEBV calculated including both DYD and YD and GEBV calculated only including DYD for the two cow populations (elite and randomly selected) and for two traits (milk yield and Somatic Cell Count (SCC) expressed in kg of milk and genetic standard deviation (for SCC) for the Normande breed

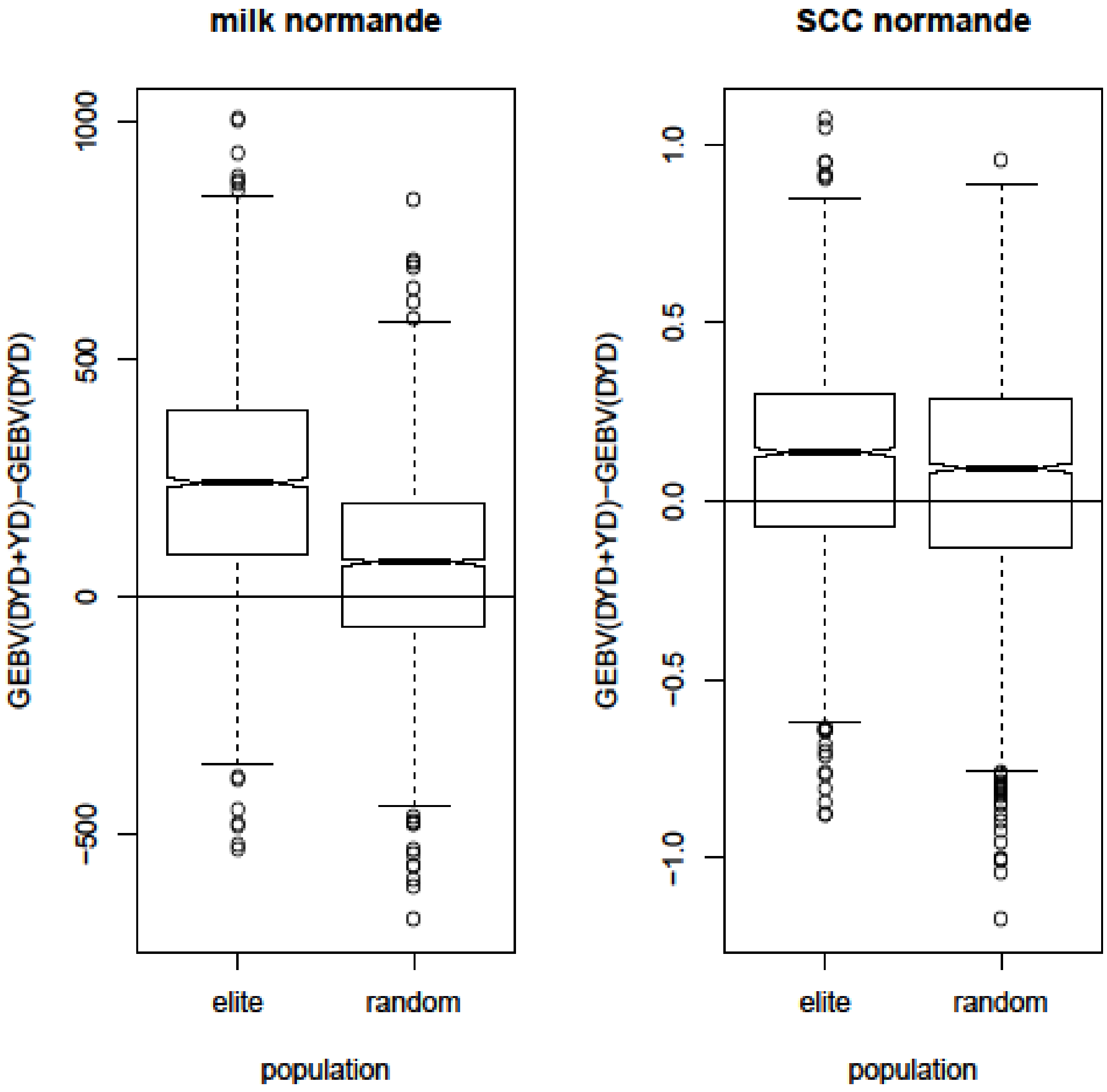
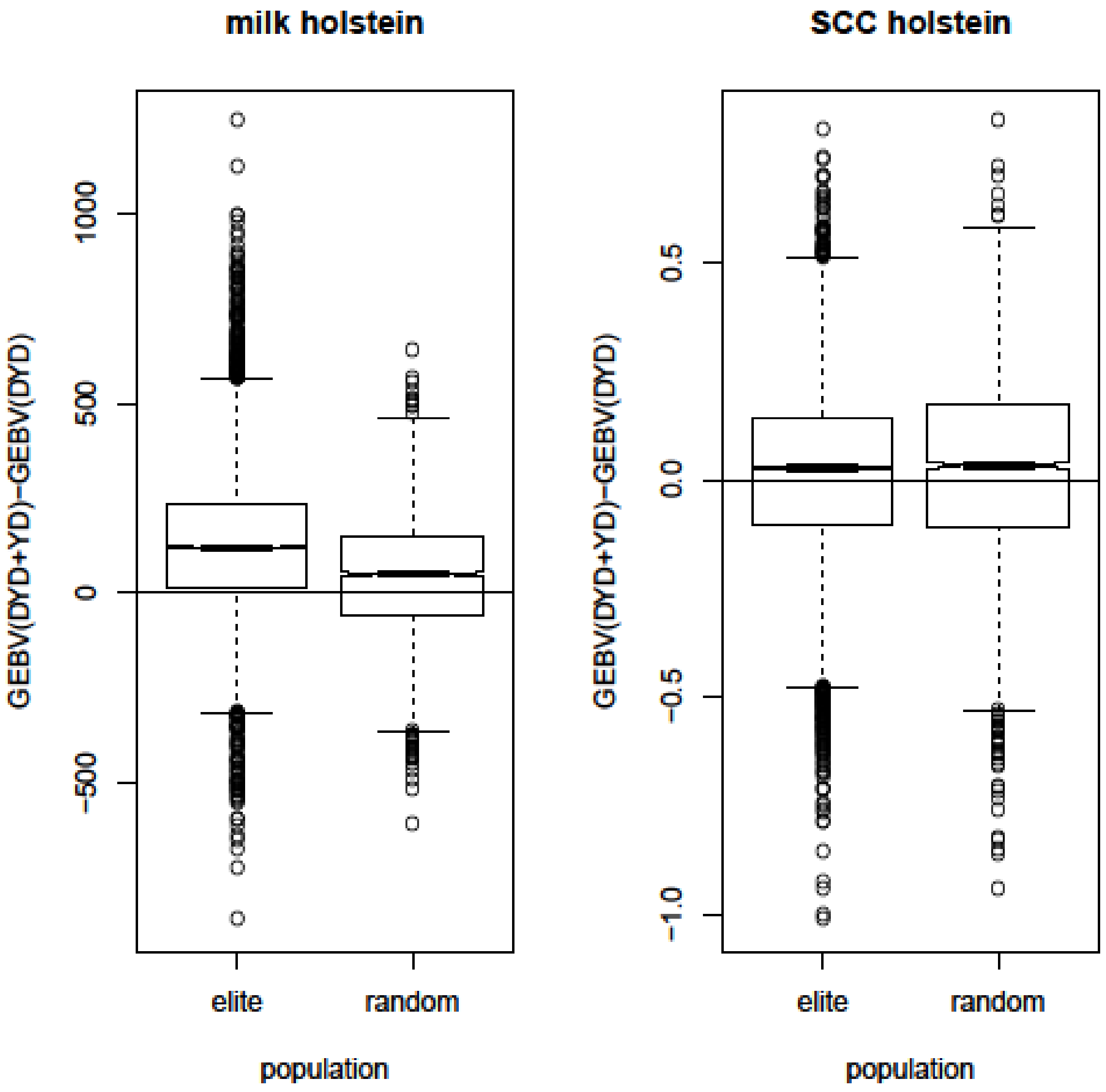


Figure 3 Dassonneville et al.

Boxplot representing the difference between GEBV calculated including both DYD and YD and GEBV calculated only including DYD for the two cow populations (elite and randomly selected) and for two traits (milk yield and Somatic Cell Count (SCC) expressed in kg of milk and genetic standard deviation (for SCC) for the Holstein breed



For each of the three breeds, the same pattern was observed; three boxplots out of 4 had a mean close to 0. The null value was included in the second quartile for SCC and the two female groups, and for milk yield only for the randomly selected group. Milk yield GEBV for the elite population clearly presented a different pattern; the box was entirely above the 0 line meaning than more than 75% of the cows had a GEBV$_{(DYD+YD)}$ higher than GEBV$_{(DYD)}$. This was not observed for SCC. It can pointed out that the median is not strictly equal to 0 for any of the 4 boxplots, though it was very close for SCC in Holstein. The elite population also presented a higher variability of differences for milk yield, especially in the Montbéliarde and Holstein breeds.

Table 4 focuses on the average difference between GEBV$_{(DYD+YD)}$ and GEBV$_{(DYD)}$. Under the suspicion of a bias induced by preferential treatment only affecting milk yield of elite cows, this difference should be 0 when looking at the randomly selected group or when considering SCC but significantly larger from 0 for the elite group for milk yield.

Table 4. Average difference between GEBV calculated including both DYD and YD and GEBV calculated only including DYD for the two cow populations (elite and randomly selected) and for two traits (milk yield and Somatic Cell Count (SCC) expressed in genetic standard deviation for the three breeds

breed	trait	elite	random
Montbéliarde	Milk yield	0.286	0.135
	SCC	0.040	0.020
Normande	Milk yield	0.409	0.117
	SCC	0.110	0.060
Holstein	Milk yield	0.168	0.060
	SCC	0.013	0.024

Across breeds, these average differences between GEBV$_{(DYD+YD)}$ and GEBV$_{(DYD)}$ were close to 0 and similar whenever randomly selected or elite groups were considered for SCC. The differences were 0.01 and 0.02 genetic standard deviation for Holstein and Montbéliarde , respectively. The elite group displayed a slightly higher value for SCC for the Normande

(0.11 instead of 0.06). Indeed, these values were again not strictly equal to 0. But for milk yield, this average difference was much larger for the elite group than for the randomly selected group. Indeed, this difference was roughly 2, 3 and 4 times greater for the Montbéliarde, Holstein, and Normande breeds, respectively. In absolute terms, the average difference observed in Normande for the elite population was 0.3 genetic standard deviation larger.

DISCUSSION

Key assumptions in our study were that females (heifers and cows) genotyped by breeding companies are elite cows and that the cows genotyped for the research project could be considered as representative of the commercial population. Indeed, each breeding company has its own strategy for bull dam selection: some genotype a wide proportion of the population and select from a broad basis while others are more selective and only genotype top cows based on their total merit index, the different breeding companies may put a different emphasis for the several traits, or the sire analysts from some cooperatives may focus on a limited number of maternal cow families (more likely to be affected by preferential treatment). For the cows genotyped in the side research project, even if the sires (constraint set on number of progeny genotyped) were the most used within each breed, our population may not be a perfect random sample of the commercial population. However, the two populations are easily identified when looking at the average cows EBV, since the elite group presented a superiority of 0.4 to 1 genetic standard deviation whatever the breed and trait. In fact, the Montbéliarde breed presented a substantially lower difference in EBV between the elite and randomly selected groups. This could be explained by the objective of the main breeding organization to genotype a large proportion of candidates in the whole population. This induced the inclusion in the “elite” population of some individuals that would not strictly fit with a selection criterion mainly based on milk yield EBV.

Milk yield is obviously a more important trait in the breeding goal, which explains why the superiority of the elite group over the other one is larger for that trait. However, the genetic superiority of the elite population was also found for SCC. This means that if a different pattern is observed for the two traits when comparing the two female populations, this would not be explained only by the genetic superiority of the elite group.

The two traits not only differ in nature (a production trait vs a health trait) but also through their heritability which is in our case 0.3 for lactation milk yield and 0.15 for SCC. This difference in heritability has an impact on the amount of information in genetic evalution which comes from an own performance for a given cow. This performance will have a larger impact on the animal's GEBV for milk yield just because of this higher heritability.

A main feature of the genomic prediction model is that polygenic effect (based on pedigree) and haplotypes effects (based on markers information) are estimated jointly. This property is favorable to properly estimate both terms, compared with blending procedures for instance. However, both effects (polygenic effect and QTL effects) may be affected if phenotypic information used (performances) is biased.

Correlations between $\text{GEBV}_{(\text{DYD+YD})}$ and $\text{GEBV}_{(\text{DYD})}$ presented a different pattern depending on whether milk yield or SCC was considered. Indeed, except for the Montbéliarde breed, the correlations were very similar for both the elite and randomly selected groups for SCC, whereas a decrease in correlation was observed for the elite group for milk yield compared to the randomly selected group (difference of up to 0.04). This is a first evidence of the existence of a bias induced by the inclusion of the own performances of genotyped cows in genomic predictions.

The correlations for SCC were higher compared to milk trait. However, as already mentioned, this can be mainly explained by the difference in heritability. Indeed, the added information related to cows' performances (YD) is less informative for a low heritability trait yielding to reduced changes in GEBV.

The Holstein reference population used was much larger than for the two other breeds. This results in markers effects which are better estimated and genomic evaluations which are more stable in Holstein. This may be the reason why, whatever the group and trait considered, the correlations were higher for this breed. This also is one possible explanation why differences observed between groups for this breed were smaller.

Differences between $\text{GEBV}_{(\text{DYD+YD})}$ and $\text{GEBV}_{(\text{DYD})}$ were computed for the 2 traits for each individual of the 2 groups. Graphical representations of these differences (figures 1 to 3) clearly showed a different pattern for milk yield for the elite group compared to the randomly selected group or the SCC situations. A large fraction of elite cows presented a positive difference for milk yield, meaning that the inclusion of their own performances in genomic predictions led to an increase of GEBV. This phenomenon was not observed for somatic cell count or for the randomly selected group where differences were almost equally distributed between positive and negative values.

The average differences observed and expressed in genetic standard deviation units confirmed that the elite group for milk yield presented different characteritics than the randomly selected group or for SCC. Admittedly, the mean values were not strictly equal to 0 for the randomly selected group or for SCC (and it is difficult to explain why). Still, the difference was quite substantial for milk yield. The mean difference was up to 0.3 genetic standard deviation (in the Normande breed) higher for the elite group than for the randomly selected group.

The elite group also presented a genetic superiority on SCC but no real difference between $\mathrm{GEBV}_{(\mathrm{DYD+YD})}$ and $\mathrm{GEBV}_{(\mathrm{DYD})}$ between female subpopulations. This means that the systematic overestimation of GEBV observed when milk yield YD are included is induced by some overestimated performances of the elite group. This is a clear evidence of a bias which affected GEBV for milk yield but not for SCC. Preferential treatment is the most immediate explanation for the source of such a bias.

Bias (potentially due to preferential treatment) was shown for milk yield for the elite group. Our findings were obtained considering the group as a whole. However, this does not mean that every single individual present in this group presented over-estimated performances. In particular, a significant proportion of the elite cows had their GEBV decreased when their own YD was included.

Wiggans et al. (2011) also demonstrated the existence of a bias in genomic evaluations when using unadjusted data for cows of the reference population. Indeed, for milk yield in Holstein, the regression coefficient (of daughters proven EBV over GEBV) was decreased, the bias was equal to 50 kg, and the realized reliability was lower. The realized reliability was calculated as the squared correlation between GEBV and deregressed proofs for bulls of the validation population. The regression coefficient is a measure of how inflated GEBV are compared to EBV. Furthermore they also observed a bias in genomic predictions equations as marker effects of the X chromosome presented a singular pattern, suggesting that females systematically behaved differently than males.

In conventional genetic evaluations some solutions were found to limit the bias due to preferential treatment (account for heterogeneous variances for instance). Now that such a bias has been demonstrated in genomic evaluations, it is needed to find methods to correct for this bias. The first solution to get rid of bias due to preferential treatment in genomic evaluations is to discard own performances, i.e. YD, of cows. It is possible to estimate direct genomic values obtained using a reference population consisting only of bulls, or to use GEBV (obtained after blending for instance) where the polygenic component only includes performances (DYD) of male relatives. However such a solution is not completely satisfactory. First, AI industry may pressure to include cows own performances even if it does not increase reliabilities of genomic evaluations. Secondly and more importantly, this solution is frustrating because it implies that a large amount of potentially valuable information is not used. Furthermore, a limited number of heifers and cows have been genotyped so far, and most of them were elite individuals. However, with the release of an efficient low density SNP chip (Boichard et al., 2012b) to genotype females at a reduced cost, one can expect that many heifers from commercial herds will be genotyped in the near future, providing a large number of genotyped cows. Obviously, for most of these commercial animals, performances are likely to be unbiased and they will build up the reference population of the future.

Another solution is to adjust (i.e. pre-correct) cows performances before their inclusion in genomic evaluations. This is the option retained by Wiggans et al. (2011) who proposed to adjust the mean and variances of the estimated Mendelian sampling term of cows so that they are similar to those of bulls. They showed interesting improvement regarding several measures related to bias of genomic breeding values and prediction equations.

However, whether they are adjusted or not, the cows' performances are such that it is not really possible to distinguish a positive Mendelian sampling from a bias due to preferential treatment. The only situation where this could be envisioned is for cows with many recorded progeny but even then, it can be suspected that these progeny may have received a preferential treatment related to their maternal origin.

Single step procedures (Misztal et al. 2009) are appealing so that non-genotyped individuals benefit from markers information of their genotyped relatives. It has also some interesting properties in terms of bias due to pre-selection of young bulls. But solutions to remove bias induced by preferential treatment (such as blending, or adjustment or mendelian sampling term) would not be possible anymore.

CONCLUSION

We compared genomic predictions obtained after evaluations either including genotyped cows with own performances in the reference population or discarding them. Results showed that when such cows belonged to the group of elite cows, their GEBV for milk yield presented a different pattern than when these cows represent a random sample of the commercial population whereas they showed similar characteristics for somatic cell count. Correlations between GEBV computed with or without cows in the reference population were lower for the elite group when milk yield was considered. A systematic over-estimation of genomic predictions was found as well when the own milk yield performances of the elite population were included in the genomic predictions. The study demonstrated that explicitly including own performances of elite females induced biased (over-estimated) genomic evaluations.

ACKNOWLEDGEMENTS

The Agence Nationale de la Recherche (ANR) and APISGENE are gratefully acknowledged for funding the AMASGEN project. The LACTOSCAN part of the Phénofinlait project (also financed by ANR and APISGENE) contributed to the genotyping of a large part of the genotyping of the randomly selected cows.

COMPETING INTERESTS

The authors declare that they have no competing interests.

REFERENCES

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. 2012. Genomic Selection in French Dairy Cattle. Animal Production Science, 52, 115-120.

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 K., Wiggans G.R. 2012. Design of a Bovine Low-Density SNP Array Optimized for Imputation. PLoS ONE, 7: e34130.

Croiseau P., Legarra A., Guillaume F., Fritz S., Baur A., Boichard D., Ducrocq V. 2011. Fine tuning genomic evaluations in dairy cattle through SNP pre-selection with the Elastic-Net algorithm. Genetics Research, 93, 409-417.

Druet T., Fritz S., Boussaha M., Ben-Jemaa S., Guillaume F., Derbala D., Zelenika D., Lechner D., Charon C., Boichard D., Gut I.G., Eggen A., Gautier M. 2008. Fine Mapping of Quantitative Trait Loci Affecting Female Fertility in Dairy Cattle on BTA03 Using a Dense Single-Nucleotide Polymorphism Map. Genetics, 178: 2227–2235.

Goddard, M.E., and B.J. Hayes. 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nature Reviews Genetics 10, 381-391

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

Lund M.S., de Ross S.W.P., 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

Misztal I., A. Legarra, I. Aguilar. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. J Dairy Sci. 92:4648-4655.

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

Van Vleck, L. D. 1987. Contemporary Groups for Genetic Evaluations. J Dairy Sci 70(11):2456-2464.

Wiggans, G.R., T.A. Cooper, P.M. VanRaden, and J.B. Cole. 2011. Technical note: Adjustment of traditional cow evaluations to improve accuracy of genomic predictions. J. Dairy Sci., 94 : 6188–6193

3.3. Strategies applied by different countries regarding preferential treatment

Suspecting a potential bias of performances of genotyped cows, several countries set up different strategies in order to properly integrate this female information when computing genomic predictions.

3-3.1. North american consortium

Canada and USA decided in 2008 to share their reference populations and to exchange genotypes of both progeny-tested bulls and cows with performances (Van Raden et al., 2009).

Right from the beginning, Canadian scientists feared that a potential bias may have a negative impact on the accuracy of genomic predictions and decided to discard genotyped cows in their reference population: only progeny-tested bulls were included in the reference population. Only bulls deregressed proofs, calculated based on reliable their daughters' average performances were used to estimate markers effects.

Initially, in the USA, it was considered that the cows were a significant and valuable fraction of the genotyped individuals with performances. So it was decided to include genotyped cows in the reference population. In other words, deregressed proofs from both progeny tested bulls and cows with own performances were used to estimate markers effects. Later, Wiggans et al. (2011) found that the difference between breeding values (PTA, predicted transmitted ability, equals to EBV/2) and direct genomic values (DGV, sum of markers effects) were centered on 0 for genotyped progeny-tested bulls but were significantly higher for cows, which indicated that cows PTA were overestimated. They proposed a method to adjust females performances (see paragraph below) before their inclusion in genomic predictions. Genomic evaluations in U.S. have been corrected since then using this method.

3-3.2. Eurogenomics consortium

In order to get more reliable genomic predictions, several European countries (gathering 4 (groups of) breeding organizations in Denmark-Finland-Sweden, France, Germany and Netherlands-Flanders decided in 2009 to join their reference population and exchange genotypes of progeny-tested bulls. More recently, they were joined by Spain (2011) and Poland (2012).

In the Netherlands (Chris Schrooten, CRV, personal communication), DGV are currently estimated using male performances only. There are no females in the training population. In a second step, the GEBV is obtained by blending their DGV with the national EBV, which is either a parent average or a performance-based EBV depending on the age and origin of the animal. The equation below is used for blending:

$$GEBV = b_1 DGV + b_2 EBV$$

b_1 and b_2 are related to reliabilities of DGV and EBV respectively (either using the EDC method or a bivariate model).

So, for females, dam performances are included through the blending step. For a male, GEBV are obtained by blending their DGV with a pedigree index based on males, or with the performance-based EBV for progeny tested bulls. The pedigree index is a weighted sum of males ancestors' breeding values and equals ½ sire + ¼ maternal grandsire + …

In the Nordic countries (Guosheng Su, Aarhus University, personal communication), a multi-step approach is implemented. In this approach, as for the Netherlands, genotyped cows are not included in the training population, but their DGV are predicted. Then the DGV are used in the blending procedure using a bivariate model. The genomic evaluation center in Nordic countries is planning to apply a single step approach (see below) soon.

The treatment of cow performances in German genomic evaluations (Zengting Liu, VIT, personal communication) is almost identical to the Dutch and Nordic situations: Cows are not included in the current genomic reference population. Only bulls with daughters form the reference population. Calculation of the German pedigree index also uses phenotypic information on the male side only, i.e., only EBV of bulls are considered. Genotyped cows get a combined genomic EBV based on their DGV and their conventional EBV. Both for male or female candidates,their GEBV is based on DGV and (male) pedigree index.

Therefore, a common approach is applied in these three (groups of) European countries: DGV are first estimated using a reference population only composed of AI buls, and then combined through a blending procedure to a pedigree index in which only males ancestors -

sire, maternal grandsire, …- are included. This strategy appears somehow similar to the one applied in Canada.

In France, as the in the U.S. performances of genotyped cows were first included. Then, since the integration of the Eurogenomic consortium, such genotyped cows were discarded from the reference population. GEBV produced in France are obtained using a male reference population, and the polygenic effect is only based on males' DYD.

3.4. Evidences of a bias in genomic predictions when performances of genotyped cows are explicitly included

3-4.1. Evidences of bias in American genomic evaluations

We have seen (section 3-1.2.) that Wiggans et al. (2011) demonstrated that bias induced by inflated cow performances had a significant impact on genomic predictions equations. The impact of such a bias was also measured at the level of genomic predictions.

EBV from progeny-tested bulls were regressed on GEBV of the same bulls (obtained at a time when no daughter information was available). The regression coefficient is a way to measure how inflated GEBV are when compared to daughter-proven breeding values (PTA or EBV). The closer this value is to 1, the less biased GEBV are. The lower the regression coefficient is, the more inflated GEBV are compared to daughters proven EBV.

In Wiggans et al. (2011, Table 4) daughter proven EBV (from 2010) of 2,975 bulls were regressed over 2006 GEBV of the same bulls. A significant difference was observed depending on whether the reference population included progeny-tested bulls only or included both bulls and cows with unadjusted performances.

For instance, for milk yield in Holstein, the regression coefficient was 0.91 with the bulls reference population and 0.86 when cows with unadjusted data were added reference population (regression coefficient were 0.94 and 0.85 respectively for fat yield, for which the difference was even more impressive).

	Reference population	
	Bulls only	Bulls + cows with unadjusted PTA
Milk yield (kg)	0.91	0.86
Fat yield (kg)	0.94	0.85

Table 4 Regression coefficient of daughter proven EBV on GEBV for two reference population (included or not genotyped cows)

3-4.2. Two kinds of biases: selected subpopulation and preferential treatment

Two kinds of bias can be considered regarding performances of genotyped cows: the first bias is related to the fact that genotyped cows correspond to a specific subset of the whole population of dairy cows: they are selected. This bias was described by Mantysaari et al. (2010) when validating national genomic evaluations. When breeding companies genotype heifers and cows in order to select bull dams, the female candidates are often already preselected, usually based on their parent average or own performances. This means that the subset of cows that are genotyped does not properly represent the whole population of milking cows.

A different situation, where genotyped cows are not a selected subpopulation is possible. Not only breeding companies genotype females candidates, farmers also have the opportunity to genotype heifers and cows. They could do so in order to sort heifers and select among them for herd replacement (see section 4-3.6.). If many commercial breeders genotype whole groups of heifers, then, once these individuals have performances, they properly represent the whole population of milking cows as they were not pre-selected before genotyping. In the near future, if females genotyped by dairy famers represent most of the genotyped cows, the impact of the bias due to the selection of genotyped individuals will be reduced.

The other bias is related to preferential treatment of elite cows. It has been described in section 3.1. Preferential treatment is related to the first source of bias: indeed, it is likely to occur in the group of genotyped cows since selected (top) cows are more prone to preferential treatment.

3-4.3. Evidences of bias in French genomic evaluations

In the American experiment (Wiggans et al., 2011), the two sources of bias were mixed. Only one population of cows (mostly represented by top individuals) was considered and the authors only looked at production traits. There was no way to differentiate the two sources of bias, and the authors showed that the biases had a strong effect on genomic predictions.

In the experimental design of our study (Article IV), and considering the limited conclusions we were able to draw from our preliminary study (section 3-1.3.), we aimed at properly characterizing the different sources of bias related to performances of genotyped cows. Bias due to the fact that the genotyped population is selected would affect the elite group for both milk yield and SCC trait in our study. Bias due to preferential treatment is likely to only affect the elite group for milk yield.

In our study a singular pattern was only observed for the elite group and for milk yield when comparing two sets of genomic breeding values: one based on a progeny-tested bulls reference population and one where genotyped cows were also added to the reference population.

3.5. Possible solutions to deal with biases in the cow population

In order to correct for the bias of performances of genotyped cows, several options have been described. Two goals are targeted: not affecting predicition equations (markers effects) and not bias GEBV of genotyped cows (and their close relatives). Some of the solutions achieve one goal but not the other. They are listed below.

3-5.1. Discard genotyped cows from the reference population

This solution is the one currently applied in Canada and in different countries of the Eurogenomics consortium. It is probably the simplest one. Only progeny-tested bulls are included in the reference population and genomic prediction equations make no use of performances of genotyped cows. While doing so, one is sure to get rid of bias due to preferential treatment. Average daughters performances (DYD or deregressed proofs) of bulls are reliable and they are not affected by performances of some cows that could be preferentially treated. Another disadvantage is the political and sociological issue which is developed later on.

- **Direct Genomic Values**

When only bulls are included in the reference population, still, there are different possibilities to compute genomic predictions for young candidates and genotyped cows. The first one consists in publishing direct genomic values only. DGV are calculated using markers information only. No polygenic effect is included and the pedigree information is not used. This solution ensures that potentially biased EBV do not impact genomic predictions. However, DGV have one main disadvantage: they are less reliable than GEBV.

- **Blending DGV and EBV to obtain GEBV**

When discarding genotyped cows from the reference population, the second possibility is to publish GEBV obtained after blending. First, DGV are calculated based on a bulls-only reference population; then, these DGV are combined with EBV obtained through classical polygenic genetic evaluations. Again, two possibilities exist. First, DGV are blended with the traditional EBV, but the EBV are calculated based on female performances (of the dam of the candidate, or of the genotyped cow herself) that may be biased. Here the potential gain regarding the bias due to preferential treatment when genotyped cows are not part of the reference population may be lost when re-introducing dam performances. Another option consists in editing EBV to obtain pedigree index only based on male ancestors (sire, maternal grandsire, …). This solution aims at combining the higher reliability of GEBV (compared to DGV) obtained after blending, with no loss in realized accuracy due to potential biases of cows performances. This is the strategy applied in several Eurogenomics countries as described above. This solutions improves the DGV by adding a polygenic component but still does not account for own performances.

In France, the genomic model includes a polygenic effect with is estimated jointly with haplotypes effects. This approach makes blending with male polygenic information useless.

- **Concerns for new traits difficult to measure**

So far, we were considering traits which have been measured for years or decades, such as milk yield or somatic cell count (or type traits or fertility). Classical evaluations based on pedigree and performances require phenotypes for dozens (or hundreds) of daughters of dozens of bulls, which are performance recorded over several generations. Some economically important traits are very difficult (and/or very expensive) to measure. This is the case for example of feed intake, some specific diseases or more recently milk fatty acids composition and methane emission. Genotyping technology and its use to obtain genomic predictions are becoming very appealing to deal with such traits. Indeed, when phenotyping is expensive and/or difficult, it is possible to both genotype and precisely phenotype a “small” number (a few thousands) of individuals and those individuals are used as a reference population. This is done once for good, although it may be necessary to update prediction equations, but this is not necessarily required at every generation or every year.

When phenotypes are difficult to measure, the reference population would only be composed of genotyped cows, no bull could be added. However, these cows should not be affected by the bias we have been dealing with so far. Indeed, they would be either sampled randomly from the commercial population, or even coming from a very few experimental herds where environmental factors are controlled.

Veerkamp et al. (2012) calculated the accuracy of GEBV theoretically achieved when including cows in the reference population for a trait such as feed intake. Bulls from the reference population had performances only for a correlated prediction trait whereas cows had performances directly on feed intake. In this situation where the heritability of the trait was supposed to be high (about 0.50) and the correlation between direct and indirect traits was relatively low, adding cows to the reference population presented a major interest in terms of accuracy of GEBV for the trait considered.

Setting up a cow reference population is also necessary in specific cases. For a breed, such as the French Tarine dairy breed with a small population and only a few AI bulls per year, genotyping cows with performances is required. Indeed, even genotyping all the bulls progeny-tested for years would not be enough to achieve a certain accuracy of genomic predictions. Another situation where genotyping cows with performances is necessary corresponds to developing countries where no breeding system has been set up yet (no performance recording, no identification, …). Fewer bulls will obtain phenotypic values (end of progeny-testing) so that update of the reference population is compromised if it only relies on males. Genotyping cows with performances could be a solution in order to implement genomic selection.

- **Political issues**

Discarding genotyped cows may rise up some concerns that we will call abusively "political issues". First of all, the AI industry, which has been financing most genotyping, may refuse not to take into account genotyped cows in the reference population which implies that their performances are not valuable for genomic evaluations, as stated by Wiggans et al. 2011: "Cow evaluations have been included in US genomic evaluations since their inception. Early studies (P. M. VanRaden, unpublished data) did not show much gain from doing so, but the industry was interested in including cow evaluations [and hoped that a way could be found to increase their value]."

The solution consisting in removing all the genotyped cows from the reference population also may appear unfair. Indeed a few potential bull dams present overestimated performances. But for the very large majority of genotyped cows for which performances are not overestimated, why this additional unbiased information should be removed ? This can also sow confusion in people's mind regarding the bias affecting performances of all cows.

Performance recording organizations consider that discarding these performances increases the risk of dairy cattle breeders giving up performances recording (Ducrocq and Santus, 2011): when considering aspects related to breeding and selection decisions on farm (and not other herd management aspects), any farmer could wonder why he should bother with performances recording when GEBV allow to sort heifers and cows according to their genetic merit, especially if the additional information corresponding to performances is not even included in the cows' genomic breeding values. For performances recording organizations, discarding genotyped cows from the reference population is synonymous to sending a wrong signal to the breeders because it could imply that the performances obtained are not reliable (since they cannot be introduced in genomic evaluations) and of limited interest, which is clearly not true.

A regular update of prediction equations is required in order to maintain high level of reliability of genomic evaluations. Adding genotyped cows with performances to the reference population may be a way to update prediction equations.

A compromise can be found when a blending approach is used and when performances of females are used to compute conventional EBV. Even though such a compromise removes some political concerns, it may still not be the optimal solution. First, it re-introduces some of the bias induced by preferential treatment in GEBV. In other words, even though it does not impact anymore the prediction equations, comparison of blended GEBV within the group of genotyped females or with non-genotyped females is not entirely fair and may cause wrong selection decisions.

- **A disappointing solution**

Discarding genotyped cows from the reference population is historically the first option. Until recently, a limited number of heifers and cows have been genotyped, and most of them are elite individuals. Discarding genotyped cows from the reference population is a relatively well adapted conservative solution during the early adoption of genomic selection: they bring very little additional information compared to thousands of progeny-tested bulls with reliable performances. Moreover, this initial female population is heavily selected, mainly composed of potential bull dams prone to preferential treatment. Buch et al. (2012) performed a theoretical demonstration that a reference population including cows would bring higher accuracy of genomic predictions. Not including thousands of genotyped cows in the reference population clearly appears as a waste of fruitful and expensive information and is theoretically damaging in terms of accuracy of GEBV.

With the release of a cheaper and efficient low density chip (article III) to reduce genotyping cost, and the worldwide adoption of genomic selection, one can expect that many heifers from commercial herds will soon be genotyped, later providing a large number of genotyped cows with supposedly unbiased performances. Such information should not be ignored.

3-5.2. ANOTHER SOLUTION: TARGET SPECIFIC COWS FOR GENOTYPING

- **Random selection among the commercial population**

In Australia, with 2,749 genotyped bulls, the bull reference population size is rather limited - compared to the North American or Eurogenomics consortia (each with about 20,000 bulls or even more). Genotyping cows with performances appears natural in order to increase the reference population size and thus to improve the reliability of genomic breeding values.

However, we have seen that there is a clear risk of bias due to preferential treatment. Pryce et al. (2012) implemented a project (the “10,000 Holstein cows project”) of genotyping randomly selected cows (with “perfect” records for all interesting traits). When adding these 10,000 cows to the reference population, they were able to increase reliability by 4 to 8 % !

Such a gain in reliability of genomic predictions could be achieved because the initial reference population was rather small, but also because the genotyped cows were properly (randomly) selected. In contrast, no gain was observed when adding cows to an already large reference population in France or in the U.S.

- **Contracted herds**

A new economical model for performance recording may be required if dairy farmers move away from traditional performance recording or for non conventional traits. If fewer and fewer dairy farmers accept to pay for performance recording of their own herd, the national breeding system will need to find a way to maintain a certain level of accuracy, in order to update prediction equations or to validate GEBV. One solution envisioned for example by Ireland could be to contract some herds in order to get some (fine) phenotypes for both new and already-recorded traits. It is even possible to select and contract a few large herds specifically devoted to large experiments and performance recording. Environmental factors may be fully controlled in such herds allowing unbiased performances. Genotyped animals with such performances should of course be included in reference populations.

3-5.3. YET ANOTHER SOLUTION: ADJUST COWS PERFORMANCES

- **Adjust Mendelian sampling terms**

Wiggans et al. (2011) proposed a method to adjust cows performances included in genomic predictions. The strategy they actually implemented on US national evaluations consisted in calculating Mendelian sampling (equal to PTA-PA) mean and variance for the cow population and to adjust these 2 terms so that they became comparable to those for the bull population. As a matter of fact, they used deregressed PTA as well as deregressed Mendelian sampling terms (see Wiggans et al. 2011 for more details). Corrected Mendelian sampling terms were then added to parent average to recreate PTA. For instance for the Holstein breed and milk yield, the variance of cow MS was reduced and the coefficient for the adjustment affecting the variance was equal 0.84. The mean was also decreased substantially (by 355 kg).

A clear effect of such an adjustment was observed on the regression of PTA on genomic predictions. We have seen above that the corresponding regression coefficient is lower when cows with unadjusted performances are included in the reference population than when the training population only included bulls. This was no longer the case with adjusted PTA for cows (see Table 5).

	Reference population		
	Bulls only	Bulls + cows with unadjusted PTA	Bulls + cows with adjusted PTA
milk yield (kg)	0.91	0.86	0.90
fat yield (kg)	0.94	0.85	0.95

Table 5 Regression coefficient of daughter proven EBV on GEBV for three reference population (depending on the performances of cows included)

Another aspect observed by Wiggans et al. (2011) was a change in realized reliability (calculated as the squared correlation between GEBV and deregressed proofs for bulls of the validation population). When cow performances were not adjusted, the loss in reliability was 3%. The value was rather high as it was in the same range as the loss in reliability when imputed 3K genotypes rather than 50K genotypes are used.

One important characteristics of this approach was that Mendelian sampling terms (and not directly deregressed proofs) were adjusted. Under the BLUP assumptions, the average Mendelian sampling terms is supposed to equal 0. With selected and potentially preferentially treated elite genotyped cows, this assumption is violated. The adjustment aims at correcting this aspect.

When the Mendelian sampling mean is reduced for the group of genotyped cows, then, PTA of these cows become comparable with those of bulls. However, within the group of genotyped cows, comparison is still unfair, since contrasts between individuals are not affected. When the Mendelian sampling variance is reduced for the group of genotyped cows, then, less emphasis is given to performances while more weight is attributed to markers effects in GEBV. This allows to partially remove the unfairness when females within the group of genotyped cows compared to each other.

- **Adjust variances directly in conventional genetic evaluations**

The approach of Wiggans et al. (2011) is based on a manipulation of PTA after the national genetic evaluation process took place. An alternative option consists in ensuring that the PTA or EBV are somewhat forced to have a more proper distribution during genetic evaluation: as stated in part 3-1.1., bias due to preferential treatment in genetic evaluations is an “old issue”. Indeed, strategies exist to account for such a bias in conventional genetic evaluations.

Accounting for heterogeneity of variances is one of these strategies. It has been described in section 3-1.1. (method described in Robert et al. 1999). If heterogeneity of variances is taken into account in conventional genetic evaluations, then DYD, YD or deregressed proofs obtained afterwards can be corrected for heterogeneous variances and somehow partially corrected for bias induced by preferential treatment. As a result, they are more homogenous.

Accounting for heterogeneity of variances is somewhat similar to pre-correct breeding values before their inclusion as deregressed proofs in genomic evaluations (solution proposed by Wiggans et al., 2011).

3-5.4. A KEY ASPECT: IDENTIFY INDIVIDUALS SUBJECT TO PREFERENTIAL TREATMENT

When performances of genotyped cows are corrected before their inclusion in genomic prediction, females are somewhat still unfairly compared to each other. Indeed, one is unable to distinguish a high performance due to a “true” positive Mendelian sampling contribution from a high performance biased by preferential treatment.

One key issue is to identify groups of individuals that are (likely to be) subject to preferential treatment. Such a group could be heavily adjusted or even discarded from the reference population. We have seen for instance that randomly selected genotyped cows could be added to the reference population while females genotyped by breeding organizations would not. Finer rules (based on the number of bull dams in the herd, biased performances detected in the past, presence in show rings or auctions, ..) may offer an even better response to this issue. Indeed, some AI companies have some knowledge on which farms are more likely to apply preferential treatment and have the tendency to avoid them. The smaller the group of animals discarded, the better the compromise about our different concerns (waste of information, political issues on performance recording, new traits).

- **Single step approaches of genetic/genomic evaluations and preferential treatment**

Another shortcoming related to genomic selection is that only a restricted sample of the breed population is genotyped. As a consequence, two distinct evaluations are usually performed: one conventional genetic evaluation based on pedigree and performances, and one genomic evaluation which uses by-products (DYD or deregressed proofs) from the conventional evaluation as input together with genotype information. This situation results in biased EBV when genomic selection is not accounted for in national genetic evaluations (see Patry and Ducrocq, 2011), as well as in tedious computation. Moreover, non-genotyped individuals do not benefit from marker information of their genotyped relatives.

The single step approach (Misztal et al. 2009) is considered as one potential solution to this problem because both conventional and genomic evaluations are performed together and GEBV are computed for all the individuals of the breed.

However, regarding bias due to preferential treatment, the single step approach does not permit to discard or manipulate potentially biased performances of genotyped cows. The reference population necessarily corresponds to all the genotyped individuals with performances, blending approaches are no longer possible (only one evaluation is run), and adjustments (as described by Wiggans et al. 2011) no longer can be implemented since they were performed between conventional and genomic evaluations.

Nevertheless, a feature described by Legarra and Ducrocq (2012) may circumvent this issue. They proposed to solve iteratively two blocks of equations resulting from the single step mixed model equations. It is possible for instance to isolate genotyped individuals that form the reference population, or synonymously, discard genotyped animals so that they are not used to compute prediction equations. Again the key consisting in being able to identify group of individuals suspected of bias. Note that this approach has never been tested yet.

CHAPTER 4 – Discussion: Genotyping females and genetic gain

In this part, we aimed at assessing the potential benefits either a breeding company or a farmer would get when genotyping females. Theoretical aspects of how to measure genetic gain are first detailed. Afterwards, several simulations studies are reviewed.

4.1. Measures of genetic gain

4-1.1. The four pathways of genetic gain (Rendel and Robertson, 1950)

In order to predict genetic gain (ΔG), one needs to quantify what is transmitted from one generation to the next. Two factors are involved: the selection intensity i, i.e. the superiority of the selected candidates measured on the selection criterion (related to the proportion of individuals selected to breed the next generation) and the accuracy r, i.e. the correlation between the selection criterion and the true breeding value (which indicates how well one can determine which individuals are "the best" ones).

$$\Delta G_1 = i \, * r$$

This value is the genetic gain obtained after one generation. To obtain an annual value, one needs to divide it by the generation interval L (average time elapsed between 2 successive generations or age of the parents when their progeny get born).

$$\Delta G_{y2} = \frac{i \, * r}{L}$$

This value is expressed in genetic standard deviation units. One may prefer to express it in units that can be measured (kg of milk for instance). To do so, the previous ratio is multiplied by the genetic standard deviation σg of the trait of interest.

$$\Delta G_{y3} = \frac{i \, * r}{L} * \sigma_g$$

Note: The heritability of the trait does not explicitly appear in the equation. For conventional genetic evaluations, it is hidden in the accuracy term r. This may no longer be the case with genomic evaluations since DYD or deregressed proofs (used as inputs in genomic evaluations), and thus GEBV associated to low heritability traits present reliabilities similar to those of moderate to highly heritable traits.

In most species, males and females are evaluated with the same genetic model, but they are not selected in the same way. All parameters i, r, and L could be different. For example, progeny test results in a longer generation interval for males than for females, a higher accuracy and a specific selection intensity. Therefore one needs to distinguish between the 2 sexes in the equation above.

In dairy cattle, it is possible to go one step further: given the sexes of both parents and the offspring, the selection intensity drastically changes. For example, on the female side, there is very limited selection of dams to breed cows, since almost all heifer calves born are required for the herd replacement. In contrast, bull dams are heavily selected, since only a few bulls are required for artificial insemination allowing a drastic selection intensity of bull dams and sires of bulls.

Rendel and Robertson (1950) described the four pathways of selection that correspond to the four possibilities for genes to be transmitted to the next generation. These four pathways are represented in Figure 9.

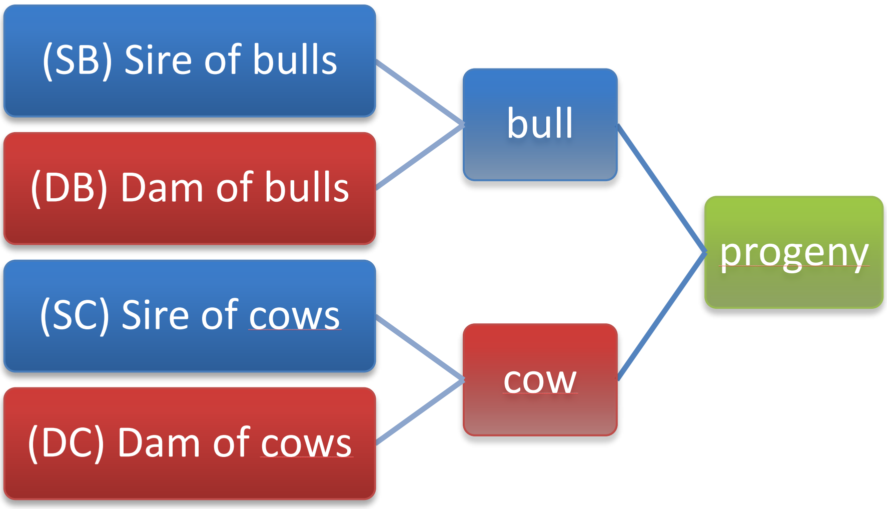


Figure 9 the 4 selection pathways as defined by Rendel and Roberston (1950)

In dairy cattle, the “sire of bulls” path (SB) is a path where a very intense selection takes place. The “dam of bulls” path (DB) or “bull dams” path is sometimes overlooked because it represents a very small fraction of the total population. However, it is essential and the selection of bull dams is conducted very carefully by breeding companies. The “sire of cows” path (SC) is the way to propagate genetic gain (mainly obtained in the male side) to the large commercial population of dairy cows. The propagation of the genetic merit is highly efficient in dairy cattle with the generalized use of artificial insemination. Finally, the “dam of cows” path (DC) corresponds to the replacement of the commercial population of dairy cows. The selection intensity applied on this path is quite low, because most of the heifer calves are needed in the herds to replace culled cows.

Rendel and Robertson (1950) also adapted the equation of yearly genetic gain to these 4 pathways and obtained the formula below:

$$\Delta G_y = \frac{i_{SB} * r_{SB} + i_{SC} * r_{SC} + i_{DB} * r_{DB} + i_{DC} * r_{DC}}{L_{SB} + L_{SC} + L_{DB} + L_{DC}}$$

ΔG_y is the annual genetic gain in genetic standard deviation units, i the selection intensity, r the accuracy and L the generation interval. SB, SC, DB and DC stand for the 4 different pathways: sire of bulls, sire of cows, dam of bulls and dam of cows respectively. This formula will be used from now on in this document.

4-1.2. Applying Rendel and Robertson’s formula to compare breeding schemes

Schaeffer (2006) described a traditional Canadian progeny-testing scheme which was typical for a breeding organization in dairy cattle in the early 2000’s. While generation intervals were more or less homogeneous across pathways, selection intensity and accuracy differ quite drastically from one pathway to another. For instance, selection was 7 times more intense and accuracy suppose to be doubled in the “sire of bulls” compared to the “dam of cows” pathways.

paths	selection %	i	accuracy	generation	i x r	ΔG_y
sire of bulls	5	2.06	0.99	6.50	2.04	
sire of cows	20	1.40	0.75	6.00	1.05	
dam of bulls	2	2.42	0.60	5.00	1.45	
dam of cows	85	0.27	0.50	4.25	0.14	
total				21.75	4.68	0.22

Table 6 the 4 pathways of selection for a progeny-testing scheme (from Schaeffer, 2006)

With such a breeding scheme, the annual genetic gain was 0.21 σ_g (genetic standard deviation).

- **Genomic selection**

Meuwissen et al. (2001) proposed several methodologies to be used once genome-wide SNP chips become available. Their simulation study showed an expected accuracy of up to 0.85. Today, one can realize that such values were somewhat optimistic because of strong underlying hypotheses (few QTL with strong effects and high LD between markers and QTL). However, this article was visionary and one of the first pillars of the setting up of genomic selection in dairy cattle.

In 2006, while SNP chip technology was still under development for livestock populations, Schaeffer carried out a study to assess potential benefits when applying genome-wide selection compared with his traditional progeny-testing scheme described above. He defined a strategy where young bulls (less than 2 year old) were heavily used for artificial insemination and culled before their progeny get performances. He used Rendel and Robertson (1950) formula to assess annual genetic gain. He also conducted a simplified study of the different additional costs and savings that such a new young bull scheme may induce. Figures were adapted to the Canadian Holstein population. Results are presented below

paths	selection %	i	accuracy	generation	i x r	ΔG_y
sire of bulls	5	2.06	**0.75**	**1.75**	**1.55**	
sire of cows	20	1.40	**0.75**	**1.75**	**1.05**	
dam of bulls	2	2.42	**0.75**	**2.00**	**1.82**	
dam of cows	85	0.27	0.50	4.25	0.14	
total				**9.75**	**4.55**	**0.47**

Table 7 the 4 pathways of selection for a genomic selection scheme (Schaeffer, 2006). Values that are changed compared to table X (progeny-testing scheme) are in bold.

When Table 7 is compared to Table 6, selection intensity was not modified for any of the 4 different pathways. Genomic evaluations were assumed to yield an accuracy of 0.75. This is much lower than the accuracy observed on the sire of bulls path in a progeny-testing scheme, in which such sires usually have second crop daughters. On the other hand, value of 0.99 accuracy achieved in the progeny-testing scheme appears very high (especially with such a generation interval). Schaeffer (2006) gave the same value for accuracy (0.75) for bulls evaluated only based on the genomic information than for progeny-tested bulls in the sire of cows path.

A substantial increase in accuracy was assumed on the dam of bulls path, because markers bring more information on the genetic merit than own performances. Dams of cows were not genotyped; this path was not modified between the two schemes.

The main change is observed on the generation interval. For the three most important pathways, sire of bulls, sire of cows, and dam of bulls, the generation interval is drastically reduced (2 to 3 times lower). This is the main reason of the strongly increased annual genetic gain when genome-wide selection is applied. The benefit of decreasing the generation interval at the sire of bulls level clearly overcomes the disadvantage of reducing the accuracy of selection. The annual genetic gain was 0.46 σ_g.

As a result, according to Schaeffer, a breeding scheme using young bulls selected based on genome-wide EBV may double the annual genetic gain and divide by 10 the related cost compared to progeny testing in Canada (the factor 10 should be taken lightly but a drastic cost reduction is possible) !

4-1.3. A STOCHASTIC SIMULATION TO ASSESS THE BENEFIT OF GENOMIC SELECTION OVER PROGENY-TESTING

Colleau et al. (2009) performed a simulation study based on figures from the main French Montbéliarde breeding organization. Three different scenarios were compared. The first one was an improvement of the traditional progeny testing scheme, where an extra step of pre-selection of young bulls based on GEBV was added before testing. The second one was a scheme where young bulls were selected and used as service sires and sires of sons (with no use of older sires). The third one was a sort of compromise, where young bulls were used for AI but some older bulls were kept and reused after their daughters obtained performances and new EBV were computed; 50% of the progeny were supposed to be born from each of the two categories. All these scenarios were simulated stochastically. In the reference scenario, 485 bull dams were selected based on EBV. In the “genomic” scenarios (AXMAX, AXMIX) 2,910 potential bull dams were genotyped and 1,455 were selected. The 2 scenarios using young bulls both allowed to almost double genetic gain, which is convincing and promising. The main difference was the evolution of inbreeding. While it slightly decreased when using only young bulls, it was almost doubled when reusing older bulls.

scenario	ΔG_y (σ_g)	ΔF_y	r_{1year}	r_{6year}
REF	0.25	0.13	0.69	0.94
AXMAX	0.46	0.10	0.61	0.99
AXMIX	0.47	0.22	0.67	0.99

Table 8 Average outputs for 3 breeding schemes (from Colleau et al., 2009). REF is a traditional progeny testing scheme with additional pre-selection step based on GEBV. AXMAX is a genomic selection scheme where only young bulls are used. AXMIX is a compromise between the 2: young bulls are used, but some older bulls are also reused after their progeny get performances. ΔG_y is the annual genetic gain (expressed in

genetic standard deviation). ΔF_y is the annual increase of inbreeding. r_{1year} and r_{6year} are the reliability of bulls obtained when they are 1 (respectively 6) year-old.

The deterministic and stochastic simulations lead to a similar conclusion: the possibility to almost double genetic gain when applying a genomic selection scheme. A few differences must be underlined: for the deterministic simulation, in the reference (progeny-testing) scheme, young candidates were not pre-selected with genomic evaluation. For the stochastic simulation, selection intensities changed between the reference and genomic schemes (the number of bull dams differed for instance).

In conclusion, maintaining the progeny-testing or using old sires for AI is useless when considering annual genetic gain.

4.2. Genotyping bull dams

4-2.1. A CRUCIAL PATHWAY FOR GENETIC IMPROVEMENT

When considering the benefits of applying genomic selection in a breeding scheme in dairy cattle, the use of young AI bulls evaluated only on their genetic markers information (i.e. with no recorded daughters yet) both as sire of bulls and sire of cows appears obvious and is often quoted as the main source of genetic gain.

Another source of genetic gain, sometimes overlooked but extremely important is the “dam to breed bulls” pathway (bull dams path). In order to demonstrate the importance of this pathway, a deterministic simulation was conducted using Scaheffer (2006) breeding schemes figures.

First of all, the bull dams pathway appears as very selective since only 2% are selected. The selection intensity and accuracy associated to the “dam of bulls” path in the initial genomic selection scheme (Table 7, Schaeffer 2006) were a bit optimistic. Indeed he considered that 2,000 females would be genotyped in order to select the best 1,000 (selection intensity would be lower than the value shown on the table). The initial 2,000 females were selected based on EBV on the whole population. This procedure is not strictly synonymous of one single step with an accuracy of 0.6 and 2% of females selected.

As stated in Table 6 and Table 7, the annual genetic gain observed in the progeny testing scheme and in the genomic selection scheme are 0.22 and 0.47 respectively. To assess the contribution of the bull dams pathway to the genetic gain in agenomic breeding scheme, we can change some parameters.

paths	selection %	i	accuracy	generation	i x r	ΔG_y
sire of bulls	5	2.06	0.75	1.75	1.55	
sire of cows	20	1.40	0.75	1.75	1.05	
dam of bulls	2	2.42	**0.60**	**5.00**	**1.45**	
dam of cows	85	0.27	0.50	4.25	0.14	
total				**12.75**	**4.18**	**0.33**

Table 9 the 4 pathways of selection for a genomic selection scheme without genotyping bull dams. Values that are different compared to Table 7 (genomic selection scheme) are in bold.

As we can see on Table 9, the annual genetic gain observed in a genomic selection scheme that would not select bull dams based on their markers information but on their own performances as in the traditional scheme would present an annual genetic gain of 0.33 (4.18/12.75). This means that only half the gain (0.33-0.21)/(0.46-0.21) achieved when a full genomic selection breeding scheme is applied is obtained when only males are selected based on their genetic markers information.

4-2.2. ISSUES RELATED TO THE USE OF YOUNG ANIMALS AS PARENTS

Schaeffer (2006) demonstrated that the main source of genetic gain when a genomic selection breeding scheme is applied was to drastically reduce the generation interval on several pathways. But it would have been possible for many years to apply a breeding scheme using young animals as parents even though the reliability of their breeding values (parent average) was lower. Applying Rendel and Robertson’s formula, the expected genetic gain could have been as large as the progeny-testing scheme but at cheaper costs (Table 10, an accuracy of 0.4 corresponds to a reliability of 0.16 for parent average, which is a "conservative" assumption).

paths	selection %	i	accuracy	generation	i x r	ΔG_y
sire of bulls	5	2.06	**0.40**	**1.75**	0.82	
sire of cows	20	1.40	**0.40**	**1.75**	0.56	
dam of bulls	2	2.42	**0.40**	**2.00**	0.97	
dam of cows	85	0.27	0.50	4.25	0.14	
total				**9.75**	**2.49**	**0.26**

Table 10 the 4 pathways of selection for a breeding scheme using young parents for parents without genotyping. Values that are changed compared to Table 6 (progeny-testing scheme) are in bold.

Such a scheme did not really emerge in dairy cattle. The reason is that both farmers and breeding companies are reluctant to use parents with low reliability. Genomic selection appears as a revolution in the field because an "acceptance threshold" in terms of reliability was achieved. Reliabilities of GEBV are somewhat lower than those of daughter proven breeding values for production traits, but the genetic superiority of young bulls overcomes the lower accuracy. This is why young bulls with genomic information only are now widely used.

A similar pattern can be observed for bull dams. A drastic selection currently occurs to determine which animals will be considered as bull dams. The owner of such a cow wants to be sure that the bull chosen to mate his cow is among the best of the breed. If a farmer just has one bull dam in his herd, he is often reluctant to take the risks to use a young bull with low reliability. Again, a sort of an "acceptance threshold" has been reached with genomic breeding values and many farmers who own top individuals now want to use young animals because of their genetic superiority.

4.3. Genotyping on farms: selecting cows to breed cows

4-3.1. Benefit at the national level and return on investment for the farmer

Pryce (2012) distinguishes two kinds of benefits of genotyping females in a farm: national benefit and the benefit farmers can get out of genotyping. At the national level, it appears very interesting that a lot of cows, properly representing the commercial population, are genotyped. Indeed, it is a way to drastically increase the reference population size, which leads to a greater accuracy of GEBV. For breeding companies, it is also interesting because genotyping a large proportion of the commercial population may highlight some very interesting heifers (or cows) that could be contracted as bull dams despite their EBV was below a threshold.

But genotyping a whole batch of heifers, even with a cheaper low density panel, is still costly. This leads to the question: “What is the return on investment from the farmer’s point of view?”

When we get back to the formula of genetic gain, and we consider Schaeffer’s figures especially for the selection intensity of the “dam of cows” path (85% of cows from one generation are dams of cows for the next generation), we can expect a rather low gain in terms of genetic improvement for the whole population. Indeed, given the low selection intensity, increasing the accuracy by genotyping cows will have a limited impact on the “dam of cows” path (compared to bull dams path). Chesnais (2011) pointed out that this path relatively contributes to only 3 to 4% of the total genetic improvement of the breed. This value is only slightly increased (5-6%) when genotyping is used.

4-3.2. Benefits of genotyping for the dairy farmer

We will review three recent studies trying to assess the benefits of genotyping by economic simulation. In Canada, Chesnais (2011) conducted a study to assess the expected net income of genotyping heifers to select the best ones for herd replacement. The net benefit corresponds to the higher economical value of the selected heifers after subtracting genotyping costs. These costs were assumed to be $47 (including chip and lab costs). The economic value of one standard deviation of LPI (Canadian total merit index) was set at $159. Chesnais compared selection based on genotypes to a situation without any selection for heifers. The net benefits were substantial (up to $7,000 for a herd of 100 milking cows), except for high replacement rates (40%) or few heifers available (related to high mortality before calving).

Pryce and Hayes (2012, Animal Production Science) also assessed the potential value of genomic technology to dairy farmers. They took the economic value of one standard deviation of APR (Australian total merit index) to be AU$80. Replacement rates ranged from 15% to 30%. Obviously, the net profit value depends on this replacement rate. Assuming genotyping costs of AU$50, the net profit of genotyping heifers to select the top 50% was AU$41. However this is in comparison to a situation without any selection. When genomic selection of replacement heifers is compared to selection based on parent average, the benefit becomes negative. They found that genotyping costs would need to be as low as AU$10 to be more profitable than selecting on parent average estimated breeding value.

Weigel et al. (2012) conducted a similar simulation study based on parameters from large American herds. They assumed genotyping costs of $40 (somewhat similar to the previous study). The value attributed to the genetic standard deviation of the total merit index (NM), which was $198 for PTA and even doubled ($396) to account for the fact that EBV equal twice the PTA ! A different genotyping strategy was applied: based on parent average, either the top, the middle, the bottom or the entire part of the group was genotyped. Whatever the replacement rate (from the top 20% to the top 80% of heifers are kept to replace culled cows), they found that the gains of selected heifer calves far exceeded prorated genotyping costs.

4-3.3. INTERESTS OF GENOTYPING COWS OTHER THAN FOR SELECTION

- **Inbreeding management**

Genomic information can be used to determine the amount of similar haplotypes two individuals share. Pryce demonstrated that using the genomic relationship matrix to control inbreeding is twice more effective than just using the pedigree relationship matrix. Indeed, the analysis is finer when thousands of markers are used to measure relationship coefficients. For instance, half sibs do not always show the expected relationship coefficient of ¼ and full sibs a coefficient of ½ when relationship is measured based on genomic information. Pryce attributed a negative value of AU$5 per annum for 1% increase of inbreeding.

- **Parentage assignment**

Genomic tools can be used to assign parents with 100% certainty as long as more than 300 SNP are genotyped (S. Fritz, UNCEIA, personal communication). In very large herds, with a lot of calves born over short periods, it is logistically difficult to identify sire and dam of a calf. In Australia, parentage assignment using microsatellites test costs AU$ 36 (this is the value used in the table below). In France and in most European countries, calving season is not as short and proper identification and parentage assignation is often mandatory: dairy farmers would probably not pay €28 to perform such a test.

	Net benefit genotyping	Net benefit pedigree
Selecting best replacements top 50%	€46.18	€76.94
Controlling inbreeding	€11.09	€5.54
Parentage	€28.11	
TOTAL	€85.38	€82.48

Table 11 Net benefits of genotyping compared to pedigree based on an Irish context (from Pryce, 2012) – genotyping costs are assumed to be €29.

To calculate benefits of selecting the best replacement heifers, Pryce (2012) assumed a standard deviation of EBI (the Irish total merit index) of €62. According to Pryce, when genotyping costs are €29, pedigree and genomic tools lead to more or less the same benefit (Table 11). And, if parentage assignment is removed from the benefit (as it would be the case for France), then genotyping present a negative return on investment. At a genotyping cost of €15, the use of genomic tool for the only purpose of selecting best replacement heifers becomes positive.

4-3.4. Discussion on the economic studies of the interest of genotyping heifers

The 3 studies measure the potential gain through the expected higher genetic merit (using reliability of GEBV) and applying the economic value of one standard deviation of total merit index (TMI). The different studies assumed similar genotyping costs, and several values for replacement rate were tested.

How to explain opposite results when considering the net benefits of genotyping heifers for the farmer?

The study of Chesnais was a bit different from the other ones because it compared selection based on genotypes to a no selection situation, which is tantamount to assume a reliability of parent average of 0 (whereas it equals 0.3 in Canada, compared to 0.6 for GEBV). Compared to the study of Pryce (2012), this means cancelling the net benefit of selecting best replacement top 50% heifers for pedigree (whereas it was €77, see Table 11, even higher than with genotyping once chip and lab costs have been deduced). Chesnais argued (personal communication) that, in practice, polygenic EBV are poorly used on farm for replacement purposes. Moreover, a herd usually consists of half-sibs or full-sibs families for which parent average present a very limited interest. The reliability of the parent average may thus not be the optimal measure of selection efficiency.

The main difference between the Australian and American studies lies on the value attributed to one standard deviation of TMI (they were AU$80 or 62€ on one hand, while it was $396 in the other one). The French economic value of TMI is between 80 to 100€. The American figure appears optimistic compared to the Australian or Irish values. Selection is more efficient when based on genotypes instead of parent average. Obviously, for a higher economic value of one standard deviation of total merit, the benefit of genotyping is increased.

To conclude, genotyping replacement heifers does not appear very interesting at the current genotyping costs (about $50) if the only purpose is the selection of the best replacement heifers. The conclusion drastically changes as soon as genotyping costs drop ($15). However, as we will see below, even at the current cost, positive interaction between genotyping and some reproduction practices may occur.

In other non-scientific studies (not shown) costs of rearing heifers were subtracted as if genotyping allowed to reduce replacement rate, whereas this is usually related to herd management practices. It is tempting, when conducting such studies, to account for benefits that are permitted by techniques different than genotyping.

4-3.5. One key aspect: the replacement rate

Replacement rate is defined by the proportion of milking cows that are culled each year. This ratio is strongly related to the proportion of available heifers required for replacement. Assuming a constant milk production and a constant number of milked cows, the higher the replacement rate, the more heifers required for replacement, and the less intense the selection occurring on this path, decreasing the expected benefits (Table 12). Increasing accuracy of breeding values through genotyping becomes valuable if an intense selection is possible. For example, with a replacement rate of 40 %, the return on investment is negative, almost all heifers are required to replace culled cows, and the gain achieved with a more accurate selection does not overcome the genotyping costs.

Replacement rate	Net income ($)
25	6870
30	4560
35	2260
40	-240

Table 12 (from Chesnais, 2011) Net income ($) when genotyping heifers depending on the replacement rate, in a herd of 100 milking cows.

4-3.6. Practical decisions in herd management favored by genotyping heifers

- **Corrective mating**

First of all, genotyping can be used to improve mating plans. Mating of heifers is usually confined in avoiding closely related bulls (to avoid inbreeding) because the farmer has no clue of the possible defects of the cow-to-be. Once breeding value is known relatively accurately for many traits, thanks to genotyping, some defects are revealed and they can be corrected through compensatory mating so that defects will not be transmitted to the progeny.

Several reproduction practices that are available in dairy cattle breeding are better used once the farmer knows more precisely the genetic level of both heifers to breed and cows. They are listed below. Some rely on the fact that genotyping can be first used to spot the top individuals:

- **Sell premium heifers**

Market of elite heifers is competitive and involves only a few players, but during auction sales, the animals with the highest GEBV are usually sold at top prices. Genotyping becomes a marketing tool in order to convince the buyer of the superiority of the individual avoiding bias induced by preferential treatment.

- **Embryo transfer**

Multiple Ovulation and Embryo Transfer consists in a hormonal treatment of the female so that several oocytes are produced during one cycle; after insemination, several embryos are expected to be produced. After a few days, these embryos are picked-up and re-implanted (sometimes after freezing) in “recipient” females. This technique allows the production of several progeny per year for one given female. This practice is heavily used by breeding companies but also by some farmers in order to multiply animals from one good family for instance.

It is obvious that the interest for such a technique consists in picking up genetically superior females so that their multiple progeny benefit from the transmitted superiority while “recipient” females present a lower genetic merit and do not genetically contribute to the next generation.

With genotyping, the farmers have access to a new tool in order to better sort heifers and cows depending on their genetic merit. The decision on which females to be picked-up and which to become “recipient” is easier thanks to the increased reliability of GEBV.

- **X sexed semen**

X-specific sexed semen is now widely available in dairy cattle. Because X spermatozoids carry 2% more DNA than Y-spermatozoids, they can be sorted, and the AI industry can provide sperm straws that guarantee 90 to 95% of female progeny. Y-specific sexed semen is also available and breeding companies can be interesting in such a tool to enhance the number of male progeny of bull dams. This technique has one main drawback: it induces a lower conception rate (about 10%). Usually, only heifers are inseminated with sexed semen because they have better fertility. Many dairy farms already have such low reproductive performances that using sexed semen is not costly effective.

Male progeny are much less valuable than heifer calves in dairy cattle. Farmers can be interested in using sexed semen in order to increase the proportion of female among the progeny. This technique is quite expensive, so it should be reserved to the best individuals. Genotyping can ensure that sexed semen is only used on the best heifers of the herd and allows to increase selection intensity.

Genotyping can also be used to spot the worse individuals.

- **No rearing**

Depending on the farming system, rearing of heifers to breed, and especially housing, can be very costly. If the replacement rate is such that all the heifers are not required to become cows, a first solution consists in not rearing a given proportion of young heifers. A Canadian study calculated the average variable cost at $1,860 plus $1,250 fixed costs per heifer.

Young heifer calves can be sold just as male young calves. The benefit does not really consist in the selling price, but in the rearing costs that are avoided.

Obviously, it is very important to properly select the individuals that will not be included the herd. It would be a pity that a top female is sold for a low price because the farmer did not know yet her high breeding value. Genotyping provides higher reliability breeding values that help to determine which animals have lower genetic merit.

- **Sell for export**

For some specific breeds, like the French Montbéliarde breed, heifers are very valuable. They can be sold in Eastern Europe for a good price. The decision process is the same as the "no rearing option" but with a much better profitability. Genotyping provides the opportunity to keep the best heifers on the farm, while others are sold abroad.

- **Crossbreeding with beef bulls**

After genotyping, the farmer gets an indication that the breeding value of one cow or one heifer is so low, that even her progeny may not be interesting for herd replacement. It is then possible to mate such a cow with a beef bull. Some beef breed such as INRA 95 have been specially selected for crossbreeding and present some interesting characteristics in terms of calving ease for example. The crossbred product (whether a male or a female) will have a valuable beef conformation. Even at an early age, the difference between a purebred dairy calf and a crossbred calf in terms of price is impressive. In France, in june 2012 (Tendances n°226, Institut de l'Elevage) the price of a 8-days dairy calf is 150 € while it is 313 € for a crossbred calf. The cow can still be milked and kept in the herd, but will not transmit her genes to the next generation while generating an interesting extra income through her crossbred progeny.

- **Special care of potentially diseased animals**

The information given by genomic tests is not confined to the TMI. GEBV are given for any trait of interest. Precise breeding values for disease-related traits such as clinical mastitis are available. These breeding values on functional traits have reliability no one could have expected for females, even with performances. Cows with very low GEBV for such traits should be specifically looked at, and preventive treatment could be used. In the near future, one can expect that traits related to other diseases become genomically evaluated so that herd management related to health trait may be based on genotypes information.

- **Combining these practices**

Many of these techniques present a positive interaction, meaning that the benefit of their combination is greater than the sum of their individual benefits. For example, sexed semen brings more potential heifers to breed so that selection can be more intense and more heifers can be sold or mated to beef bulls. The future cows have a greater genetic level while the other progeny can be sold which results in an additional income.

- **Genotyping is not necessary but helps a lot**

Obviously, all of the practices presented above could be set up even without genotyping. Indeed, it is already possible to use sexed-semen or to mate some dairy females with beef bulls. However, these techniques, even if economically interesting, are poorly used so far for several reasons including psychological aspects. Indeed, a farmer would wonder why he should take the risk not to breed some of his heifers or to mate some cows with a beef bull whereas they may be top individuals of the herd. Again, genotyping is associated with a certain reliability of breeding values which reaches an "acceptance threshold".

4-3.7. From a vicious circle to a virtuous circle on functional traits

During the last decades, genetic level for milk yield heavily increased. However, this improvement was achieved at the expense of some health traits. Indeed, reproduction traits (fertility) are negatively correlated (correlation of about -0.3) to milk yield. Moreover, functional traits present low heritability (0.02 for conception rate for instance). In conventional BLUP evaluations, even when assigned a weight in TMI equal to production traits, such low heritability traits would present lower genetic gain than production traits. So far, even with an increasing relative weight of health traits in TMI, their degradation was stopped (at best) but no improvement was observed. In dairy cattle, reproductive performances reached a very low level (as low as 30% for conception rate, meaning that 2 inseminations out of 3 failed). Such a low level does not allow any flexibility in terms of replacement rate and selection of heifers (Figure 10), which as in a vicious circle, induces even lower reproductive performances.

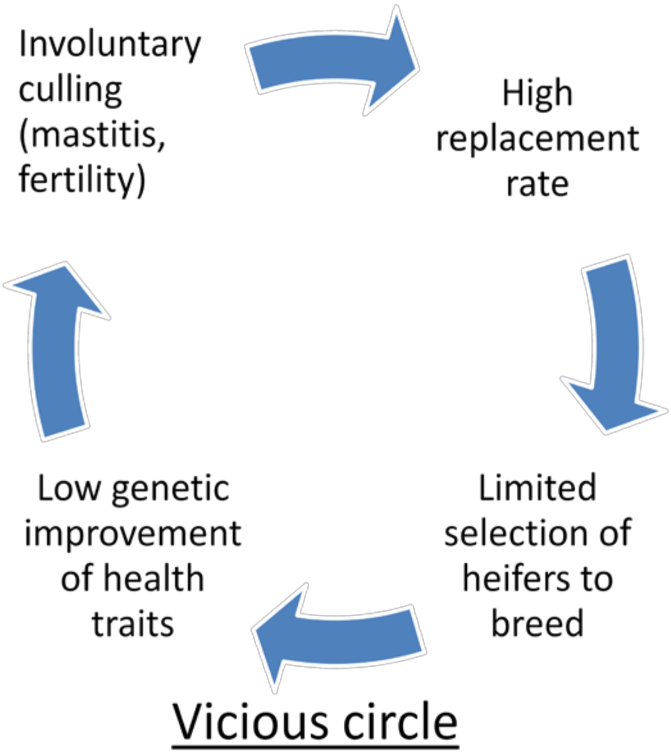

Involuntary culling (mastitis, fertility)
High replacement rate
Limited selection of heifers to breed
Low genetic improvement of health traits
Vicious circle

Figure 10 diagram representing the negative trend observed on functional traits.

Genomic selection combined with some new reproductive technologies brings some new opportunities to challenge this issue. Indeed, GEBV present similar reliabilities across traits. Selecting young AI bulls with high GEBV for health trait becomes an effective way of improving reproductive performances. Use of sexed semen currently leads to low conception rate, however 90% of the progeny is female and many more heifers to breed are available for selection. As seen on the above section, low replacement rate and high selection intensity of heifers are associated with increased benefits when heifers to breed are genotyped. A sort of virtuous circle could be set up, where a lower replacement rate allows to select heifers to breed on health traits, so that they will present a better longevity as milking cows. Additionally, mating them with young bulls selected on the same traits duplicates the effect. It is necessary to “prime the pump”, and the combination of use of sexed semen, genomically proven young bulls and the selection of heifers based on genotypes could help.

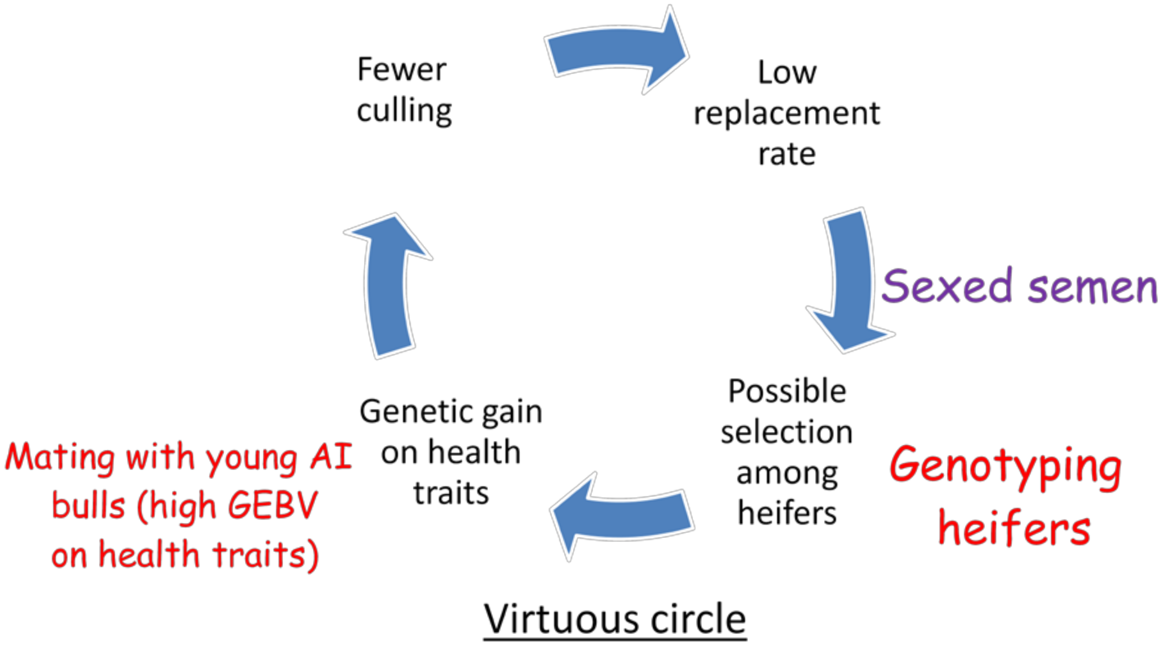

Fewer culling
Low replacement rate
Sexed semen
Possible selection among heifers
Genotyping heifers
Genetic gain on health traits
Mating with young AI bulls (high GEBV on health traits)
Virtuous circle

Figure 11 diagram representing the potential positive trend that could observed on functional traits when using genomic selection (young bulls, heifers) and reproductive technologies (sexed semen).

General conclusion

Genomic selection drastically changed the method of selection and breeding in dairy cattle where a large emphasis is put on the male pathways. This thesis addressed several questions related to the female side. First, which genotyping tool is adapted to females? Second, do performances of genotyped cows fit within the current prediction model? Last, what are the benefits of genotyping females, both for the farmer and at the breeding company level?

Imputation is the prediction of unmeasured genotypes. This technique is used to take full benefit of low density panels. Several imputation software and various possible measures of imputation accuracy were compared. The first study assessed the value of using the commercially available GoldenGate 3K SNP chip (from Illumina) in terms of imputation accuracy but it also looked at its impact on the reliability of GEBV. This study was conducted in collaboration with Aarhus University, within the Eurogenomics consortium. The allelic imputation error rate and the loss in reliability of GEBV when using imputed genotypes instead of real genotypes were moderate. While the 3K panel is interesting, some shortcomings were evidenced such as a low number of markers kept after quality control. In a second study, alternative *in silico* low density chips were described. Their imputation accuracy was compared with the initial commercial 3K product for three French dairy and beef breeds, and their concordance rate was 1 to 2.5% higher. The imputation accuracy not only depends on the number of markers, but also on MAF and spacing. A novel , fast and accurate imputation strategy based on existing software was described, which benefits from linkage disequilibrium and family information. Then, the construction of a new low density panel, adapted to many breeds and specifically dedicated to imputation, was detailed. The product is now commercialized by the Illumina company. This tool is well adapted for the genotyping of females in dairy cattle, but is also suitable for beef breeds.

A second main aspect of this thesis was to study how individual performances of genotyped cows fit within the current genomic prediction model. Two potential sources of bias exist: the genotyped cows are a selected sample of the entire population and moreover they are more likely to be affected by preferential treatment. An experimental design was set up to assess the effect of potential biases on genomic predictions based on two (elite and randomly selected) groups. Two evaluations were performed; the reference population consisted in either only progeny-tested bulls or both bulls and cows. The average difference between breeding values was significantly different from 0 for the elite group for milk yield (trait most prone to preferential treatment). Such a difference was not observed for somatic cell count (for the elite group) or for randomly selected cows (for any trait). The study demonstrated that explicitly including own milk yield performances of elite females induced biased genomic evaluations. Such a bias has two major consequences: it may affect genomic predictions equations, and it may induce overestimated breeding values for the cow and her close relatives. Different alternative solutions to properly include such performances in genomic predictions exist. One consists in discarding genotyped cows from the reference population. However this solution is disappointing because it means wasting a large amount of fruitful information. Another solution consists in adjusting cow performances as inputs of genomic evaluations. Unfortunately, comparisons between females are still partly unfair.

Finally, the benefits of genotyping heifers either by breeding companies to select bull dams or by farmers for herd management were discussed. A review of several simulations studies carried out on this topic was conducted. Selecting bulls dams based on their genotypes appears to be crucial within a breeding scheme applying genomic selection to take full advantage of the reduction of the generation interval. Indeed, it is as important as using young bulls for artificial insemination. Using genotyping tools to select heifers to replace culled cows is more controversial: different studies presented opposite results. While the benefits at the national level (increase of the reference population) are obvious, the return on investment for the famers depends on the cost of genotyping, the replacement rate as well as the economic value of the expected genetic improvement. Several herd management decisions could be facilitated when using genomic breeding values: which animals to sell, rearing only the required number of heifers, crossbreeding females with poor GEBV with beef bulls. A positive interaction exists between genomic selection within herd and several reproduction practices such as embryo transfer or use of sexed semen. Their combination may help in solving the issue that dairy cattle faces today related to the decrease of performances for health traits such as fertility.

REFERENCES

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. 2012. Genomic Selection in French Dairy Cattle. Animal Production Science, 52, 115-120.

Browning, S. R., and B. L. Browning. 2007. Rapid and accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. The American Journal of Human Genetics 81:1084-1097.

Buch, L.H., M. Kargo, P. Berg, J. Lassen, A.C. Sørensen. 2012. The value of cows in reference populations for genomic selection of new functional traits. Animal. 6(6) : 880-6.

Chesnais, J. 2011. Utiliser la génomique pour maximiser les profits des éleveurs laitiers. Proceedings of “symposium sur les bovins laitiers”, Drumondville

Colleau, J. J. 1989. Impact of the use of bovine somatotropin (BST) on dairy cattle selection. Genet. Sel. Evol. 21(4):479-491.

Colleau, J.J., S. Fritz, F. Guillaume, A. Baur, D. Dupassieux, M.Y. Boscher, L. Journaux, A. Eggen, and D. Boichard. 2009. Simulation des potentialités de la sélection génomique chez les bovins laitiers. Proceeding. Renc. Rech. Ruminants.

Daetwyler, H. D., G. R. Wiggans, B. J. Hayes, J. A. Woolliams, and M. E. Goddard. 2011. Imputation of missing genotypes from sparse to high density using long-range phasing. Genetics 189:317-327.

Druet, T., S. Fritz, M. Boussaha, S. Ben-Jemaa, F. Guillaume, D. Derbala, D. Zelenika, D. Lechner, C. Charon, D. Boichard, I. G. Gut, A. Eggen, and M. Gautier. 2008. Fine mapping of quantitative trait loci affecting female fertility in dairy cattle on BTA03 using a dense single-nucleotide polymorphism map. Genetics 178(4):2227-2235.

Druet, T., and M. Georges. 2010. A Hidden Markov Model Combining Linkage and Linkage Disequilibrium Information for Haplotype Reconstruction and Quantitative Trait Locus Fine Mapping. Genetics 184:789-798.

Druet, T., C. Schrooten, and A. P. W. de Roos. 2010. Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. J. Dairy Sci. 93(11):5443-5454.

Ducrocq, V. and E. Santus. 2011. Moving away from progeny testing schemes: consequences on conventional (inter)national evaluations. in Interbull technical workshop. Guelph, Ontario, Canada.

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

Grobet, L., L. Royo, D. Poncelet, D. Pirottin, B. Brouwers, J. Riquet, A. Schoeberlein, S. Dunner, F. Ménissier, J. Massabanda, R. Fries, R. Hanset, M.Georges. 1997. A deletion in the myostatin gene causes double-muscling in cattle, Nat. Genet. 17 7174

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

Hickey, J.M., J. Crossa, R. Babu, and G. de los Campos. 2012. Factors Affecting the Accuracy of Genotype Imputation in Populations from Several Maize Breeding Programs. Crop Science. doi: 10.2135/cropsci2011.07.0358; Published online 8 Dec. 2011.

Johnston, J., G. Kistemaker, and P.G. Sullivan. 2011. Comparison of different imputation methods. Preliminary proceedings of 2011 Interbull meeting, August 27–29, Stavanger, Norway, 7 pages.

Kong, A., G. Masson, M.L. Frigge, A. Gylfason, P. Zusmanovich, G. Thorleifsson, P.I. Olason, A. Ingason, S. Steinberg, T. Rafnar, P. Sulem, M. Mouy, F. Jonsson, U. Thorsteinsdottir, D.F. Gudbjartsson, H. Stefansson, and K. Stefansson. 2008. Detection of sharing by descent, long-range phasing and haplotype imputation. Nat Genet, 40:1068-1075

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

Lande, R., and R. Thompson. 1990. Efficiency of Marker-Assisted Selection in the Improvement of Quantitative Traits. Genetics 124(3):743-756.

Legarra, A., and V. Ducrocq. 2012. Computational strategies for national integration of phenotypic, genomic, and pedigree data in a single-step best linear unbiased prediction J.Dairy Sci., 95, 4629-4645

Mäntysaari, E., Z. Liu, and P. VanRaden. 2010. Interbull validation test for genomic evaluations. Interbull bulletin 41:17-22.

Marchini, J., B. Howie, S. Myers, G. McVean and P. Donnelly (2007) A new multipoint method for genome-wide association studies via imputation of genotypes. Nature Genetics 39: 906-913

Marchini, J., and B. Howie. 2010 Genotype imputation for genome-wide association studies. Nature Reviews Genetics 11, 499-511

Matukumalli, L. K., C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan, M. P. Heaton, J. O'Connell, S. S. Moore, T. P. L. Smith, T. S. Sonstegard, and C. P. Van Tassell. 2009. Development and Characterization of a High Density SNP Genotyping Assay for Cattle. Plos One 4(4):e5350

Meuwissen, T.H.E. and Goddard, M.E. 2001. Prediction of identity by descent probabilities from marker-haplotypes. Genetics Selection Evolution 33: 605-634.

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.

Misztal I., A. Legarra, I. Aguilar. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. J Dairy Sci. 92:4648-4655

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

Patry, C. and V. Ducrocq. 2011b. Evidence of biases in genetic evaluations due to genomic preselection in dairy cattle. J Dairy Sci 94(2):1011-1020.

Powell, R. L. and H. D. Norman. 2006. Major Advances in Genetic Evaluation Techniques. J Dairy Sci 89:1337-1348.

Pryce, J.E., and B.J. Hayes. 2012 A review of how dairy farmers can use and profit from genomic technologies. Animal Production Science 52(3) 180-184

Pryce, J.E., B.J. Hayes, and M.E. Goddard. 2012. Genotyping dairy females can improve the reliability of genomic selection for young bulls and heifers and provide farmers with new management tools. Proceeding ICAR congress, Cork.

Rendel, J.M., and A. Robertson. 1950. Estimation of genetic gain in milk yield by selection in a closed herd of dairy cattle. J. Genet., 50, p. 1.10

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

Schaeffer, L.R. (2006) Strategy for applying genome-wide selection in dairy cattle. J Anim Breed Genet 123: 218–223.

Scheet, P., and M. Stephens. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78: 629–644.

Shrimpton, A. E., and A. Robertson. 1988. The Isolation of Polygenic Factors Controlling Bristle Score in Drosophila melanogaster. II. Distribution of Third Chromosome Bristle Effects Within Chromosome Sections. Genetics 118: 445-459.

Stephens, M., Smith N.J., and Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet. 68:978-89.

Van der Werf, J.H.J., T.H.E. Meuwissen, and G. De Jong. 1994. Effects of Correction for Heterogeneity of Variance on Bias and Accuracy of Breeding Value Estimation for Dutch Dairy Cattle. J. Dairy Sci. 77: 3174–3184

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

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

VanRaden P. M., G. R. Wiggans, C. P. Van Tassell, T. S. Sonstegard, and F. Schenkel. 2009. Benefits from cooperation in genomics. Interbull Bull. 39. Interbull Centre, Uppsala, Sweden.

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

Van Vleck, L. D. 1987. Contemporary Groups for Genetic Evaluations. J Dairy Sci 70(11):2456-2464.

Veerkamp R. 2012. Selection for feed intake in dairy cattle using genomic selection. Proceeding ICAR congress, Cork.

Vinson, W. E. 1987. Potential Bias in Genetic Evaluations from Differences in Variation Within Herds. J Dairy Sci 70(11):2450-2455.

Weigel, K.A., C.P. Van Tassell, J.R. O’Connell, P.M. VanRaden, and G.R. Wiggans. 2009. Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels and population-based imputation algorithms. J. Dairy Sci., 93 : 2229–2238

Weigel, K. A., G. de los Campos, A. I. Vazquez, G. J. M. Rosa, D. Gianola, and C. P. Van Tassell. 2010. Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle. J. Dairy Sci. 93(11):5423-5435.

Weigel, K.A., P.C. Hoffman, W. Herring, andT.J. Lawlor. 2012. Potential gains in lifetime net merit from genomic testing of cows, heifers, and calves on commercial dairy farms. J. Dairy Sci. 95: 2215-2225

Wiggans, G.R., T.A. Cooper, P.M. VanRaden, and J.B. Cole. 2011. Technical note: Adjustment of traditional cow evaluations to improve accuracy of genomic predictions. J. Dairy Sci., 94 : 6188–6193

Zhang, Z., and T. Druet. 2010. Marker imputation with low-density marker panels in Dutch Holstein cattle. J. Dairy Sci. 93(11):5487-5494.

Table of illustrations

Tables

Figures