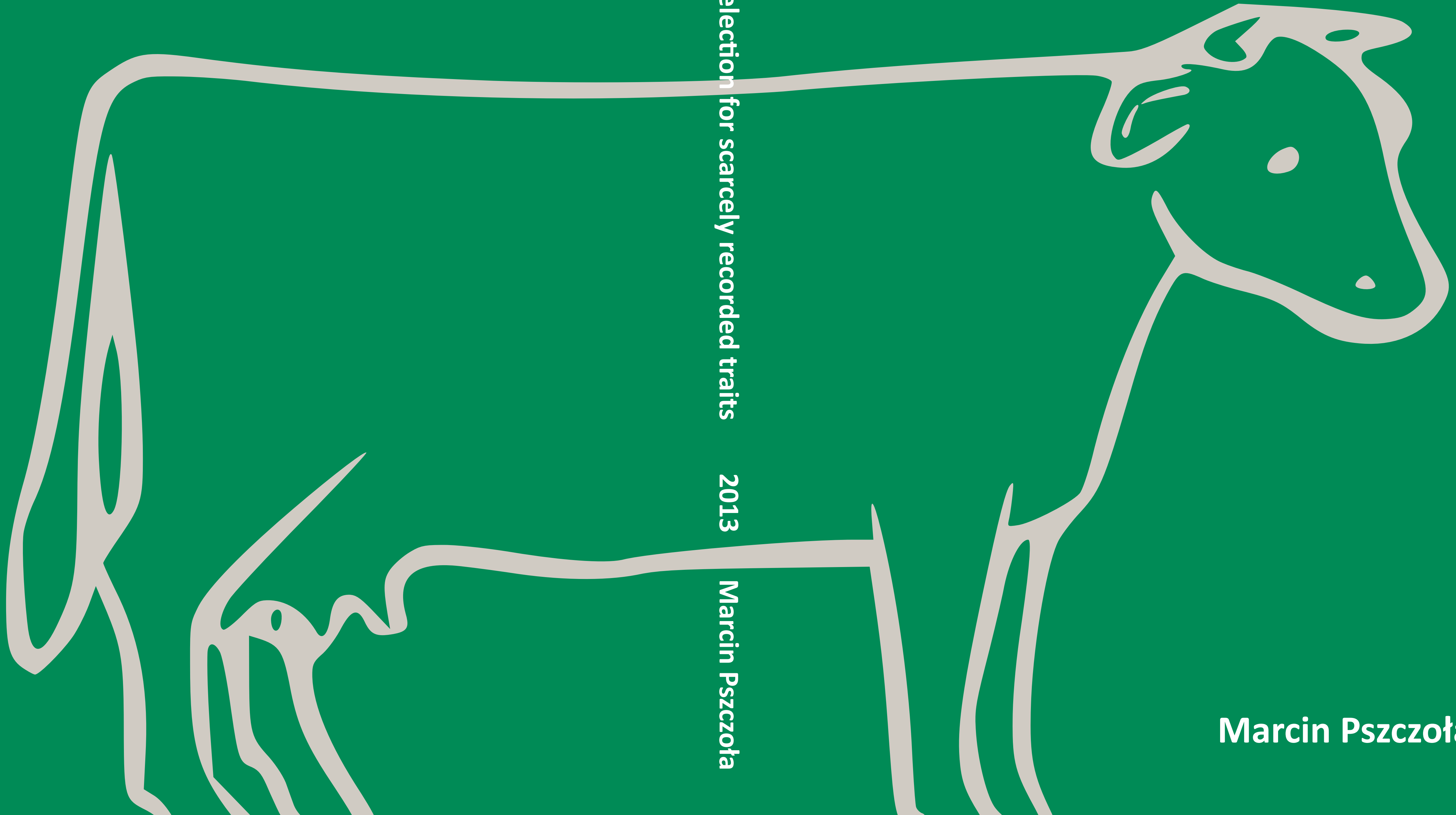
Optimizing genomic selection for scarcely recorded traits



Marcin Pszczoła

Optimizing genomic selection for scarcely recorded traits 2013 Marcin Pszczoła

Optimizing genomic selection for scarcely recorded traits

Thesis committee

Promotor
Prof. Dr J.A.M van Arendonk
Professor of Animal Breeding and Genetics
Wageningen University

Co-promotors
Dr M.P.L. Calus
Senior researcher at Animal Breeding and Genomics Centre
Wageningen UR Livestock Research

Dr T. Strabel
Professor at Department of Genetics and Animal Breeding
Poznan University of Life Sciences

Other members
Prof. Dr G. Banos, University of Edinburgh, UK
Dr D. Boichard, INRA, France
Prof. Dr J. Jensen, Aarhus University, Denmark
Prof. Dr B. J. Zwaan, Wageningen University

This research was conducted under the auspices of the Graduate School of Wageningen Institute of Animal Sciences (WIAS).

Optimizing genomic selection for scarcely recorded traits

Marcin Pszczola

Thesis
submitted in fulfillment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. Dr M.J. Kropff,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Friday November 22, 2013
at 11 a.m. in the Aula.

Pszczola, M.
Optimizing genomic selection for scarcely recorded traits,
160 pages.

PhD thesis, Wageningen University, NL (2013)
With references, with summaries in English and Dutch

ISBN 978-94-6173-766-3

Abstract

Pszczola, M. (2013). Optimizing genomic selection for scarcely recorded traits.
PhD thesis, Wageningen University, the Netherlands

Animal breeding aims to genetically improve animal populations by selecting the best individuals as parents of the next generation. New traits are being introduced to breeding goals to satisfy new demands faced by livestock production. Selecting for novel traits is especially challenging when recording is laborious and expensive and large scale recording is not possible. Genetic improvement of novel traits may be thus limited due to the small number of observations. New breeding tools, such as genomic selection, are therefore needed to enable the genetic improvement of novel traits. Using the limited available data optimally may, however, require alternative approaches and methodologies than currently used for conventional breeding goal traits. The overall objective of this thesis was to investigate different options for optimizing genomic selection for scarcely recorded novel traits. The investigated options were: (1) genotype imputation for ungenotyped but phenotyped animals to be used to enlarge the reference population; (2) optimization of the design of the reference population with respect to the relationships among the animals included in it; (3) prioritizing genotyping of the reference population or the selection candidates; and (4) using easily recordable predictor traits to improve the accuracy of breeding values for scarcely recorded traits. Results showed that: (1) including ungenotyped animals to the reference population can lead to a limited increase in the breeding values accuracy; (2) the reference population is designed optimally when the relationships within it are minimized and between the reference population and potential selection candidates relationships are maximized; (3) the main gain in accuracy when moving from traditional to genomic selection is due to genotyping the selection candidates, but preferably both reference population and selection candidates should be genotyped; and (4) including the predictor traits in the analysis when it is recorded on both reference population and selection candidates can lead to a significant increase in the selection accuracy. The key factors of successful implementation of a novel trait in a breeding scheme are: (1) maximizing accuracy of genotype prediction for ungenotyped animals to be used for updating the reference population; (2) optimizing the design of the reference population; (3) determining easy to record indicator traits that are also available on the selection candidates; (4) developing large scale phenotyping techniques; and (5) establishing strategies and policies for increasing the engagement of farmers in the recording of novel traits.

To my family

Contents

1

General introduction

1.1 Animal breeding

Animal breeding aims to genetically improve animal populations by selecting the best individuals as parents of the next generation. Best individuals are chosen by ranking the animals according to the breeding goal. The breeding goal is a set of criteria expected to be important in the next 5-10 years and therefore changes over time. Decades ago, the main breeding goal was production (Miglior *et al.*, 2005; Neeteson-van Nieuwenhoven *et al.*, 2013). Later, however, breeding goals were revised and new traits included. Presently, in some countries dairy cattle breeding goals include up to 40 commonly recorded traits (Banos, 2010). As a consequence of the increased number of traits, the relative importance of production traits in the breeding goal has dropped (Miglior *et al.*, 2005) and is predicted to reduce further in the future (Neeteson-van Nieuwenhoven *et al.*, 2013). New traits are being introduced to breeding goals to satisfy new demands faced by livestock production (Boichard and Brochard, 2012; Merks *et al.*, 2012). However, introducing a new trait and starting selection for such a trait may be difficult when the trait is novel in the sense that it has not previously been widely recorded. Selecting for such novel traits is especially challenging when recording phenotypes is laborious and expensive.

1.2 Novel traits

The potential for the genetic improvement of several novel traits is currently being investigated on the grounds that they are economically, environmentally or societally important. Because of laborious and expensive recording, for some novel traits, large scale recording in the near future is not possible and recording may therefore be limited to research herds. Two examples of novel traits that most likely will not be measured on a large scale in the near future are dry matter intake and methane emission in dairy cattle. These two traits are economically and societally important for the dairy industry, because they are both related to production efficiency and environmental footprint (Veerkamp, 1998; de Haas *et al.*, 2012a). Genetic improvement of such traits is desirable, but may be limited due to the small number of observations available. As a result, novel traits often cannot be improved directly by conventional breeding tools, as these require large numbers of observations, measured on many close relatives of each selection candidate. New breeding tools are therefore needed to enable the genetic improvement of novel traits.

1.3 Genomic selection for novel traits

Genomic selection (GS) is a new breeding tool that uses single-nucleotide polymorphism (SNP) markers spread across the genome that are expected to be in linkage disequilibrium (LD) with quantitative trait loci (QTL) influencing a trait of interest (Meuwissen *et al.*, 2001; Hayes *et al.*, 2009a). SNP markers also allow for describing relationships between the animals on a genomic rather than pedigree level (Nejati-Javaremi *et al.*, 1997; VanRaden, 2008, Yang *et al.*, 2010), which results in more accurate estimates of relationships. For example, genomic relationships allow for the differentiating of breeding values among full-sibs before their offspring's performance is known, which is not possible with pedigree information. The main benefits of introducing GS in dairy cattle breeding, therefore, are a reduced generation interval and lower number of phenotypes required for accurate breeding value estimation (Schaeffer, 2006). Because of these benefits, the uptake of GS by the dairy breeding industry has been rapid (for reviews see: Sellner *et al.*, 2007; Ibañez-Escriche and Gonzalez-Recio, 2011; Bouquet and Juga, 2012; Pryce and Daetwyler, 2012a).

To perform GS, two steps are needed. In the first step, a set of animals are genotyped and phenotyped to form the so-called reference or training population. In the reference population, phenotypic values are matched with corresponding genotypes to calibrate prediction equations. In the second step, these prediction equations are matched with genotypes of animals under evaluation, such as selection candidates, to predict their genomic breeding values. To predict these breeding values, no phenotypes of the selection candidates or their offspring are required. Usually, the selection candidates are juveniles and their genomic breeding values are much more accurate than when estimated with conventional breeding tools that rely only on parent average information (Meuwissen *et al.*, 2001).

The accuracy of GS depends on several factors. First, GS accuracy increases together with the size of the reference population (Daetwyler *et al.*, 2008; Goddard, 2009a). Second, the level of GS accuracy depends on heritability of the trait (Daetwyler *et al.*, 2008; Goddard, 2009a). Likewise, in traditional selection, higher heritability results in higher accuracy. Two other important factors affecting GS accuracy are SNP chip density and effective population size (Daetwyler *et al.*, 2008; Goddard, 2009a). In short, more dense SNP chips and a lower effective population size lead to higher accuracy. Finally, family relationships also play an important role in the accuracy of GS. Selection candidates that are closely related to the reference population will be evaluated more accurately than those distantly related (Habier *et al.*, 2007; Habier *et al.*, 2010; Wolc *et al.*, 2011).

An important feature of GS, which makes it especially interesting to apply to novel traits, is that the expensive or laborious measurements do not have to be taken on a routine scale. Even with a reference population of limited size, GS was shown to be a promising tool for starting selecting for novel traits (Calus *et al.*, 2013). As reported by the authors, genetic progress can already be made with breeding values of low accuracy. Although low accuracy of breeding values results in relatively inaccurate ranking of the individuals, at the level of the breeding program, genetic progress can be achieved. Of course, higher accuracy will lead to higher genetic progress. Using the limited available data optimally may require alternative approaches and methodology than currently used for conventional breeding goal traits.

1.4 Objective

The overall objective of this thesis was to investigate different options of optimizing genomic selection for scarcely recorded traits. The investigated options were: genotype imputation for ungenotyped but phenotyped animals to be used to enlarge the reference population; optimization of the design of the reference population with respect to the relationships among the animals included in it; prioritizing genotyping of the reference population or the selection candidates; and using easily recordable predictor traits to improve the accuracy of breeding values for scarcely recorded traits.

1.5 Thesis outline

Chapter 2 of this thesis investigated whether the accuracy of genomic selection can be improved by supplementing a small reference population by ungenotyped but phenotyped animals. A dairy cattle population was simulated from which a reference population was sampled. This reference population consisted of 1,000 phenotyped and genotyped individuals. In the subsequent scenarios, the reference population was supplemented by an additional 1,000 ungenotyped or genotyped animals. Genotypes of the ungenotyped animals were predicted based on the genotypes of their relatives and pedigree information. The accuracy of breeding values for all the scenarios were compared among each other to answer the question of whether there is a positive effect of enlarging the reference population by ungenotyped but phenotyped animals on the accuracy of genomic selection.

Chapter 3 investigated the impact of different family designs in terms of the relationships within the reference population, as well as the relationship of selection candidates to the reference population on accuracy of genomic selection. A dairy cattle population structure was simulated. Scenarios differed by the level of relationships among the animals in the reference population. Differences in predicted accuracy of breeding values were compared between scenarios. The analysis allowed for determining the optimal design of the reference population and association between relationships to the reference population and predicted breeding values accuracy.

Breeding values estimated by using genomic information are more accurate than pedigree based ones. The aim of the study as described in Chapter 4 was to investigate whether this increase is mainly due to genotyping reference or the animals evaluated. For this purpose, a simulated dataset reflecting a dairy cattle population was used. Four scenarios were considered in which genomic information on different groups of animals was available. The genomic information was available on (1) no animals; (2) reference population; (3) evaluated animals; or (4) reference population and evaluated animals. For each of the scenarios, breeding value accuracies were predicted using selection index theory.

Next, to optimize the reference population with respect to its design or size, predictor traits can be used to increase the accuracy of genomic selection for a novel trait. This option was evaluated using real data in Chapter 5 by investigating the effect of using predictor traits on the accuracy of genomic breeding values for a trait recorded on a limited cow reference population. The analyzed scenarios assumed that one or two predictor traits were available on the reference population only, or both on the reference population and the evaluated animals. The novel trait was dry matter intake and fat-protein-corrected milk yield and live weight was used as the predictor traits.

The general discussion focused on several aspects related to the genetic improvement of novel traits. First, the importance of female reference populations for novel traits was discussed. Next, the accuracy of genomic selection for novel traits at different project budgets was analyzed to indicate a break-even point between the costs of genotyping and phenotyping. Finally, foresight concerning the future challenges for selection of novel traits was given, including: development of cheap phenotyping techniques for currently difficult to measure traits, use of sequence data in the process of selecting for novel traits and farmers' participation in the phenotyping of novel traits.

2

Effect of enlarging the reference population with (un)genotyped animals on the accuracy of genomic selection in dairy cattle

M. Pszczola[1,2,3], H. A. Mulder[2] and M. P. L. Calus[2]

[1] Animal Breeding and Genomics Centre, Wageningen University, 6700 AH Wageningen, The Netherlands; [2] Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, 8200 AB Lelystad, The Netherlands; [3] Department of Genetics and Animal Breeding, Poznan University of Life Sciences, Wolynska 33, 60-637 Poznan, Poland

Journal of Dairy Science (2011) 94:431-441

Abstract

Genomic selection (GS) permits accurate breeding values to be obtained for young animals, shortening the generation interval and accelerating the genetic gain, thereby leading to reduced costs for proven bulls. Genotyping a large number of animals using high-density single nucleotide polymorphism marker arrays is nevertheless expensive, and therefore, a method to reduce the costs of GS is desired. The aim of this study was to investigate an influence of enlarging the reference population, with either genotyped animals or individuals with predicted genotypes, on the accuracy of genomic estimated breeding values. A dairy cattle population was simulated in which proven bulls with 100 daughters were used as a reference population for GS. Phenotypic records were simulated for bulls with heritability equal to the reliability of daughter yield deviations based on 100 daughters. The simulated traits represented heritabilities at the level of individual daughter performance of 0.3, 0.05, and 0.01. Three scenarios were considered in which (1) the reference population consisted of 1,000 genotyped animals, (2) 1,000 ungenotyped animals were added to the reference population, and (3) the 1,000 animals added in scenario 2 were genotyped in addition to the 1,000 animals from scenario 1. Genotypes for ungenotyped animals were predicted with an average accuracy of 0.58. Additionally, an adjustment of the diagonal elements of the genomic relationship matrix (**G**) was proposed for animals with predicted genotypes. The accuracy of genomic estimated breeding values for juvenile animals was the highest for the scenario with 2,000 genotyped animals, being 0.90, 0.79, and 0.60 for the heritabilities of 0.3, 0.05, and 0.01, respectively. Accuracies did not differ significantly between the scenario with 1,000 genotyped animals only and the scenario in which 1,000 ungenotyped animals were added and the adjustment of the **G** matrix was applied. The absence of significant increase in the accuracy of genomic estimated breeding values was attributed to the low accuracy of predicted genotypes. Although the differences were not significant, the difference between scenario 1 and 2 increased with decreasing heritability. Without the adjustment of the diagonal elements of the **G** matrix, accuracy decreased. Results suggest that inclusion of ungenotyped animals is only expected to enhance the accuracy of GS when the unknown genotypes can be predicted with high accuracy.

Key words: genomic selection, accuracy of genomic breeding value, prediction of genotypes, dairy cattle

2.1 Introduction

Different approaches of including genomic information into breeding value estimations have been presented (Meuwissen et al., 2001; Kolbehdari et al., 2007; Long et al., 2007; Muir, 2007). One of them is the genomic BLUP procedure (G-BLUP), in which equal variances are assumed for all SNP effects (Meuwissen et al., 2001). The accuracy of the obtained genomic estimated breeding values (GEBV) for juvenile animals that have no phenotypic observations is higher than the accuracy of pedigree indexes (Meuwissen et al., 2001). In the G-BLUP procedure, the genomic relationship matrix (**G**; VanRaden, 2008) can be used. Modeling SNP with equal variance is equivalent to using a genomic relationship matrix (VanRaden, 2008; Goddard, 2009a; Strandén and Garrick, 2009). The **G** matrix contains relationship coefficients among evaluated animals estimated based on genomic information. Genomic relationship coefficients are estimated with higher accuracy than when using pedigree information, because genomic information allows the capture of Mendelian sampling across the genome.

The accuracy of GEBV for juveniles increases with the number of animals included in the reference population used to estimate SNP effects (Meuwissen et al., 2001; Goddard, 2009a; Hayes et al., 2009a). Enlarging the reference population, however, implies the genotyping of additional animals, which increases the costs of GS. Enlarging the reference population with ungenotyped animals for which genotypes can be predicted might be an inexpensive method to increase the accuracy of GEBV. Gengler et al. (2007) proposed a method that treats (unknown) genotypes as (missing) phenotypes and uses the additive relationship matrix, based on pedigree information, to predict genotypes. Although this is a promising and easy to implement strategy, the eventual influence of using predicted genotypes in the reference population on the accuracy of GEBV has not yet been studied.

The objective of this study was to investigate the effect of GEBV of enlarging the reference population in a dairy cattle breeding program by adding bulls with known or predicted genotypes on the accuracy.

2.2 Materials and methods

Simulation

A dairy cattle population was simulated using similar assumptions as in Villumsen et al. (2009) and Calus et al. (2008). Data sets were simulated for daughter yield deviations (DYD) based on observations of 100 daughters of traits with high (0.3), moderate (0.05), and low (0.01) heritability. The first 1,000 generations had an effective population size of 400, consisting of 200 sires and 200 dams. All loci had alleles 1 and 2 segregating in the first generation, both with an allele frequency of 0.5. Linkage disequilibrium was established by performing random mating for the first 1,000 generations. Inheritance of parental chromosomes was in accordance with Haldane's (1919) mapping function. The mutation rate was 2×10^{-5}, where a mutation in allele 1 (2) yielded an allele 2 (1).

Generated genome length was 6 M and the genome consisted of 12 chromosomes, each 0.5 M long. This corresponds, approximately, to 20% of the actual cattle genome (Ihara et al., 2004). Marker loci (5,002) were spaced at fixed distances of 0.12 cM across the genome. After 1,000 generations of random mating, on average 4,500 markers were still segregating (i.e., on average 7.5 SNP/cM). Between 198 and 208 SNP (1 every 20 SNP) were removed from the segregating loci and used as QTL. Parameters used in the simulation are summarized in Table 2.1.

Table 2.1 Simulation parameters.

Parameter	Value		
Effective population size of the first 1,000 generations			400
Simulated genome length, M			6
Number of simulated chromosomes			12
Length of simulated chromosomes, M			0.5
Distance between adjacent markers in generation 1,008, M			0.0012
Number of SNP markers per cM			~ 7.5
Number of QTL per cM			~ 0.33
Minor allele frequency			0.29
	Trait 1	Trait 2	Trait 3
Heritability of phenotype	0.3	0.05	0.01
Heritability for DYD	0.89	0.56	0.20

After the first 1,000 generations (i.e., in generation 1,001), the population was extended to 800 individuals. In generations 1,001 to 1,008, no mutations were simulated. In generations 1,001 to 1,007, 50 males and 200 females were randomly chosen as parents of the next generation. The matings were restricted such that there were no full-sibs among the 800 offspring. For generations 1,002 to 1,008, genotypes, true breeding values (TBV), and phenotypes of the males were simulated. Generation 1,008, containing juvenile animals, was simulated with unknown phenotypes. Pedigree was stored for all animals from generations 1,002 to 1,008. The outline of the simulation is presented in Figure 2.1.

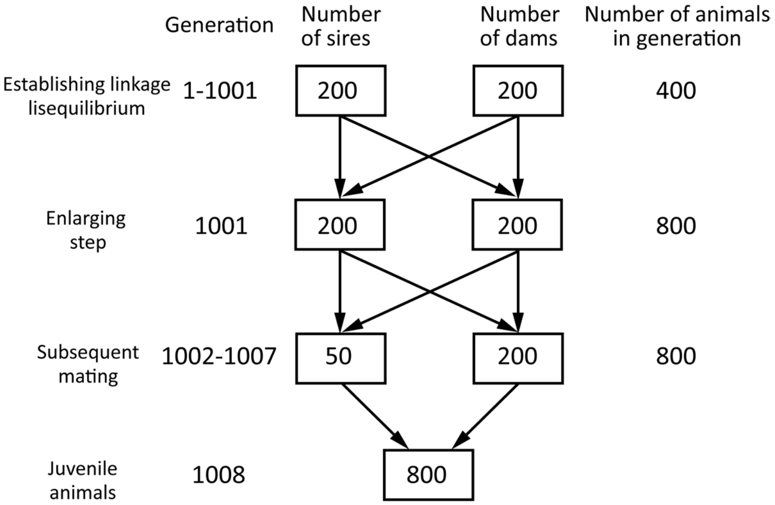


Figure 2.1 Simulation scheme.

To simulate TBV, QTL effects were randomly drawn from a normal distribution, $N(0,1)$. All QTL effects were summed per animal to obtain TBV, assuming that QTL were independent. The total variance of the TBV was calculated (denoted as σ^2_{TBV}) as a variance of TBV across all animals. Phenotypes were obtained by adding a random residual term, $N(0, {\sigma_e}^2)$, to the TBV; ${\sigma_e}^2$ was derived as:

$$\sigma_e^2 = \frac{\sigma^2_{TBV}(1-h^2)}{h^2} \quad [2.1]$$

where h^2 is the simulated heritability (see Table 2.1). Phenotypic records were simulated for bulls, with h^2 equal to the reliability of DYD. This reliability (r^2_{IH}) for progeny-tested bulls was calculated according to the formula of Mrode (2005):

$$r^2_{IH} = \frac{{}^1/_4 n h^2}{1 + {}^1/_4 (n-1) h^2} \quad [2.2]$$

where *n* is the number of daughters, which was considered to be 100. As a result, heritabilities of 0.3, 0.05, and 0.01 at the phenotypic level yielded heritabilities of 0.89, 0.56, and 0.20 at the DYD level, respectively (Table 2.1).

Scenarios

To meet the goal of this study the described simulation was performed for all 3 heritability levels, and each simulation was replicated 10 times. Three scenarios were considered with, on the one hand, different sizes of the reference population and, on the other hand, different numbers of animals with known or predicted genotypes. In the first scenario, GEBV were estimated using a reference population consisting of 1,000 genotyped sires chosen randomly from genotyped bulls within generations 1,002 to 1,007 (200 sires out of 400 available genotyped bulls) with phenotypes based on DYD.

In the second scenario, GEBV were estimated with an additional 1,000 ungenotyped bulls. Those bulls were the remaining 200 bulls per generation after choosing bulls for scenario 1 and had no offspring. A particular genotyped bull was mated only once with a particular cow, resulting in one offspring; the highest possible degree of relationship among the ungenotyped and genotyped animals was therefore half-sib or parent-offspring. The unknown genotypes were predicted using the method of regression on gene content, proposed by Gengler et al. (2007). This method treats (missing) genotypes as (unknown) phenotypes and uses the additive genetic relationship matrix (**A**) to predict them using the following model for each SNP separately:

$$gc_i = \mu_{gc} + d + e_{gc_i} \qquad [2.3]$$

where gc is the observed (missing) gene content for (un)genotyped animal *i*, μ_{gc} is an overall mean, *d* is EBV for gene content, and e_{gc} is the residual of gene content. The **A** matrix contained animals from generations 1,002 to 1,008. The heritability used in the mixed model equations was 0.99. The ASREML software (Gilmour et al., 2002) was used to solve the mixed model equations.

The third scenario was similar to the second one, except that all 2,000 bulls were considered genotyped. A traditional BLUP (scenario 4) was also performed using phenotypes for the same 2,000 bulls considered in scenario 3.

Estimation of GEBV

After each simulation, genotypes of the animals were used to create the **G** matrix according to VanRaden (2008):

$$\boldsymbol{G} = \frac{\boldsymbol{ZZ'}}{2\sum p_i(1-p_i)} \qquad [2.4]$$

where p_i is the frequency of the second allele at locus *i*, and **Z** is derived by subtracting 2 times the allele frequency expressed as a difference of 0.5; that is, $2(p_i - 0.5)$, from matrix **M** that specifies the marker genotypes for each individual as −1, 0, or 1 (VanRaden, 2008). In this study the allele frequencies p_i were considered to be 0.5; therefore, **Z** was the same as **M**. The value for p_i of 0.5 reflects the allele frequency in our simulated base population. In preliminary analysis, we found limited differences when using 0.5 instead of the allele frequency in the current population for p_i. VanRaden (2008) found that using different values for p_i hardly affects the accuracy of the GEBV. However, Aguilar et al. (2010) and Christensen and Lund (2010) found that using different allele frequencies does influence the accuracy of GEBV. Results of Aguilar et al. (2010) indicated that allele frequencies of 0.5 actually gave the highest accuracy. These different results indicate that the effect of used allele frequencies for p_i appears to be different for different data sets.

Subsequently, GEBV were estimated with G-BLUP using the following model:

$$y_j = \mu + a_j + e_j, \qquad [2.5]$$

where μ is an overall mean, a_j is an estimated breeding value and e_j is the random error term. The estimated breeding values were assumed to be distributed as $N(0, \boldsymbol{G}\sigma_a^2)$ and the residuals were assumed to be distributed as $N(0, \sigma_e^2)$. The genetic variance σ_a^2 and residual variance σ_e^2 were estimated using restricted maximum likelihood (REML) implemented in ASReml software (Gilmour et al., 2002).

In G-BLUP an inverse of the **G** matrix is required. To avoid singularities in **G**, the latter was weighted by **A** as follows: $\boldsymbol{G}\omega = \omega\boldsymbol{G} + (1 - \omega)\boldsymbol{A}$ (VanRaden, 2008), using a weighting factor (ω) of 0.99, meaning that a relatively low weight was given to the **A** matrix. The **A** matrix used to weight the **G** matrix contained only animals that were present in **G**.

Comparison of GEBV

The 10 replicates were analyzed by calculating the accuracies of GEBV, regression of simulated breeding values on GEBV, and mean squared errors of prediction (MSEP) for each group of animals. Furthermore, estimated heritabilities were compared with the simulated values.

Adjustment of diagonal elements of the G matrix

Initial analyses (Figure 2.2) showed that the diagonal elements of the **G** matrix for ungenotyped animals using predicted gene contents were much lower than 1.0. This is in conflict with the expectation of diagonal elements of a relationship matrix, which is 1 plus the inbreeding coefficient of the animal (Wright, 1922). Although the expectation of the diagonal elements of the **G** matrix is >1, the diagonal elements in the **G** matrix can be lower for some animals (VanRaden, 2008). However, in our case, the average diagonal element of **G** for ungenotyped animals was only 0.73. The low values for the diagonal elements of the **G** matrix were caused by the fact that the predicted gene content is regressed back to the mean for ungenotyped animals. This regression to the mean is due to the nature of BLUP and leads to an excess of expected heterozygotes. Both off-diagonal and diagonal elements were affected by the regression to the mean; nevertheless, in absolute terms the effect was largest on diagonal elements. The mean absolute difference of predicted and true values of the off-diagonals was only 0.03. Because the diagonal elements of **G**, based on the number of homozygous loci of an animal (VanRaden, 2008), were too low, we proposed an adjustment of the diagonal elements for ungenotyped animals. In this adjustment, the diagonal elements for ungenotyped animals are calculated, assuming no other relationships between sire and dam except between sire and maternal grandsire, as

$$g_{jj} = 1 + 0.25 * g(s_j, mgs_j) \qquad [2.6]$$

where g_{jj} is the diagonal element for the ungenotyped animal *j*, and $g(s_j, mgs_j)$ is the genomic relationship coefficient between the genotyped sire (*s*) and genotyped maternal grandsire (*mgs*) of animal *j*. Results obtained after this adjustment were denoted as an additional scenario (scenario 2a), whereas the results obtained without the adjustment were presented as scenario 2.

2.3 Results

Characteristics of the simulation

In each replicate, animals with genotypes were simulated. The average distance between adjacent loci across the whole genome, calculated in generation 1,008, was 0.13 cM (Table 2.1). The average LD between adjacent markers measured as r^2 (Hill and Robertson, 1968) was 0.41. Minor allele frequency, averaged across all marker loci in generation 1,008, was 0.29. Gene contents for 1,000 animals were predicted with an average accuracy of 0.58.

Estimated heritabilities

Estimated DYD heritabilities averaged over 10 replicates for scenarios 1, 2, 3, and 4 (Table 2.2) were always slightly higher than the simulated values. Differences among the estimated DYD heritabilities for the different scenarios, apart from scenario 2 for high and moderate heritability, were not statistically significant from the simulated value. Estimated heritability in scenario 2 for high and moderate heritability traits was significantly lower from the simulated value. The adjustment of the diagonal elements of the **G** matrix in scenario 2a resulted in estimated heritabilities not significantly different from the simulated values.

Table 2.2 Estimated heritabilities of daughter yield deviation across different scenarios and simulated heritabilities.

	High simulated heritability		Moderate simulated heritability		Low simulated heritability	
Scenario[1]	h^2	SE	h^2	SE	h^2	SE
Simulated heritability	0.89	-	0.56	-	0.20	-
1	0.90^{a}	0.04	0.58^{c}	0.01	0.22^{e}	0.05
2	0.63^{b}	0.03	0.48^{d}	0.03	0.23^{e}	0.04
2a	0.92^{a}	0.04	0.60^{c}	0.01	0.24^{e}	0.04
3	0.91^{a}	0.03	0.61^{c}	0.03	0.25^{e}	0.03
4	0.92^{a}	0.03	0.57^{c}	0.03	0.21^{e}	0.04

[a–e] Values with identical superscripts did not differ significantly among scenarios (*P > 0.05*); standard errors of 10 replicates ranged from 0 to 0.012;

[1] Scenario 1 consisted of 1,000 genotyped animals; scenario 2 consisted of 1,000 genotyped and 1,000 ungenotyped animals with unadjusted diagonal elements of the genomic relationship matrix for ungenotyped animals; scenario 2a consisted of 1,000 genotyped and 1,000 ungenotyped animals with adjusted diagonal elements of the genomic relationship matrix for ungenotyped animals; scenario 3 consisted of 2,000 genotyped animals; scenario 4 consisted of 2,000 genotyped animals analyzed with use of traditional BLUP;

[2] Standard error of estimated heritability averaged over 10 replicates.

Evaluation of G matrix

The diagonal elements of the **G** matrix, without adjusting the coefficients for predicted genotypes, were considerably lower when compared with the same coefficients for genotyped animals (Figure 2.2). The adjustment proposed in the present study resulted in diagonal elements that were closer to their expectations, as can be seen from Figure 2.3. Those coefficients were, nevertheless, still lower than for genotyped animals; the coefficients for the group of the additional animals were generally lower in the **G** matrix based on predicted gene contents than the corresponding coefficients when genotype data were available (Figure 2.3). The coefficients in the **G** matrix were higher than pedigree-based coefficients, for genotyped as well as the additional 1,000 bulls with predicted genotypes (Figures 2.4 and 2.5).

Evaluation of GEBV

The accuracy of GEBV for the first 1,000 animals for all simulated heritability levels was the highest in scenario 3 (0.96 for high, 0.88 for moderate, and 0.72 for low heritability; Table 2.3). The lowest values were obtained in scenario 2 for high heritability (0.93) and for scenario 4 for moderate (0.81) and low (0.60) heritability. The accuracy of GEBV in scenarios 1 and 2a did not differ significantly from each other ($P > 0.05$) within high and moderate heritability and was significantly different for low heritability ($P < 0.05$).

The highest accuracies for the additional group of the animals were observed in scenario 3 for high (0.96), moderate (0.87), and low (0.70) heritabilities; these estimates were significantly different from the accuracies in scenario 2, 2a, and 4 ($P < 0.05$). For all heritability levels, accuracies of scenarios 2 and 4 were significantly different from each other, whereas the accuracy in scenarios 2a and 4 were not significantly different ($P > 0.05$).

Adjusting the diagonal elements of the **G** matrix (scenario 2a) resulted in a significant increase of the accuracy of GEBV for the additional group of animals for low and moderate heritability, and for all groups of animals for high heritability when compared with the results of the scenario with no adjustment (scenario 2). For low heritability, a significant decrease was observed in the group of the first 1,000 animals.

Evaluation of GEBV

The accuracy of GEBV for the first 1,000 animals for all simulated heritability levels was the highest in scenario 3 (0.96 for high, 0.88 for moderate, and 0.72 for low heritability; Table 2.3). The lowest values were obtained in scenario 2 for high heritability (0.93) and for scenario 4 for moderate (0.81) and low (0.60) heritability. The accuracy of GEBV in scenarios 1 and 2a did not differ significantly from each other ($P > 0.05$) within high and moderate heritability and was significantly different for low heritability ($P < 0.05$).

The highest accuracies for the additional group of the animals were observed in scenario 3 for high (0.96), moderate (0.87), and low (0.70) heritabilities; these estimates were significantly different from the accuracies in scenario 2, 2a, and 4 ($P < 0.05$). For all heritability levels, accuracies of scenarios 2 and 4 were significantly different from each other, whereas the accuracy in scenarios 2a and 4 were not significantly different ($P > 0.05$).

Adjusting the diagonal elements of the **G** matrix (scenario 2a) resulted in a significant increase of the accuracy of GEBV for the additional group of animals for low and moderate heritability, and for all groups of animals for high heritability

when compared with the results of the scenario with no adjustment (scenario 2). For low heritability, a significant decrease was observed in the group of the first 1,000 animals.

Table 2.3 Accuracies (Acc.), regression coefficients (Reg.), and mean squared error of prediction (MSEP) of genomic estimated breeding values for groups of 1,000 first, additional and juvenile animals for heritability of 0.3 (0.89 daughter yield deviation) for all scenarios.

Scenario[1]	First 1,000 anim.			Additional anim.			Juvenile anim.		
	Acc.	Reg.	MSEP	Acc.	Reg.	MSEP	Acc.	Reg.	MSEP
1	0.96^{a}	1.02	6.79	-	-	-	0.84^{h}	1.01	23.09
2	0.93^{b}	1.15^{2}	12.35	0.65^{e}	1.27^{2}	45.15	0.80^{i}	1.19^{2}	28.47
2a	0.96^{a}	1.00	6.50	0.95^{f}	1.03	8.50	0.83^{h}	0.98	22.63
3	0.96^{c}	1.02	5.54	0.96^{g}	1.02	6.07	0.90^{j}	1.01	15.09
4	0.95^{d}	1.01	7.79	0.95^{f}	1.00	8.18	0.57^{k}	1.04	52.25

$^{a–k}$ Values with identical superscripts did not differ significantly (*P > 0.05*); standard errors of 10 replicates ranged from 0 to 0.02 for Acc. and Reg., and from 0.2 to 1.77 for MSEP; [1] Scenario 1 consisted of 1,000 genotyped animals; scenario 2 consisted of 1,000 genotyped and 1,000 ungenotyped animals with unadjusted diagonal elements of the genomic relationship matrix for ungenotyped animals; scenario 2a consisted of 1,000 genotyped and 1,000 ungenotyped animals with adjusted diagonal elements of the genomic relationship matrix for ungenotyped animals; scenario 3 consisted of 2,000 genotyped animals; scenario 4 consisted of 2,000 genotyped animals analyzed with use of traditional BLUP;

[2] Reg. significantly different from 1.

Evaluation of GEBV

The accuracy of GEBV for the first 1,000 animals for all simulated heritability levels was the highest in scenario 3 (0.96 for high, 0.88 for moderate, and 0.72 for low heritability; Table 2.3). The lowest values were obtained in scenario 2 for high heritability (0.93) and for scenario 4 for moderate (0.81) and low (0.60) heritability. The accuracy of GEBV in scenarios 1 and 2a did not differ significantly from each other (*P > 0.05*) within high and moderate heritability and was significantly different for low heritability (*P < 0.05*).

The highest accuracies for the additional group of the animals were observed in scenario 3 for high (0.96), moderate (0.87), and low (0.70) heritabilities; these estimates were significantly different from the accuracies in scenario 2, 2a, and 4 (*P < 0.05*). For all heritability levels, accuracies of scenarios 2 and 4 were significantly different from each other, whereas the accuracy in scenarios 2a and 4 were not significantly different (*P > 0.05*).

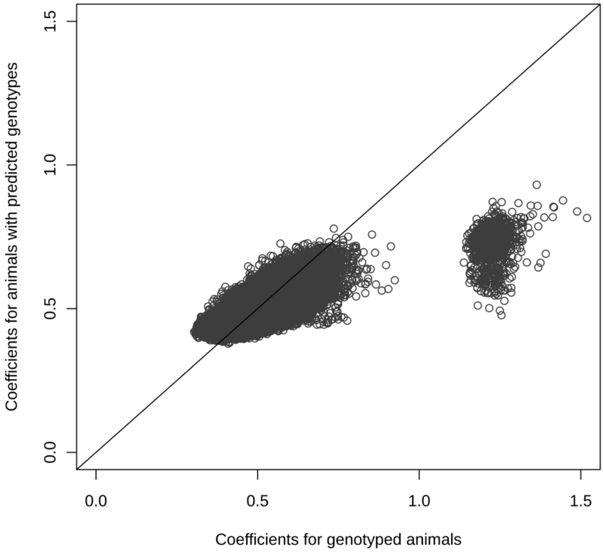


Figure 2.2 Coefficients of the genomic relationship matrix for animals with predicted genotypes plotted against the same coefficients calculated using their true genotypes without adjusting the diagonal elements.

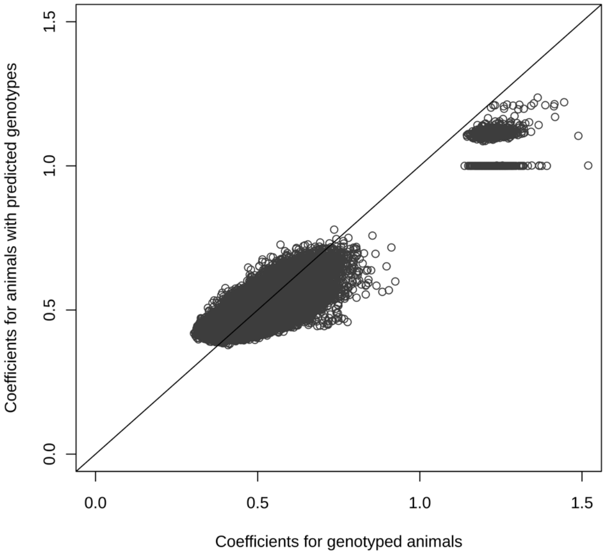


Figure 2.3 Coefficients of the genomic relationship matrix for animals with predicted genotypes plotted against the same coefficients calculated using their true genotypes with adjusting the diagonal elements.

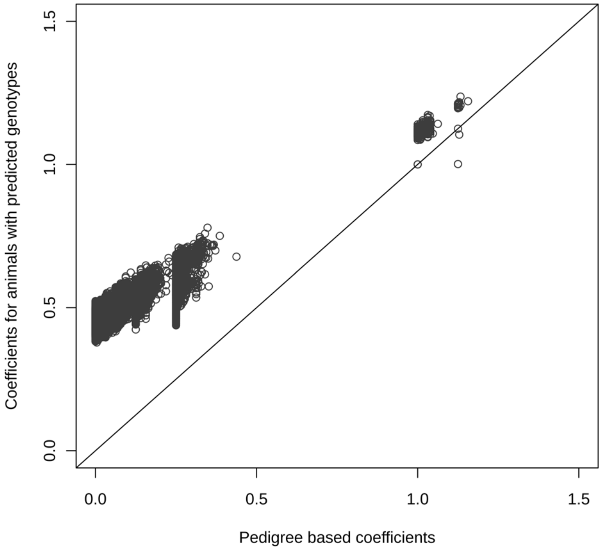


Figure 2.4 Coefficients of the genomic relationship matrix for the animals with predicted genotypes and adjusted diagonal elements of genomic relationship matrix plotted against pedigree-based relationship coefficients.

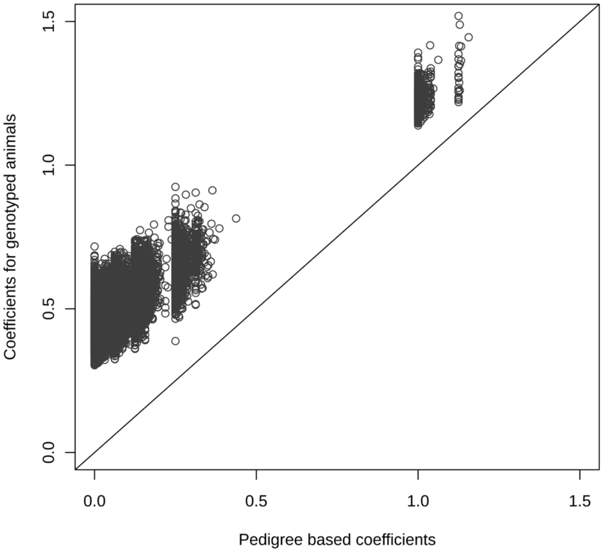


Figure 2.5 Coefficients of the genomic relationship matrix for the genotyped animals plotted against pedigree-based relationship coefficients.

Adjusting the diagonal elements of the **G** matrix (scenario 2a) resulted in a significant increase of the accuracy of GEBV for the additional group of animals for low and moderate heritability, and for all groups of animals for high heritability when compared with the results of the scenario with no adjustment (scenario 2). For low heritability, a significant decrease was observed in the group of the first 1,000 animals.

Similarly to the group of first 1,000 animals and the group of additional animals, the accuracy for the juvenile animals was the highest in scenario 3 for all 3 heritability levels: 0.90 (high), 0.79 (moderate) and 0.60 (low), and the lowest in scenario 4: 0.57 (high), 0.48 (moderate), and 0.33 (low). Accuracies reached in scenarios 1 and 2a were somewhat lower than those in scenario 3. Accuracy estimates for the first 1,000 animals in scenario 2 were significantly higher ($P < 0.05$) than those in scenario 1 for low heritability, lower for high heritability, and not significantly different from each other ($P > 0.05$) for moderate heritability. The accuracy of juvenile animals' GEBV in scenario 2 was significantly lower ($P < 0.05$) than in scenario 1 for high and not significantly different ($P > 0.05$) for the other 2 heritability levels. The adjustment of the diagonal elements of the **G** matrix (scenario 2a) resulted in significantly higher ($P < 0.05$) accuracy for all groups of the animals for high and low heritability and for the group of the 1,000 additional ungenotyped animals for moderate heritability when compared with scenario 2. The increase for the first 1,000 animals and juveniles for moderate heritability was, however, not significant ($P > 0.05$). Differences in accuracies between scenarios 1 and 2a were not significant across all heritability levels ($P > 0.05$), although the difference in accuracy between scenarios 2a and 1 tended to increase with decreasing heritability.

In summary, the accuracies of GEBV were the highest when the true genotypes of the additional 1,000 bulls were added to the reference population and were considerably lower for traditional BLUP. The accuracies of the GEBV for the animals with predicted genotypes were lower than for the genotyped ones and similar to the accuracies when applying traditional BLUP. Nevertheless, after adding animals with predicted genotypes and adjusting the diagonal coefficients of the **G** matrix, the accuracy of GEBV did not decline compared with the scenario when no animals were added. Furthermore, the accuracies of juvenile animals increased in scenario 2a compared with scenario 1 (Tables 2.4 and 2.5) for traits with moderate and low heritability, although the difference was not significantly greater than zero.

Table 2.4 Accuracies (Acc.), regression coefficients (Reg.), and mean squared error of prediction (MSEP) of genomic estimated breeding value for groups of 1,000 first, additional and juvenile animals for heritability of 0.05 (0.56 daughter yield deviation) for all scenarios.

Scenario[1]	First 1,000 anim.			Additional anim.			Juvenile anim.		
	Acc.	Reg.	MSEP	Acc.	Reg.	MSEP	Acc.	Reg.	MSEP
1	0.85^{a}	1.02	22.76	-	-	-	0.70^{g}	0.99	40.06
2	0.85^{a}	1.07[2]	22.56	0.62^{d}	1.23[2]	49.67	0.70^{g}	1.06	40.03
2a	0.85^{a}	1.01	20.93	0.79^{f}	1.12[2]	29.65	0.72^{g}	0.96	37.79
3	0.88^{b}	1.00	18.01	0.87^{e}	1.00	19.20	0.79^{h}	0.97	30.33
4	0.81^{c}	0.99	27.75	0.79^{f}	0.99	29.52	0.48^{i}	0.99	60.95

$^{a–l}$ Values with identical superscripts did not differ significantly (*P > 0.05*); standard errors of 10 replicates ranged from 0 to 0.03 for Acc. and Reg., and from 0.88 to 2.41 for MSEP; [1] Scenario 1 consisted of 1,000 genotyped animals; scenario 2 consisted of 1,000 genotyped and 1,000 ungenotyped animals with unadjusted diagonal elements of the genomic relationship matrix for ungenotyped animals; scenario 2a consisted of 1,000 genotyped and 1,000 ungenotyped animals with adjusted diagonal elements of the genomic relationship matrix for ungenotyped animals; scenario 3 consisted of 2,000 genotyped animals; scenario 4 consisted of 2,000 genotyped animals analyzed with use of traditional BLUP;

[2] Reg. significantly different from 1.

For all simulated heritability levels and across all scenarios, regression coefficients of true on estimated breeding values were close to 1 (Table 2.3, Table 2.4 and Table 2.5). For the juvenile animals, regression coefficients were generally slightly less than 1. Regression coefficients of the animals with predicted genotypes were in all cases greater than 1, indicating that the variance of their GEBV was underestimated. This is caused by the fact that estimated gene contents were shrunk back toward the mean. Regression coefficients were significantly greater than 1 in scenario 2, especially for the group of the additional animals. These differences from 1 observed in scenario 2, apart from the group of additional animals for moderate heritability, were no longer significant when the diagonal elements were adjusted (scenario 2a); MSEP ranged from 5.54 to 68.88 across all scenarios.

2.4 Discussion

The objective of this study was to investigate the effect of enlarging the reference population in a dairy cattle breeding program, by adding bulls with known or predicted genotypes, on the accuracy of GEBV. Four scenarios were evaluated that differed with regard to the number of genotyped animals and ungenotyped animals with predicted genotypes. As expected, the accuracies of GEBV for all traits were higher for the scenario with a higher number of genotyped animals. A similar trend was observed by Goddard (2009a).

Table 2.5 Accuracies (Acc.), regression coefficients (Reg.), and mean squared error of prediction (MSEP) of genomic estimated breeding value for groups of 1,000 first, additional and juvenile animals for heritability of 0.05 (0.56 daughter yield deviation) for all scenarios.

Scenario[1]	First 1,000 anim.			Additional anim.			Juvenile anim.		
	Acc.	Reg.	MSEP	Acc.	Reg.	MSEP	Acc.	Reg.	MSEP
1	0.65^a	1.03	45.72	-	-	-	0.52^i	1.01	56.14
2	0.69^b	1.02	42.04	0.51^f	1.18^2	58.19	0.54^i	0.98	54.94
2a	0.66^c	0.95	42.02	0.57^g	1.08	53.04	0.56^i	0.96	55.03
3	0.72^d	0.99	38.85	0.70^h	0.99	40.35	0.60^j	0.94	49.57
4	0.60^e	1.00	50.78	0.57^g	1.00	53.53	0.33^k	0.91^2	68.88

$^{a–k}$ Values with identical superscripts did not differ significantly (*P > 0.05*); standard errors of 10 replicates ranged from 0 to 0.07 for Acc. and Reg., and from 1.93 to 2.71 for MSEP; [1] Scenario 1 consisted of 1,000 genotyped animals; scenario 2 consisted of 1,000 genotyped and 1,000 ungenotyped animals with unadjusted diagonal elements of the genomic relationship matrix for ungenotyped animals; scenario 2a consisted of 1,000 genotyped and 1,000 ungenotyped animals with adjusted diagonal elements of the genomic relationship matrix for ungenotyped animals; scenario 3 consisted of 2,000 genotyped animals; scenario 4 consisted of 2,000 genotyped animals analyzed with use of traditional BLUP;

[2] Reg. significantly different from 1.

Accuracy of GEBV

In general, adding animals with predicted genotypes to the reference population did not significantly increase the accuracy of GEBV; however, a trend of increasing difference in accuracy between scenarios 1 and 2a with decreasing heritability was observed (Tables 2.3, 2.4, and 2.5). This trend suggests that when using traits with low heritability or less accurate phenotypic records (i.e., own performance records), scenario 2a, in which animals with predicted genotypes were added, may become beneficial.

Comparison of the estimates for BLUP (scenario 4) and G-BLUP in scenario 2a showed that, in general, accuracy increased significantly for the group of the first 1,000 animals and did not differ for the group of additional ungenotyped animals. For the juvenile animals, however, superiority of G-BLUP can be seen clearly, as also reported by others (Meuwissen et al., 2001; Schaeffer, 2006). The improvement in accuracy for juvenile animals comparing G-BLUP with traditional BLUP was apparent and similar at all heritability levels. This is in contrast to findings of others, who found that use of marker information was especially beneficial for low heritability traits (Meuwissen and Goddard, 1996; Meuwissen et al., 2001; Mulder et al., 2010).

The coefficients of the **G** matrix between animals with predicted gene contents were clearly biased as they were shrunk toward the mean. A solution could be to account for the variance of predicted gene content. The bias of the coefficients can be judged based on Figure 2.2 as the difference between the values on the *y*- and *x-axes*. Although calculation of the diagonal elements in the **G** matrix using equation [2.5] increased these coefficients on average by 0.45, Figure 2.3 shows that the variance of those adjusted coefficients is still underestimated. The adjustment of the diagonals, however, did result in a clear improvement of the accuracy of the GEBV for scenario 2a compared with scenario 2. This implies that, although parts of the **G** matrix may still be biased, our ad hoc adjustment has made the **G** matrix overall more consistent, which resulted in more accurate prediction of the GEBV. Therefore, in situations where parts of the **G** matrix are dependent on different sources of information with different levels of accuracy, such as known versus predicted gene content, it appears to be important to ensure that different parts of the matrix have similar properties. Christensen and Lund (2010) proposed a model in which, unlike the method presented here, both predicted genotypes and the variance of the prediction are included. This strategy, therefore, directly yields a more consistent matrix.

When the aim is to increase the accuracy of GEBV for a certain group of juveniles, the best strategy is probably to add animals to the reference population that are closely related to those juveniles (e.g., their parents; Habier et al., 2010). When the aim is to increase the accuracy of GEBV for juveniles throughout the population, one strategy may be to add animals that are not closely related to the reference population, and therefore add to the average relationship of the reference population to any given animal in the population. However, when animals with predicted genotypes are added to the reference population, the gain may be larger when those animals are more closely related to the reference population, because their genotypes are predicted with higher accuracy.

Figures 2.4 and 2.5 show that coefficients of the **A** matrix were lower than those of the **G** matrix, which is because of the difference in the level of inbreeding when using pedigree or genomic information. The level of inbreeding when using pedigree data was calculated using the first generation in the pedigree as the base generation, while the genomic inbreeding level is calculated using generation 1 in the simulation as base generation. Therefore, inbreeding coefficients in the **G** matrix are higher, and consequently the coefficients in the **G** matrix are higher compared with the corresponding ones in the **A** matrix.

Accuracy of genomic relationship matrix with predicted gene contents

In the present study, ungenotyped animals did not have offspring and therefore no genotype information on descendants was available. A particular bull was mated only once with a particular cow and, therefore, the highest possible degree of relationship among the ungenotyped and genotyped animals was half-sib or parent-offspring. This resulted in relatively low accuracy of predicted gene content and, thus, low accuracy of GEBV for ungenotyped and juvenile animals in scenario 2. To increase the accuracy of the prediction of genotypes, ungenotyped animals should be chosen that have genotyped offspring available. In such a scenario, ungenotyped animals could be, for instance, dams of genotyped offspring. The accuracy of predicted gene content is then expected to be 0.88 with 10 genotyped half-sib offspring, using the square root of equation [2.2] and assuming a heritability of 1.0 for gene content. In that case, superior accuracies of GEBV to the ones obtained in this study may be obtained. A disadvantage is that the phenotypic information of dams generally has a lower reliability compared with bulls.

Alternatively, if available, the additional genotypic information could be used for prediction of genotypes. The use of this additional information could lead to the maximum possible accuracy of 0.707 when both parents but no offspring are genotyped, and assuming that the heritability is 1.0 ($r = \sqrt{\boldsymbol{bG}}/\sigma_a = \sqrt{1/2}$, where both ***b*** and ***G*** contain values of ½ and $\sigma_a = 1$; Mrode, 2005). With genotyped offspring, this accuracy could be substantially higher; Gengler et al. (2008), for instance, reported an accuracy of 0.93 when Canadian Holstein data were used to evaluate the gene content prediction method.

A similar strategy to include ungenotyped animals is to combine genotypic and pedigree information in a modified relationship matrix. Such an approach was proposed by Legarra et al. (2009) and Christensen and Lund (2010), who used genomic information to enrich the **A** matrix whenever this information was available. This approach does not require explicit prediction of the unknown genotypes, but directly predicts the genomic relationships using pedigree and genomic information from relatives, and therefore provides a one-step genetic evaluation with use of genomic information. This alternative approach may be able to overcome the problems encountered due to shrinkage of estimated gene content in the method applied here. An important advantage of this approach is that breeding values can be estimated in a single step including records of all genotyped and ungenotyped animals, in contrast to our 2-step approach in which

DYD are used to predict GEBV. In the simulation presented by Christensen and Lund (2010), a higher accuracy of GEBV was obtained for the 1-step approach compared with the 2-step approach. This suggests that part of the information in the 2-step approach is lost due to errors in DYD prediction, thereby leading to lower accuracy of the GEBV.

Inversion of G matrix

G-BLUP implemented in a traditional mixed model equations procedure requires the inverted **G** matrix. When the number of animals in the dataset is large, direct inversion of **G** may not be possible due to computational difficulties. Furthermore, possible singularities in the **G** matrix makes unique inversion impossible. Singularities in the **G** matrix may appear among the animals with predicted genotypes when full-sibs are present or genotypes were predicted for the animals that do not have offspring and have unknown parents. In both cases, the method of Gengler et al. (2007) gives the same estimates for these groups of the animals. Therefore, no full-sibs were simulated among the ungenotyped animals in the present study. To deal with any further singularities, the final **G** matrix was a weighted sum of the **G** and **A** matrix following VanRaden (2008). Another method to eliminate singularities from the matrix is to modify the diagonal of the relationship matrix by adding a small number to it, for example 10^{-6} (Zhong et al., 2009).

An alternative method which does not require the inversion of the relationship matrix was recently proposed by Misztal et al. (2009a). In this method, the additive genetic relationship matrix is modified by a matrix accounting for genomic information as in Legarra et al. (2009) and then used in an asymmetric set of mixed model equations (Harville, 1976; Henderson, 1984, 1985) solved with the algorithm proposed by Misztal et al.(2009a). This method resolves the problem related to singularities in the relationship matrix by avoiding the inversion step.

Implications for animal breeding

This study confirmed that G-BLUP is more beneficial than traditional breeding value estimation. When ungenotyped animals are to be used to enlarge the reference population in order to increase accuracy of GEBV, animals should be chosen whose genotypes can be predicted with sufficiently high accuracy and have phenotypes of high quality. Fulfilling these conditions may lead to achieving high accuracy of predicted genotypes and making it worthwhile to enlarge the reference population with ungenotyped animals.

2.5 Conclusion

This study showed that inclusion of animals with predicted genotypes in the reference population did not significantly increase GEBV accuracies for juvenile animals. This lack of significance was mainly attributed to the low accuracy of predicted genotypes. Therefore, inclusion of ungenotyped animals is only expected to enhance the accuracy of GS when the unknown genotypes can be predicted with high accuracy.

2.6 Acknowledgments

Marcin Pszczola gratefully acknowledges the financial support of the Koepon Foundation (Amersfoort, the Netherlands). Han Mulder was financially supported by CRV (Arnhem, the Netherlands), Hendrix Genetics (Boxmeer, the Netherlands), IPG (Beuningen, the Netherlands), and the European Commission, within the 6th Framework project SABRE, contract No. FOOD-CT-2006-016250. The text represents the authors' views and does not necessarily represent a position of the Commission, who will not be liable for the use made of such information. The authors gratefully acknowledge the meticulous work of two anonymous reviewers who significantly improved the manuscript.

3

Reliability of direct genomic values for animals with different relationships within and to the reference population

M. Pszczola[1,2,3], T. Strabel[3], H. A. Mulder[2] and M. P. L. Calus[2]

[1] Animal Breeding and Genomics Centre, Wageningen University, 6700 AH Wageningen, The Netherlands; [2] Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, 8200 AB Lelystad, The Netherlands; [3] Department of Genetics and Animal Breeding, Poznan University of Life Sciences, Wolynska 33, 60-637 Poznan, Poland

Journal of Dairy Science (2012) 95:389-400

Abstract

Accuracy of genomic selection depends on the accuracy of prediction of single nucleotide polymorphism effects and the proportion of genetic variance explained by markers. Design of the reference population with respect to its family structure may influence the accuracy of genomic selection. The objective of this study was to investigate the effect of various relationship levels within the reference population and different level of relationship of evaluated animals to the reference population on the reliability of direct genomic breeding values (DGV). The DGV reliabilities, expressed as squared correlation between estimated and true breeding value, were calculated for evaluated animals at 3 heritability levels. To emulate difficult or expensive to measure trait, such as methane emission, reference populations were small and consisted of females with own performance records. A population reflecting a dairy cattle population structure was simulated. Four chosen reference populations consisted of all females available in the first genotyped generation. They consisted of highly (HR), moderately (MR), or lowly (LR) related animals, by selecting paternal half-sib families of decreasing size, or consisted of randomly chosen animals (RND). Of those 4 reference populations, RND had the lowest average relationship. Three sets of evaluated animals were chosen from 3 consecutive generations of genotyped animals, starting from the same generation as the reference population. Reliabilities of DGV predictions were calculated deterministically using selection index theory. Average reliabilities increased when average relationship within the reference population decreased and the highest average reliabilities were achieved for RND (e.g., from 0.53 in HR to 0.61 in RND for a heritability of 0.30). A higher relationship to the reference population resulted in higher reliability values. At the average squared relationship of evaluated animals to the reference population of 0.005, reliabilities were, on average, 0.49 (HR) and 0.63 (RND) for a heritability of 0.30; 0.20 (HR) and 0.27 (RND) for a heritability of 0.05; and 0.07 (HR) and 0.09 (RND) for a heritability of 0.01. Substantial decrease in the reliability was observed when the number of generations to the reference population increased [e.g., for heritability of 0.30, the decrease from evaluated set I (chosen from the same generation as the reference population) to II (one generation younger than the reference population) was 0.04 for HR, and 0.07 for RND. In this study, the importance of the design of a reference population consisting of cows was shown and optimal designs of the reference population for genomic prediction were suggested.

Key words: genomic selection, reference population design, reliability, direct genomic value

3.1 Introduction

Practical applications of genomic selection (GS; Meuwissen et al., 2001) are becoming more popular in animal (Hayes et al., 2009a) and plant breeding (Heffner et al., 2009; Jannink et al., 2010). Several countries use genomic information as one of the sources of information for selection in breeding programs. Genomic selection uses genome-wide SNP markers. Due to high marker density, all QTL are assumed to be in linkage disequilibrium (LD) with markers on the SNP chip (Meuwissen et al., 2001). This implies that a single SNP marker or a group of markers can be associated with QTL effects (Grapes et al., 2004, 2006; Yu et al., 2005). Based on those associations, genotypic information can be used as an additional source of information to increase the reliability of EBV, thereby increasing the accuracy of selection.

The accuracy of GS depends on 2 factors (Daetwyler et al., 2008; Goddard, 2009a): (1) the accuracy of estimated SNP effects and (2) the proportion of the genetic variance explained by the markers. The accuracy of estimated SNP effects is influenced by the size of the reference population; that is, the number of animals with genotypes and phenotypic records used to estimate SNP effects and the heritability of the considered trait. The proportion of genetic variance explained by the markers is influenced by the effective size of the considered population (N_e) and the density at which the SNP chip covers the genome. The effective population size influences the proportion of genetic variance explained by the markers. At low N_e, the number of independent segments present in the genome is expected to be lower (Goddard, 2009a). Fewer independent segments implies that fewer markers are needed to tag all segments and fewer records are needed to estimate effects of these segments (Goddard, 2010). Therefore, the accuracy of GS is expected to be higher in a population with smaller N_e than in a population with large N_e.

The reliability of direct genomic values (DGV) increases together with an increase of the reference population size, as the accuracy of estimating SNP effects increases. This was shown in theoretical studies (Goddard, 2009a; Meuwissen, 2009), simulation studies (Meuwissen et al., 2001; Pszczola et al., 2011), and real data analysis (Lund et al., 2010; for reviews, see Hayes et al., 2009a; Calus, 2010a). Furthermore, it was suggested that use of a reference population comprising animals with a wide range of phenotypes and genotypes would yield reliable predictions across the range of genotypes included in the reference population (Calus, 2010a). In addition, the family structure of the reference population may influence the reliability of GS, as shown by Pérez-Cabal et al. (2010).

The information content of phenotypes of animals in the reference population depends on their accuracy; for example, their correlation with the true breeding value of the animal. For example, in dairy cattle, bulls are usually included in the reference population using their daughter yield deviations (Gonzalez-Recio et al., 2008) or deregressed breeding values that are estimated with high reliability (Berry et al., 2009; Schenkel et al., 2009; VanRaden et al., 2009). This in turn requires numerous measurements on close relatives for those animals. Routinely recorded traits may easily meet this requirement; however, for traits difficult or expensive to measure, the number of observations may be limited and therefore daughter yield deviations (DYD) or deregressed EBV may not be available. Considering, for example, methane emission, where measuring a single observation is much more expensive than the genotyping costs per animal, it may be cost inefficient to phenotype many daughters per sire. To balance genotyping and phenotyping costs in such cases, we expect that genotyping cows that are phenotyped would be more efficient than genotyping only bulls and aggregating cow phenotypes at the bull level. In such a scenario, however, the reference population will consist of a limited number of animals with phenotypes that have lower heritability than DYD or deregressed EBV. When the reference population is small and the heritability of phenotypes is low, the reliability of predictions based on estimated SNP effects is expected to be low as well. The design of such a small reference population may be, therefore, important to maximize the reliability of predictions.

When considering an evaluated animal, a close relationship to the animals included in the reference population is expected to give a more reliable prediction (Habier et al., 2007; Legarra et al., 2008; Sonesson and Meuwissen, 2009). Optimally, all evaluated animals should have at least some closely related animals in the reference population. These observations suggest that the design of the reference population, in terms of relationships within and to the reference population, has to be considered. The design of the reference population may be especially important when the reference population consists of a limited number of individuals. With a small reference population, the reliability of DGV is expected to be low and any increase in the reliability is desired. To date, only a few results have been reported on the effect of the relationships within the reference population on the reliability of DGV for evaluated animals (Pérez-Cabal et al., 2010).

The objective of this study was to investigate the effect of various relationship levels within the reference population and the level of relationship of evaluated animals to the reference population on the reliability of DGV. The reliabilities of DGV, expressed as squared correlation between estimated and true breeding values, were calculated for the evaluated animals at 3 levels of heritability. To represent a dairy cattle breeding program aiming to initiate a reference population for a new trait, which is difficult or expensive to measure, animals included in the reference population were females and assumed to have only their own performance records. The size of the reference population was therefore assumed to be small.

3.2 Materials and methods

In this study, a dairy cattle population structure was simulated. Four reference populations with different family structures were selected. Three sets of evaluated animals with different numbers of generations to the chosen reference populations were sampled from simulated animals that were not included in 1 of the 4 reference populations. The evaluated animals were genotyped but had no phenotypes. The animals chosen for the reference populations and evaluated animals were all females. Reliabilities of DGV predictions for the evaluated animals were calculated and compared for 3 different heritability levels (0.30, 0.05, and 0.01).

Reliabilities of DGV calculation

Reliabilities of DGV predictions for all evaluated animals were calculated deterministically for 2 scenarios, without predicting the DGV themselves: (1) reliability based on pedigree relationships of the animals and their phenotypes (r_A^2); and (2) reliability based on the approach in which pedigree relationships of the animals were replaced by genomic relationships (r_G^2).

Formulas to calculate r_A^2 and r_G^2 can be derived from selection index theory (see Appendix) or from the prediction error variances of the mixed model equations used to estimate the breeding values. Values of r_A^2 were calculated for evaluated animals as

$$r_A^2 = \boldsymbol{a}\left[\boldsymbol{A} + \boldsymbol{I}\left(\frac{\sigma_e^2}{\sigma_A^2}\right)\right]^{-\mathbf{1}}\boldsymbol{a'}, \qquad [3.1]$$

where ***a*** is a vector with pedigree based relationships of an evaluated animal with the animals in the reference population; **A** is the additive relationship matrix for animals in the reference population; **I** is an identity matrix, σ_e^2 is the residual variance; and σ_A^2 is the genetic variance. The σ_e^2/σ_A^2 ratio reflects the heritability (h^2).

Calculation of r_G^2 was similar to that in [3.1] but used genomic relationship coefficients instead of pedigree-based relationships (VanRaden, 2008):

$$r_G^2 = \boldsymbol{c}\left[\boldsymbol{G} + \boldsymbol{I}\left(\frac{\sigma_e^2}{\sigma_A^2}\right)\right]^{-1}\boldsymbol{c}' \qquad [3.2]$$

where ***c*** is a vector with genomic relationships of an evaluated animal with the animals in the reference population; **G** is the genomic relationship matrix for animals in the reference population. **G** is constructed as $\frac{\boldsymbol{ZZ\prime}}{2\sum p_i(1-p_i)}$, following VanRaden (2008), where p_i is the frequency of the second allele at locus *i*, and ***Z*** is derived from genotypes of animals in the reference population, by subtracting 2 times the allele frequency expressed as a difference of 0.5, that is $2(p_i - 0.5)$, from matrix **M** that specifies the marker genotypes for each individual as −1, 0, or 1, and vector ***c*** is a column of the **C** matrix for a particular evaluated animal. The **C** matrix is created as $\frac{\boldsymbol{Z_2Z\prime}}{2\sum p_i(1-p_i)}$, where $\boldsymbol{Z_2}$ is constructed from genotypes of evaluated animals (VanRaden, 2008).

Simulation

The simulated genome was 3 M long and consisted of 3 chromosomes with a length of 1 M each. This corresponds, approximately, to 10% of the cattle genome (Ihara et al., 2004). In the first generation, monomorphic marker loci (300,000) were spaced at fixed distances of 0.001 cM across the genome.

De Roos et al. (2009) presented a simulation scheme mimicking different sizes of N_e at different stages in the historic cattle population, using inflated values for the frequency of mutation events and recombination rates. This simulation scheme was adopted in the present study. For the first phase, an historical population of 100 individuals was simulated and randomly mated for 600 generations; the frequency of mutation events was 10^{-4} per locus per generation and the recombination rate was 10^4 per M per generation. Subsequently, 200 generations were simulated, using a frequency of mutation events per locus per generation of 10^{-7} and a recombination rate per M per generation of 100. In the third phase, the number of simulated individuals was increased to 10,000 cows and 100 sires. To mimic the third phase described by de Roos et al. (2009), each sire was mated to 100 randomly chosen dams for the next 50 generations. The recombination rate was set to 1 per M per generation and mutations were stopped. In the last phase, the number of sires was reduced to 25 and the number of randomly chosen dams per sire increased to 400. This last phase was repeated for 15 generations to mimic the phase of modern cattle breed. Thereafter, another 5 generations were simulated in which the number of sires was increased to 50 and the number of offspring was, on average, 200 per sire. Pedigree was stored for the 10 last generations and genotypes for the 3 last generations. The simulation was performed according to the outline presented in Table 3.1 with parameters summarized in Table 3.2. Twenty replicates were simulated for each scenario and heritability combination.

Table 3.1 Simulation outline: number of generations and animals, mutation and recombination rates, mimicked number of generations, and simulated effective population size (N_e)[1].

No. of generations	No. of sires	No. of females	Mutation rate (per locus)	Recombination rate (per M/ gen)	Mimicked no. of generations	Simulated N_e
600	50	50	10^{-4}	10^{4}	6,000,000	~1,000,000
200	50	50	10^{-7}	100	20,000	~12,500
50	100	10,000	No mutation	1	50	~400
15	25	10,000	No mutation	1	15	~100
5	50	10,000	No mutation	1	5	~200

[1] Simulation outline adopted from de Roos *et al*. (2009)

Table 3.2 Simulation parameters.

Parameter	Value		
Simulated genome length (*M*)		3	
Number of simulated chromosomes		3	
Length of simulated chromosomes (*M*)		1	
Distance between adjacent markers in the last generation (*cM*)		~0.03	
Number of SNP markers per *cM*		~ 39.1	
Minor allele frequency		0.25	
	Trait 1	Trait 2	Trait 3
Heritability	0.30	0.05	0.01

Scenarios

Three sets of evaluated animals were chosen from the 3 consecutive generations of the genotyped animals. The evaluated set I and the reference populations were chosen from the same generation. The evaluated sets II and III were, respectively, 1 and 2 generations further away from the reference populations. Each set of animals consisted of 1,000 individuals originating from only one generation at the time. The evaluated sets consisted of animals sired by 50 sires. The 20 offspring per sire were chosen randomly from all available offspring of that sire.

Reference populations were chosen from all females that had not been chosen in the evaluated set. All the reference populations consisted of 2,000 cows. The chosen reference populations differed by their family structure. The first reference population (HR) consisted of highly related animals; the second (MR) consisted of moderately related animals; the third (LR) consisted of lowly related animals; and the fourth (RND) consisted of randomly selected animals. Differences in the level of the relationship within the reference populations were achieved by choosing paternal half-sib families of different sizes in each of the 4 scenarios. For HR, sires with at least 425 offspring were selected and from them offspring of 5 randomly chosen sires were included. For MR, offspring of 20 randomly chosen sires were included. For LR, offspring of 40 randomly chosen sires were included. For RND, 2,000 randomly chosen animals were included. In HR, MR, and LR, each sire contributed with equal number of offspring (400, 100, and 50, respectively) chosen randomly from all their available offspring. Reliabilities of DGV were calculated for each of the 3 sets of evaluated animals with use of the 4 different reference populations, using equations [3.1] and [3.2].

3.3 Results

Characteristics of the simulation

In each replicate, animals with genotypes were simulated. The average distance between adjacent segregating loci across the whole genome, calculated in the last generation, was approximately 0.03*cM* (Table 3.2). The average LD between adjacent markers, measured as r^2 (Hill and Robertson, 1968), was 0.23, and minor allele frequency averaged across all marker loci in the last generation was 0.25. The calculated allele frequencies were consistent across the reference populations. Among all the chosen reference populations, RND had the lowest average relationship.

The reference population, composed of progeny of only 5 sires (HR) on average, resulted in an average pedigree relationship of about 0.09. The average relationship in HR was higher than that in MR by about 0.04; differences in the average relationship levels among MR, LR, and RND were small (0.06 to 0.05) as shown in Table 3.3.

Reliabilities of Predictions

Use of Genomic Information

The increase in the reliabilities was clearly higher when the heritability was higher, as illustrated in Figures 3.1 and 3.2. The level of increase caused by using genomic information was similar across all reference populations (Figure 3.1). The increase in reliability caused by using genomic information, averaged across all the reference populations and evaluated sets, was 0.39, 0.13, and 0.03, respectively, for the highest, medium, and lowest heritabilities. Figure 3.3 shows that reliabilities increased when pedigree information was replaced by genomic information. The individual reliability values varied. For HR, 2 distinct groups of reliability values were found: high and low. The many more reliability values in the low group compared with the high group led to a lower mean value of reliability across individuals for HR compared with RND. In RND, no distinct groups of reliabilities were present and variance of reliabilities was much smaller.

Table 3.3 Average pedigree-based relationship within the reference population, across different scenarios averaged over all replicates.

Scenario[1]	Relationship	
	Mean	SD
HR	0.0946	0.0048
MR	0.0562	0.0020
LR	0.0497	0.0016
RND	0.0487	0.0016

[1] HR=the reference population with the highest average relationship within the reference population; MR=the reference population with moderate average relationship within the reference population; LR=the reference population with the lowest average relationship within the reference population; RND=the reference population consisted of randomly selected individuals.

Family structure of the reference population

The reliabilities increased when the average relationship within the reference population decreased (e.g., from 0.477 in HR to 0.597 in RND for heritability of 0.30) as illustrated in Figures 3.1 and 3.2 and Table 3.4. The family structure of the reference population, therefore, strongly influenced the average relationship of the reference population and therefore had an effect on reliabilities.

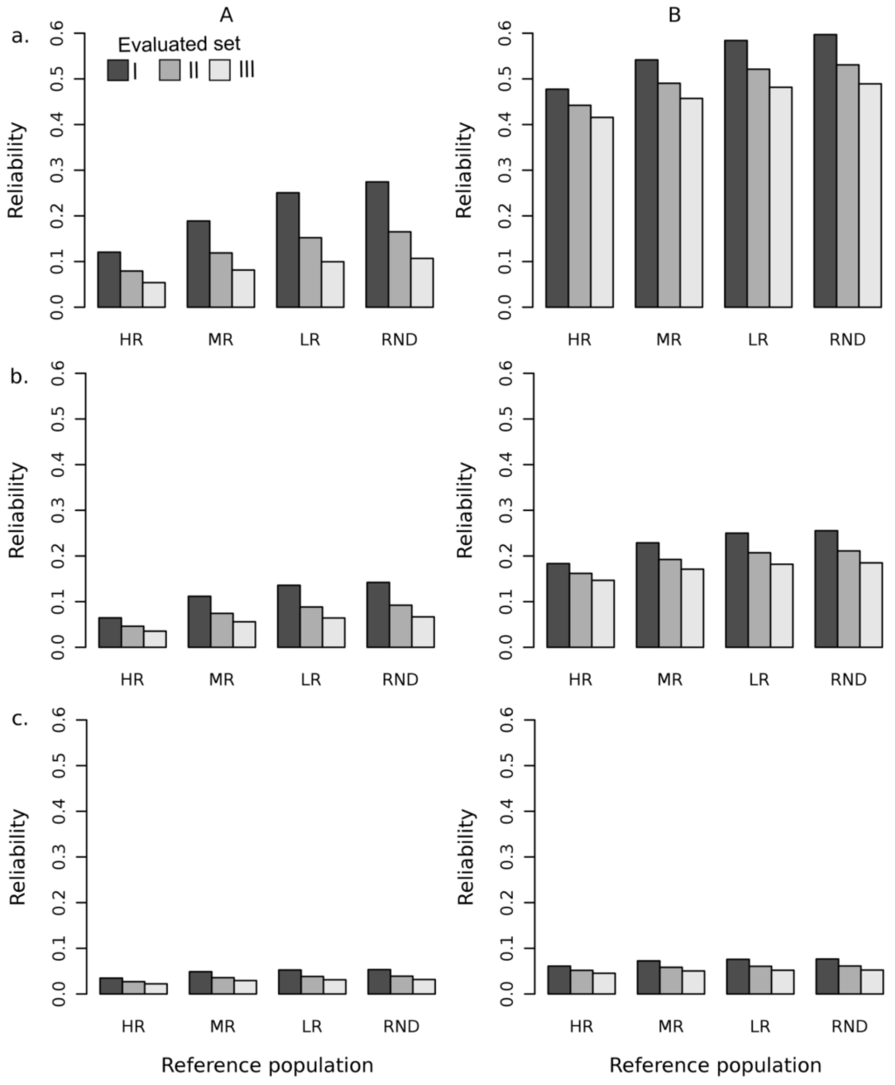


Figure 3.1 Reliability of breeding values calculated based on pedigree (A) and genomic (B) data across all simulated heritability levels (a: h^2=0.3; b: h^2=0.05; c: h^2=0.01), 4 reference populations with different family structure (highly related, HR; moderately related, MR; lowly related, LR; random selection of animals, RND), and 3 sets of evaluated animals with different distance from the reference population (I=the same generation as the animals in the reference population to III=2 generations after the reference population) and averaged over all replicates.

Table 3.4 Reliabilities obtained with use of different sources of information across all reference populations and evaluation sets for heritability of 0.30 averaged over all replicates[1].

	Evaluation set[3]					
	I		II		III	
Scenario[2]	Mean	SD	Mean	SD	Mean	SD
Pedigree data						
HR	0.121	0.001	0.079	0.004	0.054	0.002
MR	0.189	0.002	0.119	0.005	0.081	0.003
LR	0.251	0.002	0.152	0.004	0.100	0.003
RND	0.274	0.001	0.165	0.002	0.107	0.002
Genomic data						
HR	0.477	0.005	0.442	0.007	0.416	0.006
MR	0.542	0.004	0.490	0.007	0.457	0.006
LR	0.584	0.004	0.521	0.005	0.482	0.006
RND	0.597	0.004	0.531	0.004	0.489	0.005

[1] Standard deviations of 20 replicates ranged from 0.003 to 0.012;

[2] HR=the reference population with the highest average relationship within the reference population; MR=the reference population with moderate average relationship within the reference population; LR=the reference population with the lowest average relationship within the reference population; RND=the reference population consisted of randomly selected individuals;

[3] Set of evaluation animals chosen from the same generation as the reference animals (I), from offspring of the animals from the same generation as the reference population (II), and from animals 2 generations after the reference population (III).

Relationship to the reference population

Three different measures of the relationship level of the evaluated animals to the reference population were calculated: average relationship, average squared relationship, and the maximum relationship. To determine which of these measures was most closely related to the reliability, the individual reliabilities of the animals included in the evaluated set I were compared with each of those measures, as shown in Figure 3.4. The reliabilities calculated with pedigree and genomic information were the most closely related to average squared relationship. The reliabilities showed a poorer relationship to the 2 other measures. The average squared relationship, therefore, is used hereafter as a measure of the relationship to the reference population.

Figure 3.2 shows comparison of the reliabilities for HR and RND at different levels of the average squared relationship to the reference population. It can be seen that when the half-sib families in the reference population were smaller (RND), the reliabilities at the given level of the average squared relationship to the reference population were higher. Reliabilities were regressed on the average squared relationships to the reference population (Figure 3.2). Regression coefficients increased with heritability level and were considerably higher for RND than for HR. This dependency was consistent over all the replicates and was observed across all the heritability levels, as well as for all evaluated sets of animals (results not shown). Animals with a higher average squared relationship to the reference population (based on genomic information) also had higher values of the reliability. This trend was observable across all reference populations and heritability levels. For example, at the average squared relationship of 0.005, the reliability for the heritability of 0.30 was, on average, 0.49 for HR and 0.63 for RND. For the lower heritability levels, the differences between HR and RND were smaller: 0.20 for HR and 0.27 for RND at a heritability of 0.05, and 0.07 for HR and 0.09 for RND at a heritability of 0.01. Some variation in the reliability level at certain levels of the average squared relationship was found. This variation was smaller for the reference population with smaller families, as shown in Figure 3.2.

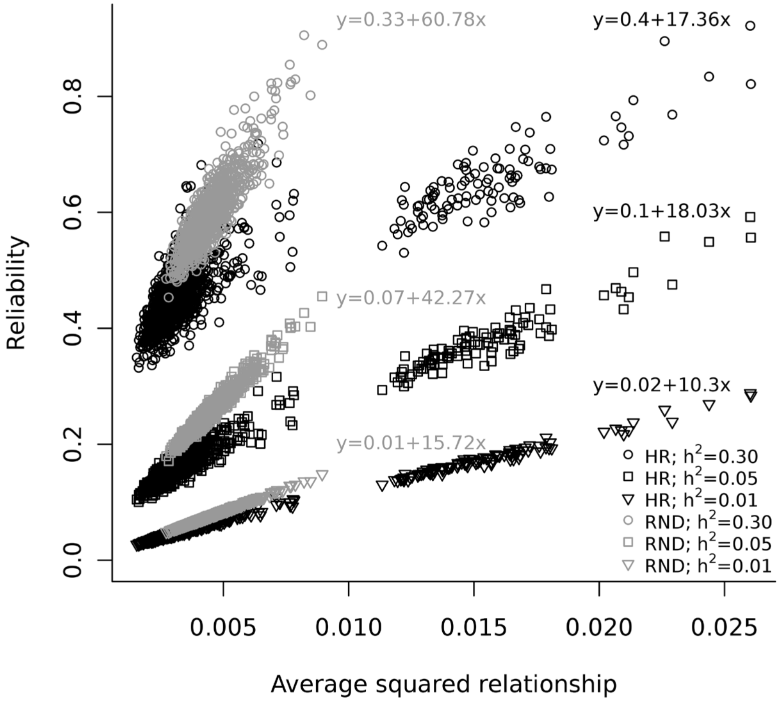


Figure 3.2 Average squared relationship (based on genomic information) to the reference population versus the reliabilities of genomic breeding values for different heritability levels and the reference populations with high (HR) and low (RND) family structure based on randomly chosen replicate.

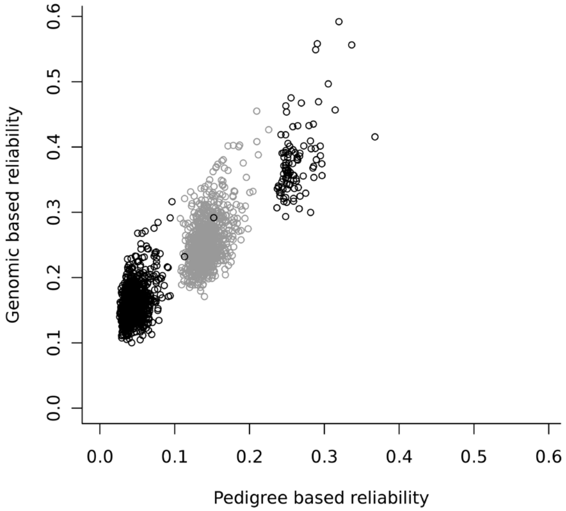


Figure 3.3 Genomic-based reliability as a function of pedigree-based reliability for high (HR; black) and low (RND; gray) family structure based on a randomly chosen replicate.

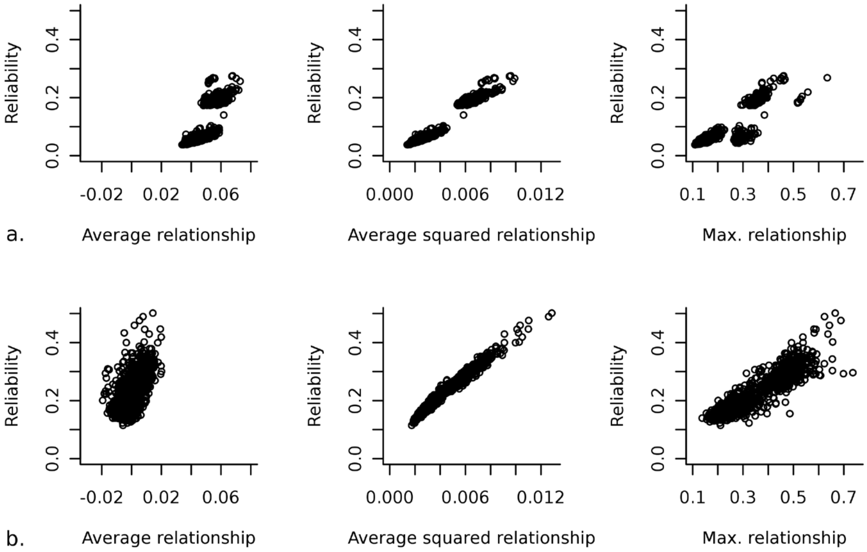


Figure 3.4 An illustration of the relationship between 3 different measures of relationship of the evaluated animals to the reliability based on pedigree (a) and genomic (b) data for the reference population with moderate family structure (MR) and heritability of 0.05, and for the set of evaluated animals originating from the same generation as the reference population (set I).

Increasing the number of generations to the chosen reference populations resulted in a decrease in reliability (Figure 3.1). This decrease in the reliability was substantial when the predictions were based on pedigree relationships. The individual reliabilities obtained for the evaluated animals from different generations and their levels of the average squared relationship to the reference population were compared (results not shown). This comparison revealed that the decrease in the reliability was mostly explained by a reduction in the average squared relationship to the reference population. The decrease tended to be larger when the average relationship within the reference population was smaller. For example, for a heritability of 0.30, the decrease in reliability due to increased number of generations in the reference population was 0.042 for HR and 0.109 for RND (from the evaluated set I to II), and 0.067 for HR and 0.167 for RND (from the evaluated set I to III; see Figure 3.1A and Table 3.4). For the remaining heritability levels, this decrease in reliability was somewhat lower, but still substantial (Figure 3.1). The decrease in reliability caused by the increased number of generations between the reference population and the evaluated animals was smaller with use of genomic data than pedigree data. For example, for heritability of 0.30, the decrease in reliability caused by an increased number of generations in the reference population was 0.035 for HR, 0.066 for RND from the evaluated set I to II and 0.061 for HR and 0.108 for RND from the evaluated set I to III (Figure 3.1B and Table 3.4). The level of reliability, moreover, decreased more slowly over the generations when genomic data were used (Figure 3.1). When using genomic data, for a heritability of 0.30, the decrease in reliability tended to be higher when the average relationship within the reference population decreased (Table 3.4). Although the decrease was larger for RND, the ranking of the scenarios remained the same.

3.4 Discussion

This study investigated the effect of relationship of evaluated animals with the reference population and various levels of relationship within the reference population on the reliability of DGV. A small reference population was assumed, consisting of animals with own performance only, to reflect a situation in which the reference population was initiated for scarcely recorded new traits. An important question is whether our results also apply to large reference populations for traits with abundant phenotypic information available. Additional analyses indicated similar trends when the reference population was twice as large (results not shown). This indicates that optimization of the reference population design, as discussed here, is indeed also applicable to larger reference populations.

Characteristics of the simulation

Our initial application of the simulation scheme of de Roos et al. (2009) resulted in a different average LD level than in the study of de Roos et al. (2009) or results from real data sets (obtained with use of the 50K SNP cattle chip). To obtain LD levels as observed in cattle populations (de Roos et al., 2008), the number of mutation events and recombination rates were increased in the first 2 phases of the simulation compared with de Roos et al. (2009). A possible reason for the initial discrepancy with results of de Roos et al. (2009) is that simulated markers were biallelic in our whole simulation process. In the study of de Roos et al. (2009), however, simulated markers were multiallelic and they were transformed to biallelic at the last stage of the simulation. The genome simulated in this study was smaller than the real genome. This smaller genome size may lead to larger sampling variance for the estimated coefficients in **G**. Differences between **A** and **G** in our study, therefore, might be somewhat different from observed in real data.

Average relationship in the reference population

The average relationship within the reference population, calculated based on at least 5 generation complete pedigrees and averaged over all replicates for MR (0.0562), LR (0.0497), and RND (0.0487) corresponded to values based on real data (Kearney et al., 2004; König and Simianer, 2006; Mrode et al., 2009). The average relationship within the reference population for HR (0.0946) was higher than the findings of the other studies because of the strong family structure of the population (i.e., all included animals were sired by only 5 individuals; Table 3.3). The HR population was somewhat similar to a daughter design with large paternal half-sib families that has typically been used in QTL mapping experiments, set up for linkage mapping of QTL (i.e., Weller et al., 1990). The HR gave the lowest DGV reliabilities, which implies that although such experimental designs may be optimal for linkage mapping, they are suboptimal for use in genomic prediction as shown in this study or in genome-wide association studies (Balding, 2006). The RND was a random sample from the simulated population, whereas the other 3 reference populations were constructed by selecting paternal half-sib families of different sizes. Therefore, the family structure of RND was weaker than that in the other 3 reference populations. This weak family structure resulted in a lower average relationship for RND. For example, in RND animals were offspring of 50 sires, whereas in LR animals were offspring of 40 sires.

Reliabilities of predictions

The reliabilities in this study were obtained in a similar way to those that can be obtained from the left-hand side of the mixed models equations. For DGV reliabilities obtained in such a way, one of the assumptions is that the markers explain all the genetic variation, whereas some part of the genetic variance is also explained by loci between markers (VanRaden, 2008). This may lead to overestimation of the reliability. Calus et al. (2009), Hayes et al. (2009b), Lund et al. (2009), and Su et al. (2010) also showed that reliabilities obtained from cross-validation are somewhat lower compared with reliabilities calculated from the prediction error variances of the mixed model equations. Therefore, in general, we can expect that reliabilities obtained in our study are slightly overestimated. An important question is whether the overestimation of DGV reliabilities is comparable across the different scenarios and reference populations. Reliabilities for the breeding values based on pedigree information were expected to be unbiased. Those gave similar patterns across the reference populations (not shown), despite the level being generally lower. Because replacing the pedigree by genomic data to some extent only adds more information, it is expected that the overestimation of the DGV reliability is mostly a scaling issue.

Table 3.5 Comparison of reliabilities obtained with deterministic formulas of Daetwyler et al. (2008) and Goddard (2009a) for all simulated levels of heritability and effective population size N_e of 100 with the reliabilities obtained in the present study averaged across all the individuals.

	Reliability		
Heritability	Present study	Daetwyler et al. (2008)	Goddard (2009a)
0.30	0.55	0.75	0.59
0.05	0.23	0.34	0.22
0.01	0.07	0.09	0.06

We also predicted reliabilities deterministically using formulas by Daetwyler et al. (2008) and Goddard (2009a) for N_e of 100 (Table 3.5). The reliabilities obtained as in Goddard (2009a) were shown to agree with the reliabilities for genomic breeding values in US and Australian Holstein-Friesians and Jerseys (Hayes et al., 2009c). These reliabilities were also in good agreement with the results obtained in the present study. Reliabilities calculated as in Daetwyler et al. (2008) were somewhat higher than those of the current study. Daetwyler et al. (2008) and Goddard (2009a) assumed unrelated animals, and therefore, animals with different levels of relationship to the reference population were examined. Examination of animals with the lowest and the highest relationship to the reference population showed that reliabilities predicted for these animals differed from the reliabilities given by the deterministic formulas. Predictions obtained in the present study also allow for assessment of the individual reliability of a particular individual accounting for the actual relationships.

Family structure of the reference population

Comparison between the genomic- and pedigree-based reliabilities presented in Figure 3.3 shows that 2 groups of evaluated animals can be distinguished for HR. One group, strongly related with the reference population, yielded the highest reliabilities. The other group, loosely related to the reference population, had the lowest reliabilities. Reliabilities for RND was more homogeneous and at an intermediate level. The highest average reliabilities in this study were achieved for the randomly composed reference population (RND). Pérez-Cabal et al. (2010) also showed that the highest reliability was obtained when animals were chosen randomly. In their study, unlike our results, RND appeared to have higher average relationship within the reference population than nonrandomly chosen animals. The differences in the average relationship in Pérez-Cabal et al. (2010), however, were small and might be due to sampling; for example, depending on the size of the population from which the reference populations were chosen.

Across different reference populations, and at the same level of the average squared relationship to the reference population, evaluated animals yielded different reliabilities, as shown in Figure 3.2. The RND population achieved the highest average reliability. The differences, despite the same average squared relationship, were because RND consisted of many small half-sib families and therefore had a low average relationship within the reference population. Related animals may partly explain the same part of variation; therefore, the theoretical maximum reliability can be achieved when all the individuals in the reference population are unrelated and their alleles are not identical-by-state. For such a case, although unrealistic, **A** and **G** in equations [3.1] and [3.2] would be diagonal and the reliability for an evaluated animal would be proportional to its sum of squared relationships to the animals in the reference population, at least when all animals are not inbred. For the reliabilities based on genomic information, according to the suggestion of Calus (2010a), animals could be chosen to represent the widest range of possible genotypes to further increase the reliability, which means minimizing the genomic relationships between animals in the reference population, such as with a randomly chosen reference population.

Relationship to the reference population

The individual reliability strongly depended on the average squared relationship to the reference population (Figure 3.2). This is in accordance with findings of Habier et al. (2010). Variation in the reliability observed at the same level of the average squared relationship could arise, for example, when the animals were similarly related to the reference population but their relatives in the reference population were differentially related to each other.

Larger numbers of generations to the reference population resulted in a decrease in the reliability, as in Habier et al. (2010). The level of genomic reliability decreased more slowly over generations than reliabilities based on pedigree data, as also reported by Wolc et al. (2011). To prevent a decrease of reliability, constant updates of the reference population with animals from more recent generations are required (e.g., animals selected from the evaluation set when their phenotypes become available; Habier et al., 2010).

Implications

This study showed that an optimally designed reference population should consist of loosely related animals. Still, evaluated animals with low relationship to the reference population had low reliability. In practical applications, for evaluated animals the reliability may be first predicted using pedigree alone. Based on the outcome, a rough estimate of the genomic reliability could be made, and a decision made as to whether the genomic prediction for this animal is reliable enough to justify the genotyping costs. When this is not the case, yet the animal itself is closely related to many other animals that may be evaluated, the animal could be considered for inclusion in the reference population. These aspects require further investigation.

The optimal design of the reference population may differ from one application to another and depend on the desired breeding strategy of the breed. For example, if an experiment intends to measure methane emission, the number of observations is limited because of high phenotyping costs. In such a case, if the goal is to predict breeding values for selection candidates only, then the reference population should consist of animals closely related to these selection candidates. An optimal solution could then be to build the reference population using progeny (i.e., daughters) of key breeding animals (i.e., sires), including only few daughters per sire. If the goal, however, is to predict breeding values (or phenotypes) for all the animals, then a reference population consisting of unrelated animals would be the most desired. Composing the most suitable reference population is therefore an optimization problem because a tradeoff exists between obtaining a low average relationship between the animals in the reference population and a high average squared relationship to the animals in the population to be evaluated.

3.5 Conclusion

This study shows the importance of optimizing the design of the reference population consisting of cows. First, the optimally designed reference population for use in genomic prediction or genome-wide association studies should have a loose family structure; that is, the average relationship within the animals included into the reference population should be low. This implies that different designs are required for reference populations than the traditional designs used for linkage mapping purposes. Second, the relationship between the reference population and the evaluated animal should be maximized. The average squared relationship appeared to be the best measure of the relationship to the reference population with respect to the reliability. Higher levels of heritability resulted in higher levels of reliabilities. To maximize reliability, relationships among animals in the reference population should be minimized and the relationships of the validation animals with the reference population should be maximized.

3.6 Acknowledgments

The authors appreciate useful comments of Johan van Arendonk (Animal Breeding and Genomics Centre, Wageningen University, Wageningen, the Netherlands) and Roel Veerkamp (Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, Lelystad, the Netherlands). Marcin Pszczola gratefully acknowledges the financial support of the Koepon Stichting (Arnhem, the Netherlands), GreenHouseMilk, and Faculty of Animal Breeding and Biology (Poznan University of Life Sciences, Poznan, Poland). The GreenHouseMilk project is financially supported by the European Commission under the Seventh Research Framework Programme, Grant Agreement KBBE-238562. This publication represents the views of the authors, not the European Commission, and the Commission is not liable for any use that may be made of the information.

3.7 Appendix

Derivation of Reliability Prediction Equations

The equations used to predict reliabilities can be derived from selection index theory. Let **P** be the (co)variance matrix between the information sources X_1 to X_n and ***g*** be the vector containing the covariances between the information sources and their true breeding values (**A***).* **P** contains the single phenotypic observations of genotyped animals in the reference population. Then, ***P*** can be formulated as follows:

$$\boldsymbol{P} = \begin{bmatrix} var(X_1) & \dots & cov(X_1, X_n) \\ \dots & \dots & \dots \\ cov(X_n, X_1) & \dots & var(X_n) \end{bmatrix} = \begin{bmatrix} \sigma_P^2 & \dots & a\sigma_A^2 \\ \dots & \dots & \dots \\ a\sigma_A^2 & \dots & \sigma_P^2 \end{bmatrix},$$

where σ_A^2 is the additive genetic variance, *a* is the coefficient indicating the proportion of genetic variance shared between the information sources, and σ_P^2 is the phenotypic variance. The vector ***g*** can be written as:

$$\boldsymbol{g} = \begin{bmatrix} cov(A, X_1) \\ \dots \\ cov(A, X_n) \end{bmatrix} = \begin{bmatrix} a\sigma_A^2 \\ \dots \\ a\sigma_A^2 \end{bmatrix}.$$

Then, the reliability of breeding values estimated with use of selection index theory can be calculated as $r_{IH}^2 = \frac{\boldsymbol{b}'\boldsymbol{g}}{a_A^2}$, where ***b*** is a vector with optimal weights for the information sources calculated as $\boldsymbol{b} = \boldsymbol{P}^{-1}\boldsymbol{g}$. Therefore, the equation for the reliability can be rewritten as:

$$r_{IH}^2 = \frac{\boldsymbol{g}'\boldsymbol{P}^{-1}\boldsymbol{g}}{a_A^2}.$$

Let ***A*** be the additive relationship matrix formulated as:

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \dots & \dots & \dots \\ a_{n1} & \dots & a_{nn} \end{bmatrix},$$

and a_A^2 equals 1, then ***P*** can be rewritten as $\left[\boldsymbol{A} + \boldsymbol{I}\left(\frac{\sigma_e^2}{\sigma_A^2}\right)\right]$, where variance ratio $\left(\frac{\sigma_e^2}{\sigma_A^2}\right)$ reflects the heritability, because $\frac{\sigma_e^2}{\sigma_A^2} = \frac{1}{h^2} - 1$, and **A** is created for the information sources (i.e., reference population); then, ***g*** will equal the additive relationship of the evaluated animal to the reference population **a**. Therefore, reliability of traditional selection can be calculated as $r_A^2 = \boldsymbol{a}\left[\boldsymbol{A} + \boldsymbol{I}\left(\frac{\sigma_e^2}{\sigma_A^2}\right)\right]^{-1}\boldsymbol{a}'$. For genomic selection, pedigree-based coefficients can be substituted by the genomic coefficients as $r_G^2 = \boldsymbol{c}\left[\boldsymbol{G} + \boldsymbol{I}\left(\frac{\sigma_e^2}{\sigma_A^2}\right)\right]^{-1}\boldsymbol{c}'$, where **G** is genomic relationship matrix for the animals in the reference population and c is the genomic-based relationship of the evaluated animal to the reference population.

4

The impact of genotyping different groups of animals on accuracy when moving from traditional to genomic selection

M. Pszczola[1,2,3], T. Strabel[3], J. A. M. van Arendonk[2] and M. P. L. Calus[1]

[1] Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, 8200 AB Lelystad, The Netherlands; [2] Animal Breeding and Genomics Centre, Wageningen University, 6700 AH Wageningen, The Netherlands; [3] Department of Genetics and Animal Breeding, Poznan University of Life Sciences, Wolynska 33, 60-637 Poznan, Poland

Journal of Dairy Science (2012) 95:5412-5421

Abstract

Compared with traditional selection, the use of genomic information tends to increase the accuracy of estimated breeding values (EBV). The cause of this increase is, however, unknown. To explore this phenomenon, this study investigated whether the increase in accuracy when moving from traditional (AA) to genomic selection (GG) was mainly due to genotyping the reference population (GA) or the evaluated animals (AG). In it, a combined relationship matrix for simultaneous use of genotyped and ungenotyped animals was applied. A simulated data set reflected the dairy cattle population. Four differently designed (i.e., different average relationships within the reference population) small reference populations and 3 heritability levels were considered. The animals in the reference populations had high, moderate, low, and random (RND) relationships. The evaluated animals were juveniles. The small reference populations simulated difficult or expensive to measure traits (i.e., methane emission). The accuracy of selection was expressed as the reliability of (genomic) EBV and was predicted based on selection index theory using relationships. Connectedness between the reference populations and evaluated animals was calculated using the prediction error variance. Average (genomic) EBV reliabilities increased with heritability and with a decrease in the average relationship within the reference population. Reliabilities in AA and AG were lower than those in GG and were higher than those in GA (respectively, 0.039, 0.042, 0.052, and 0.048 for RND and a heritability of 0.01). Differences between AA and GA were small. Average connectedness with all animals in the reference population for all scenarios and reference populations ranged from 0.003 to 0.024; it was lowest when the animals were not genotyped (AA; e.g., 0.004 for RND) and highest when all the animals were genotyped (GG; e.g., 0.024 for RND). Differences present across designs of the reference populations were very small. Genomic relationships among animals in the reference population might be less important than those for the evaluated animals with no phenotypic observations. Thus, the main origin of the gain in accuracy when using genomic selection is due to genotyping the evaluated animals. However, genotyping only one group of animals will always yield less accurate estimates.

Key words: genomic selection; reference population design; breeding value reliability; connectedness

4.1 Introduction

Genomic selection (GS) uses dense SNP marker arrays. These markers are assumed to be in linkage disequilibrium (LD) with QTL (Meuwissen et al., 2001), allowing their effects to be estimated. To estimate SNP marker effects, GS requires a set of genotyped and phenotyped animals, a so-called reference population used to evaluate genotyped animals without phenotypic information.

Another approach to evaluating animals without explicitly estimating SNP marker effects is to use SNP markers to estimate relationships between animals. Genomic relationships can capture Mendelian sampling and reveal links between animals that are seemingly unrelated through pedigree. Thus, genomic relationships are more precise, improving the connectedness between animals in a reference population and the evaluated animals. This higher connectedness reduces bias, and thus improves the genetic evaluation (Kennedy, 1981).

Based on such genomic relationships, a genomic relationship matrix (**G**) can be created using various methods (e.g., Nejati-Javaremi et al., 1997; VanRaden, 2008; Yang et al., 2010). This genomic relationship matrix can be used in the genomic BLUP procedure (G-BLUP), where **G** replaces the additive relationship matrix (**A**). Alternatively, **G** for genotyped animals can be merged with **A** for all animals, to enable simultaneous use of phenotypic information of genotyped and ungenotyped animals in genetic evaluation (e.g., Legarra et al., 2009; Misztal et al., 2009b; Aguilar et al., 2010; Christensen and Lund, 2010). This method, as shown by Christensen and Lund (2010) using the gene content prediction method (Gengler et al., 2007), provides an approximation to unseen genotypes and generates genomic EBV (GEBV) in a single step.

An important aspect of genetic improvement is the response to selection, which depends linearly on accuracy of selection. The accuracy of traditional selection depends on the availability of phenotypic information on relatives as well as the animal's own performance and on heritability of the considered trait. The accuracy of GS depends on heritability, but also on several other factors (Daetwyler et al., 2008; Goddard, 2009a). First, the size of the reference population is relevant; this is equivalent to the availability of phenotypic information on relatives and the animal's own performance in traditional selection. The larger the reference population, the higher the accuracy of predicted breeding values. Second, the lower the effective population size (N_e), the fewer independent segments in the genome, reducing the number of markers needed to tag all segments and the fewer records are needed to accurately estimate the effects of all independent segments (Goddard, 2009a). Third, the effective number of loci, dependent on the mating structure and the recombination length of the genome, affects the number of markers required to tag all potential QTL (Goddard, 2009a). Finally, the accuracy of GEBV is affected by both the relationship between the evaluated animals and the reference population (Habier et al., 2010; Wolc et al., 2011; Pszczola et al., 2012a) and its design (the relationships within the reference population; Pszczola et al., 2012a).

Choosing an appropriate design for the reference population, as suggested by Meuwissen (2009) and Calus (2010a) and shown by Pszczola et al. (2012a), may be a way to optimize current techniques. Such optimization is especially required when the number of animals in the reference population is limited. This can be a consequence of, for example, a small population size, a limited number of phenotypes available from research herds or (historic) experiments, a lack of routinely taken measurements for the considered trait, genotyping costs, or measuring a difficult or expensive trait. For instance, in the case of methane emission, measuring a single observation is very expensive because it requires sophisticated equipment.

A comparison of GS with traditional selection tools showed, especially in the case of unphenotyped juveniles, an increase in accuracy when GS was used (Meuwissen et al., 2001; Schaeffer, 2006; Pszczola et al., 2011). The origin of this increase in accuracy is unclear, given that it is unknown whether this increase is mainly due to substituting pedigree with genomic information for the reference population or for the evaluated animals. This is also an important question from a practical point of view because genomic data are sometimes unavailable for animals with valuable phenotypes. Including these animals in the reference population is possible by using approaches that combine genomic with pedigree information.

The aim of this study was to investigate whether the increase in accuracy, when moving from traditional selection to GS, is mainly due to genotyping the reference population or the evaluated animals. More specifically, it asks whether the accuracy of GEBV increases 1) when the reference population is ungenotyped while the evaluated animals are genotyped, and 2) when the evaluated animals are ungenotyped and the reference population is genotyped, and the combined relationship matrix is used. We evaluated the amount of gain by comparing, across scenarios at different heritability levels, predicted reliabilities and connectedness levels.

4.2 Materials and methods

Data

The simulated data set used in this study, reflecting a dairy cattle population, has been described in detail by Pszczola et al. (2012a). Briefly, the simulation scheme adopted from de Roos et al. (2009) mimics different sizes of N_e at different stages in the historic cattle population by using inflated values of recombination rates and frequency of mutation events. The use of inflated parameters heavily reduces the number of generations simulated at each of the stages of the historic cattle population and yields realistic LD levels (de Roos et al., 2008). To achieve LD levels for modern cattle breeds while simulating biallelic loci, this simulation scheme was further modified as described by Pszczola et al. (2012a).

The data set retained for analysis comprised pedigrees of 10 generations of randomly mated animals. Each of the first 5 generations included 25 sires mated randomly with 400 dams, and each of the last 5 generations included 50 sires mated randomly with 200 dams. The last 2 generations were composed of animals with genotypic data. Data were available from 20 replications of the simulation process.

The simulated genome was 3 M long and consisted of 3 chromosomes of 1 M each. In the last generation, the average distance between segregating loci across the whole genome was approximately 0.03 cM, the average minor allele frequency was 0.25, and the average LD, measured as r^2 (Hill and Robertson, 1968) between adjacent loci, was 0.23.

Reference populations and evaluated animals

Simulated animals were partitioned into an evaluation set and reference populations. The evaluated animals (n = 1,000) were chosen from the second genotyped generation (i.e., 20 randomly chosen progeny each of 50 randomly chosen sires). The 4 reference populations were chosen from the animals in the first genotyped generation.

Because each reference population, consisting of 2,000 cows, had a different family structure with respect to the sizes of parental half-sib families, their average relationship varied. The average relationship of the first reference population (high relationship), which consisted of progeny of 5 sires, was 0.095. The average relationship for the second reference population (medium relationship), consisting of progeny of 20 sires, was 0.056. For the third reference population (low relationship), consisting of progeny of 40 randomly chosen sires, the average relationship was 0.050. The average relationship for the fourth reference population (RND), consisting of randomly selected animals, was the lowest: 0.049.

Calculation of (G)EBV reliabilities

Reliabilities were calculated deterministically for all the evaluated animals for 4 situations: first, no animals were genotyped and the pedigree-based relationships were used (AA); second, animals in the reference population were genotyped and the evaluated animals were not (GA); third, the reference population was not genotyped whereas the evaluated animals were (AG); and fourth, all the animals were genotyped (GG).

The matrices were created as follows. Relationship matrix **H** combined pedigree and genomic data for GA and AG, following the method of Legarra et al. (2009):

$$\begin{bmatrix} \boldsymbol{A_{11}} + \boldsymbol{A_{12}A_{22}^{-1}(G - A_{22})A_{22}^{-1}A_{21}} & \boldsymbol{A_{12}A_{22}^{-1}G} \\ \boldsymbol{GA_{22}^{-1}A_{21}} & \boldsymbol{G} \end{bmatrix}, \qquad [4.1]$$

where the pedigree relationship matrix **A** was partitioned for ungenotyped (1) and genotyped (2) animals. The genomic relationship matrix, **G**, was created as $\frac{\boldsymbol{ZZ'}}{2\sum p_i(1-p_i)}$, following the method of VanRaden (2008), where p_i is the frequency of the second allele at locus *i* for which the homozygote genotype is coded as 1, and ***Z*** is derived from the genotypes of the animals in the reference population by subtracting 2 times the allele frequency expressed as a difference of 0.5 [i.e., $2(p_i - 0.5)$] from matrix **M**, which specifies the marker genotypes for each individual as −1, 0, or 1. Because for building **H**, **A** and **G** must be compatible, the **G** matrix had to be adjusted. First, the inbreeding level in **G** was scaled to the inbreeding level in **A**. This was done by calculating the pedigree inbreeding coefficients averaged across all animals $(\overline{f_p})$. Because the current population was used to calculate p_i, it was assumed that an average genomic inbreeding coefficient in **G** is zero. **G*** was then calculated following the formula derived from Wright's F-statistics, as in Powell et al. (2010):

$$\boldsymbol{G}^* = \boldsymbol{G}(1 - \overline{f_p}) + 2\overline{f_p}\boldsymbol{J},$$

where **G*** contains the relationships relative to the same base used in **A**, and **J** is a matrix of all ones. This adjustment is equivalent to what was proposed by Vitezica et al. (2011).

To create $\widehat{\mathbf{G}}$,the second step, following Yang et al. (2010), accounts for the fact that $\mathbf{G}^*$ is estimated with some error:

$$\widehat{\boldsymbol{G}} = \boldsymbol{G}^* + \boldsymbol{E} = \boldsymbol{A} + (\boldsymbol{G}^* + \boldsymbol{A}) + \boldsymbol{E},$$

where **E** is the matrix containing estimation errors for $\widehat{\mathbf{G}}$ assuming the following variances for those matrices $\mathrm{V}(\widehat{\mathbf{G}} - \mathbf{A})$=$\mathrm{V}(\mathbf{G}^* - \mathbf{A}) + \mathrm{V}(\mathbf{E})$. $\mathrm{V}(\mathbf{E})$ equals 1/N, where N is the SNP number used. To account for the sampling variance in $\widehat{\mathbf{G}}$, $\widehat{\mathbf{G}} - \mathbf{A}$ is regressed back toward **A**, and $\widetilde{\mathbf{G}}$ is

$$\widetilde{\boldsymbol{G}} = \boldsymbol{A} + \frac{(\widehat{\boldsymbol{G}}-\boldsymbol{A})\times\mathrm{V}(\boldsymbol{G}^*-\boldsymbol{A})}{[\mathrm{V}(\boldsymbol{G}^*-\boldsymbol{A})+V(\boldsymbol{E})]} = \boldsymbol{A} + \frac{(\widehat{\boldsymbol{G}}-\boldsymbol{A})\times\left[V(\widehat{\boldsymbol{G}}-\boldsymbol{A})-\frac{1}{N}\right]}{V(\widehat{\boldsymbol{G}}-\boldsymbol{A})}.$$

Because the sampling error in relationships in $\widehat{\mathbf{G}}$ depends on the value of the relationship, values for $\mathrm{V}(\widehat{\mathbf{G}} - \mathbf{A})$ were calculated separately for bins of relationships in **A**. The bins of relationships were 0 to 0.10, >0.10 to 0.25, >0.25 to 0.50, and >0.50. The last bin excluded parent-offspring pairs because their relationship is expected to be 0.5. The diagonal elements of $\widehat{\mathbf{G}}$ (i.e., self-relationships) were not regressed. The regression coefficients averaged across simulations for the 4 bins were, respectively, 0.963, 0.983, 0.991, and 0.989, which are similar to those reported in a different study based on real data (Veerkamp et al., 2011).

Deterministic predictions of GEBV reliabilities were based on formulas that can be derived either from selection index theory [for a derivation see the Appendix in Pszczola et al. (2012a)] or from the prediction error variances of the mixed model equations used to estimate the breeding values, as follows.

For AA, the reliability r_{AA}^2 was

$$\boldsymbol{a}\left[\boldsymbol{A} + \boldsymbol{I}\left(\frac{\sigma_e^2}{\sigma_a^2}\right)\right]^{-1}\boldsymbol{a}', \qquad [4.2]$$

where **a** is a vector with pedigree-based relationships of an evaluated animal with the animals in the reference population; **A** is the additive relationship matrix for animals in the reference population; **I** is an identity matrix, σ_e^2 is the residual variance; and σ_a^2 is the genetic variance. Their ratio reflects heritability (h^2).

For GA and AG, r_{GA}^2 and r_{AG}^2 were calculated as

$$\boldsymbol{h}\left[\boldsymbol{H}_{ref} + \boldsymbol{I}\left(\frac{\sigma_e^2}{\sigma_a^2}\right)\right]^{-1}\boldsymbol{h}', \qquad [4.3]$$

where **h** is a vector based on part of **H** that contains relationships of an evaluated animal with the animals in the reference population ($\mathbf{H_{ref\_eva}}$); $\mathbf{H_{ref}}$ is the combined pedigree and genomic relationship matrix for animals in the reference population. For situation GA, $\mathbf{H_{ref\_eva}} = \mathbf{A_{ref\_eva}A_{ref}^{-1}G_{ref}}$ and $\mathbf{H_{ref}} = \mathbf{G_{ref}}$ where $\mathbf{A_{ref\_eva}}$ is the part of **A** describing relationships between the reference and evaluated animals, and $\mathbf{A_{ref}}$ and $\mathbf{G_{ref}}$ are created for the genotyped reference population.

For AG, $\mathbf{H_{ref\_eva}} = \mathbf{A_{ref\_eva}A_{eva}^{-1}G_{eva}}$ and $\mathbf{H_{ref}} = \mathbf{A_{ref}} + \mathbf{A_{ref\_eva}A_{eva}^{-1}}(\mathbf{G_{eva}} - \mathbf{A_{eva}})\mathbf{A_{eva}^{-1}A_{eva\_ref}}$ where $\mathbf{A_{ref\_eva}}$ and $\mathbf{A_{eva\_ref}}$ are parts of **A** describing the relationships between the reference and evaluated animals; $\mathbf{A_{ref}}$ is part of **A** for the ungenotyped reference population; $\mathbf{A_{eva}}$ and $\mathbf{G_{eva}}$ are created for the genotyped evaluated population.

For GG, r_{GG}^2 source was as in VanRaden (2008), Goddard et al. (2011), or Pszczola et al. (2012a):

$$\boldsymbol{c}\left[\boldsymbol{G} + \boldsymbol{I}\left(\frac{\sigma_e^2}{\sigma_a^2}\right)\right]^{-1}\boldsymbol{c}', \qquad [4.4]$$

where ***c*** is a vector with the genomic relationships of an evaluated animal with the animals in the reference population. This vector is a column of the **C** matrix for a particular evaluated animal. The **C** matrix itself is created as $\frac{\boldsymbol{Z_2 Z\prime}}{2\sum p_i(1-p_i)}$ and $\mathbf{Z_2}$ is constructed from the genotypes of the evaluated and reference animals. In the GG scenario, **G** for all animals (i.e., the reference and evaluated animals) was created as $\widetilde{\mathbf{G}}$. The **C** matrix, therefore, is an off-diagonal part of $\widetilde{\mathbf{G}}$ that describes relationships between the reference population and the evaluated animals.

Connectedness

Connectedness was calculated as in Lewis et al. (1999) between the evaluated animals and animals in the reference population. Because negative relationship coefficients are present when genomic information is incorporated, the absolute values were taken; thus, the connectedness level was calculated as:

$$con = \left|\frac{PEC(\widehat{a_\iota},\widehat{a_J})}{\sqrt{PEV(\widehat{a_\iota})PEV(\widehat{a_J})}}\right|, \qquad [4.5]$$

where $\mathrm{PEV}(\widehat{\mathrm{a}_\mathrm{i}})$ is the prediction error variance of an EBV of an evaluated animal *i*, $\mathrm{PEV}(\widehat{\mathrm{a}_\mathrm{j}})$ is the prediction error variance of an EBV of a reference animal *j*, $\mathrm{PEC}(\widehat{\mathrm{a}_\mathrm{i}}, \widehat{\mathrm{a}_\mathrm{j}})$ is the prediction error covariance between the breeding values of the animals *i* and *j*. Note that $con = 0$ means that the animals are not connected. The PEV and PEC were obtained from inverted coefficient matrices of mixed model equations. For each of the evaluated animals, its average connectedness with all the animals in the reference population was calculated.

4.3 Results

Reliability

The average reliabilities of (G)EBV predictions are shown in Figures 4.1, 4.2, and 4.3 for the different heritability levels and across all the considered scenarios and reference populations. The significance of the differences between the scenarios was tested within each reference population. Almost all of these differences were highly significant (*P < 0.01*). The exceptions were differences between scenarios AA and GA for the reference population medium relationship and low relationship.

Overall, the reliabilities of the scenarios, as can be seen in Figures 4.1 to 4.3, increased with increasing heritability and with a decreasing average relationship within the reference population. In GG, when all the animals were genotyped, as was especially evident in case of the highest heritability (see Figure 4.3), the reliabilities were considerably higher than in the other cases. When the evaluated animals were genotyped and the reference population was not (AG), reliabilities were noticeably lower than those in GG. When the reference population was genotyped and the evaluated animals were not (GA), reliabilities were somewhat higher than those in AG. For example, in the case of RND for the heritability of 0.01, the average reliabilities for AA, GA, AG, and GG were, respectively, 0.039, 0.042, 0.048, and 0.052 (Figure 4.1). Differences between these scenarios increased together with an increase in the heritability level (Figures 4.1 to 4.3). Differences between the traditional breeding scheme using pedigree information (AA) and GA were always small, with ranking changing across heritability levels. For example, at the heritability of 0.9 (Figure 4.3) and with the RND reference population, AA (0.207) was slightly higher than GA (0.204), whereas at the heritability of 0.3 (Figure 4.2), AA (0.165) was slightly lower than GA (0.173). For the heritability of 0.01, the GA scenario tended to be more accurate than the AA scenario (Figure 4.1; as for example in case of RND, where GA was 0.042 and AA was 0.039) only when the average relationship within the reference was low.

Connectedness

The average connectedness level of evaluated animals with all the animals in the reference population $\overline{con}$ for different scenarios and reference populations and for a heritability of 0.3 is shown in Figure 4.4. The $\overline{con}$ ranged from 0.003 to 0.024. Differences present across designs of the reference populations were very small. The $\overline{con}$ level was lowest when the animals were not genotyped (AA; e.g., 0.004 for RND) and highest when all the animals were genotyped (GG; e.g., 0.024 for RND). Genotyping only the reference population (GA) slightly improved $\overline{con}$ when compared with AA (e.g., from 0.004 for AA to 0.005 for AG in the RND scenario). Genotyping only the evaluated animals (AG) caused a higher increase (to 0.008 for AG in the RND scenario). The $\overline{con}$ level increased when the heritability level increased (not shown; i.e., the heritability level acted as a scaling factor).

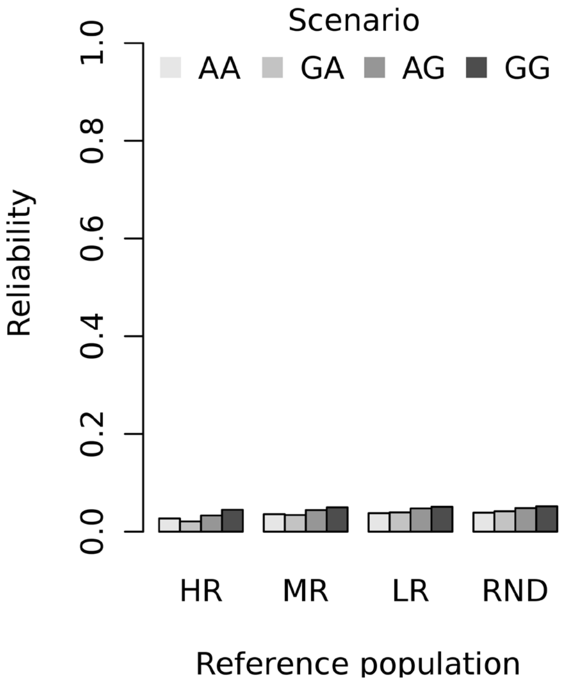


Figure 4.1 Average reliabilities of breeding values for the evaluated set of animals calculated based on scenarios in which pedigree information only is used (AA); the reference population is genotyped and the evaluated animals are ungenotyped (GA); the reference population is ungenotyped and the evaluated animals are genotyped (AG); and all animals are genotyped (GG) across 4 reference populations with different family structures (high relationship = HR; medium relationship = MR; low relationship = LR; random selection of animals = RND) for the heritability of 0.01, averaged over all replicates.

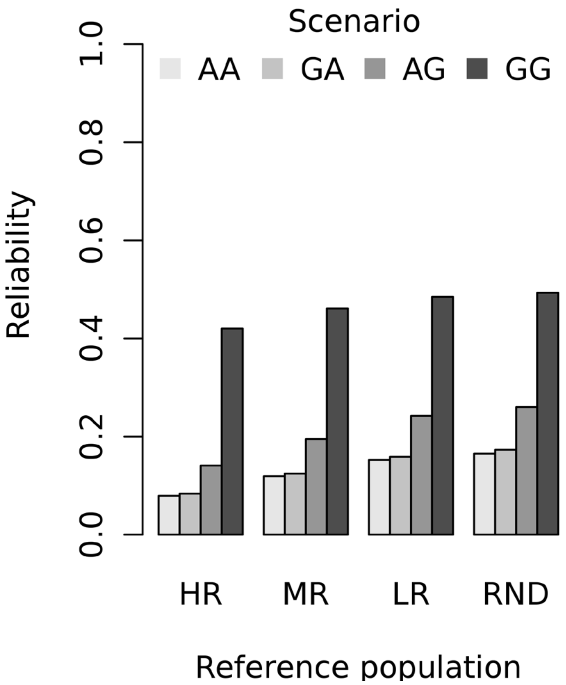


Figure 4.2 Average reliabilities of breeding values for the evaluated set of animals calculated based on scenarios in which pedigree information only is used (AA); the reference population is genotyped and the evaluated animals are ungenotyped (GA); the reference population is ungenotyped and the evaluated animals are genotyped (AG); and all animals are genotyped (GG) across 4 reference populations with different family structures (high relationship = HR; medium relationship = MR; low relationship = LR; random selection of animals = RND) for the heritability of 0.3, averaged over all replicates.

4.4 Discussion

The aim of this study was to investigate whether the increase in accuracy, when moving from traditional selection to GS, is mainly due to genotyping the reference population or the evaluated animals. In it, accuracy of selection is expressed as the reliability of (G)EBV. These reliabilities were calculated using deterministic prediction that can be derived from selection index theory or the prediction error variances of the mixed model equations.

Combined relationship matrix

In the GA and AG scenarios, the genomic-based relationships were combined with the pedigree-based relationships into the **H** matrix (see Legarra et al., 2009; Misztal et al., 2009b). Combining the pedigree-based and the genomic-based relationships may yield different breeding value reliabilities across differently created **G** matrices (Forni et al., 2011). Matrices **A** and **G** have to be on the same scale before they are combined into **H**. To ensure this, in the present study, the inbreeding level in **G** was scaled to the inbreeding level in **A** as in Powell et al. (2010), and then, using the method of Yang et al. (2010), **G** was regressed toward **A**. Further improvement of the compatibility of the 2 matrices could possibly be achieved using LD-linkage analysis methodology (Meuwissen et al., 2011); however, this option was not explored here.

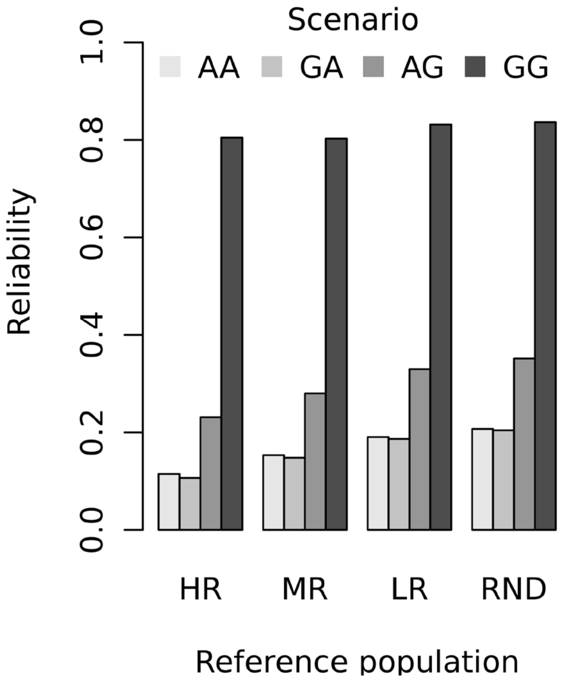


Figure 4.3 Average reliabilities of breeding values for the evaluated set of animals calculated based on scenarios in which pedigree information only is used (AA); the reference population is genotyped and the evaluated animals are ungenotyped (GA); the reference population is ungenotyped and the evaluated animals are genotyped (AG); all animals are genotyped (GG) across 4 reference populations with different family structures (high relationship = HR; medium relationship = MR; low relationship = LR; random selection of animals = RND) for the heritability of 0.9, averaged over all replicates.

Although the applied adjustments did result in a similar scale for both matrices, **G** and **A** still have different properties. This is because genomic relationships include Mendelian sampling and therefore have higher variance than do pedigree-based relationships. This difference in the matrix properties may be a reason why, in the case of GA, it was unexpected that the use of the combined relationship matrix, was only somewhat (or not) beneficial when compared with traditional BLUP (AA).

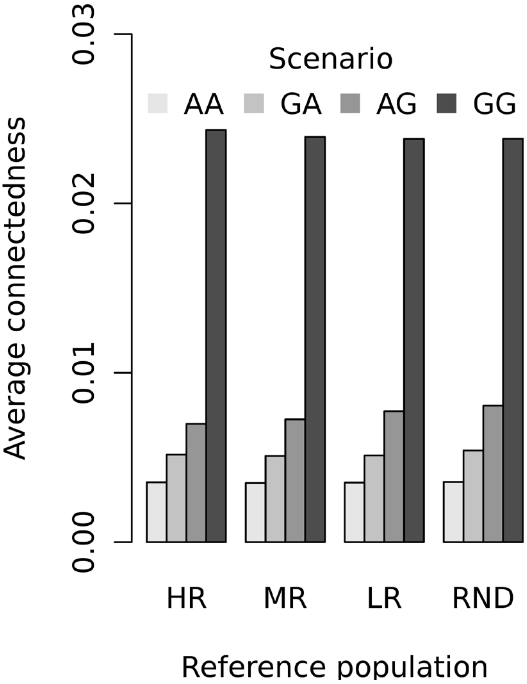


Figure 4.4 Average connectedness of evaluated animals calculated based on scenarios in which pedigree information only is used (AA); the reference population is genotyped and the evaluated animals are ungenotyped (GA); the reference population is ungenotyped and the evaluated animals are genotyped (AG); all animals are genotyped (GG) across 4 reference populations with different family structures (high relationship = HR; medium relationship = MR; low relationship = LR; random selection of animals = RND) for the heritability of 0.3, averaged over all replicates.

Scenario AG was always better than GA, which can be explained by the availability of information on Mendelian sampling. When genotypes are to be predicted from offspring to parents, this information is available, which may lead to high accuracy of the genotype prediction. For instance, if the number of genotyped offspring per ungenotyped parent is large (i.e., 30 or more), the accuracy of genotype prediction can even reach 100%, as shown by Boettcher et al. (2004). In the opposite situation (i.e., when missing genotypes are to be predicted from parents to offspring), prediction accuracy will be poorer because no information on Mendelian sampling is available; therefore, the distinction between, for example, half sibs is not possible. When both parents are genotyped, in fact, the maximum genotype prediction accuracy, measured as the correlation between the true and predicted genotypes, assuming a heritability of 1, is equal to $r = \sqrt{\boldsymbol{bG}}/\sigma_a = \sqrt{1/2}$ where both **b** and **G** contain values of 0.5 and $\sigma_a = 1$ (Mrode, 2005). It is therefore important to genotype both parents if the genotypes of their offspring are to be predicted (Calus et al., 2011a; Pszczola et al., 2011). If both parents (or more ancestors) and offspring of an animal are known, the accuracy of its genotype prediction would be substantially higher (Gengler et al., 2008). Therefore, it is expected that genotyping evaluated animals, which are usually younger than the reference animals, is more beneficial than the opposite. This is especially evident when the evaluated animals are offspring of the reference animals, as discussed above. The closest relationships available between reference and evaluated animals in our design were due to sharing the same ungenotyped sire (i.e., being half-sib family members), maternal grandsires, or other more distant relatives. Closer relationships between genotyped and ungenotyped animals in the AG and GA scenarios would increase the accuracy of predicting genomic relationships for ungenotyped animals, in a similar way as the accuracy of genotype imputation depends on relationships between genotyped and imputed animals (Mulder et al., 2012). In general, the level of accuracy in the simulated scenarios is expected to be lower than when closer relationships between reference and evaluated animals would be present. In most practical situations, however, the evaluated animals are genotyped as well as at least part of the reference animals, and stronger family links between these 2 groups may exist. Such a setup leads to considerable differences between a pedigree-based scenario and a single-step approach favoring the latter method (e.g., Aguilar et al., 2010; Christensen and Lund, 2010; Chen et al., 2011; Forni et al., 2011; Vitezica et al., 2011).

The genome simulated in this study was about 10 times smaller than the cattle genome. As indicated in Pszczola et al. (2012a), the simulated LD was close to values observed in real data (e.g., those reported by de Roos et al., 2008). To further check similarities to expectations found in real data analyses, we estimated the number of effective chromosome segments (M_e) calculated based on the formula of Daetwyler et al. (2008). Rearranging this equation yields $M_e = Nh^2/(r^2 - Nh^2)$, where N is the reference population size, h^2 is the heritability level, and r^2 is the reliability obtained from our analysis. The calculated value for M_e ranged from 352 to 437 for an h^2 of 0.9, from 617 to 829 for an h^2 of 0.3, and from 365 to 424 for an h^2 of 0.01. The lowest values were observed for reference populations with a weaker family structure. Our M_e values agreed reasonably well with the results of other studies for small cow reference populations. For example, Verbyla et al. (2010) showed that M_e calculated for a small population of Holstein-Friesian heifers for energy balance with an h^2 of 0.325 was 472.

As our results have shown, compared with traditional BLUP, genotyping only the evaluated animals achieved substantially higher reliabilities. Although the realized reliabilities will be substantially lower than if all or most of the reference and the evaluated animals are genotyped, genotyping only the evaluated animals is more beneficial than genotyping only animals with phenotypic records.

Impact of relationships

The average relationship within the reference population, as reported previously by Pérez-Cabal et al. (2010) and Pszczola et al. (2012a), affected the average reliability regardless of which animals were genotyped. When the animals in the reference population were loosely related to each other, as shown in Figures 4.1 to 4.3, the reliabilities were higher. This means that the design of the reference population is also important when genotyped and ungenotyped animals are analyzed jointly using the $\mathbf{H}$ matrix.

The availability of more precise relationship data for the reference population (GA) resulted in small differences in reliabilities compared with a situation in which none of the animals were genotyped (AA), whereas, as shown in Figures 4.1 to 4.3, genotyping the evaluated animals (AG) increased reliabilities noticeably. This may be because, from regular BLUP models (Henderson, 1985), it is known that when an animal has phenotypic records itself, the emphasis on information about relatives is reduced. Translated to the scenarios studied here, this reduced emphasis on information concerning relatives implies that very precise genomic relationships among animals in the reference population (i.e., animals with phenotypes) might be less important than those for the evaluated animals with no phenotypic observations on themselves or on their descendants.

We also considered a situation in which the evaluated animals originated from the same generation as the reference population. In this situation, the average reliabilities were higher than in the presented results, but the tendencies were similar (results not shown). For animals originating from different generations, differences in the reliability of (G)EBV were in line with those in the literature (Habier et al., 2010; Pszczola et al., 2011, 2012; Wolc et al., 2011).

Deterministic approaches can be used to approximate the reliability, given known and simple family structures of the considered population (Hayes et al., 2009d) or randomly mated population (Daetwyler et al., 2008; Goddard, 2009a). In practice, however, more complex family structures are present and animals are not mated at random, and for such a situation, deriving a general deterministic prediction formula is not trivial. On the basis of our results, however, in general one may expect to achieve higher reliabilities when the reference population is composed of many small half-sib families as opposed to a few large ones; that is, the use of an RND reference population always yielded higher average reliabilities than using a high relationship (Figures 4.1 to 4.3).

Bias of prediction

The reliability of (G)EBV can be assessed by inverting the left-hand side of the mixed model equations or, in an empirical way, by cross-validation (e.g., Calus et al., 2010b). The former method, as applied in this study, tends to give somewhat overestimated results (Calus et al., 2009; Hayes et al., 2009b; Lund et al., 2009; Su et al., 2010). One reason for this overestimation is the underlying assumption that the markers explain all the genetic variance. The genetic variance, however, is at least partially also explained by loci between the markers (VanRaden, 2008). Thus, the reliabilities presented here may be somewhat overestimated. Despite this expected slight overestimation of the reliabilities, the method used here, unlike the cross-validation method, makes an assessment of individual reliabilities possible, which is important from the practical point of view when comparing animals' breeding values. Yet even if some overestimation would be present in this study, it is expected not to affect the conclusions presented here. This is because reliabilities for the breeding values based on pedigree information are expected to be unbiased, and replacing the pedigree with genomic data, to some extent, only adds more information.

Connectedness

The reliability predictions formulas presented here are equivalent to those obtained with use of the PEV of breeding values. The PEV of differences between animals from different management units can be also used to determine connectedness (Kennedy and Trus, 1993). Genetically unconnected herds or other management units with different genetic means cannot be distinguished in the genetic evaluation; thus, the comparison of breeding values among them is biased (Kennedy, 1981). This bias would be reduced if a positive genetic covariance existed between the units (i.e., they would be connected). In other words, 2 animals are connected when the PEC between them is nonzero; the difference between their breeding values is then expected to have smaller bias than would a pair of animals with zero PEC. Using the pedigree-based relationships, PEC is nonzero only when the animals are linked through the pedigree. Seemingly unlinked animals, in fact, may be linked through distant unrecorded ancestors, as can be revealed with genomic relationships. Using genomic relationships thus improves comparisons across the management units. This was shown in Figure 4.4, where the average connectedness level in the AA scenario was considerably lower than that in GG. Small differences across differently designed reference populations may be attributed to the fact that differences in the average relationship within the reference population based on genomic data were smaller than calculated based on the pedigree data.

4.5 Conclusion

This study aimed to investigate the contribution of using genomic information on a reference population or the evaluated animals to the increase in selection accuracy. Compared with traditional selection, genotyping only the evaluated group of animals significantly increased the accuracy of the estimates, whereas genotyping only the reference population yielded minor, and sometimes unfavorable, changes in accuracy. This was attributed to the fact that the emphasis on information concerning relatives is reduced when an animal has phenotypic records. The reduced emphasis on the information of relatives implies that very precise genomic relationships among animals in the reference population are less important than those for the evaluated animals with no phenotypic observations on themselves or on their descendants. Nevertheless, although the main origin of the gain in accuracy from using GS is genotyping the evaluated animals, genotyping only one group of animals will always yield estimates that are substantially less accurate than when all the animals are genotyped. An additional benefit of using a genomic relationship matrix is reducing bias across herd, region, or country evaluations, which is demonstrated by the improved connectedness between the reference population and the evaluated animals.

4.6 Acknowledgments

Roel Veerkamp (Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, Lelystad, the Netherlands) and Han Mulder (Animal Breeding and Genomics Centre, Wageningen University, Wageningen, the Netherlands) are acknowledged for their useful comments and discussion. Linda McPhee (Linda McPhee Consulting Ltd., Cambridge, UK) is thanked for helping edit the manuscript. Marcin Pszczola gratefully acknowledges the financial support of the Koepon Stichting (Leusden, the Netherlands) and GreenHouseMilk. The GreenHouseMilk project is financially supported by the European Commission under the Seventh Research Framework Programme (Grant Agreement KBBE-238562). This publication represents the views of the authors, not the European Commission, and the Commission is not liable for any use that may be made of the information. Financial support of the National Science Centre (Krakow, Poland) is acknowledged (decision number DEC-2011/01/N/NZ9/00614).

5

Effect of predictor traits on accuracy of genomic breeding values for feed intake based on a limited cow reference population

M. Pszczola[1,2,3], R.F. Veerkamp[1,2], Y. de Haas[1], E. Wall[4], T. Strabel[3]
and M. P. L. Calus[1]

[1] Animal Breeding and Genomics Centre, Wageningen University, 6700 AH Wageningen, The Netherlands; [2] Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, 8200 AB Lelystad, The Netherlands; [3] Department of Genetics and Animal Breeding, Poznan University of Life Sciences, Wolynska 33, 60-637 Poznan, Poland; [4] Animal & Veterinary Sciences Group, SRUC, Roslin Institute Building, Easter Bush, Midlothian EH25 9RG, UK

Animal (2013) – DOI: http://dx.doi.org/10.1017/S175173111300150X

Abstract

The genomic breeding value accuracy of scarcely recorded traits is low because of the limited number of phenotypic observations. One solution to increase the breeding value accuracy is to use predictor traits. This study investigated the impact of recording additional phenotypic observations for predictor traits on reference and evaluated animals on the genomic breeding value accuracy for a scarcely recorded trait. The scarcely recorded trait was dry matter intake (DMI, n = 869) and the predictor traits were fat–protein-corrected milk (FPCM, n = 1520) and live weight (LW, n = 1309). All phenotyped animals were genotyped and originated from research farms in Ireland, the United Kingdom and the Netherlands. Multi-trait REML was used to simultaneously estimate variance components and breeding values for DMI using available predictors. In addition, analyses using only pedigree relationships were performed. Breeding value accuracy was assessed through cross-validation (CV) and prediction error variance (PEV). CV groups (n = 7) were defined by splitting animals across genetic lines and management groups within country. With no additional traits recorded for the evaluated animals, both CV- and PEV-based accuracies for DMI were substantially higher for genomic than for pedigree analyses (CV: max. 0.26 for pedigree and 0.33 for genomic analyses; PEV: max. 0.45 and 0.52, respectively). With additional traits available, the differences between pedigree and genomic accuracies diminished. With additional recording for FPCM, pedigree accuracies increased from 0.26 to 0.47 for CV and from 0.45 to 0.48 for PEV. Genomic accuracies increased from 0.33 to 0.50 for CV and from 0.52 to 0.53 for PEV. With additional recording for LW instead of FPCM, pedigree accuracies increased to 0.54 for CV and to 0.61 for PEV. Genomic accuracies increased to 0.57 for CV and to 0.60 for PEV. With both FPCM and LW available for evaluated animals, accuracy was highest (0.62 for CV and 0.61 for PEV in pedigree, and 0.63 for CV and 0.61 for PEV in genomic analyses). Recording predictor traits for only the reference population did not increase DMI breeding value accuracy. Recording predictor traits for both reference and evaluated animals significantly increased DMI breeding value accuracy and removed the bias observed when only reference animals had records. The benefit of using genomic instead of pedigree relationships was reduced when more predictor traits were used. Using predictor traits may be an inexpensive way to significantly increase the accuracy and remove the bias of (genomic) breeding values of scarcely recorded traits such as feed intake.

Key words: accuracy; bias; genomic selection; multi-trait analyses; dairy cow

5.1 Implications

Feed intake has a significant impact on the profitability and environmental footprint of livestock production. Feed intake is expensive and labour intensive to measure on individual animals, resulting in a limited number of phenotypic records. Genomic selection is suggested to be a promising selection tool to improve traits with only small numbers of available observations. In this study, an inexpensive way to increase the accuracy of genomic selection and remove the bias of (genomic) breeding values for feed intake using easily recordable predictor traits was demonstrated. This approach is successful when predictor traits are recorded on both evaluated and reference animals.

5.2 Introduction

Single-nucleotide polymorphisms (SNP) markers are used nowadays in genomic selection (Hayes *et al.*, 2009a). Genomic selection increases genetic gain compared with traditional selection tools by decreasing the generation interval without (heavily) jeopardizing the accuracy of selection (Meuwissen *et al.*, 2001, Muir, 2007, Calus, 2010a). The accuracy of genomic selection depends on the accuracy of estimated direct genomic breeding values (DGV). The DGV accuracy is affected by the effective population size by the density at which the SNP chip covers the genome and by the size of the reference population (Daetwyler *et al.*, 2008, Goddard, 2009a, Pszczola *et al.*, 2012a). Generally, a large reference population is needed to achieve high accuracy of DGV (Hayes *et al.*, 2009a). Large reference populations already exist for routinely recorded traits, for example, milk yield. However, for traits that are not routinely measured because they are difficult or expensive to measure, obtaining large numbers of phenotypic observations is problematic. Despite the availability of relatively low numbers of records, the application of genomic selection is suggested as a promising tool to start selection for difficult-to-measure traits (Calus *et al.*, 2013a).

An example of a labour-intensive and expensive-to-measure trait is feed intake, for which usually only a limited number of phenotypic records are available. The heritability of feed intake ranges from 0.16 to 0.36, showing potential to change feed intake using genetic improvement (for review see Veerkamp, 1998). As traditional selection is likely too costly to apply for feed intake, and because genotyping costs are reducing continuously, genomic selection is the expected method of choice to start genetic selection for feed intake. The limited number of phenotypic records for feed intake will, however, result in a relatively low accuracy of genomic selection for feed intake. One solution to partly overcome this effect of the limited size of the reference population is to optimize its design (Jiménez-

Montero *et al.*, 2012, Pszczola *et al.*, 2012a). Another solution, commonly used in traditional breeding programmes, is to use predictor traits. For example, somatic cell counts are used to predict udder health traits (Philipsson *et al.*, 1995), and birth weight can be used to predict calving ease or perinatal mortality (Johanson and Berger, 2003). A predictor trait has to meet several requirements. It has to be easily recordable, inexpensive to measure, heritable and, most importantly, genetically correlated with the trait of interest. Using such a predictor trait may result in an improvement of the trait of interest at a low cost. Also in genomic selection, an additional gain in accuracy can be achieved using predictor traits, as it was shown using both deterministic (Calus *et al.*, 2013a) and empirical simulations (Calus and Veerkamp, 2011c; Jia and Jannink, 2012). Both empirical studies had predictor traits measured on both the reference animals and evaluated animals. An important and unanswered question when using predictor traits in genomic selection is on which animals (reference or evaluated) the predictor traits should be recorded. Therefore, the objective of this study was to investigate the impact of using additionally recorded phenotypic observations for the predictor traits on (1) the reference, or (2) on the reference and evaluated animals, on the accuracy of breeding values for a scarcely recorded trait, using multi-trait genomic prediction models.

5.3 Materials and methods

The data used in this study originated from Teagasc, Moorepark, Ireland, the Langhill herd from the Scottish Agricultural College, United Kingdom, and from the ‘t Gen herd from Wageningen UR Livestock Research, the Netherlands. Details on the experimental treatments for different countries are described elsewhere, for Scotland (Veerkamp *et al.*, 1995, Pryce *et al.*, 1999, Coffey *et al.*, 2004, Bell *et al.*, 2011), Ireland (Horan *et al.*, 2005) and the Netherlands (Veerkamp *et al.*, 2000), and a more detailed description of the merging of the data sources and estimated variance components across the different herds can be found in a study by Banos *et al.* (2012).

Phenotypic data

The pre-corrected phenotypes used here were obtained from Veerkamp *et al.* (2012) who described in detail the steps taken to adjust the original phenotypes. In short, phenotypic measurements were corrected for the mean lactation curve, herd, nutritional treatment, milking frequency, year and month of milk test by management group, and experimental treatments using random regression. Records available for the average fat–protein-corrected milk (FPCM), live weight

(LW) and dry matter intake (DMI) of the predicted values for weeks 3 to 15 were used. The available data set comprised 1520 animals' phenotypic observations for FPCM; of these animals, 1309 had phenotypes for LW and 869 had phenotypes for DMI. Feed intake was not recorded on animals from Teagasc, Moorepark, Ireland. In total, 824 animals had pre-corrected phenotypic observations for all three traits. For descriptive statistics of the traits, see Table 5.1.

Table 5.1 Descriptive statistics of the analysed traits.

Trait	No. records	Min.	Mean	Max.	S.D.	Average reference population size[a]
FPCM (kg/d)	1 520	2.41	34.62	54.57	6.16	1 402
LW (kg/d)	1 309	241.77	520.00	699.95	43.97	1 191
DMI (kg)	869	7.06	12.39	16.76	1.60	751

FPCM = Fat-protein-corrected milk; LW = Live weight; DMI = Dry matter intake.

[a] Reference population size averaged over all cross-validation sets

Genomic data

In this study, the genotypes of 2162 animals genotyped with the Illumina BovineSNP50 BeadChip (Illumina Inc., San Diego, CA, USA) containing 54 001 SNPs were available. This included all phenotyped animals and some additional ones with no phenotypes. The additional animals were used to perform checks for Mendelian inconsistencies between pedigree and SNP data, which were performed for all genotyped parent–offspring pairs and among sibs (Calus *et al.*, 2011b). Animals with conflicting information were removed. The SNP quality control criteria were as described by Veerkamp *et al.* (2012), that is, GenCall (GC) score >0.20 and GenTrain (GT) score >0.55 for individual genotypes; call rate >95%; minor allele frequency >0.01 in each country; and no extreme deviation from Hardy–Weinberg Equilibrium (i.e. $\chi^2 < 600$). After applying the quality control of the genotypes, 36 346 SNP remained in the data set.

Models

Animal models with one, two or three traits included were fitted using ASReml 3.0 (Gilmour *et al.*, 2009) depending on how many traits were available in each scenario. The general (multi-trait) model was

$$y_{ij} = \mu_j + animal_{ij} + e_{ij},$$

where y_{ij} is the pre-corrected phenotypic record of animal *i*, μ_j is the overall mean for trait *j*, $animal_{ij}$ is the random polygenic effect of animal *i* for trait *j* and e_{ij} is a random residual for animal *i* and trait *j*. Variance components were estimated for each analysis. In pedigree analyses (P-REML), polygenic values were assumed to be normally distributed $N(0, \mathbf{A}\otimes\mathbf{G_A})$, where **A** is the numerator relationship matrix, $\mathbf{G_A}$ is the *m* × *m* polygenic covariance matrix and *m* is the number of traits in the model. In genomic-based analyses (G-REML), genomic values were assumed to be normally distributed $N(0, \mathbf{G}\otimes\mathbf{G_G})$, where **G** is the genomic relationship matrix and $\mathbf{G_G}$ is the *m* × *m* genomic covariance matrix. The **G** matrix was created with the first formula described by VanRaden (2008) using current allele frequencies. The inbreeding level in **G** was scaled to the inbreeding level in **A** and then the adjusted **G** was corrected to account for the sampling error in genomic relationships as in the study by Veerkamp *et al.* (2011).

Accuracy

The accuracy of the breeding values was calculated in two ways: using cross-validation (CV) and using prediction error variance (PEV) and, additionally, selection index theory calculations were carried out to assess the minimum and maximum expected accuracy for an animal that is unrelated to the reference population.

CV accuracy was calculated as the correlation between the estimated breeding values and the phenotypic values divided by the square root of the heritability. Accuracy was calculated for each evaluation set and then averaged over all evaluation sets. Heritabilities of the considered traits and genetic correlations between traits were estimated using pedigree and genomic relationships using a multi-trait analysis in ASReml (Gilmour *et al.*, 2009) using all available data.

The PEV-based accuracies for each animal (*i*) were calculated using standard errors of prediction (SEP) obtained from the output of the models as

$$r_i = \sqrt{1 - \frac{\boldsymbol{PEV_i}}{\sigma_A^2}}\text{, where } \boldsymbol{PEV}_i = \boldsymbol{SEP}_i^2.$$

Accuracies for the evaluated animals were calculated within the corresponding CV set and then averaged across CV sets within scenarios.

To assess the minimum and maximum expected accuracy for an animal that is unrelated to the reference population, selection index calculations were carried out for two scenarios. In the first scenario (OWN), the expected accuracy was calculated for an animal that had no relatives in the reference population and only had its own phenotypic observations available. In the second scenario (PRO), the expected accuracy was calculated for a progeny-tested bull that was assumed to be a selection candidate. This bull had very accurate breeding values for the predictor traits, instead of its own phenotypic observations. These accurate breeding values were based on 1000 offspring each with a single phenotype. In the two scenarios, three situations were distinguished, in which observations (being phenotypic records or daughter-based breeding values) were available for: (a) FPCM, (b) LW or (c) FPCM and LW. Phenotypic observations for DMI were not available.

Reference populations

Four reference populations were defined that differed by the number of traits for which phenotypic observations were available. The sizes of the different reference populations varied along with the number of phenotypic records available for the different traits (see Table 5.1). Records of DMI were assumed to be available for at least part of the animals in all the reference populations. Observations for the reference animals were available on: (1) DMI only; (2) DMI and FPCM; (3) DMI and LW; or (4) FPCM, LW and DMI.

Evaluated animals

Evaluated animals were created by defining several CV sets. Only animals from the United Kingdom and the Netherlands were included in CV sets, because only those animals had records for all traits. Animals from Ireland contributed only to the reference populations with records available for LW and FPCM. The CV sets were as in the study by De Haas *et al.* (2012b) who split the animals across genetic lines and management groups. Four CV sets were formed from the UK animals. First, animals were split based on two genetic lines: a control and a selection line. Animals in the control line have been bred to bulls with about average genetic merit for fat and protein yield, whereas the animals in the selection line were bred since 1973 to bulls with the highest genetic merit for fat and protein yield available in the United Kingdom (Veerkamp *et al.*, 1994). Second, animals were divided on the basis of two feeding strategies: one group was fed with a high-concentrate diet and the second group with a low-concentrate diet. Animals from the Netherlands composed of three groups. At the ‘t Gen farm, two groups of animals were present: a high genetic line for milk yield, that participated in the CR-Delta (Arnhem, The Netherlands) breeding programme, and a control line. Owing to its size, the high genetic line was randomly split into two sets. Each of the CV sets was assumed to be a set of evaluated animals and was analysed four times, assuming a presence of phenotypic information for: (1) none of the considered indicator traits, (2) FPCM, (3) LW or (4) FPCM and LW.

Scenarios

Of all the 16 possibilities (all combinations of four reference and four evaluated sets), nine of the most practical ones were chosen, as outlined in Table 5.2. All reference populations were used to predict all evaluation animals without using any of their records on the predictor traits to obtain baseline accuracy. The reference population for which information on all traits was available was used to evaluate all possible evaluated sets of animals (i.e. no observations, FPCM, LW and FPCM and LW). In addition, the reference population with DMI and FPCM records was used to evaluate the CV set with FPCM recorded, and the reference population with DMI and LW records was used to evaluate the CV set with LW records. All analyses were performed twice: first using pedigree relationships (P-REML) and then using genomic relationships (G-REML).

The remaining seven scenarios (i.e. out of all 16 possible combinations), where the predictor traits could be recorded only for the evaluated and not for the reference population, were only initially investigated here because of two reasons: first, having phenotypes only for the evaluated animals is rather unlikely in practice, and second, estimating variance components with only few animals in the evaluated sets proved to be difficult in the initial analyses and the obtained estimates were inaccurate.

Table 5.2 Included phenotypes (✓) in the different analyses, both for the evaluated and the reference population.

Traits recorded on reference population	Traits recorded on evaluated population			
	NO	FPCM	LW	FPCM and LW
DMI	✓			
DMI and FPCM	✓	✓		
DMI and LW	✓		✓	
DMI, FPCM and LW	✓	✓	✓	✓

FPCM = Fat-protein-corrected milk; LW = Live weight; DMI = Dry matter intake; NO = No phenotypes available.

5.4 Results

Heritability and correlation estimates

The estimated heritabilities for each trait and phenotypic and genetic correlations between the traits are shown in Table 5.3. Heritability estimates were moderate to high, ranging from 0.31 to 0.60, with the highest value observed for DMI. The strongest phenotypic and genetic correlations were between DMI and LW, whereas correlations between DMI and FPCM were somewhat weaker. In general, heritabilities and genetic correlations were somewhat lower when genomic information was used instead of pedigree information.

Table 5.3 Heritability estimates (diagonal) phenotypic (above diagonal) and genetic correlations (below diagonal) for the analysed traits estimated using pedigree and genomic data, and their approximated standard errors in parentheses.

	Data used to obtain estimates					
	Pedigree			Genomic		
	FPCM	LW	DMI	FPCM	LW	DMI
FPCM[1]	**0.36 (0.06)**	0.19 (0.03)	0.45 (0.03)	**0.31 (0.04)**	0.18 (0.03)	0.45 (0.03)
LW[2]	0.34 (0.12)	**0.48 (0.07)**	0.47 (0.03)	0.12 (0.11)	**0.41 (0.05)**	0.45 (0.03)
DMI[3]	0.32 (0.11)	0.70 (0.08)	**0.60 (0.08)**	0.24 (0.11)	0.62 (0.08)	**0.44 (0.06)**

FPCM = Fat-protein-corrected milk; LW = Live weight; DMI = Dry matter intake.

Accuracy

Genomic and pedigree-based accuracies, calculated using CV and PEV, are presented in Tables 5.4 and 5.5, respectively. When no additional traits were recorded for evaluated animals, accuracies for DMI, both CV- and PEV-based, were substantially higher for genomic analyses than for pedigree analyses (CV: maximum 0.26 for pedigree and 0.33 for the genomic analyses; PEV: maximum 0.45 and 0.52, respectively). When one or two traits were added, the differences between pedigree- and genomic-based accuracies diminished (Tables 5.4 and 5.5). With additional recording for FPCM, pedigree-based accuracies increased from 0.26 to 0.47 for CV and from 0.45 to 0.48 for PEV. Genomic-based accuracies increased from 0.33 to 0.50 for CV and from 0.52 to 0.53 for PEV. With additional recording for LW instead of FPCM, pedigree-based accuracies increased to 0.54 for CV and to 0.61 for PEV. Genomic-based accuracies increased to 0.57 for CV and to 0.60 for PEV. When both FPCM and LW were available for evaluated animals then the highest accuracy was obtained (0.62 for CV and 0.61 for PEV in pedigree analyses, and 0.63 for CV and 0.61 for PEV in genomic analyses). Recording predictor traits only for the reference population alone did not increase DMI breeding value accuracy. Recording predictor traits for both the reference population and evaluated animals significantly increased DMI breeding value accuracy and removed the bias observed when only reference animals had records.

The selection index calculations showed that the predicted accuracy for the OWN scenario, when phenotypes were available for FPCM, was 0.19. When LW instead of FPCM records was available, the predicted accuracy increased to 0.49. When both predictor traits records were used, the predicted accuracy reached the highest value of 0.50. For the PRO scenario, when phenotypes were available for

FPCM, the predicted accuracy was 0.32. When LW only and LW and FPCM phenotypes were available the predicted accuracy was 0.70.

Table 5.4 Cross-validation-based accuracies (and their standard errors) for DMI of evaluated animals for which phenotypic observations were available on different number of traits across reference populations with phenotypes available for different traits.

Traits recorded on reference population	Traits recorded on evaluated population			
	NO	FPCM	LW	FPCM and LW
		P-REML[a]		
DMI	0.24$_{(0.03)}$	.	.	.
DMI and FPCM	0.24$_{(0.03)}$	0.47$_{(0.04)}$	.	.
DMI and LW	0.26$_{(0.04)}$	.	0.54$_{(0.05)}$	.
DMI, FPCM and LW	0.26$_{(0.04)}$	0.47$_{(0.05)}$	0.54$_{(0.06)}$	0.62$_{(0.06)}$
		G-REML[b]		
DMI	0.33$_{(0.02)}$	.	.	.
DMI and FPCM	0.33$_{(0.03)}$	0.50$_{(0.06)}$	.	.
DMI and LW	0.32$_{(0.03)}$	.	0.57$_{(0.05)}$	.
DMI, FPCM and LW	0.33$_{(0.03)}$	0.50$_{(0.06)}$	0.57$_{(0.06)}$	0.63$_{(0.06)}$

FPCM = Fat-protein-corrected milk; LW = Live weight; DMI = Dry matter intake; NO = No phenotypes available.

[a] Pedigree-based analyses;

[b] Genomic-based analyses.

Table 5.5 Prediction error variance-based accuracies for DMI of evaluated animals for which phenotypic observations were available on different (number of) traits across reference populations with phenotypes available for different traits.

Traits recorded on reference population	Traits recorded on evaluated population			
	NO	FPCM	LW	FPCM and LW
		P-REML[a]		
DMI	0.43	.	.	.
DMI and FPCM	0.44	0.48	.	.
DMI and LW	0.45	.	0.61	.
DMI, FPCM and LW	0.45	0.48	0.60	0.61
		G-REML[b]		
DMI	0.51	.	.	.
DMI and FPCM	0.51	0.53	.	.
DMI and LW	0.52	.	0.60	.
DMI, FPCM and LW	0.52	0.53	0.60	0.61

FPCM = Fat-protein-corrected milk; LW = Live weight; DMI = Dry matter intake; NO = No phenotypes available.

[a] Pedigree-based analyses;

[b] Genomic-based analyses.

Bias

Slopes of regression of phenotypes on estimated breeding values are a measure of the bias in terms of the variance of estimated breeding values. Slopes greater (smaller) than one indicate underestimation (overestimation) of the variance of the estimated breeding values. The slopes ranged from 0.56 to 1.17 for pedigree and from 0.68 to 1.12 for genomic-based analyses, as shown in Table 5.6. When no observations on the evaluated animals were available and the reference animals had only DMI records, the variance of the estimates was most biased, the slope was 0.56 for pedigree analyses and 0.68 for genomic analyses. The bias was somewhat reduced when information for all traits (FPCM, LW and DMI) was added to the reference population, that is, the slopes increased to 0.66 for pedigree and 0.73 for genomic analyses. The variance of the estimated breeding values was rather unbiased when information on additional trait(s) was added both for the reference population and for the evaluated animals. This was observed, irrespective of whether FPCM, LW or both were added to the analyses; the slopes ranged from 0.99 to 1.12 for genomic and from 0.92 to 1.17 for pedigree analyses.

Re-ranking

Spearman rank correlations between the estimated breeding values were calculated to investigate whether changes in the ranking of animals occurred on the basis of their estimated breeding values across different scenarios. The correlations are shown in Appendix Tables A5.1 and A5.2 for pedigree and genomic analyses, respectively. Little re-ranking of the animals on the basis of their breeding values was present across scenarios in which no phenotypic information was available for the evaluated animals (rank correlations from 0.89 to 0.96 for pedigree and genomic analyses). Considerable re-ranking of animals was present comparing scenarios with and without information on the additional traits for evaluated animals, with rank correlations from 0.58 to 0.99. The rank correlations showed that the re-ranking of the animals coincides with the accuracy of their breeding values, that is, when the accuracy of the breeding values estimated with one method differs from the accuracy of the other method, more changes in the ranking of the animals on the basis of their breeding values are expected to occur.

5.5 Discussion

In this study, we investigated the impact of recording additional phenotypic observations for the predictor traits for reference and evaluated animals on the accuracy of breeding values for a scarcely recorded trait. Our results show that the breeding value accuracy for a scarcely recorded trait was not increased when the predictor traits were recorded only for the reference population. When the predictor traits were recorded for both the evaluated animals and the reference population, however, the breeding value accuracy increased. The above results were obtained using DMI as the scarcely recorded trait, and FPCM and LW as predictor traits for DMI.

Table 5.6 Slopes of regression of phenotype on estimated breeding values (and the standard error between cross-validation sets) for DMI of evaluated animals for which phenotypic observations were available on different traits across different reference populations.

Traits recorded on	Traits recorded on evaluated population			
reference population	NO	FPCM	LW	FPCM and LW
		P-REML[a]		
DMI	$0.56_{(0.07)}$	.	.	.
DMI and FPCM	$0.60_{(0.10)}$	$1.01_{(0.19)}$	.	.
DMI and LW	$0.63_{(0.11)}$	.	$0.92_{(0.12)}$	.
DMI, FPCM and LW	$0.66_{(0.12)}$	$1.17_{(0.21)}$	$0.96_{(0.15)}$	$1.05_{(0.15)}$
		G-REML[b]		
DMI	$0.68_{(0.11)}$	.	.	.
DMI and FPCM	$0.72_{(0.12)}$	$1.06_{(0.21)}$	.	.
DMI and LW	$0.71_{(0.13)}$	.	$0.99_{(0.12)}$	.
DMI, FPCM and LW	$0.73_{(0.13)}$	$1.08_{(0.23)}$	$1.00_{(0.15)}$	$1.12_{(0.16)}$

FPCM = Fat-protein-corrected milk; LW = Live weight; DMI = Dry matter intake; NO = No phenotypes available.

[a] Pedigree-based analyses;

[b] Genomic-based analyses.

Accuracy

Differences were shown when the predicted (selection index-based) and empirical (REML-based) accuracies for DMI were compared. When only FPCM was the predictor trait, the empirical genomic-based accuracy was lower than the corresponding predicted accuracy of the OWN and PRO scenarios. When only LW was the predictor trait, or when two predictor traits (FPCM and LW) were used, the empirical genomic-based accuracies were higher than these in OWN but lower than in PRO scenarios. The presented results show that the empirical accuracies may differ from the expectations on the basis of selection index theory. The difference between predicted and empirical accuracies may arise, for example, owing to the pedigree structure of the data used. In the theoretical calculations (OWN and PRO), a reference population of unrelated individuals was assumed, whereas the animals in the data set might actually be strongly related to each other. As shown earlier, the more the animals in the reference population are related to each other, the lower the predicted accuracy is (Pszczola *et al.*, 2012a, Pszczola *et al.*, 2012b). When the own phenotypes for the predictor traits are available (OWN), or when the correlation between the traits is weak (i.e. PRO with only FPCM recorded), utilizing information collected from relatives may increase the accuracy. The predicted accuracy of scenarios OWN and PRO LW, where no information on relatives was utilized, was indeed inferior to the empirical accuracy. Conversely, when phenotypic information is very accurate (i.e. based on numerous progeny) and the correlation between the traits is sufficient, information from the relatives has a limited impact on the accuracy. Again, this was confirmed here, as the predicted accuracy obtained from almost all the PRO scenarios was superior over the empirical accuracy. Therefore, the results of OWN and PRO scenarios, obtained with selection index theory are in agreement with the results of the REML models.

The gain in breeding value accuracy for a scarcely recorded trait is expected to be higher when the correlation with the predictor trait is stronger. In our study, LW was more strongly correlated to DMI than FPCM was. Therefore, the gain in breeding value accuracy of the scarcely recorded DMI was expected to be considerably higher when using LW as a predictor trait than when using FPCM. Contrarily, differences in accuracies between these two scenarios, as shown in Tables 5.4 and 5.5, were small. We investigated whether this unexpected difference between scenarios with either FPCM or LW can be attributed to the 211 more phenotypes available for FPCM than for LW. Analyses with equal sizes of the reference populations for the predictor traits (after removing the 211 FPCM records of the animals with no LW records), however, showed similar results (not shown) to the analyses using all FPCM observations (Table 5.4), leaving this issue unresolved. Analysing all traits jointly yielded the highest breeding value accuracy, as expected. The superiority of multi-trait genomic analyses over single trait ones was also shown earlier in stochastic (Calus and Veerkamp, 2011c, Jia and Jannink, 2012) and deterministic simulations (Calus *et al.*, 2013a), as well as in real data analysis (Aguilar *et al.*, 2011, Tsuruta *et al.*, 2011).

The expected gain in accuracy because of the use of multi-trait selection, as compared with single-trait selection, strongly depends on the value of the variance components of the traits involved (Thompson and Meyer, 1986). When the heritability of the different traits is similar and pairwise genetic and residual correlations are similar, the benefit of multi-trait applications is expected to be very limited. Large gains in accuracy are expected when pairwise genetic and residual correlations differ from each other substantially and when the predicted trait has a low heritability. Differences between the heritabilities of the analysed traits were small and diminished even more when the genomic data were used (see Table 5.3). Genetic correlations between DMI and LW were higher than between DMI and FPCM, but in both cases they were moderate. Therefore, for other traits than those analysed here, the benefit of using the predictor traits may be different. For example, differences in accuracy for various type traits, while using single- or multi-trait genomic analyses shown by Tsuruta *et al.* (2011), were up to 9%.

In our study, using two predictor traits resulted in similar accuracies for the pedigree-based and genomic-based breeding values (see Table 5.4). It should be, however, noted that this may be due to relatively low genomic-based accuracies, which was because of small number of the animals in the reference population. If the reference population would be bigger, it is likely that even with two predictor traits recorded on all the animals the genomic-based breeding value accuracy will be superior compared with the pedigree-based one.

Genomic analyses are expected to yield more accurate estimates than the pedigree ones, as shown in theoretical (Goddard, 2009a, Meuwissen, 2009) and simulation studies (Meuwissen *et al.*, 2001), as well as in most real data analyses (Lund *et al.*, 2011; for reviews see: Hayes *et al.*, 2009a, Calus, 2010a). The superiority of genomic analyses was also confirmed here for the single-trait model. The genomic-based accuracies for the single-trait model were likely superior because genomic relationships capture information on Mendelian sampling in evaluated animals and also because of increased connectedness between the reference and evaluated animals when **G** was used (Pszczola *et al.*, 2012b). Differences in accuracy were, however, moderate and nearly disappeared in multi-trait analyses. One possible explanation for the small differences between pedigree and genomic analyses might be the fact that the analysed populations had strong family links. For example, in **A**, 8351 relationships were >0.25 (excluding self-relationships), whereas in **G** this number was 18 573. Because **A** already contains numerous high relationships, apparently the added benefit of additional high relationships in **G** was limited, resulting in marginal differences between accuracies obtained from pedigree and genomic analyses. Nevertheless, in most practical cases, the family links between animals are expected to be weaker than in the data used here, and thus the differences between genomic and pedigree relationships are expected to be higher. In such cases, single- and multi-trait genomic analyses, which use more accurate relationships, are still expected to be superior over pedigree analyses. For example, Aguilar *et al.* (2011) showed an increase in accuracy of conception rate when enriching **A** with **G** and an additional substantial increase when multi-trait approach was used instead of the single-trait one. Tsuruta *et al.* (2011), who analysed linear type traits of US Holsteins, showed that, in terms of accuracy, the benefit of genomic analyses instead of pedigree was persistent across all analysed traits. Another explanation of small differences in accuracy between multi-trait pedigree and genomic scenarios is that recording own phenotypes for predictor traits on the evaluated animals gives information about the Mendelian sampling in the evaluated animals. Explaining Mendelian sampling, albeit through the genomic information, is also the most important reason why genomic selection is superior to selection using pedigree-based breeding values. This explains why, in our study there is a limited benefit to using genomic information, when predictor traits are already included in the model. Whether or not the use of genomic information, on top of using predictor traits, leads to an increase in the accuracy of breeding values depends predominantly on the genetic correlation between the predictor traits and the trait of interest (Calus and Veerkamp, 2011c).

Bias

Scenarios in which no phenotypic observations were available for the evaluated animals showed bias (Table 5.6). This observed bias appears to be the result of an incorrect scale of the variance of the estimated breeding values. Inclusion of additional trait(s) recorded on the evaluated animals caused a substantial increase in variance of the estimated breeding values and an even higher increase in co-variance between estimated breeding values and phenotypes (results not shown). Together, the changed variance of the estimated breeding values and co-variance with the phenotypes appeared to remarkably reduce the bias. This bias reduction is most likely owing to the fact that the availability of the observation(s) for the additional trait(s) helped to account for the potential heterogeneity of the data, which may have resulted in multivariate distribution of the data. The analysed data set was heterogeneous in the sense that animals originated from different countries and management systems. These factors were accounted for during pre-correction of the phenotypes with a random regression model. Using random curves for each animal should, to some extent, adjust for differences in residual variances across groups. However, it is still possible that some of the data heterogeneity was not completely removed, for example, owing to different selection strategies in particular countries or herds. Thus, differences between groups may still have been present in the data set. To additionally account for differences within the groups, the phenotypes could be adjusted such that the mean (μ) is 0 and the standard deviation (σ) is 1 within each group. This possibility was investigated by De Haas *et al.* (2012b). In their paper, despite the adjustment, accuracies between groups differed substantially, suggesting that such adjustment in fact may be too rigorous, and thus we did not use this additional adjustment here.

Data heterogeneity would not only affect bias in terms of the variance of the estimated breeding values but also the CV-based accuracy of the predicted breeding values. The CV accuracy is expected to be affected by the heterogeneity of the data because it relates estimated breeding values back to the phenotypes. In the case of heterogeneous data, the validation set may differ from the remaining animals and this may affect the bias and accuracy. To investigate whether, indeed in a case of no additional traits recorded for the evaluated animals, the data structure was not properly accounted, all the accuracies were calculated also using PEV. This method does not relate estimated breeding values back to the phenotype, but rather relies on the estimated residual variance, and thus the calculated PEV accuracies are expected to not be strongly affected by the heterogeneity of the phenotypes. Accuracies based on CV deviated from PEV accuracies when no additional observations are recorded for the evaluated animals. This suggests that, despite the pre-correction, the data (possibly) still contain heterogeneity in phenotypes. Using a multi-trait approach in general resulted in less biased estimates than in the case of the single-trait analyses, especially when additional information was available for the evaluated animals. The use of additional traits could help to reduce bias in two ways. First, by having the own phenotypes of the evaluated animals, information on their Mendelian sampling is available, as discussed previously. This helps to reduce the bias, because it enables the model to more accurately predict the ‘level’ of the breeding value of an evaluated animal, relative to the reference animals. Second, unaccounted differences between evaluated animals and the reference population can be accounted in the analysis by introducing information on the animals’ own phenotype(s) for the predictor trait(s) recorded on the evaluated animals. Thus, breeding values for the target trait are less regressed to the mean of the reference population. This regression, instead, is restricted to a certain range specified by the observed value of their own phenotype(s) for the predictor trait(s) and its correlation with the target trait. When the additional traits are recorded only for the reference population, no additional information on Mendelian sampling for the evaluated animals is provided and the regression of their breeding values for the target trait is not restricted by the own phenotype on the predictor trait(s). The additional information on Mendelian sampling for the evaluated animals and the restriction of the regression of their breeding values for the target trait resulted also in an increase of the breeding value accuracy when all the animals were phenotyped for the predictor traits. Therefore, using a multi-trait approach with additional information for all the animals allows the bias of the evaluation to be reduced by utilizing all available information.

An additional bias, which may be present in our multi-trait scenarios, can be caused by the fact that the own phenotypes for the predictor traits of evaluated animals are used in the evaluation. This bias may arise when residuals of the measured trait(s) are correlated with the evaluated animals' estimates for the target trait. If such a kind of bias was present here, CV and PEV accuracies would differ for the multi-trait scenarios, as the CV method relates back to the phenotypes, whereas PEV does not. In our study, for the scenarios in which additional traits were recorded for the evaluated animals, the differences between CV and PEV accuracies were, however, small. Thus, it appears that this type of bias, if present, is limited in the analysed data set.

Genetic parameters

The estimated heritabilities and the genetic correlations were similar to the results of Veerkamp *et al.* (2011) who used the Dutch part of the data for the analyses. The correlations between DMI, FPCM and LW obtained here, were similar to some and contradictory to other studies shown in a review of Veerkamp (1998) who attributed the differences between the results to the small sample sizes and differences in traits definitions. The same argument applies to the current study in which the estimated genetic correlations were in general higher than in the literature. Another reason of differences between the values of the genetic correlations can be differences in milk production between the groups. Some of the animals included in the analyses were bred to average, whereas other to elite bulls. Although the differences between the groups were, to some extent, accounted while pre-correcting the data, still some differences may be unaccounted. These differences may have influenced the values of genetic parameters leading to differences observed between our results and reported in other studies.

A comparison between values of the genetic parameters estimated using pedigree and genomic analyses showed that the genomic-based values are lower (see Table 5.3). This agrees with the indications of Veerkamp *et al.* (2011) who partly analysed the same data as in here. Veerkamp *et al.* (2011) indicated that differences between genomic-based and pedigree-based estimated heritabilities not only may be the result of differences in properties of the **A** v. **G** matrix, but also because G-REML may be better able to disentangle genetic and environmental effects because of using more precise relationships between animals (Lee *et al.*, 2010). Another important question is whether pedigree-based or genomic-based genetic parameters are more precise. Veerkamp *et al.* (2011) used two methodologies to estimate errors of the estimated genetic parameters. One method was approximating the error of the estimates using predicted error variances (as done also here) and the second method used a bootstrapping procedure. Whereas the first method, as in Table 5.3, showed that standard errors of pedigree-based and genomic-based are similar, the bootstrapping revealed lower standard errors for genomic-based estimates. Lower values of the genomic-based estimates of the genetic parameters, given the lower standard errors, should be closer to the true values. Such conclusion is especially valid when the data set used for estimating the genetic parameters is small and then genomic-based parameters should be used instead of the pedigree-based ones.

Optimal selection strategy for scarcely recorded traits

This study showed how DMI breeding value accuracy can be increased using predictor traits. To further increase the accuracy, several strategies can be considered. One solution could be to use more predictor traits. As shown here, three-trait analyses were more accurate than the two-trait ones, and therefore adding even more traits, which have the characteristics of a predictor trait, could further increase the accuracy. An additional strategy, when phenotyping is very difficult, is to extend the reference population by an international collaboration (de Haas *et al.*, 2012b). A similar approach is to add a larger bull reference population with more accurate phenotypes or breeding values for the predictor traits, which appears to be a successful strategy (Calus *et al.*, 2013a). Yet another approach could be to increase the reference populations for the predictor traits, while leaving the reference population for DMI unchanged. This approach, although not investigated here, is expected to be successful when the size of the reference population for the predictor traits is much larger than for the predicted trait(s) and when the genetic correlations between the predictor traits and the trait of interest are well known. In our study, the most easily accessible predictor trait was FPCM; however, it may be insufficient to use only this trait because of its relatively weak genetic correlation with DMI. Moreover, strong selection on milk production is present for many years in dairy cattle, and thus DMI is already being changed indirectly together with milk production. Therefore, the additional benefit of using FPCM as a predictor trait may be limited. Increasing the reference population for LW, however, should be more beneficial as its genetic correlation with DMI is higher. In situations with higher correlations, as shown by Calus and Veerkamp (2011c), the increase in the accuracy would be considerably higher than with weakly correlated predictor traits such as FPCM.

Combining the aforementioned approaches (i.e. using more predictor traits measured on an additional reference population of bulls with accurate phenotypic records on daughters, international collaboration to increase the reference population for the trait of interest and using multi-trait analyses) seems to be the most promising solution and could potentially yield a substantial increase in the breeding value accuracy compared with the levels obtained in this study.

5.6 Conclusion

This study showed that multi-trait genomic selection is more accurate than single-trait applications when phenotypic observations for the predictor trait(s) are recorded for both the reference population and the evaluated animals. Recording predictor traits for the reference population only did not result in an increase in accuracy for the predicted trait. LW was indicated to be a better predictor trait for DMI than FPCM. The initial benefit of using genomic instead of pedigree relationships obtained in single-trait analyses was reduced when increasingly more information from predictor traits was used in multi-trait analyses. This is because both the analyses using predictor traits and the analyses using genomic information attempted to increase the accuracy by explaining Mendelian sampling terms. The highest accuracy was achieved for multi-trait analyses with information on both predictor traits. When including the predictor trait(s) recorded for the reference and evaluated animals, the estimates were unbiased, most likely because this information helped to account for heterogeneity of the data. Therefore, using predictor traits in multi-trait genomic approach may be an inexpensive way to significantly increase the accuracy and to obtain unbiased breeding values for evaluation of scarcely recorded traits, which are expensive or difficult to measure, such as feed intake.

5.7 Acknowledgments

Marcin Pszczola acknowledges the financial support of the Foundation for Polish Science (FNP START 2013), of the Koepon Stichting (Leusden, The Netherlands) and of GreenHouseMilk. The GreenHouseMilk project is financially supported by the European Commission under the Seventh Research Framework Programme, Grant Agreement KBBE-238562. The RobustMilk project is gratefully acknowledged for providing the data, in particular, Donagh Berry (Teagasc Moorepark, Fermoy, Co. Cork, Ireland) and Mike Coffey (Scottish Rural College (SRUC), Edinburgh, UK, data collection funded by the Scottish Government). John Bastiaansen (Wageningen University, Wageningen, The Netherlands) is thanked for the initial editing of the genotype data. The RobustMilk project is financially supported by the European Commission under the Seventh Research Framework Programme, Grant Agreement KBBE-211708. This publication represents the views of the authors, not the European Commission, and the Commission is not liable for any use that may be made of the information.

5.6 Appendix

Table A5.1 Spearman rank correlations[a] between pedigree-based breeding values for DMI of evaluated animals for which phenotypic observations were available on different number of traits across reference populations with phenotypes available for different traits

Traits recorded on evaluated population		NO				FPCM		LW	
	Traits recorded on reference population	DMI	DMI and FPCM	DMI and LW	DMI, FPCM and LW	DMI and FPCM	DMI, FPCM and LW	DMI and LW	DMI, FPCM and LW
NO	DMI and FPCM	0.95							
	DMI and LW	0.92	0.89						
	DMI, FPCM and LW	0.89	0.94	0.95					
FPCM	DMI and FPCM	0.81	0.84	0.75	0.78				
	DMI, FPCM and LW	0.76	0.79	0.80	0.83	0.95			
LW	DMI and LW	0.61	0.61	0.67	0.66	0.58	0.63		
	DMI, FPCM and LW	0.58	0.63	0.64	0.69	0.59	0.65	0.98	
FPCM and LW	DMI, FPCM and LW	0.59	0.64	0.64	0.67	0.69	0.73	0.96	0.97

FPCM = Fat-protein-corrected milk; LW = Live weight; DMI = Dry matter intake; NO = No phenotypes available.

[a] Standard errors across cross-validation sets ranged from 0.002 to 0.052.

Table A5.2 Spearman rank correlations[a] between genomic-based breeding values for DMI of evaluated animals for which phenotypic observations were available on different number of traits across reference populations with phenotypes available for different traits.

Traits recorded on evaluated population		NO				FPCM		LW	
	Traits recorded on reference population	DMI	DMI and FPCM	DMI and LW	DMI, FPCM and LW	DMI and FPCM	DMI, FPCM and LW	DMI and LW	DMI, FPCM and LW
NO	DMI and FPCM	0.96							
	DMI and LW	0.94	0.92						
	DMI, FPCM and LW	0.89	0.95	0.96					
FPCM	DMI and FPCM	0.90	0.93	0.85	0.87				
	DMI, FPCM and LW	0.86	0.90	0.91	0.94	0.95			
LW	DMI and LW	0.78	0.77	0.82	0.81	0.76	0.80		
	DMI, FPCM and LW	0.74	0.78	0.79	0.82	0.76	0.81	0.98	
FPCM and LW	DMI, FPCM and LW	0.73	0.77	0.78	0.81	0.80	0.85	0.97	0.99

FPCM = Fat-protein-corrected milk; LW = Live weight; DMI = Dry matter intake; NO = No phenotypes available.

[a] Standard errors across cross-validation sets ranged from 0.002 to 0.052.

6

General discussion

6.1 Introduction

In genomic selection (GS), the requirement of collecting many phenotypes on progeny is relaxed, as opposed to conventional selection. Therefore, GS provides a new opportunity for re-considering genetic improvement of novel traits and to start selecting for them. For GS, the novel traits' phenotypes have to be collected for animals in the reference population (RP) only. Collection of the phenotypes may therefore be conducted, for example, on specialized farms, where collecting expensive or difficult data is feasible. RP for novel traits will be small compared to the routinely recorded traits. The expected limited size of RP for novel traits stresses the need for methods that optimally use all the available information. This thesis aimed to address some of the issues of optimizing GS for novel traits.

The overall objective of this thesis was to investigate different options of optimizing GS for scarcely recorded novel traits. The investigated options were: genotype imputation for ungenotyped but phenotyped animals to be used to enlarge RP; optimization of the RP design with respect to the relationships among the animals included in it; prioritizing genotyping of RP or the selection candidates; and using easily recordable predictor traits to improve the GS accuracy of the novel traits.

The general discussion will focus on several aspects related to the genetic improvement of novel traits. First, the importance of female RP will be discussed. Next, the expected GS accuracy at different project budgets will be derived. Finally, foresight into the future challenges for selection for novel traits will be given.

6.2 Optimizing reference populations for novel traits

Optimal reference population design considering relationships among animals

For novel traits, with expensive measurements, RP will remain small, resulting in low GS accuracy. To achieve the highest possible accuracy with a small RP, animals to be included in it should be chosen in an optimal way. One possibility for optimizing RP for GS is to consider relationships within the RP and between the RP and selection candidates (see, Figure 6.1). Closely related animals partly explain the same part of the genetic variation and therefore, they may also partly have similar phenotypes. When constructing RP, the goal is to capture in it as much of the usable genetic variation present in the whole population as possible. To do so, the animals in RP should be distantly related to each other but at the same time at least somehow be related to the potential selection candidates (as discussed in this thesis). Designing RP in which animals from a single population or breed are distantly related to each other will also lead to a higher average relationship between a potential selection candidate and RP, compared to when RP consists of closely related animals (e.g., from one family). Therefore, if many animals are genotyped, but only few can be phenotyped due to related costs, information on their (genomic) relationships can be utilized to assign animals for phenotyping to such that the level of relationships among them is minimized, while their relationship with the selection candidates is maximized.

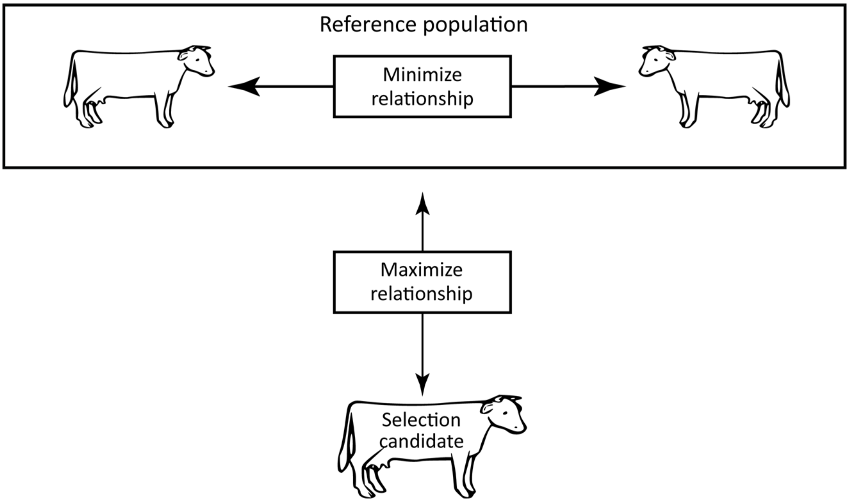


Figure 6.1 Optimal design of the reference population.

Increasing the size of the reference population by combining datasets

To create RP, a sufficient number of phenotypic records are required (Meuwissen *et al.*, 2001; Goddard and Hayes, 2009b; Hayes *et al.*, 2009a). An interesting option for obtaining a sufficient number of records for novel traits is to combine phenotypic records of multiple data sets, for instance, across different experiments, possibly also using phenotypes of historic experiments (see Veerkamp *et al.*, 2010; Berry *et al.*, 2012; Veerkamp *et al.*, 2012). To use historic phenotypes, especially together with new phenotypes (i.e., collected during recent experiments), several aspects should be taken into a consideration (Banos *et al.*, 2012). First, trait definitions have to be compared across the data sets and made compatible. Second, the data sets should be corrected for the systematic environmental factors affecting the phenotypes, especially if they were collected within different experiments. For example, differences in diet composition in nutrition trials and differences in housing conditions in management experiments should be accounted for. Third, the presence of genotype x environment (GxE) interactions (Mathur and Horst, 1994) should be investigated and if present, be accounted for in the model, for instance using multitrait model across environments. Fourth, animals for which the historic phenotypes were obtained may be at a different genetic level than those recorded currently, because of the ongoing selection for other traits genetically related to the trait of interest or different breeding strategies. Differences in genetic levels may be partly accounted for by including a year and an origin effect in the analyses.

If any of the above discussed differences between the data sets are not taken into consideration, combining different data sets carries the risk of obtaining predictions affected by a hidden data structure or data heterogeneity. This hidden data structure may result in prediction equations that predict differences between the data sets (populations or herds) instead of between the animals. The estimated breeding values will then be biased in the sense that their level will be systematically different than the true breeding values. It is important to check the pooled data to verify if the applied adjustments of the phenotypes are adequate and whether no hidden data structures were left unaccounted for. Such data checks could for instance include comparison of the distributions of the final data across the different data sources.

Combining data sets internationally may be an interesting option when only a few hundred phenotypes for a novel trait have been collected within a country; nonetheless, improving the novel trait is a global concern (e.g., reducing methane emission). Then, the data can be put together for the joint benefit of all the participating entities. The adjustment of phenotypes when combining historic and present data sets should also be applied when datasets from different locations, such as herds or countries, are combined (see, for example, Banos *et al.*, 2012). Pooling phenotypic data has been shown to increase the selection accuracy for novel traits through international collaborative projects (de Haas *et al.*, 2012b). For example, a global initiative for collecting data on dry matter intake (gDMI) was undertaken (Pryce *et al.*, 2012b). Also, mitigation of methane emission from dairy cows (de Haas *et al.*, 2012b; http://www.sruc.ac.uk/greenhousemilk) is a good example of international collaboration for improving novel traits. Yet another example was the EU FP7 project RobustMilk, which focused on the robustness of dairy cattle (Berry *et al.*, 2012; www.robustmilk.eu; Veerkamp *et al.*, 2012). The above mentioned projects combined databases of novel traits from different countries, thereby creating an international RP for increasing GS accuracy.

Whether or not combining the data sets internationally is beneficial depends on the genetic connectedness between the countries, which is present if there is an exchange of the breeding material. Such exchange allows for part of the RP from one country to be to some extent related to selection candidates from another country. Genomic data can capture Mendelian sampling and reveal links between animals that are seemingly unrelated through pedigree. Therefore, use of genomic data may somewhat reduce the problem of poor connectedness (as discussed in this thesis), but it will not solve the problem completely. The presence of a relationship between RP and animals in different countries is important for achieving high prediction accuracy for the selection candidates. Therefore, international RP has bigger potential for dairy cattle than for some other species such as sheep, where the exchange of breeding material is not extensive.

Using multi-herd data leads to an increase in the statistical power of the analysis. Using data from a single herd may limit the conclusions to specific conditions or to the farm only (Tempelman, 2009). Combining data internationally provides an opportunity for gathering multi-herd data sets for novel traits, thereby enabling the drawing of inferences and estimating the effects that are applicable across a wide range of herds or herd environments (e.g., in multiple countries).

Increasing size of the reference population by including ungenotyped animals

Genotypes of the reference animals are needed next to their phenotypes to create RP. When historic phenotypes are used to increase the RP size, it is important that DNA samples from animals on which the historic phenotypes were recorded are available. If so, genotyping is possible and RP may be increased by the historic phenotypes relatively easily. For this, no pedigree information is required, as genotypes may be used for establishing the relationships between animals. In some cases, however, DNA samples may not be available. This presents a more challenging situation, because the genotypes then have to be predicted and this does require pedigree data. One way of using ungenotyped animals in genomic evaluations is to merge pedigree-based and genomic relationships into one relationship matrix (Legarra *et al.*, 2009; Aguilar *et al.*, 2010). Although this model avoids imputation of the unobserved genotypes, it is in fact similar to imputing these genotypes by regression on gene content by utilizing pedigree information (Gengler *et al.*, 2008; Christensen and Lund, 2010; this thesis). Alternative ways for imputing genotypes of ungenotyped animals involve the use of algorithms utilizing family relationships and linkage information (see Druet and Georges, 2010; Daetwyler *et al.*, 2011; Hickey *et al.*, 2012). Imputation methods that use only linkage disequilibrium (see Li and Abecasis, 2006; Scheet and Stephens, 2006;

Howie *et al.*, 2009) are not useful for imputing ungenotyped animals, as they require that the imputed animals have at least some genotypes known.

Whichever method is used to impute the genotypes for the ungenotyped animals, it is important that genotypes are imputed accurately, as GS accuracy was shown to depend linearly on imputation accuracy (Mulder *et al.*, 2012). Increasing RP with inaccurately imputed ungenotyped animals will not increase GS accuracy as it was demonstrated in this thesis. To achieve high accuracy of genotype imputation for ungenotyped animals, genotypes of close relatives have to be available. When only parents are genotyped, the genotype imputation accuracy, based on pedigree information only, can reach at maximum 0.707. Therefore, preferably offspring of the ungenotyped animals should be genotyped to achieve the highest possible imputation accuracy. Information of genotyped offspring is important, as it reveals information on Mendelian sampling that otherwise remains unknown. For example, enlarging RP by imputed genotypes of an ungenotyped animal that had only one parent and a maternal-grand sire genotyped hardly increased the GS accuracy for selection candidates, as the imputation accuracy was poor. Calus *et al.* (2011a) and Boettcher *et al.* (2004) reported poor imputation accuracy when only few offspring of imputed individuals were available. When many genotyped offspring are available, the imputation accuracy for ungenotyped animals will be high, leading to an increase in GS accuracy. In the case of ungenotyped animals with historic phenotypes, the imputation accuracy may be low, as parents of these animals are very likely ungenotyped and most probably their progeny is not genotyped either. Therefore, ungenotyped animals with historic phenotypes are very likely not valuable for inclusion to RP. When animals with historic phenotypes were genotyped, but the experiment was undertaken very long time ago, the family links between the RP and potential selection candidates living presently may be very weak. Due to the weak family links, the benefit of including such historic phenotypic data for creating RP may also be limited. Nonetheless, such historic data sets may still be useful for estimating genetic parameters and relationships with other traits.

Females – best source of information for novel traits

As novel traits are unlikely to be recorded on a routine scale, each reference animal will at most have only few phenotyped relatives. The accuracy of breeding values (*r*) depends on the sources of information included in each phenotypic record. If this phenotypic record is based on progeny phenotypes, its accuracy can be calculated as: $r = \sqrt{\frac{{}^{1}/_{4}Nh^2}{1+{}^{1}/_{4}(N-1)h^2}}$, where *N* is the number of phenotyped progeny and h^2 is the trait heritability. The more phenotyped progeny, the higher value of the progeny-based accuracy. The accuracy of a single phenotypic measurement of an animal itself is equal to the square root of the heritability ($\sqrt{h^2}$). In case of a limited number of phenotyped progeny for each reference animal, which is quite likely to be the case for novel traits, progeny-based phenotypic records will be less accurate than own performance records, as shown in Figure 6.2. To achieve the accuracy obtained when each reference animal has a single own phenotypic record, roughly five progeny with phenotypic records are needed for low and moderate heritabilities (≤0.5). To achieve the accuracy of a single own phenotypic record for higher heritabilities (> 0.5), for each reference animal more than five progeny with phenotypic records are needed (e.g., for h^2>0.8 more than 10 phenotyped progeny are needed). This means that the use of progeny-based records requires at least 5 times more phenotypes than the use of own performance. De Roos *et al.* (2011) indicated that use of records based on 100 offspring requires ~15 to 50 times as many records compared to using own phenotypes. Thus, if the total number of phenotypic measurements is restricted, using own records of the reference animals always results in higher GS accuracy than using progeny information for the reference animals (Buch *et al.*, 2011; Van Grevenhof *et al.*, 2012).

Because own phenotype is more valuable than progeny-based phenotype and in dairy cattle many traits are expressed in females, reference populations for novel traits should be composed of females. It is important that the cows included in RP are a 'random sample' of the whole population to achieve maximum variability in relationships between the animals. In practice, however, it may be difficult to sample cows for RP randomly at the population level, because measuring a novel trait is often expensive and possible at only few locations at the time. Therefore, while choosing the animals to be phenotyped and included to RP, assuring the maximum possible variability (low relationship) among the available animals is essential.

Using females in RP for novel traits, instead of bulls with phenotyped offspring, requires more extensive genotyping and low density SNP chips could be used for this. Use of low density chips reduces the costs of genotyping, but may also result in somewhat lower prediction accuracy than higher density chips (see Habier *et al.*, 2009). To estimate the impact of using lower density chips on the predicted accuracy, we calculated the predicted accuracy for different scenarios using an adapted version of the deterministic formula of Daetwyler *et al.* (2008), where $R_G = \sqrt{\frac{N_p h^2}{N_p h^2 + M_e}}$. The reference population size (N_P) was calculated as the overall budget (*B*) divided by the sum of phenotyping (C_{phen}) and genotyping (C_{gen}) costs per animal, so that $N_p = \frac{B}{C_{phen}+C_{gen}}$. The number of independent chromosome segments (M_e) was considered to be 1,000 (Wientjes *et al.*, 2013) and h^2 was the trait heritability. The accuracy R_G was adjusted by the genotype imputation accuracy R_{imp} (from low density to 50k chip) as $R = R_G * R_{imp}$. The predicted GS accuracy was calculated for three heritability levels (0.05, 0.1, 0.6), two chip densities (low density chip and 50k chip) and four accuracies of imputing genotypes from low density to a 50k chip (0.80, 0.90, 0.95, 0.99) for different project budgets. We assumed phenotyping costs of €100, 50k chip genotyping costs (including the sample handling) of €60 and the costs of genotyping at low density (7k chip) was €30. Here, it was assumed that higher density SNP genotypes were already available for facilitating the imputation process; therefore, no additional new costs were required when the reference animals were imputed from low density to 50k.

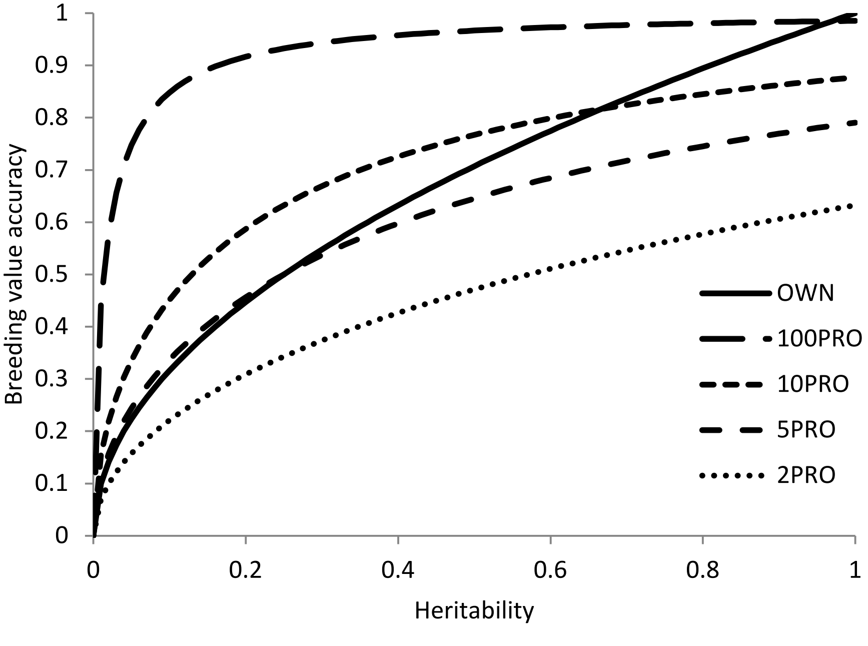


Figure 6.2 Breeding value accuracy based on own performance (OWN) and 2 (2PRO), 5 (5PRO), 10 (10PRO) or 100 (100PRO) progeny with phenotypic records, at different heritability levels.

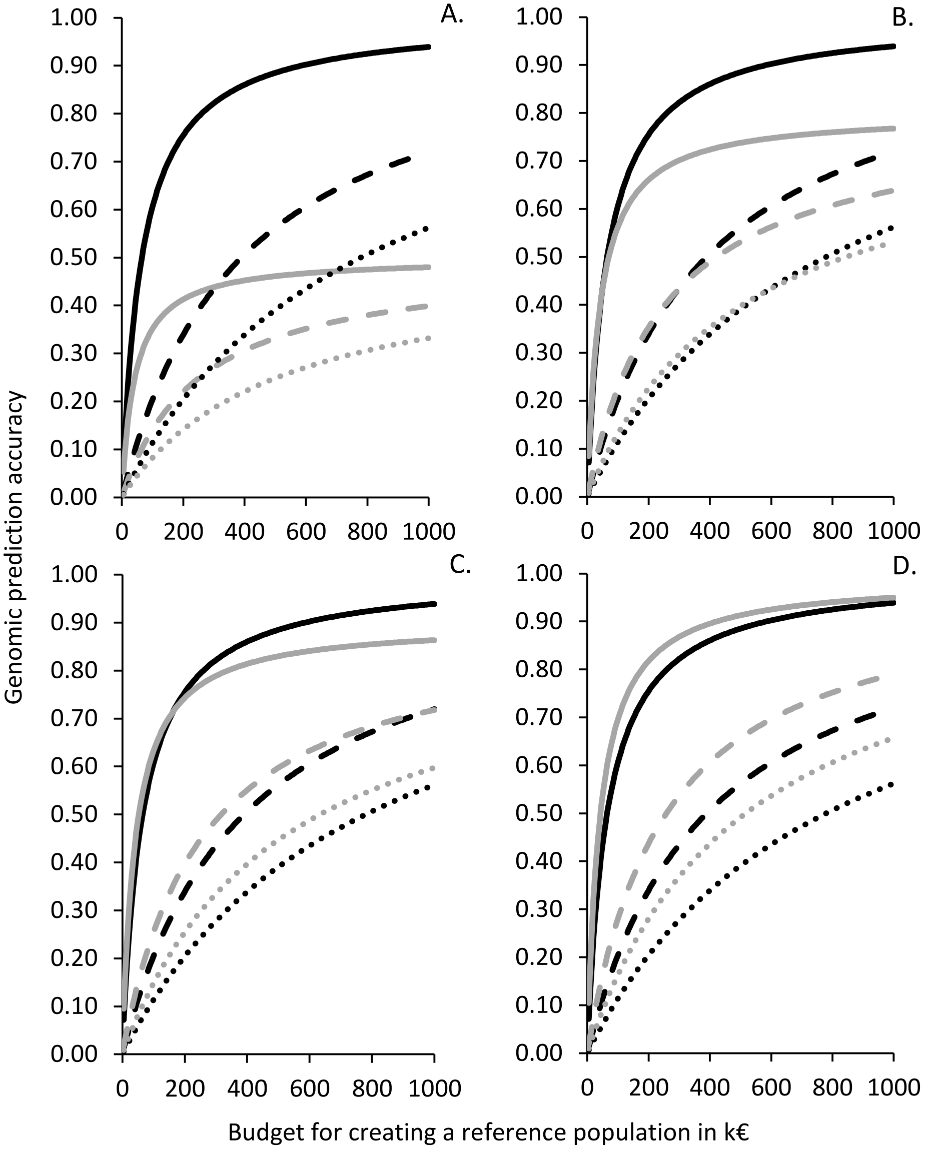


Figure 6.3 Accuracy of genomic predictions at different budgets given for an experiment using 50k chips (black lines) or low density SNP chips (grey lines) for three different heritability levels (solid line = 0.6, dashed line = 0.1 and dotted line = 0.05) and different accuracy of imputation from low to high density (A = 0.80, B = 0.90, C = 0.95,D = 0.99).

The results of the accuracy calculations showed that with sufficient accuracy of imputation from lower to higher density (>0.80), using lower density chips lead to a higher GS accuracy (see Figure 6.3). In practice, the imputation accuracy from low density to higher density may reach from 0.96 to 0.99, given that some close relatives are genotyped at higher density (Berry and Kearney, 2011, Mulder *et al.* 2012). By using a lower density chip with lower overall genotyping costs, more animals can be phenotyped and genotyped, resulting in bigger RP; as a result, higher GS accuracy is achieved. Moreover, in the future, with a further drop in genotyping costs, the phenotyped animals can be genotyped at very high density (e.g., using a 777k chip) or even be sequenced, which can possibly lead to more accurate predictions (Meuwissen *et al.*, 2001). However, at the point where re-genotyping or sequencing would be possible, the phenotyped animals may already be too distantly related to the current population to have a valuable contribution to GS accuracy. The biggest caveat of using imputation is that some animals are required to be genotyped with higher density chips to facilitate the imputation process. Nonetheless, when the imputation accuracy is high, genotyping with low density can be considered as one of the options for increasing GS accuracy for novel traits.

Joint optimization of the reference population design and its relationship with the selection candidates

It was shown in this thesis that the design of RP can be optimized by considering relationships among animals in the RP and between animals in RP and the selection candidates. Optimizing the design of the RP with respect to the relationships across generations may prove challenging. Moreover, RP should be frequently updated with new animals to maintain baseline accuracy (Habier *et al.*, 2007; Wolc *et al.*, 2011). For the novel traits where only a fraction of the population can be phenotyped, only a few animals can be used to enlarge RP.

The animals used for updating RP should be chosen in a manner that they are minimally related to the current RP and maximally related to the potential selection candidates. In the first step, RP has to be created by sampling animals from the whole population. Then, in all later rounds of selection, RP has to be updated. The animals to be used to update RP can be sampled from past selection candidates, where genotypes are already available and the additional cost is the phenotyping; alternatively, RP may be updated from the whole population. Possibly, the animals with the lowest relationship to the RP should be selected first for phenotyping and genotyping (if not yet genotyped). The updated RP will be bigger than in the previous rounds of selection. The average squared relationship between RP and selection candidate depends on the RP size, whereas another measure that approximates the GS accuracy – the sum of squared relationships – is independent from the RP size (Wientjes, personal communication, 2013). Therefore, to maximize the relationship between the updated RP and the potential selection candidates, the sum of the squared relationships between the updated RP and the potential selection candidates should be equal to or larger than in the previous selection round. Ultimately, the GS accuracy after the RP update should be at the level of the previous round of selection, or increased. A conceptual scheme of the optimizing process described above is presented in Figure 6.4.

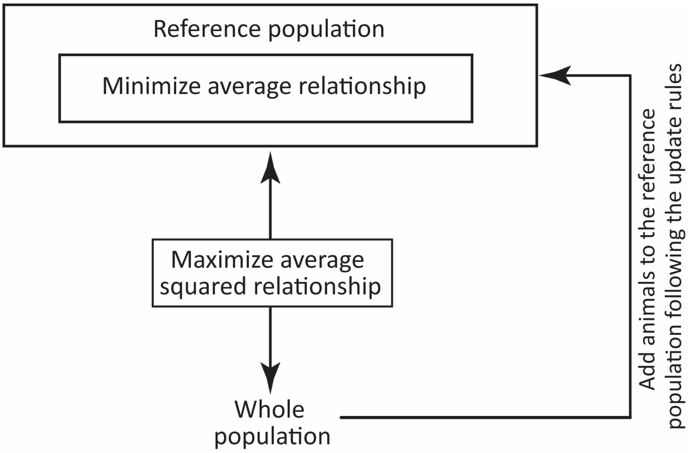

Reference population
Minimize average relationship
Maximize average squared relationship
Whole population
Add animals to the reference population following the update rules

Figure 6.4 A conceptual scheme of the process to optimize the reference population design for a novel trait across generations.

For creating a minimally related RP, the initial step for composing an RP with optimal design may be the use of clustering algorithms. For example, K-means clustering with an inverted genomic relationship matrix can be used to create a distance matrix, which could be employed for selecting the least related individuals. This option was not investigated in this thesis and requires further attention before implementation can take place. Another option is to use methods that minimize mean generalized coefficients of determination or average relationship (Rincent *et al.*, 2012). Yet another solution is to use optimization software utilized for designing breeding schemes for maintaining biodiversity. Such programs, for example, Gencont (Meuwissen, 2002), uses optimal contribution methodology to maximize breeding goal while minimizing inbreeding (i.e., choosing the least related animals as parents of the next generation). This could also be used for selection of lowly related animals to the reference population in later generations.

Use of predictor traits

Because collecting phenotypes of novel traits may prove challenging, using another supporting source of information that is (genetically) related to the novel traits, such as predictor traits, may be beneficial. Predictor trait(s) are required to be easily recordable and genetically related to the trait of interest. The additional information may be recorded on the same animals as the novel trait, or on a separate RP (Calus et al., 2013a). It is always important, however, that predictortraits are also measured on selection candidates (as discussed in this thesis). Calus and Veerkamp (2011c) compared scenarios with different genetic correlations (-0.5, 0, 0.5) between two traits (h2=0.9 and 0.6) recorded on a small RP. They reported that for an absolute value of the genetic correlation >0.5, predictions based on pedigree and with the predictor trait recorded on the selection candidates were more accurate than having only genotypic information on the selection candidates. With a genetic correlation <0.5 and having only genomic information on the selection candidates yielded more accurate predictions than having pedigree and records for the predictor trait on the selection candidates. They therefore concluded phenotyping the selection candidates for the novel trait to be beneficial, provided that the genetic correlation between the predictor and novel trait is sufficiently high. The cut-off point at which using a predictor trait is beneficial depends on the heritabilities of the traits, genetic correlations and the RP size. In Chapter 5 of this thesis, measuring the predictor trait(s) on both groups of animals (i.e., RP and selection candidates) has led to a gain in accuracy of predictions even when the genotypes were not used. These analyses used up to three traits, with heritability ranging from 0.31 to 0.44 and genetic correlations ranging from 0.12 to 0.62. The results showed that when the selection candidates are phenotyped for the predictor trait(s), additional genotyping leads to slightly more accurate predictions (i.e., pedigree-based predictions are only a little less accurate than genomic ones). The small differences between pedigree and genomic-based accuracies were probably due to a very small RP, resulting in low genomic accuracies. With bigger RP, the differences are expected to be bigger. Although it is expected that genotyping the selection candidates – next to phenotyping them for the predictor trait(s) – is advantageous, this increase may be limited if the genetic correlation of the indicator trait with the novel trait is high.

Another possibility for increasing GS accuracy with the use of predictor traits is to expand the cow RP with bull RP. Calus *et al.* (2013b) analyzed this scenario by combining cow and bull RP, where the cows were phenotyped for the novel trait of interest and the bulls had daughter yield deviations available for a trait that had a genetic correlation with the novel trait of 0.59 to 0.65. In this study, adding nearly 300 bulls to a cow reference population consisting of about 1 600 individuals caused only a limited increase in GS accuracy for the novel trait and it was concluded that a bigger set of bulls is needed (e.g., national RP) to achieve a substantial increase in GS accuracy. Although only a limited increase in GS accuracy was observed, enlarging the RP with a small number of bulls with very accurate phenotypes did increase the power for genome-wide association studies (Calus *et al.*, 2012). If the bull reference population with the correlated trait recorded is sufficiently big, then such approach may also be considered as a possible solution for increasing GS accuracy for novel traits.

6.3 Future perspectives for improving novel traits

Farmers routinely record the performance of their animals for various traits, mainly for management purposes. As a by-product, breeding values can be estimated at low cost for these routinely collected traits. Breeding values contribute to genetic improvement of the animals, leading to mutual benefits for both farmers and breeding companies. In case of novel traits, when the costs of phenotyping are high, farmers need to be motivated to participate in data recording, especially when the benefit for farmers is not obvious. The novel traits that have a more obvious benefit in terms of farm management will therefore have a higher chance to be implemented in breeding practice than others. For example, feed efficiency or fertility related traits can be easily linked with profitability of production and can thus be attractive to farmers. Other traits, like methane emission, may be less attractive to farmers. Nonetheless, only engaging farmers can enable the moving of the recording of novel traits from research herds to commercial farms, thereby increasing the possibilities to successfully implement genetic improvement of those novel traits. Several novel traits without a direct link to improving farm management therefore likely require compensation for farmers to collect the data.

Improving novel traits is a challenging issue. The most difficult part of improving novel traits is phenotyping. Obtaining reliable phenotypes for novel traits, such as methane emission or energy balance, is currently not possible on a routine basis. Emphasis therefore needs to be placed on developing cheap and reliable phenotyping technologies, which should be the focus of the breeding industry if it wants to select for novel traits. Promising examples of such technologies include the introduction of measuring methane emission using Fourier transform infrared spectroscopy, which has the potential for being applied on a larger scale as opposed to the gold standard method for measuring methane emission – respiration chambers (Lassen *et al.*, 2012), or using mid-infrared analysis to cheaply predict energy balance (McParland *et al.*, 2011). With decreasing genotyping costs, investments in phenotyping have become more attractive. Before new cheap and reliable phenotyping technologies will be available, however, the number of phenotypic records for novel traits will remain small. As it was discussed above, in such cases, own phenotypic information is more reliable than progeny-based records that are based only on a few individuals. Therefore, to start genomic selection for novel traits in dairy cattle, using a female reference population is currently the most appropriate solution.

Sequencing is being considered as a promising tool for genetic improvement of many traits, as it is believed to allow for the across-breed evaluation or identifying of causal mutations. Currently, this technique is still expensive and to fully benefit from the use of sequencing data, even more phenotypes than for a 50k chip are needed. As RP for the most of the novel traits will remain small in the near future; therefore, the sequence data is unlikely to be a source of additional increase in GS accuracy. In the future, when sequence data will be widely available and phenotyping for novel traits will become cheaper, the suitability of using sequence data for the genetic improvement of novel traits should be re-considered.

The key factors for successful implementation of selection for a novel trait in a breeding scheme are: (1) maximizing accuracy of genotype prediction for ungenotyped animals to be used for updating the reference population; (2) optimizing the design of the reference population; (3) determining easy to record indicator traits that are also available on the selection candidates; (4) developing large scale phenotyping techniques; and (5) establishing strategies and policies for increasing the engagement of farmers in the recording of novel traits.

Summary

Animal breeding aims to genetically improve animal populations by selecting the best individuals as parents of the next generation. Presently, in some countries, dairy cattle breeding goals include up to 40 commonly recorded traits. New traits are being introduced to breeding goals to satisfy new demands faced by livestock production. However, introducing a new trait and starting selection for such a trait may be difficult when the trait is novel in the sense that it has not been widely recorded before. Selecting for such novel traits is especially challenging when recording is laborious and expensive.

Because of laborious and expensive recording, for some novel traits, large scale recording will not be possible in the near future and it may therefore be limited to research herds. Therefore, despite genetic improvement of novel traits is desirable it may be limited due to the small number of observations available. Because of the small number of observations, the novel traits often cannot be improved directly by conventional breeding tools, as these require large numbers of observations, measured on many offspring of each selection candidate. New breeding tools are therefore needed to enable the genetic improvement of novel traits.

Genomic selection is a new breeding tool that uses single-nucleotide polymorphism markers spread across the genome. An important feature of genomic selection, which makes it especially interesting to apply to novel traits, is that the expensive or laborious measurements do not have to be taken on a routine scale. Using the limited available data optimally may, however, require alternative approaches and methodologies than currently used for conventional breeding goal traits.

The overall objective of this thesis was to investigate different options for optimizing genomic selection for scarcely recorded novel traits. The investigated options were: genotype imputation for ungenotyped but phenotyped animals to be used to enlarge the reference population; optimization of the design of the reference population with respect to the relationships among the animals included in it; prioritizing genotyping of the reference population or the selection candidates; and using easily recordable predictor traits to improve the accuracy of breeding values for scarcely recorded traits.

Chapter 2 of this thesis aimed to investigate whether the accuracy of genomic selection can be improved by supplementing a small reference population with ungenotyped but phenotyped animals. A dairy cattle population was simulated from which a reference population was sampled. This reference population consisted of 1,000 phenotyped and genotyped individuals. In the subsequent scenarios, the reference population was supplemented by additional 1,000 ungenotyped or genotyped animals. Genotypes of the ungenotyped animals were predicted based on the genotypes of their relatives and pedigree information. The comparison of the accuracy of breeding values among all the scenarios showed a small increase in accuracy after enlarging the reference population by ungenotyped but phenotyped animals. The increase was, however, limited what was attributed to the low genotype prediction accuracy.

Chapter 3 investigated the impact of different family designs in terms of the relationships within the reference population, as well as the relationship of selection candidates to the reference population on accuracy of genomic selection. A dairy cattle population structure was simulated. Scenarios differed by the level of relationships among the animals in the reference population. Differences in predicted accuracy of breeding values were compared between scenarios. The analyses allowed for determining the optimal design of the reference population by investigating the association between relationship to the reference population and predicted breeding values accuracy. It was demonstrated that the relationship within the reference population should be minimized, while the relationship between reference population and potential selection candidates should be maximized. Average squared relationship between reference population and a selection candidate was shown to be a good proxy for the accuracy of breeding values.

Breeding values estimated using genomic information are more accurate than pedigree based ones. The aim of the study described in Chapter 4, was to investigate whether this increase is mainly due to genotyping reference or evaluated animals. For this purpose, a simulated dataset reflecting a dairy cattle population was used. Four scenarios were considered in which genomic information was available on different groups of animals. The genomic information was available on (1) no animals; (2) reference population; (3) evaluated animals; or (4) reference population and evaluated animals. A comparison of the accuracy of breeding values predicted deterministically for all the scenarios showed that the main gain in accuracy was due to genotyping the selection candidates. Nevertheless, genotyping both reference population and selection candidates is clearly superior, indicating that both categories should be genotyped whenever possible.

In addition to optimizing the reference population with respect to its design or size, predictor traits can be used to increase the accuracy of genomic selection for a novel trait. This option was evaluated using real data in Chapter 5 for a trait recorded on a limited cow reference population. The analyzed scenarios assumed that one or two predictor traits were available on the reference population only, or both on the reference population and the evaluated animals. The novel trait was dry matter intake, while fat-protein-corrected milk yield and live weight were available as predictor traits. The analysis showed that including the predictor traits in the analysis when it is recorded on both reference population and selection candidates can lead to a significant increase in the selection accuracy. When including the predictor traits in the analysis, the added value of using genotypes became less than in single trait analysis.

The general discussion focused on several aspects related to the genetic improvement of novel traits. First, it was demonstrated that females are the most valuable source of information for novel traits in dairy cattle; therefore, reference populations for these traits should consist of females. Consequently, the importance of females as a component of the reference populations in dairy cattle breeding will increase in the future. Second, low density genotyping was shown to be a promising way for increasing genomic selection accuracy. The reason for this is that, thanks to the lower costs of genotyping, more funds can be spent on phenotyping and genotyping additional animals. This in turn results in an increase in the reference population size and by that in the increase of genomic selection accuracy. Finally, the importance of investing in phenotyping technologies and increasing the participation of farmers in the process of data collection was stressed as crucial points for enabling to move collecting of phenotypes for novel traits from research to commercial farms.

The key factors for successful implementation of selection for a novel trait in a breeding scheme are: (1) maximizing accuracy of genotype prediction for ungenotyped animals to be used for updating the reference population; (2) optimizing the design of the reference population; (3) determining easy to record indicator traits that are also available on the selection candidates; (4) developing large scale phenotyping techniques; and (5) establishing strategies and policies for increasing the engagement of farmers in the recording of novel traits.

Samenvatting

Het doel van veefokkerij is het genetisch verbeteren van een populatie dieren door het selecteren van de beste individuen als ouders voor de volgende generatie. Op dit moment worden in de melkveefokkerij in sommige landen standaard tot 40 kenmerken gemeten en meegenomen in het fokdoel. Om aan de nieuwe vragen en eisen te voldoen die aan de veehouderij worden gesteld, worden nieuwe kenmerken toegevoegd aan het fokdoel. Het invoeren van een nieuw kenmerk kan moeilijkheden geven wanneer dit voorheen niet gemeten werd. Dit is met name het geval wanneer het meten van een kenmerk arbeidsintensief en duur is.

Wanneer het meten van nieuwe kenmerken arbeidsintensief en duur is, is het in de nabije toekomst niet mogelijk om het kenmerk op grote schaal te meten. Het meten blijft dan beperkt tot de zogenaamde onderzoeksbedrijven en het aantal observaties blijft laag. De genetische vooruitgang van een kenmerk kan hierdoor beperkt blijven, ondanks dat vooruitgang wenselijk is. De huidige fokkerij-technieken zijn namelijk gebaseerd op grote aantallen observaties, gemeten aan veel nakomelingen van een selectie kandidaat. Om genetische vooruitgang op nieuwe kenmerken mogelijk te maken zijn nieuwe fokkerij-technieken nodig.

Genomic selection is een nieuwe techniek dat gebruik maakt van merkers, verspreid over het genoom. Een belangrijke eigenschap van genomic selection is dat arbeidsintensieve en dure metingen niet routinematig gedaan hoeven worden, hierdoor is deze techniek extra interessant voor nieuwe kenmerken. Voor een optimaal gebruik van het geringe aantal observaties, zijn er alternatieve methodes nodig vergeleken met de methodes die momenteel gebruikt worden voor de kenmerken in het fokdoel.

Het algemene doel van dit proefschrift was om verschillende opties te bekijken om genomic selection te optimaliseren voor nieuwe kenmerken met een gering aantal observaties. De bekeken opties waren: vergroten van de referentie populatie door genotypes te voorspellen van dieren zonder genotype maar met fenotype; optimaliseren van de samenstelling van de referentie populatie met betrekking tot de relaties tussen de dieren daarin; prioriteren van het genotyperen van dieren in de referentie populatie versus selectie kandidaten; en het gebruik van makkelijk meetbare voorspellers om de betrouwbaarheid van fokwaardes voor kenmerken met een gering aantal observaties te vergroten.

Hoofdstuk 2 van dit proefschrift had als doel om te onderzoeken of de betrouwbaarheid van genomic selection kan worden verhoogd door het toevoegen van dieren zonder genotypes maar met fenotypes aan een kleine referentie populatie. Hiervoor was een melkveepopulatie gesimuleerd, waaruit een referentie

populatie van 1.000 dieren met genotype en fenotype was gevormd. In de onderzochte scenario's werd deze referentie populatie aangevuld met 1.000 dieren met of zonder genotype. Van de dieren zonder genotype werd het genotype voorspeld op basis van de genotypes van familieleden en stamboom informatie. De betrouwbaarheid van de fokwaardes steeg licht nadat de referentie populatie was vergroot door dieren met fenotype maar zonder genotype. De stijging was echter beperkt, wat veroorzaakt werd door de lage betrouwbaarheid waarmee genotypes voorspeld werden.

In Hoofdstuk 3 werd de invloed van verschillende familie samenstellingen in termen van relaties in de referentie populatie, als ook de relatie van selectie kandidaten met de referentie populatie, op de betrouwbaarheid van genomic selection onderzocht. Hiervoor was de structuur van een melkveepopulatie gesimuleerd. De verschillende scenario's verschilden in het niveau van de relaties tussen de dieren in de referentie populatie. De betrouwbaarheid van de fokwaardes werd per scenario voorspeld en onderling vergeleken. Deze analyses onderzochten het verband tussen de relatie met de referentie populatie en de voorspelde betrouwbaarheid van de fokwaardes, om zodoende het optimale design voor de referentie populatie te bepalen. De resultaten geven aan dat de relaties in de referentie populatie geminimaliseerd moeten worden en de relaties tussen de referentie populatie en de potentiële selectie kandidaat gemaximaliseerd. De gemiddelde gekwadrateerde relatie tussen de referentie populatie en een selectie kandidaat was een goede voorspeller voor de betrouwbaarheid van de fokwaardes.

Fokwaardes voorspeld met genomic selection zijn nauwkeuriger dan fokwaardes gebaseerd op stamboom informatie. Het doel van de studie in Hoofdstuk 4 was om te onderzoeken of deze toename met name veroorzaakt wordt door het genotyperen van dieren in de referentie populatie of de selectie kandidaten. Hiervoor was een gesimuleerde dataset gebruikt, welke een melkveepopulatie beschrijft. Er waren vier verschillende scenario's gebruikt waarin genomische informatie beschikbaar was voor een andere groep dieren. De genomische informatie was beschikbaar voor (1) geen dieren; (2) referentie populatie; (3) selectie kandidaten; (4) referentie populatie en selectie kandidaten. Voor ieder scenario was de betrouwbaarheid van de fokwaardes deterministisch voorspeld en de grootste stijging in betrouwbaarheid werd veroorzaakt door het genotyperen van de selectie kandidaten. Het genotyperen van zowel de referentie populatie als de selectie kandidaten was echter duidelijk superieur, wat aangeeft dat beide groepen gegenotypeerd zouden moeten worden indien mogelijk.

Naast het optimaliseren van de referentie populatie met betrekking tot het ontwerp en de grootte, kan het gebruik van andere kenmerken als voorspellers de betrouwbaarheid voor nieuwe kenmerken verhogen. Deze optie werd bestudeerd in Hoofdstuk 5, waarvoor echte data van een kleine koeienpopulatie werd gebruikt. De bestudeerde scenario's namen aan dat één of twee voorspellers beschikbaar waren voor alleen de referentie populatie of voor zowel de referentie populatie en de selectie kandidaten. Droge stof inname werd gebruikt als nieuw kenmerk, vet en eiwit gecorrigeerde melkproductie en gewicht werden gebruikt als voorspellers. De analyses gaven aan dat het meenemen van voorspellers gemeten in zowel de referentie populatie als de selectie kandidaten kan leiden tot een significante stijging in de betrouwbaarheid van de selectie. Wanneer voorspellers werden gebruikt in de analyse, werd de toegevoegde waarde van het gebruik van genotypes lager dan in een analyse met één kenmerk.

De algemene discussie gaat in op verschillende aspecten gerelateerd aan de genetische vooruitgang van nieuwe kenmerken. Als eerste werd aangetoond dat koeien de belangrijkste informatiebron zijn voor nieuwe kenmerken in melkvee, en dat referentie populaties van deze kenmerken dus uit koeien zouden moeten bestaan. Dit benadrukt het toenemende belang van het opnemen van koeien in referentie populaties in de toekomst. Als tweede werd aangetoond dat genotyperen door middel van goedkopere chips met een lage merker dichtheid een veelbelovende manier is om de betrouwbaarheid van genomic selection te verhogen. De reden hiervoor is dat meer geld kan worden besteed aan het genotyperen en meten van kenmerken van extra dieren. Dit heeft een grotere referentie populatie, en daarmee een stijging in de betrouwbaarheid van genomic selection, als gevolg. Tot slot werd het belang om te investeren in meettechnieken en het vergroten van de deelname van boeren in het proces van dataverzameling onderstreept als cruciale punten om het verzamelen van observaties voor nieuwe kenmerken te verplaatsen van onderzoeksbedrijven naar commerciële bedrijven.

De belangrijkste factoren voor het succesvol implementeren van een nieuw kenmerk in het fokdoel zijn (1) maximaliseren van de betrouwbaarheid om genotypes te voorspellen van dieren zonder genotype voor het vergroten van de referentie populatie; (2) optimaliseren van de samenstelling van de referentie populatie; (3) definiëren van makkelijk meetbare voorspellers welke beschikbaar zijn voor de selectie kandidaten; (4) ontwikkelen van meettechnieken welke op grote schaal toegepast kunnen worden; en (5) instellen van strategieën en beleid om de betrokkenheid van boeren in het meten van nieuwe kenmerken te vergroten.

Streszczenie

Genetyczne doskonalenie zwierząt odbywa się przede wszystkim na drodze selekcji. Kierunek doskonalenia wyznacza cel hodowlany, który stanowią cechy i przypisane im wagi. W przypadku bydła mlecznego, w niektórych krajach doskonali się jednocześnie nawet ponad 40 cech, dla których obserwacje gromadzone są w populacji aktywnej, objętej kontrolą użytkowości. Wzbogacanie celów hodowlanych o nowe cechy wynika głównie ze zmian w hodowli, powodowanych dostosowywaniem się do nowych warunków rynkowych i chęcią poprawy ekonomiki produkcji, konsekwencji występowania korelacji genetycznych pomiędzy cechami czy dążeniem do ograniczenia negatywnego wpływu produkcji zwierzęcej na środowisko naturalne. Uwzględnienie nowej cechy w hodowli nie jest łatwe, szczególnie gdy nie jest ona objęta kontrolą użytkowości, a jej pomiar jest kosztowny i pracochłonny. Niewielka liczba obserwacji zgromadzona od małej populacji przekłada się na niską dokładność oceny i ograniczoną możliwość uzyskania pożądanego postępu genetycznego przy wykorzystaniu selekcji tradycyjnej. U bydła mlecznego postęp genetyczny opierał się na selekcji buhajów, do oceny których konieczne było pozyskanie obserwacji od wielu córek, co było procesem kosztownym i czasochłonnym. Stąd aktualne jest poszukiwanie rozwiązań i narzędzi, które umożliwią doskonalenie nowych cech.

Selekcja genomowa otwiera nowe możliwości w doskonaleniu zwierząt. Wykorzystuje markery podstawień jednonukleotydowych (SNP) zebrane na panelach SNP. Markery SNP to dodatkowe źródło informacji, pozwalające przeprowadzić ocenę wartości hodowlanej. W selekcji genomowej kontrolą użytkowości obejmuje się grupę zwierząt zwaną populacją referencyjną. Większa populacja referencyjna przekłada się na wyższą dokładność oceny. Osobniki oceniane nie muszą posiadać informacji fenotypowej, stąd możliwe jest przyspieszenie selekcji młodych osobników. Selekcja genomowa pozwala na bardziej efektywne wykorzystanie ograniczonej liczby obserwacji, co otwiera możliwości doskonalenia nowych cech o trudnych pomiarach. Jako stosunkowo nowa metoda, wymaga ona poznania mechanizmów, na których się opiera oraz poszukiwania sposobów jej optymalizacji.

Głównym celem badań przedstawionych w niniejszej pracy było poszukiwanie możliwości optymalizacji selekcji genomowej dla doskonalenia nowych cech, dla których dostępna jest niewielka liczba obserwacji. W kolejnych rozdziałach pracy rozważono wpływ użycia czterech metod optymalizacyjnych na dokładność oceny wartości hodowlanych w selekcji genomowej. Przeanalizowane rozwiązania to: (1) powiększenie populacji referencyjnej o osobniki niezgenotypowane, których genotypy zostały oszacowane na podstawie informacji rodzinowej; (2) optymalizacja struktury populacji referencyjnej pod kątem spokrewnienia zwierząt ją tworzących; (3) genotypowanie tylko osobników w populacji referencyjnej lub kandydatów selekcyjnych; oraz (4) wykorzystanie cech wskaźnikowych o łatwych pomiarach.

Pierwszy rozdział wprowadza do tematyki selekcji genomowej ze szczególnym naciskiem na aspekty związane z wykorzystaniem jej do doskonalenia nowych cech. W drugim rozdziale pracy, w oparciu o badania symulacyjne, przedstawiono możliwość poprawy dokładności selekcji genomowej poprzez powiększenie populacji referencyjnej o zwierzęta niezgenotypowane. Osobniki wchodzące w skład populacji referencyjnej zostały wybrane losowo spośród wszystkich zwierząt. Populacja referencyjna składała się z 1 000 zgenotypowanych zwierząt. W dalszych wariantach, populacja ta została powiększona o 1 000 zgenotypowanych lub 1 000 niezgenotypowanych zwierząt. Do predykcji genotypów wykorzystano informację markerową od krewnych oraz informację rodzinową. Wykazano niewielki wzrost dokładności oceny, uzyskany dzięki powiększeniu populacji referencyjnej o zwierzęta niezgenotypowane. Ten stosunkowo niewielki wzrost dokładności wynikał prawdopodobnie z niskiej dokładności predykcji genotypów.

W trzecim rozdziale oceniono wpływ spokrewnienia zwierząt w ramach populacji referencyjnej oraz pomiędzy tą populacją, a kandydatami selekcyjnymi na dokładność selekcji genomowej. W tym celu przeprowadzono badania symulacyjne odwzorowujące strukturę populacji charakterystyczną dla bydła mlecznego. Przeprowadzone symulacje różniły się poziomem spokrewnienia zwierząt włączonych do populacji referencyjnej. Obliczono dokładność oceny dla kandydatów selekcyjnych charakteryzujących się różnym stopniem spokrewnienia z populacją referencyjną. Analizy pozwoliły na określenie optymalnej struktury rodzinowej tej populacji. Najwyższą dokładność oceny uzyskano dla populacji referencyjnej o najniższym średnim spokrewnieniu oraz dla osobników silnie spokrewnionych z tą populacją. Maksymalizacja dokładności oceny wymaga więc minimalizowania spokrewnienia w ramach populacji referencyjnej oraz maksymalizowania spokrewnienia tej populacji z potencjalnymi kandydatami selekcyjnymi. Dodatkowo dowiedziono, że średni kwadrat spokrewnień pomiędzy populacją referencyjną, a kandydatem selekcyjnym jest dobrym indykatorem dokładności oceny wartości hodowlanej.

W czwartym rozdziale pracy starano się odpowiedzieć na pytanie, czy wzrost dokładności oceny przy przejściu z selekcji tradycyjnej na genomową wynika głównie z wykorzystania informacji genotypowej dla populacji referencyjnej, czy może dla osobników ocenianych. W oparciu o dane symulacyjne przeanalizowano cztery warianty dostępności informacji genomowej, zakładając jej brak lub obecność dla: populacji referencyjnej, osobników ocenianych lub obu tych grup. Porównanie oszacowanych wartości hodowlanych wykazało, iż dla wzrostu dokładności w selekcji genomowej ważniejszy jest genotyp zwierząt ocenianych, aniżeli populacji referencyjnej. Niemniej jednak, zdecydowany wzrost dokładności oceny wartości hodowlanej w selekcji genomowej w porównaniu do tradycyjnej, został osiągnięty tylko gdy zgenotypowane były obydwie grupy zwierząt.

Przyjęto, że obok optymalizacji rozmiaru oraz struktury populacji referencyjnej, dokładność selekcji genomowej dla cech o trudnych pomiarach, może zostać zwiększona, dzięki wykorzystaniu cech wskaźnikowych. Teza ta została zweryfikowana w rozdziale piątym niniejszej pracy. Wykorzystany zbiór danych składał się z obserwacji dokonanych na krowach. Użyto jednej lub dwóch cech wskaźnikowych, które były dostępne tylko dla populacji referencyjnej lub zarówno dla populacji referencyjnej jak i zwierząt ocenianych. Nową trudną w ocenie cechą było pobranie suchej masy, a cechami wskaźnikowymi: wydajność mleczna poprawiona o zawartość tłuszczu i białka oraz waga przyżyciowa. Analizy pozwoliły określić, iż użycie cech wskaźnikowych dostępnych zarówno dla populacji referencyjnej jak i zwierząt ocenianych może prowadzić do znacznego wzrostu dokładności oceny. Dodatkowo wykazano, że przy wykorzystaniu informacji o cechach wskaźnikowych przewaga selekcji genomowej nad selekcją tradycyjną jest mniejsza niż przy braku wykorzystania tych cech.

W dyskusji skupiono się na kilku aspektach związanych z doskonaleniem nowych cech o trudnych pomiarach. Po pierwsze, wykazano iż samice to najcenniejsze źródło informacji fenotypowej dla tych cech, a zatem populacje referencyjne dla nich tworzone powinny się składać z samic posiadających obserwacje własne. Można więc przewidywać wzrost znaczenia informacji pozyskanych od samic w doskonaleniu bydła mlecznego. Następnie, wykazano, iż genotypowanie z wykorzystaniem paneli SNP o mniejszej gęstości (tańszych w zastosowaniu) może być skuteczną metodą poprawy dokładności selekcji, pozwala to bowiem przeznaczyć środki zaoszczędzone na genotypowaniu na pozyskanie większej liczby obserwacji fenotypowych prowadząc do powiększenia populacji referencyjnej. Wskazano również na konieczność przeprowadzenia inwestycji mających na celu opracowanie nowoczesnych technologii pomiarowych oraz potrzebę zwiększenia zaangażowania hodowców w proces zbierania informacji, by zapewnić szerszą pulę pozyskanych obserwacji dla nowych cech.

W podsumowaniu, za kluczowe czynniki pozwalające na skuteczne wdrożenie nowych cech do nowoczesnych programów hodowlanych uznano: (1) maksymalizację predykcji genotypów niezgenotypowanych zwierząt mających powiększyć populację referencyjną; (2) optymalizację struktury populacji referencyjnej; (3) poszukiwanie cech wskaźnikowych o łatwych pomiarach, dostępnych również dla osobników ocenianych; (4) opracowanie technik umożliwiających rutynowe pomiary nowych obserwacji fenotypowych; oraz (5) zwiększenie zaangażowania hodowców w proces pozyskiwania informacji.

References

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

Aguilar, I., I. Misztal, S. Tsuruta, G.R. Wiggans and T. J. Lawlor. 2011. Multiple trait genomic evaluation of conception rate in Holsteins. J. Dairy Sci. 94:2621-2624.

Balding, D.J. 2006. A tutorial on statistical methods for population association studies. Nat. Rev. Genet. 7:781-791.

Banos, G. 2010. Past, present and future of international genetic evaluations of dairy bulls. Proc. 9th World Congress on Genetics Applied to Livestock Production (WCGALP), Leipzig, Germany.

Banos, G., M.P. Coffey, R.F. Veerkamp, D.P. Berry, and E. Wall. 2012. Merging and characterising phenotypic data on conventional and rare traits from dairy cattle experimental resources in three countries. Animal 6:1040-1048.

Bell, M.J., E. Wall, G. Russell, G. Simm, and A.W. Stott. 2011. The effect of improving cow productivity, fertility, and longevity on the global warming potential of dairy systems. J. Dairy Sci. 94:3662-3678.

Berry, D.P., J.F. Kearney, and B.L. Harris. 2009. Genomic selection in Ireland. Interbull Bull. 39:29-34.

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

Berry, D.P., J.W.M. Bastiaansen, R. F. Veerkamp, S. Wijga, E. Wall, B. Berglund, and M.P.L. Calus. 2012. Genome-wide associations for fertility traits in Holstein-Friesian dairy cows using data from experimental research herds in four European countries. Animal 6:1206-1215.

Boettcher, P.J., G. Pagnacco, and A. Stella. 2004. A Monte Carlo approach for estimation of haplotype probabilities in half-sib families. J. Dairy Sci. 87:4303-4310.

Boichard, D. and M. Brochard. 2012. New phenotypes for new breeding goals in dairy cattle. Animal 6:544-550.

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

Buch, L.H., M. Kargo, P. Berg, J. Lassen, and A.C. Sørensen. 2011. The value of cows in reference populations for genomic selection of new functional traits. Animal 6:880-886.

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

Calus, M.P.L., H.A. Mulder, K. L. Verbyla, and R.F. Veerkamp. 2009. Estimating reliabilities of genomic breeding values. Interbull Bull. 40:198–201.

Calus, M.P.L. 2010a. Genomic breeding value prediction: methods and procedures. Animal 4:157-164.

Calus, M.P.L., H.A. Mulder, and R.F. Veerkamp. 2010b. Comparison of Reliabilities of Direct Genomic Values. Interbull Bull. 41:25–28.

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

Calus, M.P.L., H.A. Mulder, and J.W.M. Bastiaansen. 2011b. Identification of Mendelian inconsistencies between SNP and pedigree information of sibs. Genet. Sel. Evol. 43:34.

Calus, M.P.L., and R.F. Veerkamp. 2011c. Accuracy of multi-trait genomic selection using different methods. Genet. Sel. Evol. 43:26.

Calus, M.P.L., Y. de Haas, and R.F. Veerkamp. 2012. Genomic prediction for new traits combining cow and bull reference populations. Proc. 63rd Annual Meeting of the European Federation of Animal Science, Bratislava, Slovakia, p. 90.

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

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

Chen, C.Y., I. Misztal, I. Aguilar, A. Legarra, and W.M. Muir. 2011. Effect of different genomic relationship matrices on accuracy and scale. J. Anim. Sci. 89:2673-2679.

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

Coffey, M.P., G. Simm, J.D. Oldham, W G. Hill, and S. Brotherstone. 2004. Genotype and diet effects on energy balance in the first three lactations of dairy cows. J. Dairy Sci. 87:4318-4326.

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

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.

de Haas, Y., J.J. Windig, M.P.L. Calus, J. Dijkstra, M. de Haan, A. Bannink, and R.F. Veerkamp. 2012a. Genetic parameters for predicted methane production and potential for reducing enteric emissions through genomic selection. J. Dairy Sci. 94:6122-6134.

de Haas, Y., M.P.L. Calus, R.F. Veerkamp, E. Wall, M.P. Coffey, H.D. Daetwyler, B.J. Hayes, and J.E. Pryce. 2012b. Improved accuracy of genomic prediction for dry matter intake of dairy cattle from combined European and Australian data sets. J. Dairy Sci. 95, 6103-6112.

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

de Roos, A.P.W., B.J. Hayes, and M.E. Goddard. 2009. Reliability of genomic predictions across multiple populations. Genetics 183:1546-1553.

de Roos, A.P.W. 2011. Genomic selection in dairy cattle. PhD Thesis. Wageningen University.

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.

Falconer, D.S. and T.F.C. Mackay. 1996. Introduction to quantitative genetics. 4th ed. Longman, New York, NY.

Forni, S., I. Aguilar, and I. Misztal. 2011. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genet. Sel. Evol. 43:1.

Gengler, N., P. Mayeres, and M. Szydlowski. 2007. A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. Animal 1:21-28.

Gengler, N., S. Abras, C. Verkenne, S. Vanderick, M. Szydlowski, and R. Renaville. 2008. Accuracy of prediction of gene content in large animal populations and its use for candidate gene detection and genetic evaluation. J. Dairy Sci. 91:1652-1659.

Gilmour, A.R., B.J. Gogel, B.R. Cullis, S.J. Welham, and R. Thompson. 2002. ASReml user guide, release 1.0. VSN International Ltd, Hemel Hempstead, UK.

Gilmour, A.R., B.J. Gogel, B.R. Cullis, S. J. Welham, and R. Thompson. 2009. ASReml user guide release 3.0. VSN International Ltd, Hemel Hempstead, UK.

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

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

Goddard, M.E. 2010. Genomic selection in farm animal species - Lessons learnt and future perspectives. Proc. 9th World Congress on Genetics Applied to Livestock Production (WCGALP), Leipzig, Germany.

Goddard, M.E., B.J. Hayes, and T.H.E. Meuwissen. 2011. Using the genomic relationship matrix to predict the accuracy of genomic selection. J. Anim. Breed. Genet. 128:409-421.

González-Recio, O., D. Gianola, N. Long, K.A. Weigel, G.J. M. Rosa, and S. Avendano. 2008. Nonparametric methods for incorporating genomic information into genetic evaluations: An application to mortality in broilers. Genetics 178:2305-2313.

Grapes, L., J.C.M. Dekkers, M.F. Rothschild, and R.L. Fernando. 2004. Comparing linkage disequilibrium-based methods for fine apping quantitative trait loci. Genetics 166:1561-1570.

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

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

Habier, D., R.L. Fernando, and J. C. M. Dekkers. 2009. Genomic selection using low-density marker panels. Genetics 182:343-353.

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

Haldane, J. 1919. The combination of linkage values and the calculation of distances between the loci of linked factors. J. Genet. 8:299-309.

Harville, D. 1976. Extension of the Gauss-Markov theorem to include the estimation of random effects. Ann. Stat. 4:384-395.

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

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

Hayes, B.J., H.D. Daetwyler, P.J. Bowman, G. Moser, B. Tier, R. Crump, M. Khatkar, H. Raadsma, and M.E. Goddard. 2009c. Accuracy of genomic selection: Comparing theory and results. Proc. Assoc. Advmt. Anim. Breed. Genet. 18:34-37.

Hayes, B.J., P.M. Visscher, and M.E. Goddard. 2009d. Increased accuracy of artificial selection by using the realized relationship matrix. Genet. Res. (Camb.) 91:47–60.

Heffner, E.L., M.E. Sorrells, and J-L. Jannink. 2009. Genomic selection for crop improvement. Crop Sci. 49:1-12.

Henderson, C.R. 1984. Applications of linear models in animal breeding. University of Guelph, Guelph, Ontario, Canada.

Henderson, C.R. 1985. Best linear unbiased prediction using relationship matrices derived from selected base populations. J. Dairy Sci. 68:443-448.

Hickey, J., B. Kinghorn, B. Tier, J.H.J. van der Werf, and M. Cleveland. 2012. A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. Genet. Sel. Evol. 44:9.

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

Horan, B., P. Dillon, D.P. Berry, P.O'Connor, and M. Rath. 2005. The effect of strain of Holstein Friesian, feeding system and parity on lactation curves characteristics of spring-calving dairy cows. Livest. Prod. Sci. 95:231-241.

Howie, B.N., P. Donnelly, and J. Marchini. 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS genetics 5:e1000529.

Ibañez-Escriche, N., and O. Gonzalez-Recio. 2011. Promises, pitfalls and challenges of genomic selection in breeding programs. Spanish Journal of Agricultural Research 9:404-413.

Ihara N, Takasuga A, Mizoshita K, Takeda H, Sugimoto M, Mizoguchi Y, Hirano T, Itoh T, Watanabe T, Reed K.M, Snelling W.M, Kappes SM, Beattie C.W, Bennett G.L, Sugimoto Y. 2004. A comprehensive genetic map of the cattle genome based on 3802 microsatellites. Genome Res. 14:1987-1998.

Jannink, J-L., A. J. Lorenz, and H. Iwata. 2010. Genomic selection in plant breeding: from theory to practice. Brief. Funct. Genomics 9:166-177.

Jia, Y. and J-L. Jannink. 2012. Multiple trait genomic selection methods increase genetic value prediction accuracy. Genetics 192:1513-1522.

Jiménez-Montero, J. A., O. González-Recio, and R. Alenda. 2012. Genotyping strategies for genomic selection in small dairy cattle populations. Animal 6:1216-1224.

Johanson, J.M., and P.J. Berger. 2003. Birth weight as a predictor of calving ease and perinatal mortality in Holstein cattle. J. Dairy Sci. 86:3745-3755.

Kearney, J.F., E. Wall, B. Villanueva, and M.P. Coffey. 2004. Inbreeding trends and application of optimized selection in the UK Holstein population. J. Dairy Sci. 87:3503-3509.

Kennedy, B. W. 1981. Bias and mean square error from ignoring genetic groups in mixed model sire evaluation. J. Dairy Sci. 64:689-697.

Kennedy, B. W., and D. Trus. 1993. Considerations on genetic connectedness between management units under an animal model. J. Anim. Sci. 71:2341-2352.

Kolbehdari, D., L.R. Schaeffer, and J. A.B. Robinson. 2007. Estimation of genome-wide haplotype effects in half-sib designs. J. Anim. Breed. Genet. 124:356-361.

König, S. and H. Simianer. 2006. Approaches to the management of inbreeding and relationship in the German Holstein dairy cattle population. Livest. Sci. 103:40-53.

Lassen, J., P. Løvendahl, and J. Madsen. 2012. Accuracy of noninvasive breath methane measurements using Fourier transform infrared methods on individual cows. J. Dairy Sci. 95:890-898.

Lee S. H., M.E. Goddard, P.M. Visscher, and J.H.J. van der Werf. 2010. Using the realized relationship matrix to disentangle confounding factors for the estimation of genetic variance components of complex traits. Genet. Sel. Evol. 42:22.

Legarra, A., C. Robert-Granie, E. Manfredi, and J-M. Elsen. 2008. Performance of genomic selection in mice. Genetics 180:611-618.

Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. J. Dairy Sci. 92:4656-4663.

Lewis, R.M., R.E. Crump, G. Simm, and R. Thompson. 1999. Assessing connectedness in across-flock genetic evaluations. Proc. Br. Soc. Anim. Sci. Br. Soc. Anim. Sci., Penicuik, UK.

Li, Y. and G.R. Abecasis. 2006. Mach 1.0: rapid haplotype reconstruction and missing genotype inference. Am. J. Hum. Genet. S 79:2290.

Long, N., D. Gianola, G.J. M. Rosa, K.A. Weigel, and S. Avendaño. 2007. Machine learning classification procedure for selecting SNPs in genomic selection: Application to early mortality in broilers. J. Anim. Breed. Genet. 124:377-389.

Lund, M.S., G. Su, U.S. Nielsen, and G.P. Aamand. 2009. Relation between accuracies of genomic predictions and ancestral links to the training data. Interbull Bull. 40:162–166.

Lund, M.S., A.P.W. de Roos, A.G. de Vries, T. Druet, V. Ducroq, S. Fritz, F. Guillaume, B. Guldbrandtsen, Z. Liu, R. Reents, C. Schrooten, F. Seefried, and G. Su. 2010. Improving genomic prediction by EuroGenomics collaboration. Proc. 9th World Congress on Genetics Applied to Livestock Production (WCGALP), Leipzig, Germany.

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, F. Seefried, and G.

Su 2011. A common reference population from four European Holstein populations increases reliability of genomic predictions. Genet. Sel. Evol. 43:43.

Mathur, P.K. and P. Horst. 1994. Methods for evaluating genotype x□environment interactions illustrated by laying hens. J. Anim. Breed. Genet. 111:265-288.

McParland, S., G. Banos, E. Wall, M.P. Coffey, H. Soyeurt, R.F. Veerkamp, and D.P. Berry. 2011. The use of mid-infrared spectrometry to predict body energy status of Holstein cows. J. Dairy Sci. 94:3651-3661.

Merks, J.W.M., P.K. Mathur, and E. F. Knol. 2012. New phenotypes for new breeding goals in pigs. Animal 6:535-543.

Meuwissen, T.H.E., and M.E. Goddard. 1996. The use of marker haplotypes in animal breeding schemes. Genet. Sel. Evol. 28:161-176.

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.

Meuwissen, T.H.E. 2002. Gencont: an operational tool for controlling inbreeding in selection and conservation schemes. Proc. 7th World Congress on Genetics Applied to Livestock Production (WCGALP), Montpellier, France,

Meuwissen, T.H.E. 2009. Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. Genet. Sel. Evol. 41:35.

Meuwissen, T.H.E., T. Luan, and J. A. Woolliams. 2011. The unified approach to the use of genomic and pedigree information in genomic evaluations revisited. J. Anim. Breed. Genet. 128:429-439.

Miglior, F., B.L. Muir, and B.J. Van Doormaal. 2005. Selection indices in Holstein cattle of various countries. J. Dairy Sci. 88:1255-1263.

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

Misztal, I., I. Aguilar, and A. Legarra. 2009b. Single-step national evaluation using phenotypic, full pedigree and genomic information. J. Dairy Sci. 92:4648–4655.

Mrode, R. 2005. Linear models for the prediction of animal breeding values. CABI, Wallingford, UK.

Mrode, R., J. F. Kearney, S. Biffani, M.P. Coffey, and F. Canavesi. 2009. Short communication: Genetic relationships between the Holstein cow populations of three European dairy countries. J. Dairy Sci. 92:5760-5764.

Muir, W. M. 2007. Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. J. Anim. Breed. Genet. 124:342-355.

Mulder, H.A., T.H.E. Meuwissen, M.P.L. Calus, and R.F. Veerkamp. 2010. The effect of missing marker genotypes on the accuracy of gene-assisted breeding value estimation: A comparison of methods. Animal 4:9-19.

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

Neeteson-van Nieuwenhoven, A-M., P. Knap, and S. Avendaño. 2013. The role of sustainable commercial pig and poultry breeding for food security. Animal Frontiers 3:52-57.

Nejati-Javaremi, A., C. Smith, and J.P. Gibson. 1997. Effect of total allelic relationship on accuracy of evaluation and response to selection. J. Anim Sci. 75:1738-1745.

Pérez-Cabal, M.A., A.I. Vazquez, D. Gianola, G.J. M. Rosa, and K.A. Weigiel. 2010. Accuracy of genomic predictions in USA Holstein Cattle from different training-testing designs. Proc. 9th World Congress on Genetics Applied to Livestock Production (WCGALP), Leipzig, Germany.

Philipsson. J, G. Ral, and B. Berglund. 1995. Somatic cell count as a selection criterion for mastitis resistance in dairy cattle. Livest. Prod. Sci. 41:195-200.

Powell, J.E., P.M. Visscher, and M. E. Goddard. 2010. Reconciling the analysis of IBD and IBS in complex trait studies. Nat. Rev. Genet. 11:800-805.

Pryce, J.E., B.L. Nielsen, R.F. Veerkamp, and G. Simm. 1999. Genotype and feeding system effects and interactions for health and fertility traits in dairy cattle. Livest. Prod. Sci. 57:193-201.

Pryce, J.E. and H.D. Daetwyler. 2012a. Designing dairy cattle breeding schemes under genomic selection: a review of international research. Anim. Prod. Sci. 52:107-114.

Pryce, J.E., Y. de Haas, B.J. Hayes, M.P. Coffey, and R.F. Veerkamp. 2012b. Genomic selection for feed efficiency in dairy cattle - a complex objective. Proc. 63rd Annual Meeting of the European Federation of Animal Science, Bratislava, Slovakia., p. 22: 21.

Pszczola, M., H.A. Mulder, and M.P.L. Calus. 2011. Effect of enlarging the reference population with (un)genotyped animals on the accuracy of genomic selection in dairy cattle. J. Dairy Sci. 94:431-441.

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

Pszczola, M., T. Strabel, J.A.M. van Arendonk and M.P.L. Calus 2012b. The impact of genotyping different groups of animals on accuracy when moving from traditional to genomic selection. J. Dairy Sci. 95:5412-5421.

Rincent, R., D. Laloë, S. Nicolas, T. Altmann, D. Brunel, P. Revilla, V.M. Rodriguez, J. Moreno-Gonzales, A.E. Melchinger, E. Bauer, C-C. Schön, N. Meyer, C. Giauffret, C. Bauland, P. Jamin, J. Laborde, H. Monod, P. Flament, A. Charcosset, and L. Moreau. 2012. Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of Maize inbreds (Zea mays L.). Genetics 2:715-728.

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. The Am. J. Hum. Genet. 78:629-644.

Schenkel, F.S., M. Sargolzaei, G. Kistemaker, G. Jansen, P. Sullivan, B. Van Doormaal, P. VanRaden, and G. Wiggans. 2009. Reliability of genomic evaluation of Holstein cattle in Canada. Interbull Bull. 39:51-58.

Sellner, E.M., J.W. Kim, M.C. McClure, K.H. Taylor, R.D. Schnabel, and J.F. Taylor. 2007. Board-invited review: Applications of genomic information in livestock. J. Anim. Sci. 85:3148-3158.

Sonesson, A.K. and T.H.E. Meuwissen. 2009. Testing strategies for genomic selection in aquaculture breeding programs. Genet. Sel. Evol. 41:37.

Strandén, I., and D.J. Garrick. 2009. Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. J. Dairy Sci. 92:2971–2975.

Su, G., B. Guldbrandtsen, V.R. Gregersen, and M.S. Lund. 2010. Preliminary investigation on reliability of genomic estimated breeding values in the Danish Holstein population. J. Dairy Sci. 93:1175-1183.

Tempelman, R.J. 2009. Invited review: Assessing experimental designs for research conducted on commercial dairies. J. Dairy Sci. 92:1-15.

Thompson, R. and K. Meyer. 1986. A review of theoretical aspects in the estimation of breeding values for multi-trait selection. Livest. Prod. Sci. 15:299-313.

Tsuruta, S., I. Misztal, I. Aguilar, and T.J. Lawlor. 2011. Multiple-trait genomic evaluation of linear type traits using genomic and phenotypic data in US Holsteins. J. Dairy Sci. 94:4198-4204.

Van Grevenhof, E., J.A.M. Van Arendonk, and P. Bijma. 2012. Response to genomic selection: The Bulmer effect and the potential of genomic selection when the number of phenotypic records is limiting. Genet. Sel. Evol. 44:26.

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

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

Veerkamp, R.F., G. Simm, J.D. Oldham. 1994. Effects of interaction between genotype and feeding system on milk-production, feed-intake, efficiency and body tissue mobilization in dairy cows. Livest. Prod. Sci. 39:229-241

Veerkamp, R.F., G.C. Emmans, A.R. Cromie, and G. Simm. 1995. Variance components for residual feed intake in dairy cows. Livest. Prod. Sci. 41:111-120.

Veerkamp, R. F. 1998. Selection for economic efficiency of dairy cattle using information on live weight and feed Intake: A review. J. Dairy Sci. 81:1109-1119.

Veerkamp, R.F., J.K. Oldenbroek, H.J. Van Der Gaast and J.H.J van der Werf. 2000. Genetic correlation between days until start of luteal activity and milk yield, energy balance, and live weights. J. Dairy Sci. 83:577-583.

Veerkamp, R.F., H.A. Mulder, and M.P.L. Calus. 2010. Estimation of heritability for dairy traits combining pedigree with dense SNPs information on some animals. Proc. 9th World Congress of Genetics Applied to Livestock Production (WCGALP), Leipzig, Germany.

Veerkamp, R.F., H.A. Mulder, R. Thompson, and M.P.L. Calus. 2011. Genomic and pedigree-based genetic parameters for scarcely recorded traits when some animals are genotyped. J. Dairy Sci. 94:4189-4197.

Veerkamp, R.F., M.P. Coffey, D.P. Berry, Y. de Haas, E. Strandberg, H. Bovenhuis, M.P.L. Calus, and E. Wall. 2012. Genome-wide associations for feed utilisation complex in primiparous Holstein-Friesian dairy 6 cows from experimental research herds in four European countries. Animal 6:1738-1749.

Verbyla, K.L., M.P.L. Calus, H.A. Mulder, Y. de Haas, and R.F. Veerkamp. 2010. Predicting energy balance for dairy cows using high-density single nucleotide polymorphism information. J. Dairy Sci. 93:2757-2764.

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

Vitezica, Z.G., I. Aguilar, I. Misztal, and A. Legarra. 2011. Bias in genomic predictions for populations under selection. Genet. Res. (Camb.) 93:357–366.

Weller, J.I., Y. Kashi, and M. Soller. 1990. Power of daughter and granddaughter designs for determining linkage between marker loci and quantitative trait loci in dairy cattle. J. Dairy Sci. 73:2525-2537.

Wientjes, Y.C.J., R.F. Veerkamp, and M.P.L. Calus. 2013. The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. Genetics 193:621-631.

Wolc, A., J. Arango, P. Settar, J.E. Fulton, N.P. O'Sullivan, R. Preisinger, D. Habier, R. Fernando, D.J. Garrick, and J.C.M. Dekkers. 2011. Persistence of accuracy of genomic estimated breeding values over generations in layer chickens. Genet. Sel. Evol. 43:23.

Wright, S. 1922. Coefficients of inbreeding and relationship. Am. Nat. 56:330-338.

Yang, J., B. Benyamin, B.P. McEvoy, S. Gordon, A.K. Henders, D.R. Nyholt, P.A. Madden, A.C. Heath, N.G. Martin, G.W. Montgomery, M.E. Goddard, and P.M. Visscher. 2010. Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42:565-569.

Yu, K., J. Xu, D. C. Rao, and M. Province. 2005. Using tree-based recursive partitioning methods to group haplotypes for increased power in association studies. Ann. Hum. Genet. 69:577-589.

Zhong, S., J.C.M. Dekkers, R.L. Fernando, and J-L. Jannink. 2009. Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: A barley case study. Genetics 182:355-364.

Acknowledgements

This thesis was a collaborative project between the Department of Genetics and Animal Breeding of Poznan University of Life Sciences in Poland and Animal Breeding and Genomics Centre of Wageningen UR in The Netherlands.

I would like to thank my supervising committee for their support and guidance during the subsequent steps of my PhD study. I am particularly grateful to Johan van Arendonk, Mario Calus and Tomasz Strabel, who approved and enabled the idea of a collaborative project to come through. Mario thank you for your help and efforts to improve my scientific skills what has started already in my MSc studies and continued during my PhD. Tomasz you have initiated in me the idea of a scientific career in Animal Breeding and helped a lot while going through its next steps. Without your input I would have not even thought about starting a PhD study. I appreciate it very much. Han Mulder, thank you for very detailed and useful remarks to my papers. Roel Veerkamp thank you for you critical attitude helping to push my thoughts in the right direction. Johan thank you for your always very insightful comments.

Thanks to the co-authors of the papers included in my thesis: Yvette de Haas and Eileen Wall for interesting and valuable discussions.

Ewa, thank you for your friendship and hospitality when I needed a place to stay for some time, for myself and my family or when I needed an advice. You and Jasiu were always helpful whenever I was in need. Yvonne, thanks for your companionship, interesting discussions and also for trusting me that I or my kids would not destroy your apartment and leaving it in our hands for a little while. Both of you, thanks of taking the responsibility of being my paranymphs. Rianne, thanks for the urgent-taxi-service in the most unexpected moment and being a good office mate. Myrthe, thanks for sharing with me the office for most of my PhD. Aniek, thank you for discussions we had over the coffee. Beata and Paulina, thank you for helping me dealing with various issues in Poland when I was not able to be there physically and for your companionship when I was there.

To researchers and my colleagues at both Universities, not mentioned above, thank you for your friendship, helpfulness and creating the atmosphere that I enjoyed a lot while working on my thesis. I would like to give a special recognition to: Amabel, André, Bart, Christos, Coralia, Dirkjan, Egbert, Gabriel, Gus, Hamed, Henk, Heyun, Ilse, Ina, Irene, Jack, Jan, John, Katrijn, Krista, Lucia, Mahlet, Marcos, Marike, Marinus, Marzieh, Mathijs, Nancy, Naomi, Panya, Piter, Sandrine, Sipke-Joost, and Susan from Wageningen UR; and Marek Świtoński, Dorota Cieślak-Lechniak, Zbigniew Sobek, Maciej Szydłowski, and Tomasz Szwaczkowski as well as to Ala, Ania, Dagna, Ewelina, Iza, Joanna, Kamila, Katarzyna, Magda, Marysia, Monika,

Monika, Sylwia, Zosia, Jarek, Kuba, Maciej, Mariusz, Piotr, and Sebastian from Poznan University of Life Sciences. Also, I would like to acknowledge the administrative staff members for all the help that I received from them. I thank Ada, Carina, Lisette, Monique and Stasia. All the others unnamed here that helped me during my study – thank you.

I would like to acknowledge members of scientific network created within a Marie Curie Initial Training Network – GreenHouseMilk. Eillen, Ellodie, Hélène, Karolina, Yvette, Donagh, Frank, Laurence, Nic, Nicolas, Phuong, and Purna thank you for lots of valuable interactions during our meetings. Special thanks to Yvette for coordinating the project.

I am grateful to Mike Grossman for changing my view on presenting the scientific knowledge which have had a very big impact on the style of presenting my papers. The attendance in your workshop has led to winning several 'best presentation' awards – thank you.

Special thanks go to my family for their patience and support given to me over the last years. Mam and Dad, thank you for your guidance received through all my life. Magda and Ola, my dear sisters, thank you for your companion during my childhood. Ania, my wife and best friend, thank you for being always there for me and giving me all the love. Zosia and Szymon, my beloved children, thank you for your cheerfulness needed so much during demanding time of working on my PhD.

Dear reader, thank you for taking an effort to read this thesis. I hope that you enjoyed it and consider the time spent on the lecture as valuable.

Marcin

Curriculum Vitae

Marcin Pszczoła was born on the 13th of October 1984 in Śrem, Poland. He was raised in Borowo. In 2003 he graduated from high school in Kościan and started his BSc study in Animal Husbandry at Faculty of Animal Sciences of Poznań University of Life Sciences. During his BSc studies, he worked as technical assistant at Department of Genetics and Animal Breeding of Poznań University of Life Sciences. After his graduation in 2007, he started with his MSc study in Animal Sciences at Wageningen University with specialization 'Animal Breeding and Genetics'. He performed a minor thesis under supervision of Prof. Dr Ignacy Misztal at University of Georgia, USA. In this research, he investigated the effects of heat stress on fertility of U.S. Holstein cattle. His major thesis was performed under supervision of Dr Mario Calus and Dr Han Mulder at Wageningen UR Livestock Research and focused on genomic selection for small reference populations. This work initiated further research realized within his PhD which started in 2009. In his PhD, Marcin worked on optimizing genomic selection for scarcely recorded novel traits and the results of his study are presented in this thesis. Marcin's PhD was a collaborative project between Poznan University of Life Sciences and Wageningen UR Livestock Research.

List of publications

Refereed

Pszczola, M., I. Aguilar, I. Misztal. 2009. Short communication: Trends for monthly changes in days open in Holsteins. J. Dairy Sci. 29:4689-4696.

Pszczola, M., H.A. Mulder, M.P.L. Calus. 2011. Effect of enlarging the reference population with (un)genotyped animals on the accuracy of genomic selection in dairy cattle. J. Dairy Sci. 94: 431-441.

Mucha, S., M. Pszczola, T. Strabel, A. Wolc, P. Paczyńska, M. Szydłowski. 2011. Comparison of analyses of the QTLMAS XIV common dataset. II: QTL analysis. BMC Proc. 5(Suppl 3):S1.

Pszczola, M., T. Strabel, A. Wolc, S. Mucha, M. Szydłowski. 2011. Comparison of analyses of the QTLMAS XIV common dataset. I: genomic selection. BMC Proc. 5(Suppl 3):S1.

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

Pszczola, M., T. Strabel, J.A.M. van Arendonk, M.P.L. Calus. 2012. The impact of genotyping different groups of animals on accuracy when moving from traditional to genomic selection J. Dairy Sci. 95:5412-5421.

Zeng, J., M. Pszczola, A. Wolc, T. Strabel, R.L. Fernando, D.J. Garrick, J.C.M. Dekkers. 2012. Genomic breeding value prediction and QTL mapping of QTLMAS2011 data using Bayesian and GBLUP methods. BMC Proc. 6(Suppl 2):S7.

Calus, M.P.L., Y. de Haas, M. Pszczola, R.F. Veerkamp. 2013. Predicted accuracy of and response to genomic selection for new traits in dairy cattle. Animal. 7(2): 183-191.

Pszczola, M., R.F. Veerkamp, Y. de Haas, E. Wall, T. Strabel, M.P.L. Calus. 2013. Effect of predictor traits on accuracy of genomic breeding values for feed intake based on a limited cow reference population. Animal. doi:10.1017/S175173111300150X.

Conference contributions

Pszczola, M., T. Strabel. 2007. Dobór par do rozrodu - analiza wyników kojarzeń wspomaganych komputerowo. XV Szkoła Zimowa, 25–31.03, Zakopane, Poland.

Pszczola, M., T. Strabel. 2007. Dobór par do kojarzeń praktyczną metodą doskonalenia genetycznego cech kończyn bydła mlecznego. III Poznańskie Forum Zootechniczno-Weterynaryjne, 7-9.09, Poznan, Poland.

Pszczola, M., I. Aguilar, I. Misztal. 2009. Trends for monthly changes in days open in Holsteins. Joint Annual Meeting of ADSA-CSAS-ASAS, 12-16.07, Quebec Canada.

Pszczola, M., H.A. Mulder, M.P.L. Calus. 2010. The accuracy of genomic selection using (un)genotyped animals to enlarge the reference population. 9th Word Congress on Genetics Applied to Livestock Production. 1-6.08.2010. Leipzig, Germany.

Pszczola, M., H.A. Mulder, M.P.L. Calus. 2010. Wpływ powiększenia populacji referencyjnej o niezgenotypowane osobniki na dokładność genomowej oceny wartości hodowlanej. III Polish Congress of Genetics. 12-15.09. Lublin, Poland.

Pszczola, M., A. Wolc, S. Mucha, M. Szydłowski, M.A. Wietrzykowski, A. Borowska, T. Strabel. 2011. Estimation of breeding values and detection of QTL in QTL-MAS 2011 dataset using GBLUP and Bayesian approaches. 15th QTL-MAS Workshop. 19-20.05. Rennes, France.

Pszczola, M., T. Strabel, H.A. Mulder, M.P.L. Calus. 2011. Design of the reference population affects the reliability of genomic selection. 62nd Annual Meeting of the European Federation of Animal Science. 29.08-2.09. Stavanger, Norway.

Pszczola, M., T. Strabel, H.A. Mulder, M.P.L. Calus. 2012. Reliability of genomic selection for different reference population designs. BSAS Annual Meeting, Healthy Food from Healthy Animals. 24-25.04. Notthingam, UK.

Pszczola, M., R.F. Veerkamp, Y. de Haas, T. Strabel, M.P.L. Calus. 2012. Predictor traits improve accuracy of genomic breeding values for scarcely recorded traits. 4th International Conference on Quantitative Genetics. 17-22.06. Edinburgh, UK.

Pszczola, M., T. Strabel, J.A.M. Van Arendonk, M.P.L. Calus. 2012. Reference population designs affects reliability of selection for (un)genotyped animals. 63rd Annual Meeting of the EAAP. 27.08-31.08. Bratislava, Slovakia.

Pszczola, M., R.F. Veerkamp, Y. de Haas, T. Strabel, M.P.L. Calus. 2012. Multivariate genomic prediction improves breeding value accuracy for scarcely recorded traits. 63rd Annual Meeting of the EAAP. 27.08-31.08. Bratislava, Slovakia.

Calus, M.P.L., Y. de Haas, M. Pszczola, R.F. Veerkamp. 2012. Genomic selection for new traits: optimal prediction and reference population design. 63rd Annual Meeting of the EAAP. 27.08-31.08. Bratislava, Slovakia.

Pszczola, M., T. Strabel, J.A.M. van Arendonk, M.P.L. Calus. 2012. Reliability of genomic breeding values at different reference population's designs when some or all animals are genotyped. ADSA - AMPA - ASAS - CSAS - WSASAS Joint Annual Meeting. 15-19.07. Phoenix, USA.

Pszczola, M., T. Strabel, M.P.L. Calus. 2012. Ways to increase accuracy of genomic breeding values when the reference population size is restricted. 25th International Conference „Genetic Days". 18 – 20.09.2012. Wroclaw, Poland.

Pszczola, M., M.P.L. Calus, T. Strabel. 2013. Nowe cechy w doskonaleniu bydła mlecznego. XXI Szkoła Zimowa Hodowców Bydła. 11-15.03. Zakopane, Poland.

Pszczola, M., T. Strabel, R.F. Veerkamp, H.A. Mulder, J.A.M. van Arendonk, M.P.L. Calus. 2013. Required increase in training set to keep accuracy of genomic selection constant across generations. 64th Annual Meeting of the EAAP. 26-30.08. Nantes, France.

Training and supervision plan

The basic package (3 ECTS)	year
WIAS Introduction Course	2009
WGS course Ethics and Philosophy in Life Sciences	2011
Scientific exposure (26 ECTS)	
International conferences (13 ECTS)	
6th European Poultry Genetics Symposium, Bedlewo, Poland	2009
9th WCGALP,Leipzig,Germany	2010
Interbull Meeting, Stavanger, Norway	2011
62nd Annual Meeting EAAP, Stavanger, Norway	2011
25th International Conference “Genetic Days”, Wroclaw, Poland	2012
BSAS, Nottingham, UK	2012
4th International Conference on Quantitative Genetics, Edinburgh, UK	2012
63rd Annual Meeting of the EAAP, Bratislava, Slovakia	2012
64th Annual Meeting EAAP, Nantes, France	2013
Conference Greenhouse Gases & Animal Agriculture, Dublin, Ireland	2013
Seminars and workshops (2 ECTS)	
Developments in genome-wide evaluation, Wageningen, the Netherlands	2009
QTL-MAS Workshop, Poznan, Poland	2010
QTL-MAS Workshop, Rennes, France	2011
Mini symposium on Advanced Genetics, Wageningen, the Netherlands	2012
WIAS Science Day	2012
Presentations (11 ECTS)	
QTL-MAS Workshop in Poznan, Poland (oral)	2010
9th WCGALP, Leipzig, Germany (oral)	2010
QTL-MAS Workshop in Rennes, France (oral)	2011
62nd EAAP, Stavanger, Norway (poster - 1st price winner)	2011
WIAS Science Day, Wageningen, the Netherlands (oral)	2012
BSAS, Nottingham, UK (oral - 1st price winner)	2012
63rd EAAP, Bratislava, Slovakia (poster)	2012
63rd EAAP, Bratislava, Slovakia (oral)	2012
25th International Conference “Genetic Days” (oral - 1st price winner)	2012
Conference Greenhouse Gases & Animal Agriculture, Dublin, Ireland (poster)	2013
64th EAAP, Nantes, France (oral)	2013

In-depth studies (9 ECTS)

Disciplinary and interdisciplinary courses (6 ECTS)

Quantitative Genetics of Selection Response, Wageningen, the Netherlands	2012
Genomic Selection in Livestock, Wageningen, the Netherlands	2011
Statistical Learning Methods For DNA-based prediction of complex traits, Wageningen, the Netherlands	2011
Advanced methods and algorithms in animal breeding with focus on genomic selection, Wageningen, the Netherlands	2012

Advanced statistics courses (2 ECTS)

Advanced R for Genetic Analysis, Edinburgh, UK	2012
Markov chain Monte Carlo for Genetics, Edinburgh, UK	2012

PhD students' discussion groups (1 ECTS)

Journal Club (at Poznan University of Life Sciences)

Professional skills support courses (5 ECTS)

Techniques for writing and presenting scientific papers	2010
Writing for an academic publication	2012
Interpersonal Communication for PhD Students	2013

Research skills training (6 ECTS)

Preparing own PhD research proposal	2009
External training period, 1 month in SAC, Edinburgh, UK	2012

Didactic skills training (10 ECTS)

Lecturing (8 ECTS)

Course on selection theory for MSc students at Poznan Univ. of Life Sci. (135hrs)	2009/ 2010
Course on selection theory for MSc students at Poznan Univ. of Life Sci. (90hrs)	2012/ 2013

Preparing course material (2 ECTS)

Materials for Linux course in Polish	2009

Management skills training (2 ECTS)

Organization of seminars and courses

QTL-MAS Workshop, Poznan, Poland	2010

Education and training total: 60 ECTS

Colophon

The research described in this thesis was financially supported by the European Commission within a Marie Curie Initial Training Network project called GreenHouseMilk (KBBE-238562), Foundation for Polish Science (START 2013, FNP), the Koepon Stichting (The Netherlands), National Science Centre, Poland (DEC-2011/01/N/NZ9/00614), and Faculty of Animal Breeding of Poznan University of Life Sciences (Poland).
The author also declares that he is a scholarship holder within the project "Scholarship support for PhD students specializing in majors strategic for Wielkopolska's development", Sub-measure 8.2.2 Human Capital Operational Programme, co-financed by European Union under the European Social Fund.

The printing of this thesis was funded by the Animal Breeding and Genomics Centre, Wageningen UR, the Netherlands and the GreenHouseMilk project (KBBE-238562).

The cover of this thesis was designed by Marcin Pszczoła.

This thesis was printed by GVO drukkers & vormgevers B.V. | Ponsen & Looijen, Ede, the Netherlands.