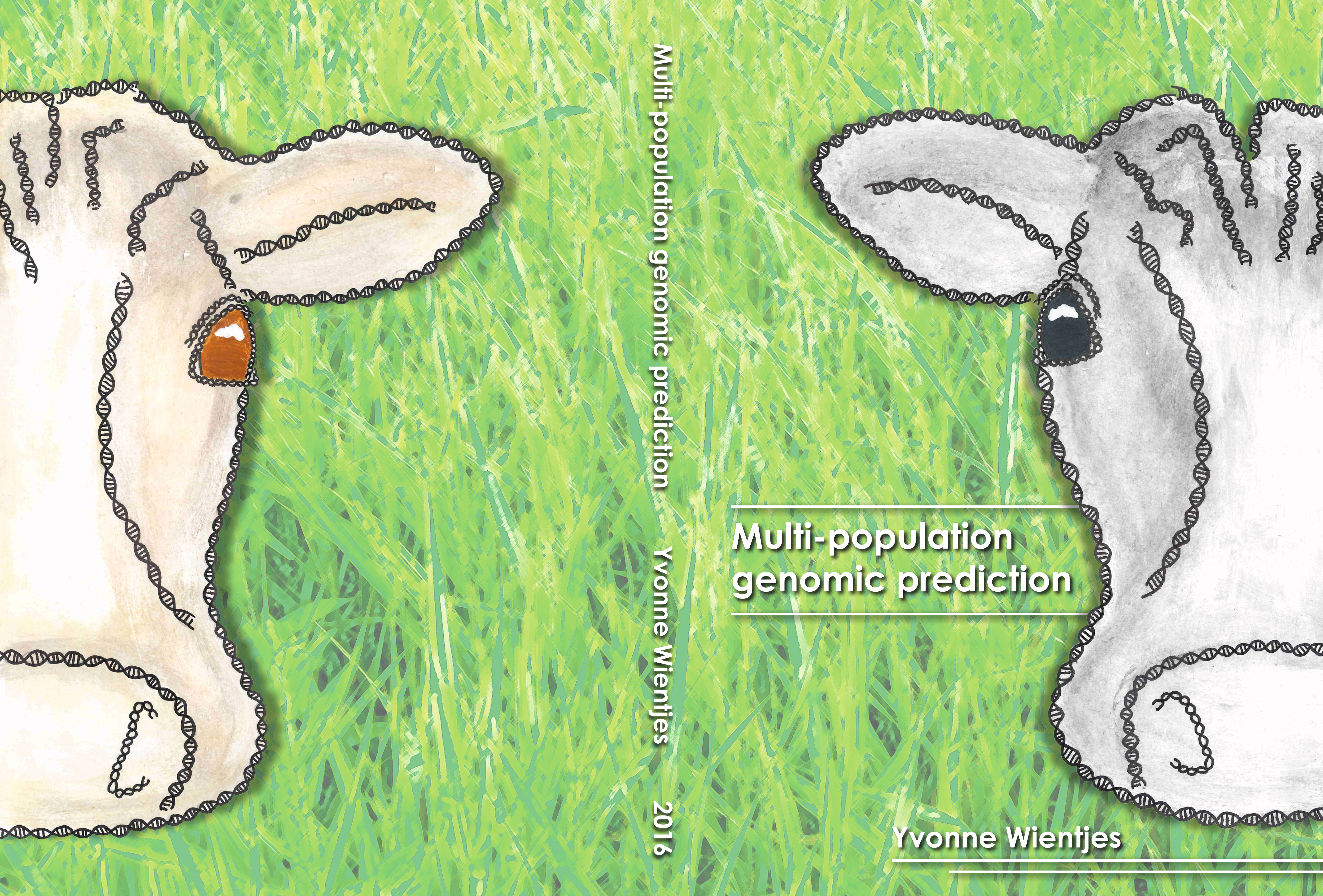


Multi-population genomic prediction Yvonne Wientjes 2016

Multi-population genomic prediction

Yvonne Wientjes

Multi-population genomic prediction

Yvonne C.J. Wientjes

Thesis committee

Promotor
Prof. Dr R.F. Veerkamp
Professor of Numerical Genetics and Genomics
Wageningen University

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

Other members
Prof. Dr M.C.M. De Jong, Wageningen University
Prof. Dr T.H.E. Meuwissen, Norwegian University of Life Sciences, Norway
Prof. Dr M.S. Lund, Aarhus University, Denmark
Dr H.D. Daetwyler, Department of Economic Development, Jobs, Transport, and Resources, Australia

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

Multi-population genomic prediction

Yvonne C.J. Wientjes

Thesis

submitted in fulfillment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Friday 22 January 2016
at 4 p.m. in the Aula.

Wientjes, Y.C.J.
Multi-population genomic prediction
268 pages.

PhD thesis, Wageningen University, Wageningen, NL (2016)
With references, with summaries in English and Dutch

ISBN 978-94-6257-619-3

Abstract

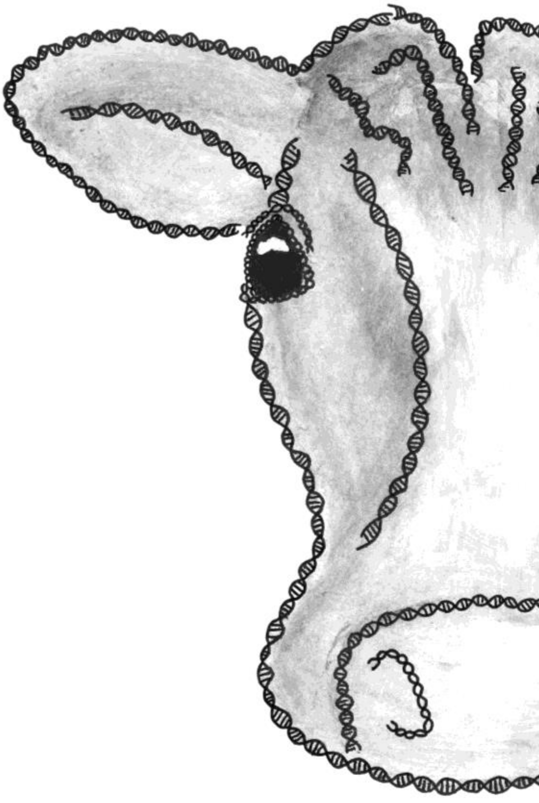
Wientjes, Y.C.J. (2016). Multi-population genomic prediction. PhD thesis, Wageningen University, the Netherlands

In genomic selection, genotype information is used to select the genetically best animals to produce the next generation. To identify the best animals, genotypes for single-nucleotide polymorphisms (SNPs) of selection candidates are combined with SNP effects, estimated in a reference population containing individuals with known phenotypes and SNP genotypes, to estimate genomic breeding values. For numerically small populations, the size of the reference population is often limited, which results in a low accuracy of genomic prediction. Enlarging the reference population by adding individuals from another population is an attractive approach to increase the accuracy. This thesis aimed to investigate the accuracy of multi-population genomic prediction, by 1) investigating the effect of different factors on the accuracy, and 2) deriving deterministic equations to predict the accuracy. Results show that the level of family relationships between reference and selection individuals has a higher effect on the accuracy of genomic prediction than the strength of linkage disequilibrium (LD) between quantitative trait loci (QTL) and SNPs. The accuracy of across-population genomic prediction is proportional to the genetic correlation between the populations. The consistency of multi-locus LD across populations can be calculated using selection index theory, and is highly related to the accuracy of across-population genomic prediction. The SNPs close to a QTL have a higher consistency in LD with the QTL across populations, which indicates that focusing on those SNPs could potentially improve the accuracy. It was also demonstrated that QTL properties, such as allele frequency pattern and distribution of allele substitution effects, are key parameters determining the accuracy of single- and multi-population genomic prediction. Moreover, two deterministic equations to predict the accuracy of multi-population genomic prediction were derived. The first equation is based on genomic relationships and was able to accurately predict the accuracy. The second equation is using population parameters, such as the number of effective chromosome segments across populations and the genetic correlation between populations, and can accurately predict the accuracy when the proportion of the genetic variance in the selection candidates captured by the SNPs in the reference population is known. Using this equation, it was shown that combining populations in one reference population can increase the accuracy when populations are closely related, the initial reference population is small, and a large number of animals is added.

Contents

CHAPTER 1

GENERAL INTRODUCTION



1.1 Dairy cattle breeding

The general aim of animal breeding is to improve the performance of future generations by selecting the genetically best animals in the current generation to produce the next generation. For many years, the genetically best animals were identified based on estimated breeding values, calculated using the performance of the animals themselves, their relatives, or a combination of both. In dairy cattle, for example, selection of bulls was based on the performance records of many daughters, which was first implemented in Denmark, but soon applied in other countries as well (Johansson 1960). The process of collecting performance records, known as phenotypes, is time-consuming and sometimes costly, especially for traits that are expressed late in life or that are difficult to measure. Moreover, not all traits can be measured on the selection candidates themselves, since some traits are only expressed in one sex or can only be measured after slaughtering the animal. Those factors influence the accuracy of identifying the genetically best animals, thereby affecting the genetic improvement of future generations.

The DNA carried by an animal determines whether an animal has a high genetic merit or not. Therefore, the possibility to use DNA information in selection strategies to select the genetically best animals was investigated since the second half of the last century (e.g., Smith 1967; Soller 1978). At first, only information of a few markers related to quantitative trait loci (QTL), i.e., the regions on the DNA that affect a quantitative trait, associated with a specific trait was used to estimate breeding values for that trait. Those breeding values were used for selecting the best animals, a strategy that is known as marker-assisted selection (e.g., Fernando and Grossman 1989; Dekkers 2004). In dairy cattle, marker-assisted selection was first introduced in breeding programs in the beginning of this century in France and Germany (Boichard *et al.* 2002; Bennewitz *et al.* 2003). Most quantitative traits are influenced by many QTL with small effects, making it very difficult to identify all QTL affecting a trait. Therefore, marker-assisted selection only had a limited effect on the genetic improvement of populations (Dekkers 2004).

In the beginning of this century, Meuwissen *et al.* (2001) proposed to use thousands of genome-wide markers simultaneously, regardless of the correlation to the trait, for the prediction of genomic estimated breeding values (GEBVs) to select the best animals, a strategy known as genomic selection. In contrast to marker-assisted selection, genomic selection does not require that all QTL are identified. In dairy cattle, genomic selection has the potential to double the genetic improvement per year, for a review see Pryce and Daetwyler (2012) or Bouquet and Juga (2013). This increase in genetic improvement is mainly a result of an increase in accuracy of estimating breeding values for young animals for which accurate phenotypes are not yet available (e.g., Calus *et al.* 2008; VanRaden 2008). The higher accuracy for those young animals results in the possibility to start selecting the best animals at a younger age and, thereby, reduces the generation interval (Schaeffer 2006). Due to the high potential of genomic selection, it is currently successfully implemented in bull breeding programs of Holstein Friesian dairy cattle populations worldwide, in which indeed a doubling of the genetic gain is obtained (Patry 2015).

1.2 Genomic prediction

The process of calculating GEBVs, that are used in genomic selection to select the best selection candidates, is known as genomic prediction. The GEBVs are calculated based on genotypes of markers, spread across the whole genome, and estimated effects of those markers, estimated simultaneously in a reference population containing individuals with both phenotypes and marker genotypes (Meuwissen *et al.* 2001). Single-nucleotide polymorphism (SNP) markers are commonly used for genomic prediction, based on the assumption that the SNPs are in linkage disequilibrium (LD) with the QTL influencing the trait. Therefore, the SNPs can be used to explain the QTL variation. The stronger the LD between SNPs and QTL, the more accurate the SNPs can explain the QTL variation, and the higher the accuracy of genomic prediction (Calus *et al.* 2008; Solberg *et al.* 2008). Besides the strength of LD between SNPs and QTL, the accuracy of genomic prediction also depends on the size of the reference population, i.e., the number of individuals with known phenotypes and genotypes used for estimating SNP effects. The larger the size of the reference population, the higher the accuracy of estimating SNP effects and the higher the accuracy of genomic prediction (e.g., Meuwissen *et al.* 2001; Daetwyler *et al.* 2008; VanRaden *et al.* 2009). Moreover, the accuracy is higher for individuals that are more closely related to the reference population (Habier *et al.* 2007; Habier *et al.* 2010).

The accuracy of genomic prediction also varies with the model used to estimate breeding values. At the moment, the commonly used models can roughly be divided in two different types; genomic best linear unbiased prediction (GBLUP) models and Bayesian variable selection models. The original GBLUP model, as described by Meuwissen *et al.* (2001), assumes that all SNPs explain an equal amount of the genetic variance, so basically assumes an infinitesimal model, and uses a genomic relationship matrix to estimate breeding values. A Bayesian variable selection model accommodates for some SNPs explaining a larger part of the genetic variance compared to other SNPs (Meuwissen *et al.* 2001). Therefore, a subset of SNPs is eligible to have a large effect and the other SNPs only have a small or no effect on the trait. In general, the accuracy of both models is very comparable, unless the trait is mainly influenced by a few QTL with a large effect that can be explained by a subset of the SNPs, which results in an advantage of the Bayesian variable selection model (Daetwyler *et al.* 2010).

1.3 Multi-population genomic prediction

For numerically small populations, establishing a reference population with a sufficient size is impossible, which limits the accuracy of genomic prediction for those populations. This might result in a lower rate of genetic improvement in those populations compared to numerically larger populations. Enlarging the reference population of a numerically small population by adding individuals from other populations is an attractive approach to increase the accuracy of genomic prediction. The value of a reference individual from another population might, however, be lower than the value of a reference individual from the same population, due to differences between the populations. At least four possible differences between populations are known that can influence the value of individuals from another population, which are described hereafter.

The first possible difference between populations is the difference in LD pattern. In a different population, the QTL might be in high LD with another SNP or the linkage phase between SNP and QTL might be reversed. Different studies have shown differences in the pattern of LD across different populations (e.g., Heifetz *et al.* 2005; Gautier *et al.* 2007; De Roos *et al.* 2008). At short distances on the genome, however, the consistency in LD is found to be reasonably high (Andreescu *et al.* 2007; De Roos *et al.* 2008; Zhou *et al.* 2013). Therefore, a high SNP density, with around 300,000 SNPs evenly spread across the genome in cattle (De Roos *et al.* 2008), is suggested to be able to overcome the differences in LD pattern between populations.

The second possible difference between populations is the difference in allele frequencies of both QTL and SNPs. In an extreme case, QTL might only segregate in one of the populations (Kemper *et al.* 2015a), indicating that another population is not going to improve the prediction of that specific QTL. When the SNPs surrounding that population-specific QTL are segregating in both populations, the apparent effect of the SNPs might be different across the populations. Moreover, the QTL that explain a large part of the genetic variance are most accurately estimated and most important for genomic prediction (Daetwyler *et al.* 2008). The genetic variance explained by a QTL depends on the size of the effect as well as on its allele frequency. Due to differences in QTL allele frequencies between populations, differences exist in the genetic variance explained by a QTL. When the genetic variance explained by a QTL is very low in one population, this population is not going to greatly improve the accuracy of estimating the effect of this QTL. Therefore, the benefit of adding another population can be expected to depend on the differences in allele frequencies of QTL between the populations.

The third possible difference between populations is the difference in allele substitution effects of QTL. Due to the presence of non-additive effects in combination with differences in allele frequencies, the average allele substitution effects might be different across populations (Falconer and Mackay 1996; Huang *et al.* 2012). An example of a gene with population-specific effects in dairy cattle is DGAT1 (*diacylglycerol O-acyltransferase 1*), for which the effect on milk yield of the causal mutation in Jersey and Fleckvieh was found to be only 80% (Spelman *et al.* 2002) and 70% (Thaller *et al.* 2003), respectively, of the effect in Holstein Friesian cattle. This shows that the effects estimated in one population cannot directly be used in another population. The correlation between allele substitution effects of QTL across populations is commonly referred to as the genetic correlation between the populations (Bohren *et al.* 1966; Falconer and Mackay 1996).

The fourth possible difference between populations is the level of family relationships, which is much lower, or even non-existing, between populations than within populations. This indicates that adding individuals from another population to the reference population does not increase the relatedness between selection candidates and reference population. Since the accuracy of genomic prediction is much higher for individuals that are more closely related to the reference population (Habier *et al.* 2007; Habier *et al.* 2010), adding unrelated individuals to the reference population has a smaller impact on the accuracy than adding related individuals from the same population.

At the start of this thesis, the effect of each of the four possible differences between populations on the accuracy of multi-population genomic prediction, where different populations are combined in the reference population, was not quantified. Therefore, it was difficult to realistically predict the potential to increase the accuracy of genomic prediction by adding information from other populations.

1.4 Expected potential of multi-population genomic prediction at the start of this thesis

The potential of multi-population genomic prediction was first investigated using simulations. Those simulation studies have shown that already at a low SNP density, it is beneficial for the accuracy to combine populations that separated only a few generations ago and have a highly consistent LD pattern (De Roos *et al.* 2009; Ibánẽz-Escriche *et al.* 2009). By increasing the SNP density, an increase in accuracy could be observed when combining populations that separated more than 300 (De Roos *et al.* 2009) or 550 (Ibánẽz-Escriche *et al.* 2009) generations ago. In another simulation study, using real genotypes and simulated phenotypes from very diverse cattle breeds, no benefit of combining populations for genomic prediction was observed (Kizilkaya *et al.* 2010). This was suggested to be a result of the used SNP density, which was too low to find a consistent LD pattern across the populations. Based on the results of those simulation studies, it was expected that combining information from different populations is an effective way to increase the accuracy of genomic prediction, provided that the marker density is high enough to find a consistent LD pattern between QTL and SNPs across populations.

In a study using real data, it was shown that combining four closely related Holstein Friesian populations from different European countries resulted in an average increase in accuracy of 10% compared to an analysis within country (Lund *et al.* 2011). Combining closely related breeds with only a small number of genotyped animals each, like the Danish, Swedish, and Finnish Red dairy cattle breeds, in one reference population resulted in an average increase in accuracy of 14% compared to single-breed genomic evaluation (Brøndum *et al.* 2011). The benefit of combining the distantly related Holstein Friesian and Jersey cattle breeds was lower, and sometimes even a decrease in accuracy was observed (Hayes *et al.* 2009; Harris and Johnson 2010; Pryce *et al.* 2011). Furthermore, extending an Australian Holstein Friesian and Jersey reference population with a Fleckvieh population from Germany and Austria did not result in an increase in accuracy of genomic prediction (Pryce *et al.* 2011). The benefit was slightly larger when a Bayesian variable selection model was used compared to a GBLUP model, although the benefit was generally low for both models (Hayes *et al.* 2009; Pryce *et al.* 2011). This is probably due to the possibility in the Bayesian model to give a higher weight to the SNPs with a consistent LD with the QTL across breeds compared to GBLUP. Those SNPs are in general located closer to the QTL and also have a higher LD with the QTL within breed (Hayes *et al.* 2009). Genotypes of only ~50,000 SNPs were used in all the mentioned studies using real data, which might indicate that the SNP density was too low to find a consistent LD phase between the investigated populations. Therefore, this low SNP density was suggested to be the reason for the relatively low benefit of combining populations found in empirical studies compared to simulation studies, and a higher SNP density was expected to be able to increase the potential of combining populations (Pryce *et al.* 2011).

So, at the start of this thesis four years ago, a high potential of combining populations for genomic prediction was expected based on the results described above. Even combining populations from different breeds was expected to result in an increase in accuracy, provided that the marker density was high enough to find a consistent linkage phase between the populations (>300,000 SNPs in cattle; De Roos *et al.* 2008).

1.5 Recent studies regarding multi-population genomic prediction

In the last four years, a lot of research has focused on multi-population genomic prediction. In dairy cattle, the studies can roughly be divided in studies combining populations from the same breed from different countries and studies combining populations from different breeds.

The first group of studies focused on combining populations from the same breed from different countries, for example by combining different Holstein Friesian populations (De Haas *et al.* 2012; VanRaden *et al.* 2012; Zhou *et al.* 2013; De Haas *et al.* 2015; Haile-Mariam *et al.* 2015), Jersey populations (Haile-Mariam *et al.* 2015; Wiggans *et al.* 2015), and Brown Swiss populations (Zumbach *et al.* 2010; Jorjani *et al.* 2011). In general, those studies showed a higher accuracy of genomic prediction when populations were combined in one reference population compared to using a within-country reference population, both using 50,000 SNPs (De Haas *et al.* 2012; Zhou *et al.* 2013; Haile-Mariam *et al.* 2015; Wiggans *et al.* 2015) and using 777,000 SNPs (VanRaden *et al.* 2012; De Haas *et al.* 2015). The highest accuracies were obtained when a multi-trait model was used, in which the same trait in the different countries was modelled as a different trait to account for factors like genotype by environment interactions, differences in trait definitions and differences in measurement method of the trait across countries (De Haas *et al.* 2012; De Haas *et al.* 2015). Moreover, the increase in accuracy was more pronounced for the population with the lowest number of genotyped individuals (De Haas *et al.* 2015; Wiggans *et al.* 2015), and for individuals that were least related to the reference population from the country itself (Haile-Mariam *et al.* 2015).

The second group of studies focused on combining populations from different breeds. In general, a lower benefit was obtained for combining populations from different breeds than for combining populations from different countries, see also Lund *et al*. (2014) for a review. Combining the closely related Nordic Red breeds resulted in a higher increase in accuracy (Zhou *et al.* 2014a) compared to combining the more distantly related Holstein Friesian breed with either Nordic Red breeds (Zhou *et al.* 2014b), different French cattle breeds (Karoui *et al.* 2012; Hozé *et al.* 2014a), Ayrshire breed (Chen *et al.* 2014), or the Jersey breed (Erbe *et al.* 2012; Olson *et al.* 2012), for which the increase in accuracy was almost negligible. In contrast to the expectations, the use of a higher density SNP chip (777,000 versus 50,000 SNPs) only resulted in a slight increase in the benefit of combining populations from different breeds (Erbe *et al.* 2012; Hozé *et al.* 2014a; Kemper *et al.* 2015b). Those results found in dairy cattle are in agreement with the results found in beef cattle and other livestock species. In beef cattle, for example, the increase in accuracy obtained by combining different breeds was low or negative as well (Chen *et al.* 2013; Kachman *et al.* 2013; Boerner *et al.* 2014), even when a high-density (777,000 SNPs) SNP chip was used (Bolormaa *et al.* 2013). In sheep, combining different breeds in one reference population had either no or a negative effect on the accuracy of genomic prediction (Legarra *et al.* 2014; Moghaddar *et al.* 2014), even when crossbred individuals that partly originated from the same breed were added (Moghaddar *et al.* 2014). In chicken, the effect on the accuracy of combining different lines was also shown to be absent or at most limited (Simeone *et al.* 2012; Calus *et al.* 2014; Huang *et al.* 2014), even when closely related lines were combined (Calus *et al.* 2014; Huang *et al.* 2014).

The different studies showed that the increase in accuracy was slightly higher for the breed with the lowest number of genotyped individuals (Hozé *et al.* 2014a), and for the individuals that were least related to the reference population from their own breed (Hozé *et al.* 2014a; Zhou *et al.* 2014b). Some studies showed higher accuracies of multi-population genomic prediction when a Bayesian variable selection model was used compared to a GBLUP model (Erbe *et al.* 2012; Bolormaa *et al.* 2013; Zhou *et al.* 2014a; Kemper *et al.* 2015b). Other studies, however, showed higher accuracies when a GBLUP model was used (Chen *et al.* 2013; Calus *et al.* 2014). For all studies, the differences in accuracies obtained with both models were generally small.

Different methods have been proposed to account for the differences in SNP effects across breeds, that are a result of differences in allele substitution effects of QTL and differences in LD between SNPs and QTL. Karoui *et al.* (2012) and Olson *et al.* (2012), for example, proposed a multi-trait GBLUP model, where the same trait in different breeds was modelled as a different, but correlated trait. The genetic correlation was generally estimated with a high standard error (Karoui *et al.* 2012; Huang *et al.* 2014) and the benefit of using a multi-trait model in combination with the estimated genetic correlation was low (Karoui *et al.* 2012; Huang *et al.* 2014; Legarra *et al.* 2014). Assuming a genetic correlation of 0.3 between breeds, however, resulted in a slightly higher accuracy of multi-breed genomic prediction (Olson *et al.* 2012). Chen *et al*. (2014) introduced a multi-task Bayesian variable selection approach, in which the breeds are combined to select the SNPs to be included in the model with a large effect, but SNP effects were estimated within breed. This multi-task Bayesian variable selection approach was shown to be able to increase the accuracy of genomic prediction for a small breed, even in situations where pooling the data resulted in a decrease in accuracy compared to within-population genomic prediction (Chen *et al.* 2014). Another approach that was investigated is to give a higher weight to the SNPs that explain a large part of the genetic variance in another population in the model of estimating SNP effects. This approach was shown to be able to slightly increase the accuracy of genomic prediction for a numerically small population (Brøndum *et al.* 2012; Hozé *et al.* 2014b; Khansefid *et al.* 2014).

Altogether, those recent findings in literature indicate that the expectation that combining distantly related breeds in one reference population can be beneficial as long as the marker density is high enough, was too optimistic. In the last four years, more and more information became available indicating that also other differences between populations, like differences in allele substitution effects of QTL and the presence of population-specific QTL, have to be taken into account for multi-population genomic prediction.

1.6 Predicting the accuracy of genomic prediction

Since the accuracy of predicting breeding values determines the response to selection, it is important to be able to predict the accuracy of genomic prediction before individuals are genotyped to be able to optimize breeding programs. As described in the previous paragraphs, different factors can affect the accuracy of multi-population genomic prediction, but the effect of each of those factors was not quantified at the start of this thesis. Therefore, it was difficult to realistically predict the potential accuracy of multi-population genomic prediction. A few different equations were, however, available to estimate the accuracy of genomic prediction within a population (e.g., Daetwyler *et al.* 2008; VanRaden 2008; Goddard 2009; Daetwyler *et al.* 2010). One type of equation can be derived both from selection index theory and from prediction error variance of the mixed model equations. This equation uses the genomic relationships within the reference and between reference and selection individuals to estimate the accuracy of genomic prediction (VanRaden 2008). Since genomic relationships between selection and reference individuals are needed, this equation cannot be used to predict the accuracy before individuals are genotyped. Another type of equation is using population parameters, such as the heritability of the trait, the number of individuals in the reference population and the effective number of chromosome segments (Daetwyler *et al.* 2008; Daetwyler *et al.* 2010). When estimates of the input parameters are available, the equation can be used to predict the accuracy before individuals are genotyped. All of the aforementioned equations can, however, only be used when breeding values are estimated for individuals from the same population as the reference population.

1.7 Objective and outline of this thesis

The overall objective of this thesis was to investigate the accuracy of multi-population genomic prediction in dairy cattle. This overall objective was divided in two sub-objectives. The first sub-objective was to investigate the effect of different factors on the accuracy of genomic prediction, such as the absence of close family relationships and differences across populations in LD patterns, allele frequencies, and allele substitution effects. The second sub-objective was to derive deterministic equations to predict the accuracy of multi-population genomic prediction.

In **Chapter 2**, the effect of absence of close family relationships between reference and selection individuals on the accuracy of genomic prediction was investigated. Moreover, it was investigated if deterministic prediction equations for the accuracy of genomic prediction, developed assuming populations of unrelated individuals, could be used to predict the accuracy in a population with a complex family structure. In **Chapter 3**, two deterministic equations to estimate the accuracy of across-population genomic prediction were derived. Furthermore, the effect of genetic correlations between populations lower than 1 and the number of QTL underlying the trait on across-population genomic prediction accuracy was investigated. The focus of this chapter was across-population genomic prediction, where the population of the selection candidates is not included in the reference population. In **Chapter 4**, the consistency of multi-locus LD across populations and its relationship with the accuracy of across-population genomic prediction was investigated. Here, it was expected that multi-locus LD was a better predictor for the potential of combining populations than consistency of LD between neighboring loci, since the effect of a QTL is distributed among a number of SNPs in genomic prediction models. **Chapter 5** studied the effect of QTL properties, such as allele frequency pattern and distribution of allele substitution effects, on accuracy of multi-breed genomic prediction. The objective of **Chapter 6** was to develop and validate a deterministic equation to predict the accuracy of genomic prediction when multiple populations are combined in the reference population.

The general discussion in **Chapter 7** of this thesis discusses five different topics, related to the studies in the earlier Chapters. As a first topic, the potential of multi-population genomic prediction is discussed by considering different scenarios, such as combining populations from the same breed from different countries, closely related breeds, or distantly related breeds. As a second topic, the impact of the model used to estimate GEBVs on the accuracy of multi-population genomic prediction is discussed. As a third topic, the possibility to estimate the genetic correlation based on SNP information is discussed. As a fourth topic, the relation between different measures for the consistency of LD across populations, namely the effective number of chromosome segments and the consistency in multi-locus LD, influencing the accuracy of multi-population genomic prediction is discussed. As a fifth topic, research directions for multi-population genomic prediction are discussed, focusing on the use of sequence data in genomic prediction, the identification and use of significant regions across populations, and the potential of including non-additive effects in genomic prediction models.

1.8 References

Andreescu, C., S. Avendano, S. R. Brown, A. Hassen, S. J. Lamont*, et al.*, 2007 Linkage disequilibrium in related breeding lines of chickens. Genetics 177: 2161-2169.

Bennewitz, J., N. Reinsch, H. Thomsen, J. Szyda, F. Reinhart*, et al.*, 2003 Marker assisted selection in German Holstein dairy cattle breeding: Outline of the program and marker assisted breeding value estimation. Proc. 54th Annu. Mtg. Eur. Assoc. Anim. Prod., Wageningen Academic Publishers, Rome.

Boerner, V., D. J. Johnston and B. Tier, 2014 Accuracies of genomically estimated breeding values from pure-breed and across-breed predictions in Australian beef cattle. Genet. Sel. Evol. 46: 61.

Bohren, B. B., W. G. Hill and A. Robertson, 1966 Some observations on asymmetrical correlated responses to selection. Genet. Res. 7: 44-57.

Boichard, D., S. Fritz, M.-N. Rossignol, M. Y. Boscher, A. Malafosse*, et al.*, 2002 Implementation of marker-assisted selection in French dairy cattle. Proc. 7th World Congr. Genet. Appl. Livest. Prod., INRA, Montpellier.

Bolormaa, S., J. E. Pryce, K. Kemper, K. Savin, B. J. Hayes*, et al.*, 2013 Accuracy of prediction of genomic breeding values for residual feed intake and carcass and meat quality traits in Bos taurus, Bos indicus, and composite beef cattle. J. Anim. Sci. 91: 3088-3104.

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

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

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

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. Huang, A. Vereijken, J. Visscher, J. Ten Napel*, et al.*, 2014 Genomic prediction based on data from three layer lines: a comparison between linear methods. Genet. Sel. Evol. 46: 57.

Chen, L., F. Schenkel, M. Vinsky, D. Crews and C. Li, 2013 Accuracy of predicting genomic breeding values for residual feed intake in Angus and Charolais beef cattle. J. Anim. Sci. 91: 4669-4678.

Chen, L., C. Li, S. Miller and F. Schenkel, 2014 Multi-population genomic prediction using a multi-task Bayesian learning model. BMC Genet. 15: 53.

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

De Haas, Y., M. P. L. Calus, R. F. Veerkamp, E. Wall, M. P. Coffey*, et al.*, 2012 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 Haas, Y., J. E. Pryce, M. P. L. Calus, E. Wall, D. P. Berry*, et al.*, 2015 Genomic prediction of dry matter intake in dairy cattle from an international data set consisting of research herds in Europe, North America, and Australasia. J. Dairy Sci. 98: 6522-6534.

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: 1545-1553.

Dekkers, J. C. M., 2004 Commercial application of marker- and gene-assisted selection in livestock: Strategies and lessons. J. Anim. Sci. 82 E-Suppl: E313-328.

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

Falconer, D. S. and T. F. C. Mackay, 1996 *Introduction to quantitative genetics*. Pearson Education Limited, Harlow.

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

Gautier, M., T. Faraut, K. Moazami-Goudarzi, V. Navratil, M. Foglio*, et al.*, 2007 Genetic and haplotypic structure in 14 European and African cattle breeds. Genetics 177: 1059-1070.

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

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

Haile-Mariam, M., J. E. Pryce, C. Schrooten and B. J. Hayes, 2015 Including overseas performance information in genomic evaluations of Australian dairy cattle. J. Dairy Sci. 98: 3443–3459.

Harris, B. L. and D. L. Johnson, 2010 Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. J. Dairy Sci. 93: 1243-1252.

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

Heifetz, E. M., J. E. Fulton, N. O'Sullivan, H. Zhao, J. C. M. Dekkers*, et al.*, 2005 Extent and consistency across generations of linkage disequilibrium in commercial layer chicken breeding populations. Genetics 171: 1173-1181.

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

Hozé, C., S. Fritz, F. Phocas, D. Boichard, V. Ducrocq*, et al.*, 2014b Genomic evaluation using combined reference populations from Montbéliarde and French Simmental breeds. Proc. 10th World Congr. Genet. Appl. Livest. Prod., ASAS, Vancouver.

Huang, H., J. J. Windig, A. Vereijken and M. P. Calus, 2014 Genomic prediction based on data from three layer lines using non-linear regression models. Genet. Sel. Evol. 46: 75.

Huang, W., S. Richards, M. A. Carbone, D. Zhu, R. R. H. Anholt*, et al.*, 2012 Epistasis dominates the genetic architecture of Drosophila quantitative traits. Proc. Nat. Acad. Sci. U. S. A. 109: 15553-15559.

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

Johansson, I., 1960 Progeny testing methods in Europe. J. Dairy Sci. 43: 706-713.

Jorjani, H., J. Jakobsen, M. A. Nilforooshan, E. Hjerpe, B. Zumbach*, et al.*, 2011 Genomic evaluation of BSW populations, InterGenomics: Results and Deliverables. Interbull Bull. 43: 5-8.

Kachman, S. D., M. L. Spangler, G. L. Bennett, K. J. Hanford, L. A. Kuehn*, et al.*, 2013 Comparison of molecular breeding values based on within- and across-breed training in beef cattle. Genet. Sel. Evol. 45: 30.

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

Kemper, K. E., B. J. Hayes, H. D. Daetwyler and M. E. Goddard, 2015a How old are quantitative trait loci and how widely do they segregate? J. Anim. Breed. Genet. 132: 121-134.

Kemper, K. E., C. M. Reich, P. J. Bowman, C. J. Vander Jagt, A. J. Chamberlain*, et al.*, 2015b Improved precision of QTL mapping using a nonlinear Bayesian method in a multi-breed population leads to greater accuracy for across-breed genomic predictions. Genet. Sel. Evol. 47: 29.

Khansefid, M., J. E. Pryce, S. Bolormaa, S. P. Miller, Z. Wang*, et al.*, 2014 Estimation of genomic breeding values for residual feed intake in a multibreed cattle population. J. Anim. Sci. 92: 3270-3283.

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

Legarra, A., G. Baloche, F. Barillet, J. Astruc, C. Soulas*, et al.*, 2014 Within- and across-breed genomic predictions and genomic relationships for Western Pyrenees dairy sheep breeds Latxa, Manech, and Basco-Béarnaise. J. Dairy Sci. 97: 3200-3212.

Lund, M. S., S. P. W. De Roos, A. G. De Vries, T. Druet, V. Ducrocq*, et al.*, 2011 A common reference population from four European Holstein populations increases reliability of genomic predictions. Genet. Sel. Evol. 43: 43.

Lund, M. S., G. Su, L. Janss, B. Guldbrandtsen and R. F. Brøndum, 2014 Invited review: Genomic evaluation of cattle in a multi-breed context. Livest. Sci. 166: 101-110.

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.

Moghaddar, N., A. A. Swan and J. H. Van der Werf, 2014 Comparing genomic prediction accuracy from purebred, crossbred and combined purebred and crossbred reference populations in sheep. Genet. Sel. Evol. 46: 58.

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

Patry, C., 2015 How international collaboration fostered an efficient use of the genomics for a reliable cattle breeding. Proc. Interbull Ind. Meeting, Interbull Centre, Verden.

Pryce, J. E., B. Gredler, S. Bolormaa, P. J. Bowman, C. Egger-Danner*, et al.*, 2011 Short communication: Genomic selection using a multi-breed, across-country reference population. J. Dairy Sci. 94: 2625-2630.

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

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

Simeone, R., I. Misztal, I. Aguilar and Z. G. Vitezica, 2012 Evaluation of a multi-line broiler chicken population using a single-step genomic evaluation procedure. J. Anim. Breed. Genet. 129: 3-10.

Smith, C., 1967 Improvement of metric traits through specific genetic loci. Anim. Prod. 9: 349-358.

Solberg, T. R., A. K. Sonesson, J. A. Woolliams and T. H. E. Meuwissen, 2008 Genomic selection using different marker types and densities. J. Anim. Sci. 86: 2447-2454.

Soller, M., 1978 The use of loci associated with quantitative effects in dairy cattle improvement. Anim. Prod. 27: 133-139.

Spelman, R. J., C. A. Ford, P. McElhinney, G. C. Gregory and R. G. Snell, 2002 Characterization of the DGAT1 gene in the New Zealand dairy population. J. Dairy Sci. 85: 3514-3517.

Thaller, G., W. Krämer, A. Winter, B. Kaupe, G. Erhardt*, et al.*, 2003 Effects of DGAT1 variants on milk production traits in German cattle breeds. J. Anim. Sci. 81: 1911-1918.

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*, et al.*, 2009 Invited review: Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92: 16-24.

VanRaden, P. M., K. Olson, D. Null, M. Sargolzaei, M. Winters*, et al.*, 2012 Reliability increases from combining 50,000-and 777,000-marker genotypes from four countries. Interbull Bull. 46: 75-79.

Wiggans, G. R., G. Su, T. A. Cooper, U. S. Nielsen, G. P. Aamand*, et al.*, 2015 Short communication: Improving accuracy of Jersey genomic evaluations in the United States and Denmark by sharing reference population bulls. J. Dairy Sci. 98: 3508-3513.

Zhou, L., X. Ding, Q. Zhang, Y. Wang, M. S. Lund*, et al.*, 2013 Consistency of linkage disequilibrium between Chinese and Nordic Holsteins and genomic prediction for Chinese Holsteins using a joint reference population. Genet. Sel. Evol. 45: 7.

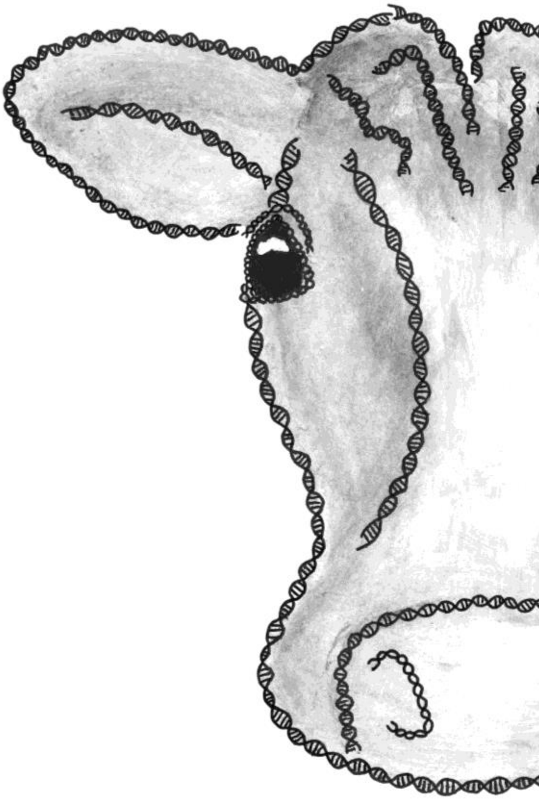
Zhou, L., B. Heringstad, G. Su, B. Guldbrandtsen, T. Meuwissen*, et al.*, 2014a Genomic predictions based on a joint reference population for the Nordic Red cattle breeds. J. Dairy Sci. 97: 4485-4496.

Zhou, L., M. S. Lund, Y. Wang and G. Su, 2014b Genomic predictions across Nordic Holstein and Nordic Red using the genomic best linear unbiased prediction model with different genomic relationship matrices. J. Anim. Breed. Genet. 131: 249-257.

Zumbach, B., H. Jorjani and J. Dürr, 2010 Brown Swiss genomic evaluation. Interbull Bull. 42: 44-51.

CHAPTER 2

THE EFFECT OF LINKAGE DISEQUILIBRIUM AND FAMILY RELATIONSHIPS ON THE RELIABILITY OF GENOMIC PREDICTION



Y.C.J. WIENTJES[1,2]

R.F. VEERKAMP[1,2]

M.P.L. CALUS[1]

[1] ANIMAL BREEDING AND GENOMICS CENTRE,
WAGENINGEN UR LIVESTOCK RESEARCH,
6700 AH WAGENINGEN, THE NETHERLANDS
[2] ANIMAL BREEDING AND GENOMICS CENTRE,
WAGENINGEN UNIVERSITY,
6700 AH WAGENINGEN, THE NETHERLANDS

GENETICS (2013) 193:621-631

Abstract

Although the concept of genomic selection relies on linkage disequilibrium (LD) between quantitative trait loci and markers, reliability of genomic predictions is strongly influenced by family relationships. In this study, we investigated the effects of LD and family relationships on reliability of genomic predictions and the potential of deterministic formulas to predict reliability using population parameters in populations with complex family structures. Five groups of selection candidates were simulated taking different information sources from the reference population into account: 1) allele frequencies; 2) LD pattern; 3) haplotypes; 4) haploid chromosomes; 5) individuals from the reference population, thereby having real family relationships with reference individuals. Reliabilities were predicted using genomic relationships among 529 reference individuals and their relationships with selection candidates and with a deterministic formula where the number of effective chromosome segments (M_e) was estimated based on genomic and additive relationship matrices for each scenario. At a heritability of 0.6, reliabilities based on genomic relationships were 0.002±0.0001 (allele frequencies), 0.022±0.001 (LD pattern), 0.018±0.001 (haplotypes), 0.100±0.008 (haploid chromosomes) and 0.318±0.077 (family relationships). At a heritability of 0.1, relative differences among groups were similar. For all scenarios, reliabilities were similar to predictions with a deterministic formula using estimated M_e. So, reliabilities can be predicted accurately using empirically estimated M_e and level of relationship with reference individuals has a much higher effect on the reliability than linkage disequilibrium *per se*. Furthermore, accumulated length of shared haplotypes is more important in determining the reliability of genomic prediction than the individual shared haplotype length.

Key words: genomic prediction, linkage disequilibrium, family relationships, reliability, effective chromosome segments

2.1 Introduction

Currently, it is feasible in most plant and animal breeding programs to genotype individuals at low costs for many thousands of single-nucleotide polymorphisms (SNPs) spread across the whole genome. With a sufficiently large reference population containing individuals with phenotypes and genotypes, SNP effects can be estimated. Subsequently, estimated SNP effects and an individual's genotype for each SNP can be used for genomic prediction of breeding values. Selection based on those genomic breeding values is called genomic selection (Meuwissen *et al.* 2001) and this method has high potential both in animal (e.g., Hayes *et al.* 2009a) and plant breeding (e.g., Heffner *et al.* 2009; Jannink *et al.* 2010). Many studies demonstrated higher reliabilities for direct genomic breeding values compared to breeding values based on pedigree information only, especially for juvenile individuals without phenotypic information (e.g., Meuwissen *et al.* 2001; Calus *et al.* 2008; VanRaden 2008).

The response to genomic selection relies on linkage disequilibrium (LD) between specific alleles of SNPs and quantitative trait loci (QTL) (Meuwissen *et al.* 2001); the stronger the LD, the higher the reliability of genomic predictions (Calus *et al.* 2008; Solberg *et al.* 2008). Since LD between QTL and SNPs will decrease over generations, reliability of genomic prediction is expected to decrease without re-estimating SNP effects in more recent generations (Muir 2007). However, the observed decrease in reliability of genomic predictions over generations following the generation in which SNP effects are estimated is higher than the expected decrease due to the decay of LD between SNPs and QTL alone (Habier *et al.* 2007; Habier *et al.* 2010). This higher decrease in reliability is a result of decreasing family relationships (i.e., all non-zero additive genetic relationships) over generations of the selection candidates with the reference population, indicating that SNPs used for genomic selection not only capture LD between SNP and QTL, but capture family relationships among individuals as well (Habier *et al.* 2007; Gianola *et al.* 2009; Habier *et al.* 2010). Indeed, several studies already showed higher reliabilities for genomic predictions when selection candidates were more closely related to the reference population (e.g., Meuwissen 2009; Habier *et al.* 2010; Makowsky *et al.* 2011).

Separating effects of LD and family relationships on the reliability of genomic predictions is difficult because LD and family relationships are entangled. The extent of LD in a population is related with effective population size (N_e) (Sved 1971); the lower the N_e, the higher the kinship level among individuals and the higher the extent of LD (Falconer and Mackay 1996). Besides that, LD can differ between families within breed (Dekkers 2004) and differs even more between diverged populations or breeds (De Roos *et al.* 2008; De Roos *et al.* 2009). A high marker density may enable achievement of similar LD between markers and QTL across breeds (De Roos *et al.* 2008), however, family relationships are still absent. Thus far, little is known about the effect of LD in situations without family relationships on the reliability of genomic predictions.

Deterministic formulas for predicting reliability of genomic prediction using population and trait parameters, which can be used before data on selection candidates are collected, are derived by Daetwyler *et al.* (2008) and Goddard (2009). Both formulas assume that selection candidates are unrelated to individuals from the reference population. Hayes *et al.* (2009d) applied the formula of Goddard (2009) to individuals that were related to the reference population, however, only simple family structures were used, such as selection candidates with full-sibs, half-sibs or double first cousins in the reference population. A deterministic method for predicting the reliability of genomic prediction that accounts for any type of family structure, by using all relationships among animals in a population, was derived by VanRaden (2008). However, the method of VanRaden (2008) uses genotypes of selection candidates and reference individuals to predict individual reliabilities instead of population parameters to predict the average reliability for a population. Therefore, this formula can be applied only after genotypic data are collected on selection candidates in contrast to the previous two deterministic formulas (Goddard *et al.* 2011). Family structures occur in real data and, so far, possibilities of applying deterministic formulas based on population parameters to predict reliability of genomic prediction are limited in such situations.

The first objective of this study was to examine the effects of LD and family relationships on the reliability of genomic predictions. The second objective of this study was to investigate whether deterministic prediction formulas for the reliability of genomic prediction based on population parameters can be used in real data sets with a complex family structure between selection candidates and individuals in the reference population. This article is organized as follows; first, we start by describing a real reference population set and the different sets of selection candidates simulated based on information from the reference population. Thereafter, the different methods to predict the reliabilities of the selection candidates are explained. Finally, results are presented and discussed.

2.2 Materials and methods

In this study, reliability of genomic prediction was predicted for five scenarios with simulated genotypes for selection candidates and using a reference population composed of real individuals with genotypic information. To create differences in LD and family relationships among the five scenarios, genotypes for the selection candidates were simulated using allele frequency, LD pattern, haplotypes, chromosomes, or family relationships from the reference population (Table 2.1). Finally, reliability of genomic prediction for each of the five scenarios was determined using two methods, namely those presented by: 1) VanRaden (2008), which explicitly accounts for family relationships between selection candidates and reference individuals, and 2) Daetwyler *et al.* (2008), where we aimed to account for family relationships by using an alternative way to estimate one of the parameters. For the last scenario, reliability was also empirically evaluated using observed phenotypic data and leave-one-out cross-validation.

Table 2.1 Overview of the information from the reference population used in the simulations of the different scenarios.

	Allele frequencies	**LD-pattern**	**Haplotypes**	**Chromosomes**	**Family relationships**
FREQ	X				
LD	X	X			
HAP	X	X	X		
CHR	X	X	X	X	
FAM	X	X	X	X	X

2.2.1 Reference population

The reference population consisted of 529 genotyped Holstein Friesian cows from the Netherlands. The cows were genotyped using the Illumina 50K SNP chip (Illumina, San Diego, CA), containing 54,001 SNPs. During a quality check, performed on a larger data set including those 529 cows, SNPs with a GCscore ≤0.2, a GTscore ≤0.55, a call rate ≤95%, a minor allele frequency ≤1%, deviating from Hardy-Weinberg equilibrium (X^2 ≥600), and SNPs that could not be assigned to a location on one of the chromosomes or were assigned to the X chromosome using the UMD3.0 bovine genome assembly from the University of Maryland were deleted. Individuals with Mendelian inconsistencies (Calus *et al.* 2011) between SNP data and pedigree in genotyped parent-offspring pairs and among sibs were removed. The software package Beagle (Browning and Browning 2007) was used to simultaneously phase the SNP data and impute any missing genotypes due to low call rates using the larger data set. One of the SNPs from each SNP pair with very high LD (i.e., r^2 >0.99) within the population of 529 individuals was deleted as well, to avoid problems of non-positive definite matrices during the analyses. Finally, 35,002 SNPs remained for the purpose of the study.

The data set used in this study contained many close family relationships. In total, the population contained 117 mother-daughter pairs, 48 full-sib families with on average 2.27 individuals per family, 69 paternal half-sib families with on average 7.23 individuals per family and 65 maternal half-sib families with on average 2.65 individuals per family.

2.2.2 Simulation of selection candidates

In this study, five different scenarios were considered in which genotypes of 529 selection candidates for 35,002 SNPs were simulated, using either the allele frequency, LD pattern, haplotypes, chromosomes, or family relationships from the reference population. The deterministic equations used to predict the individual reliabilities only used genotype information and considered variance components, so no phenotypes were simulated for the selection candidates. The last scenario was an exception to this, where we also used observed phenotypes for an empirical evaluation of the reliability.

2.2.2.1 FREQ

The first scenario (FREQ) simulated selection candidates using only allele frequencies of the reference population to show the potential reliability of genomic prediction in the absence of LD and family relationships. This scenario allocated genotypes to the simulated individuals with probabilities calculated by using the observed allele frequencies in the reference population, assuming that the loci were independent and that the population was in Hardy-Weinberg equilibrium.

2.2.2.2 LD

The second scenario (LD) used allele frequency and LD pattern between the SNPs of the reference population to simulate selection candidates, resulting in the potential reliability due to LD in the absence of family relationships. Only the 50 surrounding SNPs of a certain SNP were taken into account. To achieve this, a multivariate normal distribution was simulated by drawing one random number per SNP for each individual from a standard normal distribution, i.e., $N(0,1)$. Those random numbers were multiplied with the Cholesky decompositions of the correlation matrices between the SNPs per chromosome from the reference population. Whenever this correlation matrix was not positive definite, it was bended following Jorjani *et al.* (2003). The correlation matrices were calculated from the phased allelic data and represent LD, i.e., the square of those values is the well-known LD measure r^2 (Hill and Robertson 1968).

The random numbers drawn from the multivariate normal distribution were translated into genotypes by calculating two cut-off values on the normal distribution for each SNP using the allele frequency (p_i) of the reference population: 1) a cut-off value with an area of size $(1\text{-}p_i)^2$ to the left of it, and 2) a cut-off value with an area of size $(p_i)^2$ to the right of it. When the random number was below the first cut-off value (above the second cut-off value), the genotype of the individual for that SNP was set to -1 (1). When the random number was in between the two cut-off values, which was the case for a proportion of $2p_i(1\text{-}p_i)$ of the individuals, the genotype was set to 0.

2.2.2.3 HAP

Two individuals coming from the same population are expected to share some haplotypes, even if they do not share a common ancestor in the recent past. In this third scenario (HAP), the reliability due to sharing haplotypes with individuals in the reference population was investigated. The number of haplotypes used was equal to the number of effective chromosome segments, M_e, present in the reference population (estimation of M_e is explained later). For simplicity, all haplotypes were assumed to have an equal length in base pairs, although in reality haplotype length depends on LD structure of the genome. For each haplotype, 1058 (529*2) haploid copies were present in the reference population. Simulating selection candidates was done by randomly drawing two copies per haplotype from those 1058 copies and combining them across haplotypes to form the genome of the simulated individual. The number of haploid haplotypes shared between a simulated individual and a specific reference individual was divided equally over the 529 reference individuals. Note that this scenario is a theoretical scenario and used as an intermediate between the LD and FAM scenario.

2.2.2.4 CHR

VanRaden (2009) suggested a hypothetical scenario in which individuals are created by combining the best chromosomes present in a population to further increase the genetic progress. Although, e.g., chromosome substitution lines exist in mice by successive backcrossing of inbred lines (Nadeau *et al.* 2000; Singer *et al.* 2004), the scenario suggested by VanRaden (2009) is currently not feasible in practice for most animal and plant species. The reliability of those hypothetical individuals was investigated in this fourth scenario (CHR). As an alternative to picking the best chromosomes, we simulated individuals by randomly picking chromosomes from the reference population. Selection candidates in this scenario were in general simulated in the same way as in the HAP scenario, but instead of haplotypes, haploid chromosomes were used. The maximum number of haploid chromosomes shared between a simulated individual and a reference individual was restricted to one.

2.2.2.5 FAM

For this last scenario (FAM), instead of simulating genotypes of selection candidates, genotypes of real individuals were used to include family relationships. Each of the selection candidates had at least one genomic relationship of at least 0.125 with one of the individuals in the reference population, which is equal to the relationship of an individual with its great-grandparent. Reliabilities for this scenario were predicted by deleting each individual once from the reference population and using the remaining 528 individuals as reference population. This approach is also known as leave-one-out cross-validation and the effect due to differences of the composition of the reference population by one individual on the reliability is expected to be negligible.

For an empirical evaluation of the reliability of genomic prediction in this scenario, pre-corrected phenotypes on milk production were used. For all 529 cows used as selection candidate and reference individual, pre-corrected phenotypes were available. A detailed description of the pre-correction is given by Veerkamp *et al.* (2012).

All scenarios were set up such that allele frequencies across simulated selection candidates were expected to be similar to the allele frequencies observed in the reference population. Inspection of the simulated data showed that this was indeed the case.

2.2.3 Predicting reliability

Reliabilities were predicted in all scenarios using two different deterministic methods at a heritability of 0.1 and 0.6. One of the deterministic methods was also used to study the effect of the size of the reference population on the magnitude of effects of LD versus family relationships on the reliability of genomic prediction.

Besides both deterministic methods, reliabilities were also predicted using milk production phenotypes in the FAM scenario. For a good comparison of the empirical and deterministic predicted reliabilities, the estimated heritability for milk production based on the empirical data was used as well to predict the reliability of genomic prediction in the FAM scenario using the deterministic methods.

2.2.3.1 VanRaden (2008)

The first method to predict reliability was derived by VanRaden (2008) and predicted reliability of genomic prediction separately for each selection candidate as:

$$r_{VR}^2 = \mathbf{c}'\left[\mathbf{G}+\mathbf{I}\left(\frac{\sigma_e^2}{\sigma_a^2}\right)\right]^{-1}\mathbf{c}\,, \qquad (2.1)$$

in which **c** is a vector of genomic relationships of the selection candidate with each of the individuals in the reference population, **G** is the genomic relationship matrix of the reference population, **I** is an identity matrix, σ_e^2 is the residual variance and σ_a^2 is the additive genetic variance. The heritability (h^2) of the trait is reflected by $\frac{1-h^2}{h^2}=\frac{\sigma_e^2}{\sigma_a^2}$.

The genomic relationship matrix is calculated as $\mathbf{G}=\frac{\mathbf{XX}'}{n}$ (Yang *et al.* 2010), in which *n* is the number of SNPs. The **X** matrix contains standardized genotypes calculated as $x_{ij}=\frac{g_{ij}-2(p_i-0.5)}{\sqrt{2p_i(1-p_i)}}$, in which g_{ij} codes the genotype at SNP locus *i* for individual *j* as -1 for a homozygote, 0 for the heterozygote and 1 for the opposite homozygote and p_i is the allele frequency of the second allele at locus i (for which the homozygote genotype is coded 1). Subtraction of $2(p_i-0.5)$ from the genotype code sets the average value of the estimated allele effects per locus to zero. Division by $\sqrt{2p_i(1-p_i)}$ results in unbiased estimates of the relationships among individuals using **XX**’. Diagonal elements were calculated in the same way as off-diagonal elements, following Goddard *et al.* (2011) and Meuwissen *et al.* (2011).

Another common approach is to calculate **G** as $\frac{\mathbf{ZZ}'}{2\sum p_i(1-p_i)}$, in which **Z** is calculated as $g_{ij}-2(p_i-0.5)$ (e.g., VanRaden 2008; Legarra *et al.* 2009). This approach gives less weight to alleles with a low allele frequency, resulting in a weighted **G**. Meuwissen *et al.* (2011) suggested that the approach of Yang *et al.*

(2010), i.e., $\mathbf{G} = \frac{\mathbf{XX'}}{n}$, would result in the best, unweighted, estimate of **G** when a high proportion of loci with low minor allele frequencies are used. Therefore, the approach of Yang *et al.* (2010) was used to calculate **G** in this study.

The vector including genomic relationships of the selection candidate with each of the individuals in the reference population is computed as $\mathbf{c} = \frac{\mathbf{x}_2\mathbf{X'}}{n}$ (VanRaden 2008; Yang *et al.* 2010). In this calculation, **X** is the **X** matrix of the reference population and $\mathbf{x}_2$ is the **X** matrix of the selection candidates, which becomes a vector when only one selection candidate at a time is evaluated. Similarly, **c** becomes a vector as well.

The calculated **G** and **c** are biased, because **G** and **c** are based on a sample of segregating loci from the whole genome of an individual (Powell *et al.* 2010; Goddard *et al.* 2011). For an unbiased estimate of **G** (i.e., $\hat{\mathbf{G}}$), we assume that (Yang *et al.* 2010):

$$\hat{\mathbf{G}} = \mathbf{G} + \mathbf{E} = \mathbf{A} + (\mathbf{G} - \mathbf{A}) + \mathbf{E}, \tag{2.2}$$

in which **E** is a matrix with error terms due to sampling of the SNPs from the genome. The variances for those matrices are $Var(\hat{\mathbf{G}} - \mathbf{A}) = Var(\mathbf{G} - \mathbf{A}) + Var(\mathbf{E})$ in which $Var(\mathbf{E})$ is equal to $\frac{1}{n}$.

The unbiased $\hat{\mathbf{G}}$ was calculated by regressing **G** back to **A** as (Yang *et al.* 2010; Goddard *et al.* 2011):

$$\hat{\mathbf{G}} = \mathbf{A} + b(\mathbf{G} - \mathbf{A}), \tag{2.3}$$

in which

$$b = \frac{Var(\mathbf{G} - \mathbf{A})}{\left[Var(\mathbf{G} - \mathbf{A}) + Var(\mathbf{E})\right]} = \frac{Var(\hat{\mathbf{G}} - \mathbf{A}) - \frac{1}{n}}{Var(\hat{\mathbf{G}} - \mathbf{A})}. \tag{2.4}$$

The sampling error on the elements in $\hat{\mathbf{G}}$ depends on the level of family relationships, which is accounted for by calculating the regression coefficient *b* separately for bins of family relationships in **A** (0-0.10, >0.10-0.25, >0.25-0.50 and >0.50) with calculated *b*'s of respectively 0.973, 0.976, 0.990 and 0.997. All parent-offspring relationships were expected to be 0.5 and those relationships were excluded from the regression. Besides that, only off-diagonal elements were regressed.

Elements of **c** were regressed back to **A** as well, resulting in unbiased $\hat{\mathbf{c}}$. For the FAM scenario, the regression for **c** was done in the same way as for **G**, because $\hat{\mathbf{c}}$

was directly obtained from $\hat{\mathbf{G}}$. For the other scenarios, all family relationships between selection and reference individuals were zero, resulting in an **A** matrix where all elements were zero. Therefore the regression coefficient used for regressing **c** reduced to $b = Var(\mathbf{C}) \Big/ \left(Var(\mathbf{C}) + \frac{1}{n} \right)$, in which **C** is a matrix containing all **c** vectors with genomic relationships between selection and reference individuals.

2.2.3.2 Daetwyler et al. (2008)

The second formula for predicting the reliability of genomic predictions was derived by Daetwyler *et al.* (2008):

$$r_D^2 = \frac{N_p h^2}{N_p h^2 + N_g}, \tag{2.5}$$

in which h^2 is the heritability of the trait, N_p is the number of individuals in the reference population, and N_g is the number of independent loci underlying the trait. Assumptions underpinning this equation were: 1) loci are independent, 2) all loci have an effect, and 3) there are no family relationships between selection candidates and reference population. To account for the fact that segregating loci in real population are not independent, N_g was replaced by M_e in our study, as suggested by Daetwyler *et al.* (2008; 2010). Estimation of M_e is explained later. The formula of Daetwyler *et al.* (2008) provides one reliability that applies to the whole group of selection candidates, whereas r_{VR}^2 provides a single reliability for each selection candidate.

2.2.3.3 Impact of reference population size

The size of the reference population affects reliability of direct genomic values and, therefore, may also affect the magnitude of the effect of LD versus family relationships on the reliability. For this reason, we predicted the reliability using the formula of Daetwyler *et al.* (2008) for all five scenarios with different reference population sizes, ranging from 100 to 60,000 individuals. Heritability and M_e were assumed to be constant across different sizes of the reference population, reflecting a situation where reference individuals and selection candidates are a representative sample of the whole population.

2.2.3.4 Empirical estimation

In the FAM scenario, reliability of genomic prediction was empirically evaluated using pre-corrected phenotypes on milk production. Genomic breeding values for milk production were calculated for all individuals using a GBLUP-model in ASReml (Gilmour *et al.* 2009) and leave-one-out cross-validation. The GBLUP-model used the same genomic relationship matrix as used for the deterministic prediction of the reliabilities and explicitly estimated variances for the trait in the model. The average reliability across all individuals in the reference population was calculated as the squared correlation between the phenotypes and the genomic breeding values, divided by the heritability, as explained in Verbyla *et al.* (2010). The heritability for this trait was estimated from the same GBLUP model when all 529 reference individuals were included.

2.2.4 Estimating M_e

The M_e was estimated for each scenario using the genomic relationship matrix and the additive genetic relationship matrix. Only for the last scenario, FAM, we estimated M_e based on the estimated N_e as well, because this was the only scenario with a generation structure.

*2.2.4.1 Based on the **G** and **A** matrix*

Goddard *et al.* (2011) showed that the variance of off-diagonal elements of **G** for unrelated individuals, all having expected values of zero, is about equal to the average of r_{LD}^2 (i.e., $\bar{r}_{LD}^2$) as a measure of LD over all pairs of loci. This $\bar{r}_{LD}^2$, and therefore the variance of **G** as well, is related with M_e as $M_e = \frac{1}{\bar{r}_{LD}^2} = \frac{1}{Var(\mathbf{G})}$. For related individuals, we can use $\mathbf{D} = \mathbf{G} - \mathbf{A}$, in which **G** is the genomic relationship matrix and **A** the additive genetic relationship matrix, where the expected values for all elements of **D** are zero. This suggests that $Var(\mathbf{D})$ is related to $\bar{r}_{LD}^2$ over all pairs of loci and, therefore, that M_e for a specific population with related individuals can be estimated as:

$$M_e = \frac{1}{Var(\mathbf{D})}. \tag{2.6}$$

In the formula for calculating **D**, **G** should contain the genomic relationships between reference individuals and selection candidates (Goddard *et al.* 2011). Following our earlier notation, here we use the $\hat{\mathbf{C}}$ matrix, containing all $\hat{\mathbf{c}}$ vectors with the relationships between selection and reference individuals. For the FAM scenario, **A** was calculated based on the pedigree. In the other scenarios, individuals were simulated without family relationships with the reference individuals and therefore lacked pedigree information. For those scenarios, additive genetic relationships between selection candidates and reference individuals were assumed to be zero.

2.2.4.2 Based on N_e

For the FAM scenario, M_e was also estimated based on N_e. In this study, we used the two most frequently used formulas, namely $M_e = \frac{2N_eL}{\ln(4N_eL)}$ (Goddard 2009) and $M_e = 2N_eL$ (Hayes *et al.* 2009d). In those formulas, L was the genome size that was assumed to be 31.6 M (Ihara *et al.* 2004). The required value for N_e was estimated for the reference population. For each t generations back, N_e is correlated with a mean r_{LD}^2 (i.e., $\bar{r}_{LD}^2$) as a measure of LD over a chromosome segment with length $c = \frac{1}{2t}$ (Hayes *et al.* 2003), in which c is the length of the chromosome segment in morgans. All r_{LD}^2 of SNP intervals in between the chromosome segment length using $(t-0.1)$ and $(t+0.1)$ and assuming 1 cM = 1 Mb were averaged to calculate $\bar{r}_{LD}^2$, which is used to estimate N_e following $\bar{r}_{LD}^2 = \frac{1}{4N_ec+1}$ (Sved 1971). For t the values 1-5 were used and the final N_e of the population was estimated as the mean N_e over those last 5 generations.

2.3 Results

2.3.1 Reliabilities of the different scenarios

The different scenarios showed predicted reliabilities of 0.002 ± 0.0001 (FREQ), 0.022 ± 0.001 (LD), 0.018 ± 0.001 (HAP), 0.100 ± 0.008 (CHR) and 0.318 ± 0.077 (FAM) using the formula of VanRaden (2008) at a heritability of 0.6 (rel_VR; Figure 2.1A). This indicates that reliability of selection candidates that share only allele frequencies with the reference population was almost zero. Adding the LD pattern or haplotype information as information source used for simulating selection candidates slightly increased the reliability. Using chromosomes from the reference population to simulate selection candidates showed an increase in reliability of about 0.1. Adding family relationships between selection candidates and reference individuals resulted in a relatively high increase in reliability compared to the other scenarios (an increase of >0.3 compared to the FREQ scenario and >0.2 compared to the CHR scenario). So, the average reliabilities of genomic predictions increased by simulating selection candidates using an increasing amount of information from the reference population and this increase was highest when family relationships were added as an information source.

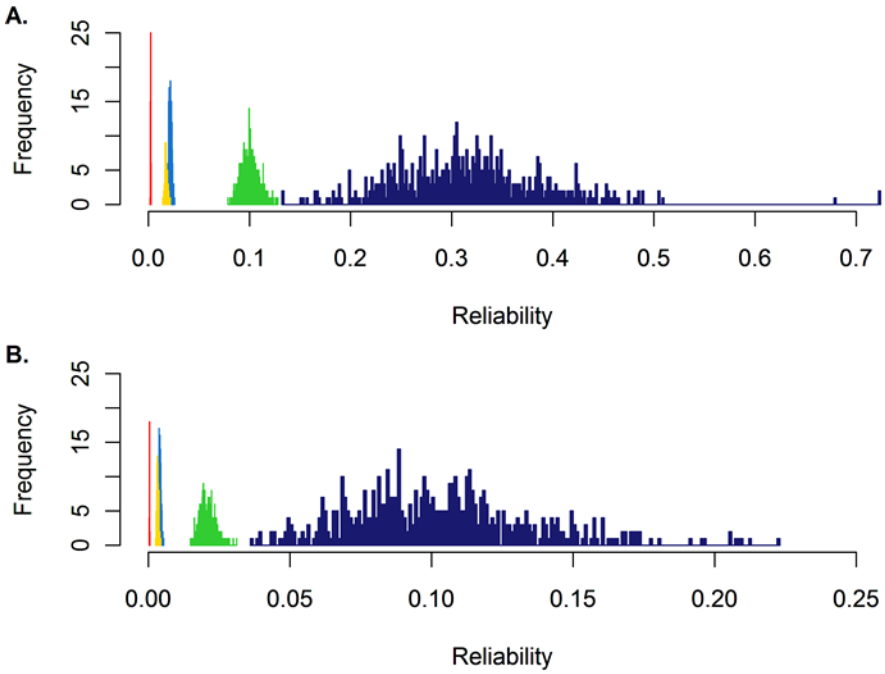


Figure 2.1 Histograms depicting distributions of reliabilities of genomic preditions using a reference population of 529 genotyped individuals at a heritability of 0.6 (A) and 0.1 (B) over the five different scenarios using different information sources from the reference population (from left to right): 🟥 = Selection candidates simulated based on allele frequency of the reference population (FREQ); 🟨 = Selection candidates simulated based on 837 haplotypes of equal length segregating in the reference population (HAP); 🟦 = Selection candidates simulated based on LD pattern of the reference population (LD); 🟩 = Selection candidates simulated based on haploid chromosomes segregating in the reference population (CHR); ■ = Individuals from the reference population (FAM).

Next to the increase in reliability when more information from the reference population was used to simulate selection candidates, variation in reliability among selection candidates increased as well (Figure 2.1A). Especially the variation in the FAM scenario, using family relationships between selection candidates and reference individuals, was high compared to the other scenarios and the reliabilities in that scenario ranged from 0.13 to 0.72. The distributions of the reliabilities overlapped between the LD and HAP scenario. For the other scenarios, the distributions were not overlapping.

For all scenarios, rel_VR was lower at a heritability of 0.1 compared to a heritability of 0.6, but relative differences between and standard deviations of reliabilities within groups were similar to those observed at a heritability of 0.6 (Figure 2.1B).

2.3.2 Applying the formula of Daetwyler *et al.* (2008) to populations with a complex family structure

Another method used to predict reliability of genomic prediction is the formula of Daetwyler *et al.* (2008). A disadvantage of this formula is the inability to predict reliabilities for populations with a complex family structure. In this study, this disadvantage was overcome by estimating M_e in the formula based on the genomic and additive genetic relationship matrix. At the same heritability, reliabilities predicted with the formula of Daetwyler *et al.* (2008), denoted as rel_D hereafter, were in good agreement with rel_VR presented before, being: 0.003 (FREQ), 0.027 (LD), 0.021 (HAP), 0.129 (CHR) and 0.275 (FAM; Table 2.2). Those predicted rel_D values at a heritability of 0.6 were almost equal to rel_VR for the FREQ scenario and the difference was highest for the FAM scenario (0.043). At a heritability of 0.1, predicted rel_D and rel_VR were equal for the FREQ and LD scenario and the maximum difference was 0.044 (FAM).

The formula of Daetwyler *et al.* (2008) was also applied to study the effect of size of the reference population on the magnitude of effects of LD versus family relationships on the reliability of genomic prediction. Reliabilities at a heritability of 0.6 of all five scenarios using different sizes of the reference population are shown in Figure 2.2. For the FAM scenario, reliability shows a steep marginal increase by increasing reference population size at small initial sizes of the reference population. At reference population sizes of about 5000-10,000, when reliability approaches the maximum reliability of 1, the marginal increase in reliability starts to decline. For the LD scenario, the marginal increase is more gradual; so less steep at small sizes of the reference population and more steep at bigger sizes of the reference population. The increase in reliability is, however, still higher at small initial sizes of the reference population compared to bigger sizes. For the CHR, the pattern is in between the ones from the FAM and LD scenario, and for the HAP scenario, the pattern is more or less the same as for the LD scenario. For the FREQ scenario, the increase in reliability is almost linear across the considered range of reference population sizes. Those results indicate that the effect of LD versus family relationship does indeed depend on the size of the reference population.

2.3.3 Empirical estimation

In the FAM scenario, empirical estimation of the reliability using leave-one-out cross-validation for milk production resulted in an estimated reliability of 0.291. At the heritability estimated for milk production in this data set (0.56), the FAM scenario showed a rel_VR of 0.305 and rel_D of 0.261. So, both deterministic predictions were very close to the empirically estimated reliability.

Table 2.2 Comparison of average reliabilities of genomic predictions at different heritabilities for five different scenarios obtained with the deterministic formulas of VanRaden (2008) (rel_VR) and Daetwyler *et al.* (2008) (rel_D), using the estimated number of effective chromosome segments (M_e).

h^2	Scenario	M_e[a]	Rel_VR	Rel_D
0.6	**FREQ**	122116	0.002	0.003
0.6	**LD**	11458	0.022	0.027
0.6	**HAP**	14627	0.018	0.021
0.6	**CHR**	2139	0.100	0.129
0.6	**FAM**	837	0.318	0.275
		805[b]		0.283
		7774[c]		0.039
0.1	**FREQ**	122116	0.0004	0.0004
0.1	**LD**	11458	0.004	0.005
0.1	**HAP**	14627	0.003	0.004
0.1	**CHR**	2139	0.021	0.024
0.1	**FAM**	837	0.104	0.059
		805[b]		0.062
		7774[c]		0.007

[a] M_e estimated as $M_e = \frac{1}{Var(\mathbf{D})}$;

[b] M_e estimated as $M_e = \frac{2N_eL}{\ln(4N_eL)}$ (Goddard 2009);

[c] M_e estimated as $M_e = 2N_eL$ (Hayes *et al.* 2009d).

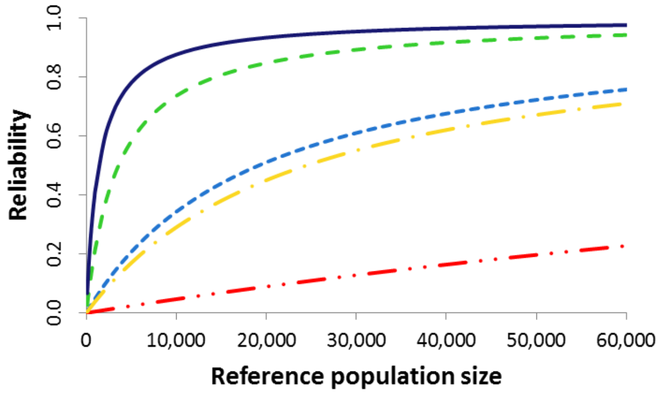


Figure 2.2 Predicted reliability of genomic prediction, at a heritability of 0.6 and different sizes of the reference population, obtained with the deterministic formula of Daetwyler *et al.* (2008) for the five different scenarios using different information sources from the reference population: — · · = Selection candidates simulated based on allele frequency of the reference population (FREQ); - - - - - = Selection candidates simulated based on LD pattern of the reference population (LD); — · — = Selection candidates simulated using 837 haplotypes of equal length segregating in the reference population (HAP); — — — = Selection candidates simulated based on haploid chromosomes segregating in the reference population (CHR); ——— = Individuals from the reference population (FAM).

2.3.4 Calculating N_e and M_e

The N_e of the reference population was estimated to be 123 and this value was used to approximate the M_e of the FAM scenario using two different formulas. The first formula, $M_e = \frac{2N_eL}{\ln(4N_eL)}$ (Goddard 2009), resulted in almost the same M_e as based on the genomic and additive genetic relationship matrix and, therefore, predicted reliability using this value was in good agreement with rel_VR and rel_D (Table 2.2). The second formula, $M_e = 2N_eL$ (Hayes *et al.* 2009d), showed an almost 10 times higher value for M_e, resulting in a much lower predicted reliability compared to rel_VR and rel_D.

2.3.5 Genomic relationship versus reliability

Since the reliability predicted with the formula of VanRaden (2008) was predicted separately for each individual, it was possible to evaluate the relation between genomic relationship and reliability. Average squared genomic relationship, which was found to be an accurate indicator of reliability in the study of Pszczola *et al.* (2012), also showed a high correlation with reliability in our study (Figure 2.3); the higher the average squared relationship with the reference population, the higher the reliability of genomic prediction. Fitting a linear regression line through the data presented in Figure 2.3A resulted in a model R^2 ranging from 0.51 to 0.60 (FREQ=0.57, LD=0.54, HAP=0.58, CHR=0.60, FAM=0.51) at a heritability of 0.6. The mean and variance of the average squared genomic relationship within a scenario were both affected by the relationship with the reference population, i.e., using more information from the reference population to simulate the selection candidates resulted in a higher mean and variance of the average squared genomic relationship.

The relation between average squared relationships and reliability at heritability values of 0.1 and 0.6 was very similar (Figure 2.3B). Nevertheless, average squared relationship predicted the reliabilities more accurately at a heritability of 0.1, with a R^2 of the regression model ranging from 0.92 to 0.94 (FREQ=0.92, LD=0.92, HAP=0.92, CHR=0.94, FAM=0.93).

2.4 Discussion

2.4.1 Effect of LD and family relationships on reliability

The first aim of this study was to investigate the effects of LD and family relationships on the reliability of direct genomic values. The results indicate that family relationships between selection candidates and reference population can have a large effect on the reliability of genomic predictions compared to linkage disequilibrium *per se*.

The difference in reliability between selection candidates distantly and closely related to the reference population in our study was >0.5 at a heritability of 0.6. For breeding practices, it is therefore advisable to predict reliability for each selection candidate individually. However, it should be noted that both the general level and the variation of relationships within the data set used in our study was high, and the reference population was small. In data sets used for breeding practices, the difference in relationships among selection candidates may be lower and the size of the reference population may be higher, resulting in smaller differences in reliability.

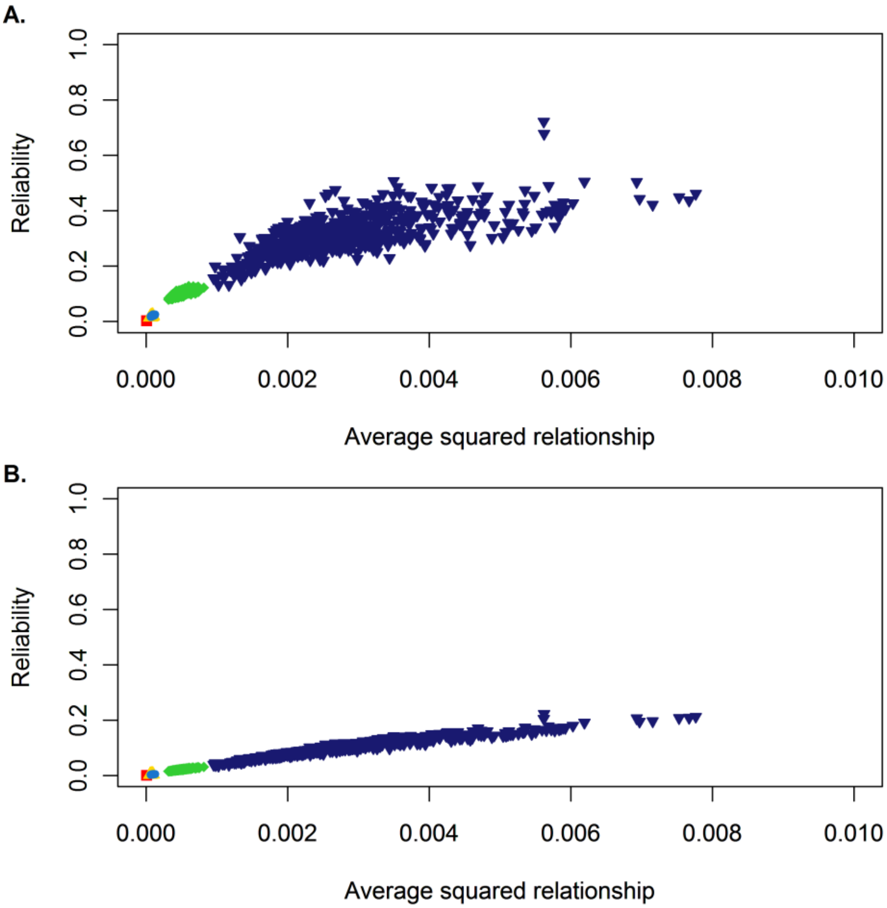


Figure 2.3 Average squared relationships to the reference population versus the reliability of genomic predictions at a heritability of 0.6 (A) and 0.1 (B) for the five different scenarios using different information sources from the reference population (from left to right): ■ = Selection candidates simulated based on allele frequency of the reference population (FREQ); ▲ = Selection candidates based on 837 haplotypes of equal length segregating in the reference population (HAP); ● = Selection candidates simulated based on LD pattern of the reference population (LD); ♦ = Selection candidates simulated based on haploid chromosomes segregating in the reference population (CHR); ▼ = Individuals from the reference population (FAM).

The size of the reference population influences the relative effect of LD and family relationships on the reliability of genomic prediction; small reference populations result in a higher effect of family relationships compared to LD, and larger reference populations result in a higher effect of LD on reliability. Those results are in agreement with the results of Clark *et al.* (2012), who stated that the effect of family relationships is reduced at an increasing size of the reference population. Size of the reference population combined with the high general level of relationships between selection candidates and reference individuals in our study also explains at least part of the difference between our results and results of Habier *et al.* (2007), who found that less than half of the reliability of a population one generation younger than the reference population, including both parents, was due to family relationships.

Both deterministic approaches used in this study to predict the reliability of genomic prediction are based on a genomic relationship matrix. The genomic relationship matrix is quite consistent over different numbers of SNPs, with a correlation >0.98 when anywhere between ~10,000 and 40,000 SNPs are used to set up the matrix (Rolf *et al.* 2010). Therefore, the conclusions of our study are supposed to be independent from the number of SNPs used to set up the genomic relationship matrix, provided that at least 10,000 SNPs are used.

The reliabilities achieved in the LD and HAP scenario are very similar. This indicates that most of the information coming from the considered haplotypes in the HAP scenario coincides with the information captured by the LD pattern in our data. Decreasing the number of haplotypes, and thereby increasing the haplotype length, will result in a higher additional amount of information captured in the HAP scenario compared to the LD scenario. The most extreme scenario of haplotypes in terms of their length is represented by the CHR scenario, which showed a considerably higher reliability than LD and HAP.

Length of haplotypes identical by descent between two individuals is related to the number of generations diverged from the common ancestor (Chapman and Thompson 2003; Browning 2008). The length of chromosome segments shared between individuals is, therefore, expected to be correlated with the level of family relationships between individuals (Sved 1971; VanRaden *et al.* 2011); and also with the reliability of genomic prediction. The results in our study do not completely agree with these expectations. In the CHR scenario, simulated individuals shared whole un-recombined chromosomes with the reference population. The genomic relationship and reliability was, however, lower than achieved in the FAM scenario, where individuals had shorter haplotypes in common with reference individuals. In the CHR scenario, selection candidates had only one long haplotype in common with any one reference individual; while in the FAM scenario, more shorter haplotypes were shared between a selection candidate and the same reference individual resulting in a higher relationship due to a higher accumulated length of shared haplotypes and, therefore, a higher reliability of genomic prediction. Moreover, this indicates that reliabilities of individuals composed of the best chromosomes present in a population, assuming this would be possible without going through the usual process of meiosis and recombination, as suggested by VanRaden (2009) and Cole and VanRaden (2011), may be substantially lower compared to individuals that have some degree of family relationship to one or more reference individuals. So, accumulated length of shared haplotypes between selection candidates and individuals in the reference population is more important than individual length of shared haplotypes.

2.4.2 Predicting the reliability for populations with a complex family structure

The second aim of this article was to investigate whether deterministic prediction formulas for the reliability of genomic prediction using population parameters can be used in situations with a complex family structure between selection candidates and the reference population. The results show that the formula of Daetwyler *et al.* (2008), using M_e estimated based on the difference between genomic and additive genetic relationship matrices, yields similar predicted reliabilities for populations with a complex pedigree structure as using the formula of VanRaden (2008) and a cross-validation method based on observed phenotypes.

The formula of VanRaden (2008) can be used to predict the reliability of genomic prediction for populations with a complex family structure. Previous studies that performed an empirical evaluation of the formula of VanRaden (2008), which is equal to predicting the reliability based on the prediction error variance as shown by Strandén and Garrick (2009), in general overestimated the reliability (Hayes *et al.* 2009b; Lund *et al.* 2009; Thomasen *et al.* 2012). This overestimation can be reduced by regressing the genomic relationship matrix back to the additive genetic relationship matrix calculated from pedigree information (Goddard *et al.* 2011). In our study, using such regressed genomic relationship matrix resulted in good agreement between the reliability predicted with the formula of VanRaden (2008) and the empirically estimated reliability.

Previous empirical evaluations of the formula of Daetwyler *et al.* (2008) all showed good agreement between empirically and deterministically derived reliabilities (Hayes *et al.* 2009c; Clark *et al.* 2012; Pryce *et al.* 2012). This formula assumes that selection candidates and reference individuals are unrelated. In our study, family structure between reference and selection individuals was taken into account in the prediction of M_e. Agreement between empirically estimated reliability and the reliabilities predicted with the formulas of VanRaden (2008) and Daetwyler *et al.* (2008) shows that the formula of Daetwyler *et al.* (2008) can also be applied to populations with a complex family structure, by using a value for M_e that represents the family structure in the population.

The M_e estimated as $2N_eL$ (Hayes *et al.* 2009d) was much higher, resulting in an unrealistically low reliability, compared to the M_e and reliability estimated with $M_e = \frac{1}{Var(\mathbf{G}-\mathbf{A})}$. The other formula used to estimate M_e, $M_e = \frac{2N_eL}{\ln(4N_eL)}$ (Goddard 2009), resulted in a similar value for M_e as using $M_e = \frac{1}{Var(\mathbf{G}-\mathbf{A})}$, indicating that the reliabilities of genomic prediction using $M_e = \frac{1}{Var(\mathbf{G}-\mathbf{A})}$ were similar to those using $M_e = \frac{2N_eL}{\ln(4N_eL)}$ in the formula of Daetwyler *et al.* (2008).

2.4.3 Implications

Currently, more and more research is focused on the use of multi-breed or multi-line reference populations to enable genomic selection for smaller breeds or lines. Compared to within-breed genomic prediction, reliability of across-breed predictions may be lower due to differences in allele frequencies, LD pattern, and haplotypes among breeds (e.g., De Roos *et al.* 2008; Pryce *et al.* 2010; Goddard 2012) and because family relationships among full-bred individuals of different breeds are absent (VanRaden *et al.* 2011). In addition, breed-specific allele effects might exist (Spelman *et al.* 2002; Thaller *et al.* 2003), which further reduces the reliability of genomic prediction for multi-breed populations.

A high marker density is expected to increase the consistency of LD between SNPs and QTL across breeds and the corresponding reliability (De Roos *et al.* 2008; Ibánẽz-Escriche *et al.* 2009). The problem of different allele frequencies and breed-specific allele effects can, however, not be solved by a higher marker density. Therefore, the expected reliability using a reference population of another breed is supposed to be lower than the reliability in the LD scenario in our study. Estimating M_e for such scenarios, as shown in this study for populations with a complex family structure, is a potential starting point for predicting the reliability for those multi-breed population structures.

2.5 Conclusion

In conclusion, our results showed that the level of family relationships between selection candidates and the reference population has a higher effect on the reliability of direct genomic values than linkage disequilibrium *per se*. Furthermore, accumulated length of shared haplotypes across a reference individual and a selection candidate are more important in determining the reliability of genomic prediction than individual length of shared haplotypes. And finally, existing deterministic formulas using population parameters can accurately predict the reliability of genomic prediction using reference populations with complex family structures by estimating the number of effective chromosome segments based on genomic and additive genetic relationship matrices.

2.6 Acknowledgments

The authors are thankful for useful comments from Chris Schrooten and Henk Bovenhuis. The authors acknowledge financial support of Coöperatie Rundvee Verbetering B.V. (CRV, Arnhem, The Netherlands).

2.7 References

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

Browning, S. R., 2008 Estimation of pairwise identity by descent from dense genetic marker data in a population sample of haplotypes. Genetics 178: 2123-2132.

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 and J. W. M. Bastiaansen, 2011 Identification of Mendelian inconsistencies between SNP and pedigree information of sibs. Genet. Sel. Evol. 43: 34.

Chapman, N. H. and E. A. Thompson, 2003 A model for the length of tracts of identity by descent in finite random mating populations. Theor. Popul. Biol. 64: 141-150.

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

Cole, J. B. and P. M. VanRaden, 2011 Use of haplotypes to estimate Mendelian sampling effects and selection limits. J. Anim. Breed. Genet. 128: 446-455.

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

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: 1545-1553.

Dekkers, J. C. M., 2004 Commercial application of marker- and gene-assisted selection in livestock: Strategies and lessons. J. Anim. Sci. 82 E-Suppl: E313-328.

Falconer, D. S. and T. F. C. Mackay, 1996 *Introduction to quantitative genetics*. Pearson Education Limited, Harlow.

Gianola, D., G. De Los Campos, W. G. Hill, E. Manfredi and R. L. Fernando, 2009 Additive genetic variability and the Bayesian alphabet. Genetics 183: 347-363.

Gilmour, A. R., B. Gogel, B. Cullis, R. Thompson, D. Butler*, et al.*, 2009 *ASReml user guide release 3.0*. VSN International Ltd, Hemel Hempstead.

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

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.

Goddard, M. E., 2012 Uses of genomics in livestock agriculture. Anim. Prod. Sci. 52: 73-77.

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., 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.

Hayes, B. J., P. M. Visscher, H. C. McPartlan and M. E. Goddard, 2003 Novel multilocus measure of linkage disequilibrium to estimate past effective population size. Genome Res. 13: 635-643.

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. 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*, et al.*, 2009c Accuracy of genomic selection: comparing theory and results. Proc. Assoc. Advmt. Anim. Breed. Genet., Barossa Valley, South Australia.

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

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

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

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

Ihara, N., A. Takasuga, K. Mizoshita, H. Takeda, M. Sugimoto*, et al.*, 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.

Jorjani, H., L. Klei and U. Emanuelson, 2003 A simple method for weighted bending of genetic (co)variance matrices. J. Dairy Sci. 86: 677-679.

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

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.

Makowsky, R., N. M. Pajewski, Y. C. Klimentidis, A. I. Vazquez, C. W. Duarte*, et al.*, 2011 Beyond missing heritability: Prediction of complex traits. PLoS Genet. 7: e1002051.

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., 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.

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.

Nadeau, J. H., J. B. Singer, A. Matin and E. S. Lander, 2000 Analysing complex genetic traits with chromosome substitution strains. Nat. Genet. 24: 221-226.

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

Pryce, J. E., M. Haile-Mariam, K. Verbyla, P. J. Bowman, M. E. Goddard*, et al.*, 2010 Genetic markers for lactation persistency in primiparous Australian dairy cows. J. Dairy Sci. 93: 2202-2214.

Pryce, J. E., J. Arias, P. J. Bowman, S. R. Davis, K. A. Macdonald*, et al.*, 2012 Accuracy of genomic predictions of residual feed intake and 250-day body weight in growing heifers using 625,000 single nucleotide polymorphism markers. J. Dairy Sci. 95: 2108-2119.

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

Rolf, M. M., J. F. Taylor, R. D. Schnabel, S. D. McKay, M. C. McClure*, et al.*, 2010 Impact of reduced marker set estimation of genomic relationship matrices on genomic selection for feed efficiency in Angus cattle. BMC Genet. 11: 24.

Singer, J. B., A. E. Hill, L. C. Burrage, K. R. Olszens, J. Song*, et al.*, 2004 Genetic dissection of complex traits with chromosome substitution strains of mice. Science 304: 445-448.

Solberg, T. R., A. K. Sonesson, J. A. Woolliams and T. H. E. Meuwissen, 2008 Genomic selection using different marker types and densities. J. Anim. Sci. 86: 2447-2454.

Spelman, R. J., C. A. Ford, P. McElhinney, G. C. Gregory and R. G. Snell, 2002 Characterization of the DGAT1 gene in the New Zealand dairy population. J. Dairy Sci. 85: 3514-3517.

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

Sved, J. A., 1971 Linkage disequilibrium and homozygosity of chromosome segments in finite populations. Theor. Popul. Biol. 2: 125-141.

Thaller, G., W. Krämer, A. Winter, B. Kaupe, G. Erhardt*, et al.*, 2003 Effects of DGAT1 variants on milk production traits in German cattle breeds. J. Anim. Sci. 81: 1911-1918.

Thomasen, J. R., B. Guldbrandtsen, G. Su, R. F. Brøndum and M. S. Lund, 2012 Reliabilities of genomic estimated breeding values in Danish Jersey. Animal 6: 789-796.

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

VanRaden, P. M., 2009 Future animal improvement programs applied to global populations. Interbull Bull. 40: 247-251

VanRaden, P. M., K. M. Olson, G. R. Wiggans, J. B. Cole and M. E. Tooker, 2011 Genomic inbreeding and relationships among Holsteins, Jerseys, and Brown Swiss. J. Dairy Sci. 94: 5673-5682.

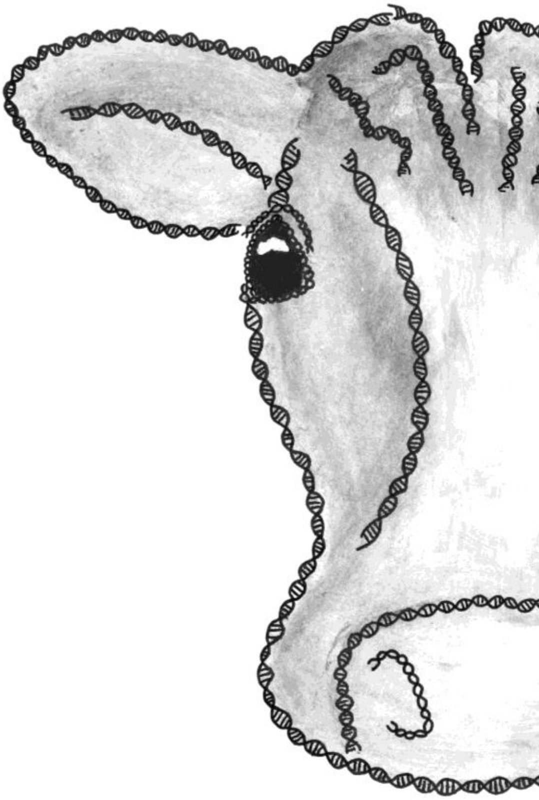
Veerkamp, R. F., M. P. Coffey, D. P. Berry, Y. De Haas, E. Strandberg*, et al.*, 2012 Genome-wide associations for feed utilisation complex in primiparous Holstein-Friesian dairy cows from experimental research herds in four European countries. Animal 6: 1738-1749.

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.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders*, et al.*, 2010 Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42: 565-569.

CHAPTER 3

EMPIRICAL AND DETERMINISTIC ACCURACIES OF ACROSS-POPULATION GENOMIC PREDICTION



Y.C.J. WIENTJES[1,2]

R.F. VEERKAMP[1,2]

P. BIJMA[1]

H. BOVENHUIS[1]

C. SCHROOTEN[3]

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,
6700 AH WAGENINGEN, THE NETHERLANDS
[3] CRV BV, 6800 AL ARNHEM, THE NETHERLANDS

GENETIC SELECTION EVOLUTION (2015) 47:5

Abstract

Background: Differences in linkage disequilibrium and in allele substitution effects of quantitative trait loci (QTL) may hinder genomic prediction across populations. Our objective was to develop a deterministic formula to estimate the accuracy of across-population genomic prediction, for which reference individuals and selection candidates are from different populations, and to investigate the impact of differences in allele substitution effects across populations and of the number of QTL underlying a trait on the accuracy.

Methods: A deterministic formula to estimate the accuracy of across-population genomic prediction was derived based on selection index theory. Moreover, accuracies were deterministically predicted using a formula based on population parameters and empirically calculated using simulated phenotypes and a GBLUP (genomic best linear unbiased prediction) model. Phenotypes of 1033 Holstein Friesian, 105 Groningen White Headed and 147 Meuse-Rhine-Yssel cows were simulated by sampling 3000, 300, 30 or 3 QTL from the available high-density single-nucleotide polymorphism (SNP) information of three chromosomes, assuming a correlation of 1.0, 0.8, 0.6, 0.4, or 0.2 between allele substitution effects across breeds. The simulated heritability was set to 0.95 to resemble the heritability of deregressed proofs of bulls.

Results: Accuracies estimated with the deterministic formula based on selection index theory were similar to empirical accuracies for all scenarios, while accuracies predicted with the formula based on population parameters overestimated empirical accuracies by ~25 to 30%. When the between-breed genetic correlation differed from 1, i.e., allele substitution effects differed across breeds, empirical and deterministic accuracies decreased in proportion to the genetic correlation. Using a multi-trait model, it was possible to accurately estimate the genetic correlation between the breeds based on phenotypes and high-density genotypes. The number of QTL underlying the simulated trait did not affect the accuracy.

Conclusions: The deterministic formula based on selection index theory estimated the accuracy of across-population genomic predictions well. The deterministic formula using population parameters overestimated the across-population genomic prediction accuracy, but may still be useful because of its simplicity. Both formulas could accommodate for genetic correlations between populations lower than 1. The number of QTL underlying a trait did not affect the accuracy of across-population genomic prediction using a GBLUP method.

Key words: genomic prediction, accuracy, across-population genomic prediction, prediction equation

3.1 Background

For genomic prediction, a reference population that consists of individuals with phenotypes and marker genotypes is used to estimate marker effects and to predict breeding values for another group of genotyped individuals, called selection candidates. The accuracy of predicting breeding values for selection candidates within one population is influenced by the level of linkage disequilibrium (LD) between markers, i.e., single-nucleotide polymorphisms (SNPs), and quantitative trait loci (QTL) that influence the trait, and by the level of family relationships (Daetwyler *et al.* 2008; VanRaden 2008; Zhong *et al.* 2009; De los Campos *et al.* 2013). Across populations, there are differences in LD, allele frequencies (De Roos *et al.* 2008; Zhong *et al.* 2009; De los Campos *et al.* 2012), and allele substitution effects of QTL (Spelman *et al.* 2002; Thaller *et al.* 2003), and close family relationships between individuals of different populations are absent. Therefore, the potential accuracy of predicting breeding values when the predicted population is not included in the reference population is likely to be limited. Indeed, in dairy cattle breeding, several empirical studies showed that the potential of using information across breeds was limited (e.g., Hayes *et al.* 2009; Pryce *et al.* 2011; Schrooten *et al.* 2013). The concept of combining individuals of different breeds in cattle is essentially similar to combining individuals from different lines in other animal and plant species (e.g., Ibáñẽz-Escriche *et al.* 2009; Zhong *et al.* 2009; Simeone *et al.* 2012) or from different subpopulations in humans (e.g., De los Campos *et al.* 2012; De los Campos *et al.* 2013) because close family relationships are absent and the extent of LD is limited across breeds, lines, and subpopulations.

A higher marker density may increase the consistency in LD phase across populations, since at short distances (5 to 30 kb) LD phases are conserved across populations (De Roos *et al.* 2008). However, several empirical studies showed that an increase in marker density resulted only in a small increase in accuracy using multiple populations in the reference population (Harris *et al.* 2011; Erbe *et al.* 2012). This small effect of marker density on accuracy indicates that other factors are also important, such as differences in segregating QTL or in the effect of QTL across populations due to differences in genetic background between populations (Spelman *et al.* 2002; Thaller *et al.* 2003). *DGAT1* (diacylglycerol O-acyltransferase 1) is one example of a gene with different effects across populations in dairy cattle. Allele substitution effects of a QTL in the *DGAT1* locus on milk yield and fat yield have been found to be on average 0.8 and 0.5 times, respectively, as large in Jersey than in Holstein Friesian populations in New Zealand (Spelman *et al.* 2002) and 0.7 and 1.2 times, respectively, as large in Fleckvieh than in Holstein Friesian populations in Germany (Thaller *et al.* 2003). Since the SNP that was analyzed is considered to be the causal polymorphism, which rules out incomplete LD, these results demonstrate that large differences in allele substitution effects can exist across populations.

Another factor that may affect accuracy of genomic prediction across populations is the number of QTL underlying the trait. For genomic prediction based on one population, accuracy is shown to be independent of the number of QTL underlying the trait when a genomic best linear unbiased prediction method (GBLUP) is used (Daetwyler *et al.* 2010; Clark *et al.* 2011), at least in situations for which there are no QTL that explain an extremely large part of the genetic variance. However, those studies only looked at the effect of the number of QTL on accuracy of genomic prediction within one population and not across populations.

For genomic prediction within one population, different deterministic formulas have been proposed to calculate the accuracy (Daetwyler *et al.* 2008; VanRaden 2008). The formula of Daetwyler *et al.* (2008) uses population and trait parameters, i.e., size of the reference population, heritability and number of effective chromosome segments. If the number of effective chromosome segments is calculated from the variation of genomic relationships around their expectations based on pedigree information, the formula of Daetwyler *et al.* (2008) can also be applied for populations with a complex family structure (Wientjes *et al.* 2013). The formula of VanRaden (2008) can be derived both from selection index theory and prediction error variance of the mixed model equation and it estimates the accuracy using the relationships within the reference population and between selection candidates and the reference population. Hayes *et al.* (2009) showed that applying the formula based on prediction error variance in multi-population situations without rescaling the genomic relationships across populations resulted in overestimation of the accuracy. This indicates that formulas for estimating the accuracy of genomic prediction using multiple populations need further investigation to define the best way to calculate genomic relationships across populations.

The first objective of this study was to develop a deterministic formula to estimate the accuracy of across-population genomic prediction. The second objective was to investigate the effect of differences in allele substitution effects of QTL across populations on accuracy of across-population genomic prediction. The last objective was to investigate the effect of the number of QTL underlying a trait on accuracy of across-population genomic prediction. Two deterministic formulas were evaluated and empirical accuracies were calculated using simulated phenotypes based on real genotypes from three cattle breeds representing different populations. Phenotypes were simulated using different correlations between allele substitution effects across breeds and different numbers of QTL underlying the trait. The reason for simulating the phenotypes of the individuals was to be able to investigate the actual effects of differences in allele substitution effects of QTL across populations and of the number of QTL by changing one factor at a time without changing the other factors, which would not be possible with real data.

3.2 Methods

3.2.1 Across-population genomic prediction

For genomic prediction based on one population, breeding values are predicted for individuals using a reference population of individuals from the same population. In most genomic prediction models, the QTL effects that underlie the traits of interest are assumed to be additive (e.g., Meuwissen *et al.* 2001). For across-population genomic prediction, breeding values are predicted for individuals using a reference population of individuals from one or more different populations. Due to differences in allele frequencies across populations, the presence of non-additive effects can result in differences in allele substitution effects of QTL (Falconer and Mackay 1996). Therefore, the models used for across-population genomic prediction should include non-additive effects or allow for differences in allele substitution effects across populations. Since it is difficult to accurately estimate non-additive effects (e.g., Wittenburg *et al.* 2011; Su *et al.* 2012), assuming additive gene action and, at the same time, allowing for differences in allele substitution effects may be a good first step and is the focus of this study. The correlation between allele substitution effects across populations can be considered as the genetic correlation between the populations (Bohren *et al.* 1966; Falconer and Mackay 1996).

Based on the assumption of additive QTL effects and using selection index theory, the breeding value of individual *i* of population *A* can be predicted using reference population *B* as:

$$\hat{a}_{A_i} = \mathbf{b}'_{AB}\,\mathbf{y}_B = Cov(a_{A_i}, \mathbf{y}_B)\left[Var(\mathbf{y}_B)\right]^{-1}\mathbf{y}_B\,, \tag{3.1}$$

where $\hat{a}_{A_i}$ is the predicted breeding value of individual *i* of population *A*, $\mathbf{b}_{AB}$ is a n_B x1 vector with partial regression coefficients of breeding values of population *A* on phenotypes of population *B*, $\mathbf{y}_B$ is a n_B x1 vector with phenotypes corrected for fixed effects of individuals from population *B*, a_{A_i} is the true breeding value of individual *i* of population *A*, and n_B is the number of individuals in reference population *B*.

The covariance between the true breeding value (TBV) of individual *i* of population *A* and the phenotypes of individuals from population *B* is:

$$Cov(a_{A_i},\mathbf{y}_B)=Cov(a_{A_i},\mathbf{a}_B+\mathbf{e}_B)=Cov(a_{A_i},\mathbf{a}_B)+Cov(a_{A_i},\mathbf{e}_B), \qquad (3.2)$$

where $\mathbf{a}_B$ is a n_Bx1 vector with TBV of individuals from population *B* and $\mathbf{e}_B$ is a n_Bx1 vector with environmental effects of individuals from population *B*. In an additive model $Cov(\mathbf{a},\mathbf{e})=0$, Equation 3.2 reduces to:

$$Cov(a_{A_i},\mathbf{y}_B)=Cov(a_{A_i},\mathbf{a}_B)=r_{G_{AB}}\sigma_{a_A}\sigma_{a_B}\mathbf{g}'_{A_i,B}, \qquad (3.3)$$

where $r_{G_{AB}}$ is the genetic correlation between population *A* and population *B*, σ_{a_A} and σ_{a_B} are the genetic standard deviations in populations *A* and *B*, respectively, $\mathbf{g}_{A_i,B}$ is a n_Bx1 vector with genomic relationships between individual *i* of population *A* and reference individuals of population *B*.

Under the assumption that SNPs are representative of QTL, i.e., that characteristics such as allele frequency are the same for SNPs and QTL, resulting in usable LD between SNPs and QTL, a genomic relationship matrix based on SNPs can be used to represent the relationships between breeding values of the individuals. To calculate the genomic relationships, covariances between the individuals of both populations need to be calculated. The mathematical definition of a covariance, $Cov(x,y)=E[(x-\bar{x})(y-\bar{y})]$, indicates that both components are corrected for their own mean. For the genomic relationships, this can be achieved by correcting the SNP genotypes of the individuals using the allele frequencies of their own population. Thus, the genotype of individual *i* from population *j* at locus *k*, g_{ijk}, is standardized as $x_{ijk}=\dfrac{g_{ijk}-2p_{jk}}{\sqrt{2p_{jk}\left(1-p_{jk}\right)}}$, where p_{jk} is the allele frequency of population *j* at locus *k*, and the standardized genotypes are used to calculate the genomic relationship matrices using the method of Yang *et al.* (2010), which will be described later.

Hence, Equation 3.1 can be written as:

$$\hat{a}_{A_i}=r_{G_{AB}}\sigma_{a_A}\sigma_{a_B}\mathbf{g}'_{A_i,B}\left[Var(\mathbf{y}_B)\right]^{-1}\mathbf{y}_B. \qquad (3.4)$$

This expression for the estimated breeding value (EBV) will subsequently be used in the next section to derive the accuracy.

3.2.1.1 Deterministic accuracy of across-population genomic prediction based on selection index theory

The general formula to calculate the accuracy of prediction of a breeding value is (Falconer and Mackay 1996):

$$r_{A_i} = \frac{Cov(\hat{a}_{A_i}, a_{A_i})}{\sqrt{Var(a_{A_i})Var(\hat{a}_{A_i})}}\ . \qquad (3.5)$$

In single-population situations, it is well known that $Cov(\hat{a}_{A_i}, a_{A_i}) = Var(\hat{a}_{A_i})$ (Falconer and Mackay 1996). This is also correct for across-population genomic prediction, as shown in the Appendix. Therefore, the expression for the accuracy of across-population genomic prediction reduces to:

$$r_{A_i} = \sqrt{\frac{Cov(\hat{a}_{A_i}, a_{A_i})}{Var(a_{A_i})}}\ . \qquad (3.6)$$

The covariance between the predicted and true breeding value of individual *i* of population *A* can be calculated as (see Appendix):

$$Cov(\hat{a}_{A_i}, a_{A_i}) = r^2_{G_{AB}}\, \sigma^2_{a_A}\, \sigma^2_{a_B}\, \mathbf{g}'_{A_i,B}\left[Var(\mathbf{y}_B)\right]^{-1} \mathbf{g}_{A_i,B}\,. \qquad (3.7)$$

Hence:

$$r_{A_i} = \sqrt{\frac{r^2_{G_{AB}}\, \sigma^2_{a_A}\, \sigma^2_{a_B}\, \mathbf{g}'_{A_i,B}\left[Var(\mathbf{y}_B)\right]^{-1} \mathbf{g}_{A_i,B}}{\sigma^2_{a_A}}} = r_{G_{AB}} \sqrt{\sigma^2_{a_B}\, \mathbf{g}'_{A_i,B}\left[Var(\mathbf{y}_B)\right]^{-1} \mathbf{g}_{A_i,B}}\ . \quad (3.8)$$

Equation 3.8 contains the variance of the phenotypes of individuals from population *B*, which can be written as:

$$Var(\mathbf{y}_B) = Cov(\mathbf{y}_B, \mathbf{y}_B) = Var(\mathbf{a}_B) + Var(\mathbf{e}_B) = \mathbf{G}_B\, \sigma^2_{a_B} + \mathbf{R}_B\, \sigma^2_{e_B}\,, \qquad (3.9)$$

where $\mathbf{G}_B$ is the n_B x n_B genomic relationship matrix of reference individuals of population *B*, $\sigma^2_{a_B}$ is the genetic variance in population *B*, $\mathbf{R}_B$ is a n_B x n_B standardized matrix that describes the correlations between environmental effects of individuals from population *B*, and $\sigma^2_{e_B}$ is the environmental variance in population *B*. Substituting Equation 3.9 into Equation 3.8 results in:

$$r_{A_i} = r_{G_{AB}} \sqrt{\mathbf{g}'_{A_i,B}\left[\mathbf{G}_B + \mathbf{R}_B \frac{\sigma^2_{e_B}}{\sigma^2_{a_B}}\right]^{-1} \mathbf{g}_{A_i,B}}\ . \qquad (3.10)$$

3.2.1.2 Deterministic accuracy of across-population genomic prediction using multiple populations in the reference population based on selection index theory

Equation 3.10 is valid when there is only one reference population. However, it may be interesting to combine reference populations to predict breeding values for individuals from another population. Based on a combined reference population from two populations, i.e., population *B* and *C*, the breeding value for a selection candidate *i* of population *A* can be predicted as:

$$\hat{a}_{A_i} = \begin{bmatrix}\mathbf{b}'_{AB} & \mathbf{b}'_{AC}\end{bmatrix}\begin{bmatrix}\mathbf{y}_B \\ \mathbf{y}_C\end{bmatrix} = Cov\left(a_{A_i}, \begin{bmatrix}\mathbf{y}_B \\ \mathbf{y}_C\end{bmatrix}\right)\left(Var\begin{bmatrix}\mathbf{y}_B \\ \mathbf{y}_C\end{bmatrix}\right)^{-1}\begin{bmatrix}\mathbf{y}_B \\ \mathbf{y}_C\end{bmatrix}, \qquad (3.11)$$

where $\mathbf{b}_{AC}$ is a n_Cx1 vector with partial regression coefficients of breeding values of individuals from population *A* on phenotypes of population *C*, $\mathbf{y}_C$ is a n_Cx1 vector with phenotypes corrected for fixed effects of individuals from population *C*.

Following Equation 3.3, the covariance between the TBV of individual *i* of population *A* and the phenotypes of individuals from population *B* and *C* is:

$$Cov\left(a_{A_i}, \begin{bmatrix}\mathbf{y}_B \\ \mathbf{y}_C\end{bmatrix}\right) = \begin{bmatrix}r_{G_{AB}}\sigma_{a_A}\sigma_{a_B}\mathbf{g}'_{A_i,B} & r_{G_{AC}}\sigma_{a_A}\sigma_{a_C}\mathbf{g}'_{A_i,C}\end{bmatrix}, \qquad (3.12)$$

where $r_{G_{AC}}$ is the genetic correlation between population *A* and population *C*, σ_{a_C} is the genetic standard deviation in population *C*, and $\mathbf{g}_{A_i,C}$ is a n_Cx1 vector of genomic relationships between individual *i* of population *A* and reference individuals of population *C*.

Hence, Equation 3.11 can be written as:

$$\hat{a}_{A_i} = \begin{bmatrix}r_{G_{AB}}\sigma_{a_A}\sigma_{a_B}\mathbf{g}'_{A_i,B} & r_{G_{AC}}\sigma_{a_A}\sigma_{a_C}\mathbf{g}'_{A_i,C}\end{bmatrix}\left(Var\begin{bmatrix}\mathbf{y}_B \\ \mathbf{y}_C\end{bmatrix}\right)^{-1}\begin{bmatrix}\mathbf{y}_B \\ \mathbf{y}_C\end{bmatrix}. \qquad (3.13)$$

In this situation, Equation 3.6 can also be used to calculate the accuracy. The covariance between the predicted and true breeding value of individual *i* of population *A* based on a reference population of individuals from population *B* and *C* is:

$$Cov(\hat{a}_{A_i}, a_{A_i}) = Cov\left(\begin{bmatrix}r_{G_{AB}}\sigma_{a_A}\sigma_{a_B}\mathbf{g}'_{A_i,B} & r_{G_{AC}}\sigma_{a_A}\sigma_{a_C}\mathbf{g}'_{A_i,C}\end{bmatrix}\left(Var\begin{bmatrix}\mathbf{y}_B \\ \mathbf{y}_C\end{bmatrix}\right)^{-1}\begin{bmatrix}\mathbf{y}_B \\ \mathbf{y}_C\end{bmatrix}, a_{A_i}\right)$$

$$= \begin{bmatrix}r_{G_{AB}}\sigma_{a_A}\sigma_{a_B}\mathbf{g}'_{A_i,B} & r_{G_{AC}}\sigma_{a_A}\sigma_{a_C}\mathbf{g}'_{A_i,C}\end{bmatrix}\left(Var\begin{bmatrix}\mathbf{y}_B \\ \mathbf{y}_C\end{bmatrix}\right)^{-1}\begin{bmatrix}r_{G_{AB}}\sigma_{a_A}\sigma_{a_B}\mathbf{g}_{A_i,B} \\ r_{G_{AC}}\sigma_{a_A}\sigma_{a_C}\mathbf{g}_{A_i,C}\end{bmatrix}. \qquad (3.14)$$

Using this expression in Equation 3.6, the accuracy of genomic prediction becomes:

$$r_{A_i} = \sqrt{\begin{bmatrix} r_{G_{AB}} \sigma_{a_B} \mathbf{g}'_{A_i,B} & r_{G_{AC}} \sigma_{a_C} \mathbf{g}'_{A_i,C} \end{bmatrix} \left(Var\begin{bmatrix} \mathbf{y}_B \\ \mathbf{y}_C \end{bmatrix} \right)^{-1} \begin{bmatrix} r_{G_{AB}} \sigma_{a_B} \mathbf{g}_{A_i,B} \\ r_{G_{AC}} \sigma_{a_C} \mathbf{g}_{A_i,C} \end{bmatrix}} . \qquad (3.15)$$

The (co-)variances of the phenotypes of the reference individuals of population *B* and *C* in Equation 3.15 can be written as:

$$Var\begin{bmatrix} \mathbf{y}_B \\ \mathbf{y}_C \end{bmatrix} = \begin{bmatrix} Var(\mathbf{y}_B) & Cov(\mathbf{y}_B, \mathbf{y}_C) \\ Cov'(\mathbf{y}_B, \mathbf{y}_C) & Var(\mathbf{y}_C) \end{bmatrix} . \qquad (3.16)$$

The variance of the phenotypes within one population follows from Equation 3.9. The covariance of the phenotypes across the two populations is:

$$Cov(\mathbf{y}_B, \mathbf{y}_C) = Cov(\mathbf{a}_B + \mathbf{e}_B, \mathbf{a}_C + \mathbf{e}_C) = Cov(\mathbf{a}_B, \mathbf{a}_C) = r_{G_{BC}} \sigma_{a_B} \sigma_{a_C} \mathbf{G}_{BC} . \qquad (3.17)$$

Combining Equations 3.9, 3.16, and 3.17 yields:

$$Var\begin{bmatrix} \mathbf{y}_B \\ \mathbf{y}_C \end{bmatrix} = \begin{bmatrix} \mathbf{G}_B \sigma^2_{a_B} + \mathbf{R}_B \sigma^2_{e_B} & r_{G_{BC}} \sigma_{a_B} \sigma_{a_C} \mathbf{G}_{BC} \\ r_{G_{BC}} \sigma_{a_B} \sigma_{a_C} \mathbf{G}'_{BC} & \mathbf{G}_C \sigma^2_{a_C} + \mathbf{R}_C \sigma^2_{e_C} \end{bmatrix} . \qquad (3.18)$$

Substituting this result into Equation 3.15 yields

$$\begin{aligned} r_{A_i} &= \sqrt{\begin{bmatrix} r_{G_{AB}} \sigma_{a_B} \mathbf{g}'_{A_i,B} & r_{G_{AC}} \sigma_{a_C} \mathbf{g}'_{A_i,C} \end{bmatrix} \begin{bmatrix} \mathbf{G}_B \sigma^2_{a_B} + \mathbf{R}_B \sigma^2_{e_B} & r_{G_{BC}} \sigma_{a_B} \sigma_{a_C} \mathbf{G}_{BC} \\ r_{G_{BC}} \sigma_{a_B} \sigma_{a_C} \mathbf{G}'_{BC} & \mathbf{G}_C \sigma^2_{a_C} + \mathbf{R}_C \sigma^2_{e_C} \end{bmatrix}^{-1} \begin{bmatrix} r_{G_{AB}} \sigma_{a_B} \mathbf{g}_{A_i,B} \\ r_{G_{AC}} \sigma_{a_C} \mathbf{g}_{A_i,C} \end{bmatrix}} \\ &= \sqrt{\begin{bmatrix} r_{G_{AB}} \mathbf{g}'_{A_i,B} & r_{G_{AC}} \mathbf{g}'_{A_i,C} \end{bmatrix} \begin{bmatrix} \mathbf{G}_B + \mathbf{R}_B \dfrac{\sigma^2_{e_B}}{\sigma^2_{a_B}} & r_{G_{BC}} \mathbf{G}_{BC} \\ r_{G_{BC}} \mathbf{G}'_{BC} & \mathbf{G}_C + \mathbf{R}_C \dfrac{\sigma^2_{e_C}}{\sigma^2_{a_C}} \end{bmatrix}^{-1} \begin{bmatrix} r_{G_{AB}} \mathbf{g}_{A_i,B} \\ r_{G_{AC}} \mathbf{g}_{A_i,C} \end{bmatrix}} . \qquad (3.19) \end{aligned}$$

Although Equation 3.19 is derived for across-population genomic prediction, this formula can also be applied to estimate the accuracy of multi-population genomic prediction for which one of the reference populations is the population of the selection candidates. Moreover, it is interesting to note that when one population is included in the reference population and selection candidates are from the same population as the reference individuals, Equation 3.19 becomes equivalent to the expression derived by VanRaden (2008).

3.2.1.3 Deterministic accuracy of across-population genomic prediction based on population parameters

In general, the accuracy with which an effect is predicted equals the square root of the proportion of variance explained by the effect. The accuracy of a sire's EBV based on progeny information, for example, equals $\sqrt{\frac{\frac{1}{4}\sigma_a^2}{\frac{1}{4}\sigma_a^2+(\sigma_p^2-\frac{1}{4}\sigma_a^2)/n}}$, where the numerator is the variance due to the sire, and the denominator the variance of the average of *n* progeny (Falconer and Mackay 1996). In the same way, when each chromosome segment explains an amount of variance equal to σ_a^2/M_e, in which M_e is the effective number of chromosome segments (Goddard *et al.* 2011), the accuracy of the predicted segment effect equals:

$$r=\sqrt{\frac{\sigma_a^2/M_e}{\sigma_a^2/M_e+\sigma_p^2/N_p}}\,, \qquad (3.20)$$

where σ_p^2 is the phenotypic variance and N_p is the size of the reference population. In the denominator, it is assumed that a single segment explains very little variance, so that $\sigma_p^2-\sigma_a^2/M_e \approx \sigma_p^2$. When the accuracy is the same for all effective segments, this is also the accuracy of genomic prediction. Multiplying both numerator and denominator of Equation 3.20 by N_pM_e/σ_p^2 yields a simple expression for the accuracy of genomic prediction for all selection candidates of the same population:

$$r_P=\sqrt{\frac{N_ph^2}{N_ph^2+M_e}}\,, \qquad (3.21)$$

where h^2 is the heritability of the trait. This result was originally derived by Daetwyler *et al.* (2008; 2010), but with a more complex derivation.

For within-population genomic prediction, M_e follows from Goddard *et al.* (2011):

$$M_e=\frac{1}{Var(\mathbf{G}_{RP_{ij}}-\mathbf{A}_{RP_{ij}})}\,, \qquad (3.22)$$

where $\mathbf{G}_{RP_{ij}}$ is the genomic relationship between individuals *i* and *j* from the reference population, $\mathbf{A}_{RP_{ij}}$ is the corresponding pedigree relationship, and the variance is taken over all pairs *ij* in the reference population. For across-population genomic prediction, we propose the following analogy:

$$M_e=\frac{1}{Var(\mathbf{G}_{RP_i,SK_j}-\mathbf{A}_{RP_i,SK_j})}\,, \qquad (3.23)$$

in which the index RP_i,SK_j refers to reference individual *i* and selection candidate *j*, and the variance is taken over all the pair-wise relationships between reference individuals and selection candidates. As explained by Goddard *et al.* (2011), the expectation of the genomic relationships for unrelated animals should be 0. This can be achieved by using population-specific allele frequencies to rescale the genotypes for setting up $\mathbf{G}_{RP_i,SK_j}$, as explained before for the expression based on selection index theory.

For across-population genomic prediction, the genetic correlation between populations has to be taken into account, because it limits the part of the genetic variance in the selection candidates that can be explained by the reference population. Therefore, the genetic correlation between the reference population and the selection candidates, $r_{G_{RP,SK}}$, was incorporated into Equation 3.21, giving:

$$r_P = r_{G_{RP,SK}} \sqrt{\frac{N_p h^2}{N_p h^2 + M_e}} \ . \qquad (3.24)$$

3.2.2 Simulations

3.2.2.1 Genotypes

Genotypes were available for 1285 dairy cows from the Netherlands that originated from three breeds (1033 Holstein Friesian (HF), 105 Groningen White Headed (GWH), and 147 Meuse-Rhine-Yssel (MRY)). All individuals were pure-bred animals since at least 87.5% of their genes originated from one of the three breeds.

Individuals from the breeds GWH and MRY were genotyped with the Illumina BovineHD Beadchip (777k, Illumina, San Diego, CA). Quality controls consisted in removing genotypes with a GenCall (GC) score lower than 0.2, SNPs with a call rate smaller than 95% in one of the breeds and SNPs with an unknown map position or located on the sex chromosomes. The HF individuals were genotyped with the Illumina BovineSNP50 Beadchip (50k, Illumina, San Diego, CA), and imputed to high-density (777k) using a reference population of 3150 HF individuals as described by Pryce *et al.* (2014). Quality control consisted in removing SNPs with a call rate smaller than 95% or with an unknown map position or located on the sex chromosomes. After editing the imputed genotypes, the mean Beagle R^2 value, which reflects the accuracy of imputation, was equal to 0.96 across imputed loci, which indicates that imputation was highly accurate.

Loci for which the genotypes passed the quality control of both the HF dataset and the combined GWH and MRY dataset were retained in the entire dataset. From this entire dataset, SNPs with a minor allele frequency equal to or lower than 0.5%, SNPs for which only two genotypes were observed, and SNPs in complete LD (r^2 = 1) with an adjacent SNP were removed. To increase the power of accurately estimating genomic breeding values, arbitrarily, we took only three chromosomes, namely chromosomes 13, 23 and 28 that contained about 10% of the remaining high-density SNPs into account. According to the literature, the LD pattern of those chromosomes is comparable to the LD pattern of the entire cattle genome (McKay *et al.* 2007; Khatkar *et al.* 2008). After editing, a total of 31,503 SNPs remained across the three chromosomes.

3.2.2.2 Simulation of phenotypes

Phenotypes of the individuals were simulated using different scenarios with two variables i.e., 1) the number of QTL underlying the simulated trait and 2) the correlation between allele substitution effects of the QTL underlying the simulated trait in the different populations, i.e., the genetic correlation between populations (Bohren *et al.* 1966; Falconer and Mackay 1996). From the 31,503 SNPs available after editing, 5000 were randomly selected to become candidate QTL, regardless of the chromosome. In each replicate, the actual QTL with an effect on the trait were randomly sampled from those candidate QTL. The remaining (31,503 – 5000 =) 26,503 SNPs composed the group of markers used in all analyses. Using this approach allowed us to keep the set of markers constant across all replicates but still made it possible to randomly select the QTL from the group of candidate QTL within each replicate. The numbers of QTL underlying the simulated trait were equal to 3000 (~10% of all SNPs), 300 (~1%), 30 (~0.1%) or 3 (~0.01%).

The allele substitution effects of QTL were sampled from a multinormal distribution with mean 0 and standard deviation 1, assuming a correlation of 1, 0.8, 0.6, 0.4, or 0.2 between the allele substitution effects across all three pairs of breeds. This was simulated by sampling random numbers from a normal distribution with mean 0 and standard deviation 1 and multiplying those numbers with the Cholesky decomposition of the covariance matrix between the allele substitution effects of the breeds.

For each of the individuals, the TBV was calculated by multiplying the simulated allele substitution effects with the genotypes of the 3000, 300, 30, or 3 QTL coded as 0, 1, and 2. Only additive effects and no dominance effects or epistatic interactions were simulated, therefore, the effects were summed over all QTL. Finally, TBV of all individuals of the three breeds were rescaled to a mean of 0 and variance of 1 across breeds. By rescaling the TBV in this way, their mean and variance were the same for each replicate and for the different numbers of QTL, which indicates that when the number of QTL was higher, each QTL explained a smaller part of the variance.

Allele frequencies for simulated QTL (sampled from the SNPs) differed for each of the three breeds, resulting in differences in average TBV between the breeds. To simulate environmental effects for each individual assuming equal heritability for the three breeds, TBV were first adjusted by subtracting the average TBV of the individual's breed before the genetic variance across TBV was calculated. Thereafter, the environmental effect per individual was sampled for the three breeds from a normal distribution with mean 0 and variance $\left(\frac{1}{h^2}-1\right)$ *(variance of TBV corrected for mean TBV within breed). For each individual, the phenotype was calculated as the sum of its TBV and the randomly sampled environmental effect. Note that the within-breed TBV means were only subtracted from the TBV to calculate the environmental variance, the TBV itself, and therefore the phenotypes as well, still included the within-breed TBV mean.

For each scenario, simulations were replicated 100 times using a heritability of 0.95 to simulate phenotypes in each of the three breeds and for each number of QTL underlying the trait. A high heritability of 0.95 was chosen to increase the achieved accuracies and to make the differences in accuracies between the different scenarios more pronounced for the size of reference population used. In dairy cattle breeding, a heritability of 0.95 can be achieved by using deregressed proofs of bulls for a trait with a heritability of 0.25 based on 285 daughters, following (Mrode and Thompson 2005):

$$r=\sqrt{\frac{nh^2}{nh^2+(4-h^2)}}\,, \tag{3.25}$$

where r is the accuracy for a sire's breeding value, n is the number of daughters of that sire, and h^2 is the heritability of the trait.

3.2.2.3 Scenarios to evaluate accuracy of genomic prediction

Mean accuracy of genomic prediction was empirically and deterministically evaluated for five different scenarios. The first scenario, i.e., the base scenario, which represented single-population genomic prediction, used HF animals as reference population and selection candidates. In the other scenarios, the reference population consisted of one or two populations and breeding values were predicted for individuals from another population, which means that across-population genomic prediction was applied (Table 3.1). For the across-population scenarios, the reference population was the same for all selection candidates of a specific population. In the scenario with HF individuals both as reference population and selection candidates, the deterministic accuracies (Equations 3.19

and 3.24) were calculated for a single HF individual using a reference population consisting of all remaining HF individuals. The empirical accuracy was calculated using 20-fold cross-validation, where in each replicate, individuals were randomly divided in 20 equally-sized groups using each group once as selection candidates and the remaining 19 as reference population.

Table 3.1 Overview of the breeds used in the different reference populations and as selection candidates.

	Reference population		Predicted individuals	
Scenario	**Breed(s)**	**Nb of individuals**	**Breed**	**Nb of individuals**
Base	HF	1032 / 981-982[a]	HF	1 / 51-52[a]
1	HF	1033	GWH	105
2	HF + MRY	1180	GWH	105
3	HF	1033	MRY	147
4	HF + GWH	1138	MRY	147

[a]Deterministic formulas used leave-one-out cross-validation, empirical calculations used 20 fold cross-validation using 20 groups of 51 or 52 individuals due to computational reasons; HF = Holstein Friesian; GWH = Groningen White Headed; MRY = Meuse-Rhine-Yssel.

3.2.2.4 Empirical accuracy based on simulated phenotypes

For the empirical estimation of the accuracy, a GBLUP-model type, called GREML, was run in ASReml (Gilmour *et al.* 2009). This GREML model used a genomic relationship matrix (**G**) and simulated phenotypes based on 3000, 300, 30 or 3 QTL underlying the simulated trait. In this model, breed was included as a fixed effect. This model is termed GREML, because it has the same features as the commonly known GBLUP model, however variances were not assumed to be known but were estimated simultaneously with the breeding values using REML. Accuracy was calculated for each population as the correlation between EBV from this model and TBV. Since simulated phenotypes were different per replicate, averages and standard errors of empirical accuracies were calculated across replicates.

The **G** matrix used in GREML contained all reference individuals and selection candidates and was calculated based on the method of Yang *et al.* (2010); $\mathbf{G}_{SNPs} = \frac{\mathbf{XX}'}{n}$. In this equation, *n* represents the number of SNPs (26,503) and the **X** matrix contains standardized genotypes (one locus per column) of each individual (one individual per row). For the empirical estimation of the accuracy, standardized genotypes were calculated as $x_{ij} = \frac{g_{ij} - 2p_j}{\sqrt{2p_j(1-p_j)}}$, where g_{ij} codes the genotype for individual *i* at marker locus *j* as 0, 1 and 2, and p_j is the allele frequency at marker locus *j* for the second allele averaged over all breeds. To calculate the average allele frequency per locus, the allele frequency per locus was calculated per breed and thereafter averaged over the three breeds, with an equal weight for each of the breeds. In that way, average allele frequency is not dominated by the breed with the largest number of genotyped individuals. Note that for each scenario, the $\mathbf{G}_{SNPs}$ matrix contained only the reference individuals and selection candidates (and the SNPs segregating in that group), so four different $\mathbf{G}_{SNPs}$ matrices were calculated that contained 1) all HF individuals (26,486 SNPs), 2) all HF and GWH individuals (26,500 SNPs), 3) all HF and MRY individuals (26,498 SNPs), and 4) all HF, GWH and MRY individuals (26,503 SNPs).

In the calculation of $\mathbf{G}_{SNPs}$, allele frequencies of the current population were used, which means that the current population was used as the base population. This indicates that the inbreeding level in $\mathbf{G}_{SNPs}$ differed from the inbreeding level in the pedigree-based relationship matrix, **A**, and that $\mathbf{G}_{SNPs}$ and **A** were not compatible. To rescale the inbreeding level in $\mathbf{G}_{SNPs}$ to the inbreeding level of **A**, the following adjustment was made to within-breed genomic relationships (Powell *et al.* 2010):

$$\mathbf{G}^*_{SNPs} = \left(1 - \overline{F_b}\right)\mathbf{G}_{SNPs} + 2\overline{F_b}\,\mathbf{J}, \tag{3.26}$$

where F_b was the average inbreeding coefficient of all individuals of breed *b* based on the pedigree and **J** was a matrix filled with ones.

Due to only three chromosomes being selected for this study and due to sampling variance of the SNPs on the chip, $E(\mathbf{G}\,|\,\mathbf{G}^*_{SNPs})$ is not $\mathbf{G}^*_{SNPs}$ (Powell *et al.* 2010; Goddard *et al.* 2011). Therefore, we regressed the $\mathbf{G}^*_{SNPs}$ matrix back to the **A** matrix, which is the additive genetic relationship matrix based on the pedigree, following Yang *et al.* (2010) and Goddard *et al.* (2011):

$$\hat{\mathbf{G}} = \mathbf{A} + b\left(\mathbf{G}^*_{SNPs} - \mathbf{A}\right), \tag{3.27}$$

where

$$b = \frac{Var\left(\mathbf{G}^*_{SNPs} - \mathbf{A}\right)}{\left[Var\left(\mathbf{G}^*_{SNPs} - \mathbf{A}\right)\right] + Var(\mathbf{E})} = \frac{Var\left(\hat{\mathbf{G}} - \mathbf{A}\right) - 1/n}{Var\left(\hat{\mathbf{G}} - \mathbf{A}\right)}. \tag{3.28}$$

Since the level of family relationships influences the sampling error on the elements in **G**, the regression coefficient *b* was calculated separately for bins of family relationships in **A** (0-0.10, >0.10-0.25, >0.25-0.50 and >0.5) within each breed and for each combination of breeds. Across-breed relationships were indeed 0 in **A**, so in that case $Var\left(\hat{\mathbf{G}}-\mathbf{A}\right)$ approximately reduced to $Var\left(\hat{\mathbf{G}}\right)$. Parent-offspring relationships and self-relationships were not or hardly affected by sampling error and therefore excluded from the regression. The regression coefficient *b* was always above 0.95, and, in most cases, even above 0.99. Therefore, the effect of regressing the **G** matrix back to the **A** matrix was limited.

The inbreeding level in **A** depends on the depth of the pedigree, which indicates that different pedigree depths across populations can cause differences in inbreeding levels across the populations. To remove these differences in pedigree depth, the pedigree was cut off at seven generations for all individuals. Based on the pedigree, small relationships between some animals of the different breeds occurred, with a maximum relationship of 0.035 between HF and GWH, 0.034 between HF and MRY, and 0.029 between GWH and MRY. These relationships resemble more or less the relationship between an individual and one of its ancestors five generations back.

3.2.2.5 Deterministic accuracies of genomic prediction

For each scenario, accuracies of genomic prediction were deterministically derived using the two methods explained before; one method based on selection index theory (Equation 3.19) and one method based on population parameters (Equation 3.24). It is interesting to note that the formula based on selection index theory provides a single accuracy for each selection candidate, while the formula using population parameters provides an accuracy that applies to all selection candidates of the same population. Both deterministic methods calculate the accuracy based on genomic relationships and do not use phenotypes. Since the subset of SNPs was constant across all replicates and scenarios with different numbers of QTL, only one accuracy was calculated that applied to all replicates and numbers of QTL. Therefore, it was not possible to calculate standard errors across replicates for the deterministic accuracies.

3.2.2.6 Estimating genetic correlations between populations

In this simulation study, the genetic correlation between populations was known. In studies using real data, this is usually not the case and the genetic correlation needs to be estimated from the data. We investigated how accurate the genetic correlations between HF and GWH, and between HF and MRY are estimated using a multi-trait model in ASReml (Gilmour *et al.* 2009) in which the same trait in different breeds was treated as different traits. Within the multi-trait model, the same **G** matrix was used as in the GBLUP model, the environmental correlation was set to 0 and genetic and environmental variances of GWH and MRY animals were fixed at the simulated values, because the small number of animals in those breeds made it difficult to estimate variance components reliably.

3.3 Results

3.3.1 Differences between populations

In this study, accuracy of genomic prediction was evaluated by using genotypes of three cattle breeds. In cases where allele substitution effects were equal across breeds, differences in accuracy between single- and across-breed genomic predictions were due to differences in allele frequencies, relationships and LD pattern across breeds. The correlation between allele frequencies of all 26,503 SNPs was 0.67 for HF and GWH, 0.73 for HF and MRY, and 0.65 for GWH and MRY. Correlations of allele frequencies of SNPs and candidate QTL across breeds were similar.

Based on pedigree information, there were few differences in average relationships between breeds with average relationships of 0.0004 between HF and GWH (ranging from 0 to 0.035), 0.0004 between HF and MRY (ranging from 0 to 0.034), and 0.0005 between GWH and MRY (ranging from 0 to 0.029). Based on genotype data, differences in average relationships across breeds became more pronounced, with average relationships of -0.084 between HF and GWH (ranging from -0.194 to +0.115), -0.050 between HF and MRY (ranging from -0.151 to +0.125), and -0.098 between GWH and MRY (ranging from -0.184 to +0.088).

3.3.2 Equal allele substitution effects across populations

Accuracies of genomic prediction are shown in Figure 3.1 for scenarios with equal allele substitution effects for the three breeds. Figure 3.1 shows that standard errors for all empirically calculated accuracies were small. Since both deterministic accuracies did not use replicates, there are no standard errors across replicates. However, the method based on selection index theory estimates accuracy per individual and this accuracy depended on the relationships of the selection candidate with the reference individuals. For each scenario, standard errors of the accuracy were calculated over all individuals and were equal to (mean and standard errors) 0.934 ± 0.001 (base scenario), 0.467 ± 0.006 (scenario 1), 0.492 ± 0.006 (scenario 2), 0.437 ± 0.003 (scenario 3), and 0.458 ± 0.003 (scenario 4).

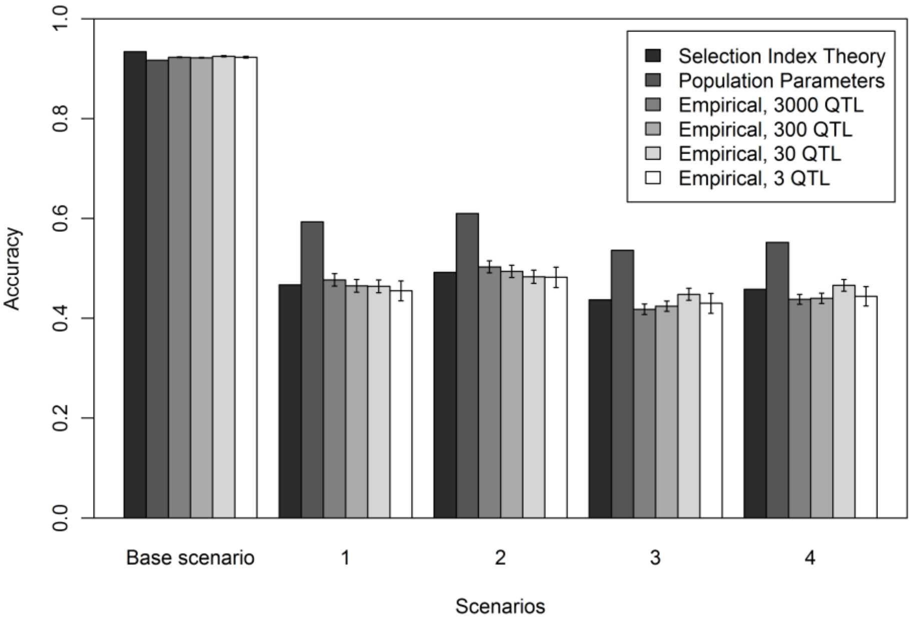


Figure 3.1 Empirical and deterministic accuracies of genomic prediction (± standard error) with a heritability of 0.95 and using equal allele substitution effects of the QTL underlying the simulated trait in the three breeds for five different scenarios; Base = reference HF (Holstein Friesian) population, selection candidates HF; 1 = reference population HF, selection candidates GWH (Groningen White Headed); 2 = reference population HF and MRY (Meuse-Rhine-Yssel), selection candidates GWH; 3 = reference population HF, selection candidates MRY; 4 = reference population HF and GWH, selection candidates MRY.

Accuracies for the base scenario, for which breeding values of HF individuals were predicted using a reference population of HF individuals, were very high (> 0.9). Empirically derived accuracies were the same for the different numbers of QTL underlying the trait, which indicates that the number of QTL did not affect empirical accuracy in single-breed genomic prediction. With both deterministic methods, accuracies were in good agreement with the empirically-derived accuracies.

Accuracies with the other four scenarios, for which across-breed genomic prediction was applied, were much lower than those with the base scenario, but still ranged from 0.4 to 0.5. In each scenario, empirical accuracies using different numbers of QTL underlying the trait were very similar, which indicates that there is no effect of number of QTL on empirical accuracy. As with single-breed genomic prediction, estimated accuracies based on selection index theory were in good agreement with empirical accuracies for all four scenarios of across-breed genomic prediction. The deterministic prediction formula using population parameters overestimated empirical accuracies by about 25%.

Empirical accuracies as well as deterministic accuracies were slightly higher for selection candidates from breed GWH than for those from breed MRY. For both breeds, empirical and deterministic accuracies slightly increased when the other breed was added to the HF reference population, thus maintaining a near constant difference in accuracy between GWH and MRY individuals.

3.3.3 Different allele substitution effects across populations

Accuracies of genomic prediction are shown in Figure 3.2 for scenarios with a correlation of allele substitution effects across breeds equal to A) 0.8, B) 0.6, C) 0.4, or D) 0.2. Standard errors for the empirical accuracies were low as with scenarios with equal allele substitution effects across breeds. The average estimated accuracies based on selection index theory and the variances across all individuals decreased for each scenario, the reduction being proportional to the correlation between allele substitution effects across populations.

As expected, deterministic and empirical accuracies were about equal to the accuracies obtained with equal allele substitution effects across breeds multiplied by the correlation between allele substitution effects. Empirical accuracies across the different numbers of QTL underlying the trait were again very similar, although those obtained with the 3-QTL scenario seemed to differ slightly from the other scenarios. This is in agreement with the much higher standard error across the replicates obtained with the 3-QTL scenario than with the 3000-, 300- or 30-QTL scenarios.

As in scenarios with equal allele substitution effects across breeds, accuracies obtained with the formula based on selection index theory were in good agreement with empirical accuracies. This indicates that this formula can be used to estimate the accuracy even when the genetic correlation between populations differs from 1. The formula using population parameters overestimated empirical accuracies by about 25% to 30%, regardless of the genetic correlation between breeds.

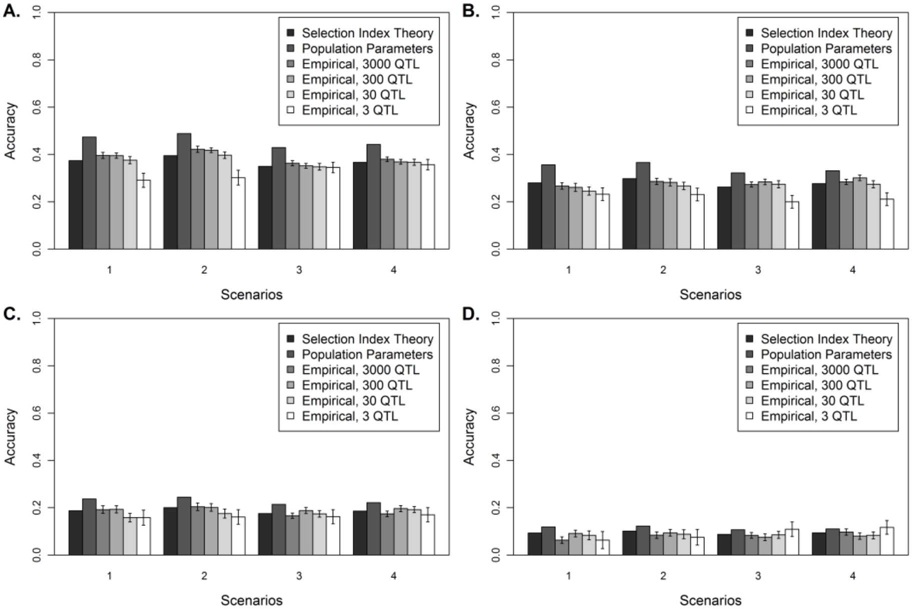


Figure 3.2 Empirical and deterministic accuracies of genomic prediction (± standard error) at a heritability of 0.95 using a correlation of (A) 0.8, (B) 0.6, (C) 0.4, or (D) 0.2 between allele substitution effects of the QTL underlying the simulated trait in the different breeds for four different scenarios. 1 = reference population HF (Holstein Friesian), selection candidates GWH (Groningen White Headed); 2 = reference population HF and MRY (Meuse-Rhine-Yssel), selection candidates GWH; 3 = reference population HF, selection candidates MRY; 4 = reference population HF and GWH, selection candidates MRY.

3.3.4 Estimated genetic correlations between populations

Estimated genetic correlations are shown in Table 3.2 for the different scenarios. When the simulated genetic correlation was 1, the genetic correlations between the breeds were slightly underestimated and ranged from 0.85 to 0.92. When the simulated genetic correlation was different from 1, estimated and simulated genetic correlations between the breeds were in good agreement for the 3000-, 300- and 30-QTL scenarios. The estimated genetic correlation for the 3-QTL scenario was generally much lower than the simulated value, which is in agreement with the results found for the empirical accuracies and is probably due to the higher sampling error on the correlation in this scenario.

Table 3.2 Simulated and estimated genetic correlations (standard errors across replicates) between the populations.

Populations	Simulated genetic correlation	Estimated genetic correlation (s.e.)							
		3000 QTL		300 QTL		30 QTL		3 QTL	
HF - GWH	**1.0**	0.91	(0.01)	0.92	(0.01)	0.89	(0.01)	0.86	(0.02)
HF - GWH	**0.8**	0.79	(0.02)	0.79	(0.01)	0.77	(0.02)	0.56	(0.05)
HF - GWH	**0.6**	0.61	(0.02)	0.60	(0.02)	0.57	(0.03)	0.53	(0.05)
HF - GWH	**0.4**	0.47	(0.02)	0.51	(0.03)	0.44	(0.03)	0.31	(0.06)
HF - GWH	**0.2**	0.19	(0.03)	0.22	(0.03)	0.20	(0.04)	0.16	(0.07)
HF - MRY	**1.0**	0.89	(0.01)	0.89	(0.01)	0.91	(0.01)	0.85	(0.02)
HF - MRY	**0.8**	0.81	(0.02)	0.78	(0.02)	0.81	(0.02)	0.69	(0.04)
HF - MRY	**0.6**	0.61	(0.02)	0.69	(0.02)	0.62	(0.02)	0.46	(0.05)
HF - MRY	**0.4**	0.44	(0.02)	0.45	(0.03)	0.44	(0.03)	0.28	(0.06)
HF - MRY	**0.2**	0.24	(0.02)	0.25	(0.03)	0.24	(0.04)	0.24	(0.06)

HF = Holstein Friesian; GWH = Groningen White Headed; MRY = Meuse-Rhine-Yssel.

3.4 Discussion

3.4.1 Deterministic accuracy of across-population genomic prediction

The first objective of this study was to develop a deterministic formula to investigate the accuracy of across-population genomic prediction. Our study as other previous studies (VanRaden 2008; Clark *et al.* 2012; Wientjes *et al.* 2013) shows that the formula based on selection index theory (Equation 3.19) and the formula using population parameters (Equation 3.24) can accurately estimate the accuracy of genomic prediction within one population using relationship matrices. By setting up across-population genomic relationship matrices based on population-specific allele frequencies, it was also possible to accurately estimate the accuracy of across-population genomic prediction based on selection index theory. The application of the prediction formula using population parameters, as described in our study, overestimated the empirical accuracy for across-population genomic prediction in all scenarios by about 25 to 30%.

The genetic correlation in the deterministic formulas accounts for differences in allele substitution effects across populations. These differences may also lead to differences in genetic variances across populations, i.e., heterogeneous variances. For example, among populations, the genetic variance tends to be larger for the population with the highest mean for a given trait (Legates 1962; Boldman and Freeman 1990). In addition, differences in allele frequencies across populations may also lead to heterogeneous variances; for example, a QTL may only segregate in one of the populations, which results in differences in the genetic variance explained by that QTL across populations although the actual allele substitution effects could be the same. Moreover, environmental variances may be different across populations when deregressed proofs of bulls are used as phenotypes, since the heritability of those proofs depends on the number of daughters of the bull, which can differ across populations. Heterogeneous variances across populations, which are not properly accounted for, may affect bias and accuracy of EBV. The deterministic formula based on selection index theory can take those heterogeneous variances into account as well, in contrast to the application of the formula based on populations parameters described here. Makgahlela *et al.* (2013) empirically showed that accuracies of multi-breed genomic prediction can be increased by accounting for those heterogeneous variances across breeds in a multi-trait random regression model (Makgahlela *et al.* 2013; Strandén and Mäntysaari 2013).

The genomic relationship matrix used in the deterministic formulas was calculated based on population-specific allele frequencies. Harris and Johnson (2010) already mentioned that differences in allele frequencies should be taken into account to calculate genomic covariances and relationships between individuals of different populations. Not using population-specific allele frequencies results in average genomic relationships across populations different from 0 (Karoui *et al.* 2012), large differences in average diagonal elements across populations (Harris and Johnson 2010; Simeone *et al.* 2012) and overestimation of the accuracies (Hayes *et al.* 2009). In our study, using population-specific allele frequencies resulted in average genomic relationship close to 0, i.e., equal to 0.00003 with a standard deviation of 0.023 between HF and GWH, and 0.00003 with a standard deviation of 0.020 between HF and MRY.

The deterministic formula based on selection index theory (Equation 3.19) estimated the accuracy of across-population genomic prediction accurately for all scenarios. With a genetic correlation of 0.8, 0.6, 0.4, or 0.2, empirical and deterministic accuracies were respectively 80%, 60%, 40%, or 20% of the accuracies achieved with a genetic correlation of 1. This indicates that the deterministic formula can be used to estimate genetic correlations between populations (but does not provide information about the mechanism underlying this correlation); for example when the empirical accuracy is only 60% of the accuracy estimated assuming a genetic correlation of 1, the actual genetic correlation between populations is expected to be 0.6. Using this deterministic formula to estimate the genetic correlation between populations can be especially attractive when only one of the populations has a small number of genotyped individuals.

Overestimation of accuracies with the formula using population parameters for the across-population scenarios is probably due to the inability of the SNPs to capture all the genetic variance in the selection candidates (Daetwyler 2009; Erbe *et al.* 2013), which is an underlying assumption of this formula. The empirical accuracy was about 80% of the predicted accuracy, both when GWH individuals or MRY individuals were used as selection candidates. This indicates that only 80% of the genetic variance in the selection candidates was captured by the markers in the reference population, due to differences in LD and allele frequencies of QTL between the reference population and the selection candidates. This proportion of the genetic variance in the selection candidates captured by SNPs in the reference population is the maximum accuracy of genomic prediction for those populations based on the used SNP chip (Daetwyler 2009).

By using an estimation of the genetic variance in the validation population that can be captured by SNPs in the reference population, the formula based on population parameters becomes a useful formula to predict the accuracy of across-population genomic prediction. This formula is very simple to use and can assess expected accuracies before individuals are genotyped. However, an important question remains regarding which values to use for M_e and the genetic correlation. In this study, M_e were estimated based on the variation in genomic relationships between reference and selection individuals around their expectations based on pedigree information. Similarly to the single-population scenario, M_e of the across-population scenarios were estimated based on the relationships across population. Using this approach, an M_e of about 1800 was estimated when GWH individuals were used as selection candidates, and 2400 when MRY individuals were used as selection candidates, both when HF individuals were used as reference population. Since only 10% of the genome was taken into account, this M_e should be multiplied by 10 to get the actual M_e across those populations. In a previous study, an M_e of 11,500 was obtained when reference individuals and selection candidates shared allele frequencies and LD patterns and of 122,000 when reference individuals and selection candidates shared only allele frequencies (Wientjes *et al.* 2013). Across breeds, allele frequencies are different, but LD patterns may be partly the same, therefore, M_e across breeds was indeed to fall within the values of those groups. This suggests that perhaps an M_e of about 20,000 could be used to predict the accuracy of across-population genomic prediction for closely related cattle breeds and an M_e of about 40,000 or more for more distantly related cattle breeds.

The actual genetic correlation between populations, which is needed in the prediction formula, is in practice not known and depends on the traits and populations of interest. However, we showed that this genetic correlation can be estimated quite accurately using a multi-trait model and high-density genotypes. Thus, it may be possible to estimate this genetic correlation in a limited number of animals and to use it to predict the accuracies of genomic selection for different scenarios.

3.4.2 Empirical accuracies of genomic prediction

The second objective of this study was to investigate the effect of differences in allele substitution effects of QTL between populations, i.e., genetic correlations that differ from 1, on accuracy of across-population genomic prediction. Our results showed that genetic correlations between populations that are smaller than 1 resulted in a reduced accuracy of across-population genomic prediction that is proportional to the genetic correlation.

In this study, it was assumed that SNPs are representative of QTL, i.e., that SNPs and QTL have the same characteristics. Regarding this assumption, we know that for most complex traits, QTL minor allele frequencies are expected to be low (Goddard and Hayes 2009; Yang *et al.* 2010; Kemper and Goddard 2012). However, the SNPs on the chip were selected to have an intermediate allele frequency (Matukumalli *et al.* 2009), resulting in ascertainment bias of these SNPs. These differences in allele frequencies indicate that, in practice, QTL and SNPs have other characteristics, thereby reducing LD between QTL and SNPs in empirical studies. In our study, QTL were selected from the SNPs on the chip, which did not completely cover the range of expected allele frequencies of the actual QTL. Therefore, LD between QTL and SNPs may be overestimated, which results in higher accuracies of genomic prediction. In a future study, we will investigate the effect of different QTL allele frequencies on the accuracy of multi-population genomic prediction using loci with different allele frequencies and representative of the whole genome.

Another assumption used in this study was that the trait of interest was only influenced by additive effects. Due to the existence of non-additive effects, the average effects of allele substitution depend on the QTL allele frequencies (Falconer and Mackay 1996), and might therefore be different across populations. In this study, different effects were considered by simulating genetic correlations between populations that differed from 1. In general, empirical studies use additive models for across-population genomic prediction and provide much lower accuracies than those obtained in this study for a genetic correlation of 1 (e.g., Hayes *et al.* 2009; Pryce *et al.* 2011). This suggests that either SNPs do not represent QTL or that non-additive effects are important for the traits of interest in empirical studies, or a combination of both, which is important biological information.

In this study, genetic correlations between populations of 1, 0.8, 0.6, 0.4, and 0.2 were used to simulate phenotypes. Our results showed that genetic correlations between populations can be estimated quite accurately from the data using a multi-trait model. To date, this was done only in a few empirical studies (Karoui *et al.* 2012; Legarra *et al.* 2014). Karoui *et al.* (2012) reported estimated genetic correlations between French dairy cattle breeds that ranged from 0 (fertility; Montbéliarde – Normande) to 0.79 (milk; Montbéliarde – Holstein), with only two out of nine estimated genetic correlations above 0.6. These empirical results show that genetic correlation between populations can differ from 1 and depends on the trait of interest.

Results of this study clearly show that genetic correlation between populations is an important parameter for across-population genomic prediction. The true genetic correlation between populations is not influenced by differences in LD between QTL and SNPs. It is worth noting that apart from differences in allele substitution effects, the genetic correlation can also differ from 1 because of different QTL for the same trait. In terms of accuracy, the value of the genetic correlation is important and not the underlying cause of this genetic correlation. In fact, the genetic correlation specifies the maximum accuracy that can be obtained with across-population genomic prediction, provided that the reference population is very large and the number of SNPs is large enough to find a consistent linkage phase across populations.

3.4.3 Effect of number of QTL

The third objective of this study was to investigate the effect of the number of QTL underlying a trait on accuracy of across-population genomic prediction, which was studied using a GBLUP method. The results showed that changing the number of QTL without changing any other parameter had no effect on the accuracy.

In the case of genomic prediction within one population, different studies have already shown that accuracies of genomic prediction using GBLUP do not depend on number of QTL underlying the trait (Daetwyler *et al.* 2010; Clark *et al.* 2011). If variable selection models were used for genomic prediction, higher numbers of QTL resulted in lower accuracies (Coster *et al.* 2010; Daetwyler *et al.* 2010; Clark *et al.* 2011). One of these studies also showed that variable selection models have an advantage over GBLUP when the number of QTL is below M_e in genomic prediction within one population (Daetwyler *et al.* 2010). In across-population situations, M_e is much larger than within one population (Wientjes *et al.* 2013), which suggests that, in those situations, it will be easier to have a number of QTL smaller than M_e and, thus it is expected that the use of variable selection models will be beneficial.

3.5 Conclusions

The deterministic formula based on selection index theory, that was derived in this study, can accurately estimate the accuracy of across-population genomic prediction by using population-specific allele frequencies to set-up genomic relationship matrices. Another deterministic formula using population parameters overestimates the accuracy of across-population genomic prediction, because the SNPs in the reference population cannot capture all of the genetic variance in the selection candidates. However, this formula may still be useful because of its simplicity, and is expected to be much more accurate when the proportion of genetic variance in the selection candidates is known with reasonable accuracy and included in the formula. Moreover, the results of this study show that differences in allele substitution effects across populations reduce the accuracy of across-population genomic prediction, with a proportion equal to the correlation between allele substitution effects across populations. The number of QTL underlying a trait does not affect the accuracy of across-population genomic prediction when a GBLUP method is used.

3.6 Acknowledgements

This study was financially supported by Breed4Food (KB-12-006.03-005-ASG-LR), a public-private partnership in the domain of animal breeding and genomics, and CRV (Arnhem, The Netherlands). The RobustMilk project and the National Institute of Food and Agriculture (NIFA) are acknowledged for providing the 50k genotypes of the HF cows, and the gDMI consortium is acknowledged for imputing those to 777k genotypes. The Dutch Milk Genomics Initiative and the project 'Melk op Maat', funded by Wageningen University (the Netherlands), the Dutch Dairy Association (NZO, Zoetermeer, the Netherlands), the cooperative cattle improvement organization CRV BV (Arnhem, the Netherlands), the Dutch Technology Foundation (STW, Utrecht, the Netherlands), the Dutch Ministry of Economic Affairs (The Hague, the Netherlands) and the Provinces of Gelderland and Overijssel (Arnhem, the Netherlands), are thanked for providing the 777k genotypes of the GWH and MRY cows. The authors acknowledge Myrthe Maurice-Van Eijndhoven for collecting the data of the GWH and MRY cows, and the herd owners for their help in collecting the data.

3.7 Appendix

Proving that $Cov(\hat{a}_{A_i}, a_{A_i}) = Var(\hat{a}_{A_i})$ is correct for across-population genomic prediction

The covariance between the predicted and true breeding value of individual i of population A using a reference population of population B is:

$$\begin{aligned}
&Cov(\hat{a}_{A_i}, a_{A_i})\\
&= Cov(r_{G_{AB}} \sigma_{a_A} \sigma_{a_B} \mathbf{g'}_{A_i,B} [Var(\mathbf{y}_B)]^{-1} \mathbf{y}_B, a_{A_i})\\
&= r_{G_{AB}} \sigma_{a_A} \sigma_{a_B} \mathbf{g'}_{A_i,B} [Var(\mathbf{y}_B)]^{-1} Cov(\mathbf{y}_B, a_{A_i})\\
&= r_{G_{AB}} \sigma_{a_A} \sigma_{a_B} \mathbf{g'}_{A_i,B} [Var(\mathbf{y}_B)]^{-1} Cov(\mathbf{a}_B, a_{A_i})\\
&= r_{G_{AB}} \sigma_{a_A} \sigma_{a_B} \mathbf{g'}_{A_i,B} [Var(\mathbf{y}_B)]^{-1} \mathbf{g}_{A_i,B} r_{G_{AB}} \sigma_{a_A} \sigma_{a_B}\\
&= r^2_{G_{AB}} \sigma^2_{a_A} \sigma^2_{a_B} \mathbf{g'}_{A_i,B} [Var(\mathbf{y}_B)]^{-1} \mathbf{g}_{A_i,B} \ . \qquad \text{(A3.1)}
\end{aligned}$$

The variance of the predicted breeding value of individual i of population A using a reference population of population B is:

$$\begin{aligned}
&Var(\hat{a}_{A_i}) = Var(r_{G_{AB}} \sigma_{a_A} \sigma_{a_B} \mathbf{g'}_{A_i,B} [Var(\mathbf{y}_B)]^{-1} \mathbf{y}_B)\\
&= r_{G_{AB}} \sigma_{a_A} \sigma_{a_B} \mathbf{g'}_{A_i,B} [Var(\mathbf{y}_B)]^{-1} Var(\mathbf{y}_B) \times [Var(\mathbf{y}_B)]^{-1} \mathbf{g}_{A_i,B} r_{G_{AB}} \sigma_{a_A} \sigma_{a_B}\\
&= r^2_{G_{AB}} \sigma^2_{a_A} \sigma^2_{a_B} \mathbf{g'}_{A_i,B} [Var(\mathbf{y}_B)]^{-1} \mathbf{g}_{A_i,B} . \qquad \text{(A3.2)}
\end{aligned}$$

Combining Equation A3.1 and A3.2, results in:

$$Cov(\hat{a}_{A_i}, a_{A_i}) = Var(\hat{a}_{A_i}) . \qquad \text{(A3.3)}$$

3.8 References

Bohren, B. B., W. G. Hill and A. Robertson, 1966 Some observations on asymmetrical correlated responses to selection. Genet. Res. 7: 44-57.

Boldman, K. G. and A. E. Freeman, 1990 Adjustment for heterogeneity of variances by herd production level in dairy cow and sire evaluation. J. Dairy Sci. 73: 503-512.

Clark, S. A., J. M. Hickey and J. H. J. Van Der Werf, 2011 Different models of genetic variation and their effect on genomic evaluation. Genet. Sel. Evol. 43: 18.

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

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

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., 2009, *Genome-wide evaluation of populations*. PhD thesis: Animal Breeding and Genomics Centre, Wageningen, Wageningen University, Wageningen, NL

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

De los Campos, G., Y. C. Klimentidis, A. I. Vazquez and D. B. Allison, 2012 Prediction of expected years of life using whole-genome markers. PLoS ONE 7: e40964.

De los Campos, G., A. I. Vazquez, R. Fernando, Y. C. Klimentidis and D. Sorensen, 2013 Prediction of complex human traits using the genomic best linear unbiased predictor. PLoS Genet. 9: e1003608.

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.

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

Erbe, M., B. Gredler, F. R. Seefried, B. Bapst and H. Simianer, 2013 A function accounting for training set size and marker density to model the average accuracy of genomic prediction. PLoS ONE 8: e81046.

Falconer, D. S. and T. F. C. Mackay, 1996 *Introduction to quantitative genetics*. Pearson Education Limited, Harlow.

Gilmour, A. R., B. Gogel, B. Cullis, R. Thompson, D. Butler*, et al.*, 2009 *ASReml user guide release 3.0*. VSN International Ltd, Hemel Hempstead.

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

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.

Harris, B. L. and D. L. Johnson, 2010 Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. J. Dairy Sci. 93: 1243-1252.

Harris, B. L., F. E. Creagh, A. M. Winkelman and D. L. Johnson, 2011 Experiences with the Illumina high density Bovine BeadChip. Interbull Bull. 44: 3-7.

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

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

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

Kemper, K. E. and M. E. Goddard, 2012 Understanding and predicting complex traits: Knowledge from cattle. Hum. Mol. Genet. 21: R45-R51.

Khatkar, M. S., F. W. Nicholas, A. R. Collins, K. R. Zenger, J. A. L. Cavanagh*, et al.*, 2008 Extent of genome-wide linkage disequilibrium in Australian Holstein-Friesian cattle based on a high-density SNP panel. BMC Genom. 9: 187.

Legarra, A., G. Baloche, F. Barillet, J. Astruc, C. Soulas*, et al.*, 2014 Within-and across-breed genomic predictions and genomic relationships for Western Pyrenees dairy sheep breeds Latxa, Manech, and Basco-Béarnaise. J. Dairy Sci. 97: 3200-3212.

Legates, J. E., 1962 Heritability of fat yields in herds with different production levels. J. Dairy Sci. 45: 990-993.

Makgahlela, M. L., E. A. Mäntysaari, I. Strandén, M. Koivula, U. S. Nielsen*, et al.*, 2013 Across breed multi-trait random regression genomic predictions in the Nordic Red dairy cattle. J. Anim. Breed. Genet. 130: 10-19.

Matukumalli, L. K., C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan*, et al.*, 2009 Development and characterization of a high density SNP genotyping assay for cattle. PLoS ONE 4: e5350.

McKay, S. D., R. D. Schnabel, B. M. Murdoch, L. K. Matukumalli, J. Aerts*, et al.*, 2007 Whole genome linkage disequilibrium maps in cattle. BMC Genet. 8: 74.

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.

Mrode, R. A. and R. Thompson, 2005 *Linear models for the prediction of animal breeding values*. CABI Publishing, Wallingford.

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

Pryce, J. E., B. Gredler, S. Bolormaa, P. J. Bowman, C. Egger-Danner*, et al.*, 2011 Short communication: Genomic selection using a multi-breed, across-country reference population. J. Dairy Sci. 94: 2625-2630.

Pryce, J. E., J. Johnston, B. J. Hayes, G. Sahana, K. A. Weigel*, et al.*, 2014 Imputation of genotypes from low density (50,000 markers) to high density (700,000 markers) of cows from research herds in Europe, North America, and Australasia using 2 reference populations. J. Dairy Sci. 97: 1799-1811.

Schrooten, C., G. C. B. Schopen and P. Beatson, 2013 Across-breed genomic evaluation based on BovineHD genotypes, and phenotypes of bulls and cows. Proc. 64th EAAP annual meeting Wageningen Academic Publishers, Wageningen, Nantes.

Simeone, R., I. Misztal, I. Aguilar and Z. G. Vitezica, 2012 Evaluation of a multi-line broiler chicken population using a single-step genomic evaluation procedure. J. Anim. Breed. Genet. 129: 3-10.

Spelman, R. J., C. A. Ford, P. McElhinney, G. C. Gregory and R. G. Snell, 2002 Characterization of the DGAT1 gene in the New Zealand dairy population. J. Dairy Sci. 85: 3514-3517.

Strandén, I. and E. A. Mäntysaari, 2013 Use of random regression model as an alternative for multibreed relationship matrix. J. Anim. Breed. Genet. 130: 4-9.

Su, G., O. F. Christensen, T. Ostersen, M. Henryon and M. S. Lund, 2012 Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. PLoS ONE 7: e45293.

Thaller, G., W. Krämer, A. Winter, B. Kaupe, G. Erhardt*, et al.*, 2003 Effects of DGAT1 variants on milk production traits in German cattle breeds. J. Anim. Sci. 81: 1911-1918.

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

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.

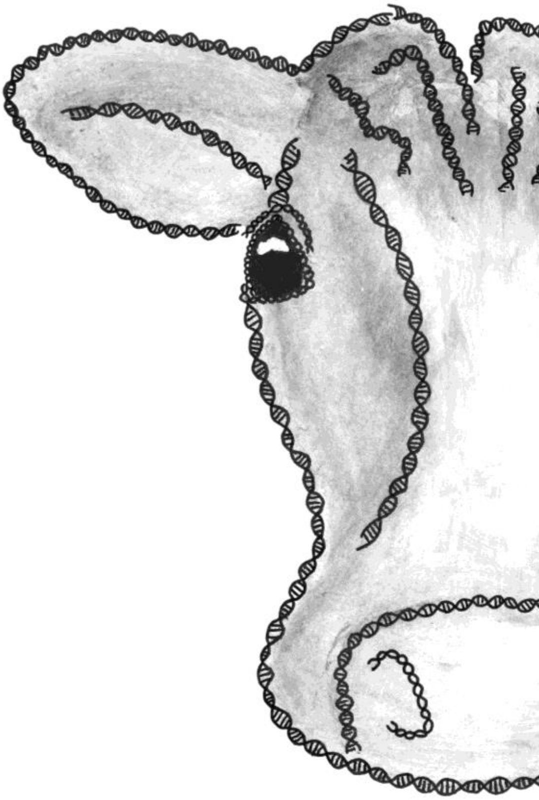
Wittenburg, D., N. Melzer and N. Reinsch, 2011 Including non-additive genetic effects in Bayesian methods for the prediction of genetic values based on genome-wide markers. BMC Genet. 12: 74.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders*, et al.*, 2010 Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42: 565-569.

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.

CHAPTER 4

USING SELECTION INDEX THEORY TO ESTIMATE CONSISTENCY OF MULTI- LOCUS LINKAGE DISEQUILIBRIUM ACROSS POPULATIONS



Y.C.J. WIENTJES[1,2]

R.F. VEERKAMP[1,2]

M.P.L. CALUS[1]

[1] ANIMAL BREEDING AND GENOMICS CENTRE,
WAGENINGEN UR LIVESTOCK RESEARCH,
6700 AH WAGENINGEN, THE NETHERLANDS
[2] ANIMAL BREEDING AND GENOMICS CENTRE,
WAGENINGEN UNIVERSITY,
6700 AH WAGENINGEN, THE NETHERLANDS

BMC GENETICS (2015) 16:87

Abstract

Background: The potential of combining multiple populations in genomic prediction is depending on the consistency of linkage disequilibrium (LD) between SNPs and QTL across populations. We investigated consistency of multi-locus LD across populations using selection index theory and investigated the relationship between consistency of multi-locus LD and accuracy of genomic prediction across different simulated scenarios. In the selection index, QTL genotypes were considered as breeding goal traits and SNP genotypes as index traits, based on LD among SNPs and between SNPs and QTL.

Methods: The consistency of multi-locus LD across populations was computed as the accuracy of predicting QTL genotypes in selection candidates using a selection index derived in the reference population. Different scenarios of within- and across-population genomic prediction were evaluated, using all SNPs or only the four neighboring SNPs of a simulated QTL. Phenotypes were simulated using different numbers of QTL underlying the trait. The relationship between the calculated consistency of multi-locus LD and accuracy of genomic prediction using a GBLUP type of model was investigated.

Results: The accuracy of predicting QTL genotypes, i.e., the measure describing consistency of multi-locus LD, was much lower for across-population scenarios compared to within-population scenarios, and was lower when QTL had a low minor allele frequency compared to QTL randomly selected from the SNPs. Consistency of multi-locus LD was highly correlated with the realized accuracy of genomic prediction across different scenarios and the correlation was higher when QTL were weighted according to their effects in the selection index instead of weighting QTL equally. By only considering neighboring SNPs of QTL, accuracy of predicting QTL genotypes within population decreased, but it substantially increased the accuracy across populations.

Conclusions: Consistency of multi-locus LD across populations is a characteristic of the properties of the QTL in the investigated populations and can provide more insight in underlying reasons for a low empirical accuracy of across-population genomic prediction. By focusing in genomic prediction models only on neighboring SNPs of QTL, multi-locus LD is more consistent across populations since only short-range LD is considered, and accuracy of predicting QTL genotypes of individuals from another population is increased.

Key words: Multi-locus LD, consistency of LD, genomic prediction, across-population genomic prediction, accuracy, selection index theory

4.1 Background

In genomic prediction, marker information is used to predict breeding values for selection candidates based on estimated marker effects in a reference population consisting of individuals with phenotypes and marker genotypes. The accuracy of predicting genomic breeding values depends on the size of the reference population, the heritability of the trait, and on the level of family relationships between the reference population and selection candidates (e.g., Daetwyler *et al.* 2008; Habier *et al.* 2010; Wientjes *et al.* 2013). Moreover, the accuracy is influenced by the level of linkage disequilibrium (LD), i.e., non-random associations, between the single-nucleotide polymorphism (SNP) markers and quantitative trait loci (QTL) influencing the trait of interest (Meuwissen *et al.* 2001). The higher the level of LD, the more accurate breeding values can be predicted for the selection candidates (Goddard 2009). Therefore, the consistency of linkage phase between SNPs and QTL across populations has been suggested to be an important factor determining the success of across- and multi-population genomic prediction (De Roos *et al.* 2009; Hayes *et al.* 2009). Within a population, the level of LD between a QTL and a SNP depends on the effective population size, the recombination rate, the distance between the QTL and SNP on the genome, and the difference in allele frequency between the QTL and SNP (Hill and Robertson 1968). Several studies showed different LD patterns across different cattle (Gautier *et al.* 2007; De Roos *et al.* 2008), chicken (Heifetz *et al.* 2005; Andreescu *et al.* 2007), pig (Veroneze *et al.* 2013) and human (Sawyer *et al.* 2005) populations. In different livestock species, however, the consistency of linkage phase across populations is found to be reasonable high at short distances on the genome (Andreescu *et al.* 2007; De Roos *et al.* 2008; Zhou *et al.* 2013), and depending on the degree of relatedness between the populations; the higher the relatedness between the populations, the higher the consistency of LD (Andreescu *et al.* 2007).

The studies investigating the consistency of LD across populations focused on the LD between two loci. However, genomic prediction models trained within populations are expected to use more than one SNP to capture the genetic variance explained by one QTL (Erbe *et al.* 2012). Hayes *et al.* (2007) for example showed a substantial increase in the proportion of the QTL variance captured by the SNPs when going from haplotypes based on 2 SNPs per haplotype to 4 SNPs per haplotype and from 4 SNPs per haplotype to 6 SNPs per haplotype. Moreover, the proportion of the QTL variance explained by haplotypes with more than 2 SNPs was higher than the proportion that could be explained by the SNP in highest LD with the QTL (Hayes *et al.* 2007). Also for fine mapping QTL, the use of haplotypes consisting of multiple SNPs is shown to be beneficial compared to using one SNP at a time (Meuwissen and Goddard 2000; Grapes *et al.* 2006; Calus *et al.* 2009). This indicates that SNPs in less strong LD with the QTL might be helpful in genomic prediction, and linear combinations of several linked SNPs form the within-population prediction equation. Therefore, a measure of multi-locus LD, compared to the average LD between two adjacent loci, might be better able to explain the contribution of LD to the accuracy of genomic prediction. This might especially be important for situations with multiple populations, because the consistency of LD across populations is decreasing more rapidly at increasing distances on the genome (Gautier *et al.* 2007; De Roos *et al.* 2008; Abasht *et al.* 2009).

The first objective of this study was to investigate the consistency of multi-locus LD across different populations using selection index theory. The consistency of multi-locus LD is one of the components of the accuracy of genomic prediction, therefore, the second objective was to investigate the relationship between consistency of multi-locus LD and accuracy of genomic prediction across different simulated within- and across-population genomic prediction scenarios. Three different cattle breeds with real SNP genotype information were used to represent different populations. Phenotypes of the individuals were simulated by sampling QTL from the SNPs, such that the actual QTL genotypes influencing the phenotypes were known.

4.2 Methods

4.2.1 Prediction accuracies

4.2.1.1 Using selection index theory to predict QTL genotypes

In this study, the consistency of multi-locus LD across different populations is investigated using selection index theory (Smith 1936; Hazel and Lush 1942; Hazel 1943), which is equivalent to multiple regression of the QTL genotypes on the SNP genotypes. In the selection index calculations, a regression equation to predict the QTL genotypes (i.e., the breeding goal traits) using SNP genotypes (i.e., the index traits) was derived in population *A* and the accuracy of this equation to predict the QTL genotypes in population *B* was investigated. This approach is different from other studies investigating the consistency of LD across populations (e.g., Gautier *et al.* 2007; De Roos *et al.* 2008; Zhou *et al.* 2013), where the consistency of LD was calculated using the correlation of the LD measure *r* between two single loci across populations. The advantage of our selection index method is that a measure is obtained of explaining the QTL genotypes using the information of multiple SNPs instead of a single SNP.

In population *A*, a selection index can be derived to predict the QTL genotype for a single individual using all SNP genotypes of that same individual, following:

$$I_i = \mathbf{b'}_A \, \mathbf{x}_i \, , \tag{4.1}$$

in which I_i forms the selection index for individual *i*, $\mathbf{b}_A$ is a vector containing regression coefficients on the SNP genotypes to predict I_i, and $\mathbf{x}_i$ is a vector containing all SNP genotypes of individual *i*.

Rather than predicting I_i, the aim is to predict the aggregated genotype including all QTL:

$$H_i = \mathbf{v'g}_i \, , \tag{4.2}$$

in which H_i is the aggregate genotype of individual *i*, **v** is a vector with weighting factors for each of the QTL genotypes and $\mathbf{g}_i$ is a vector containing the genotype for each QTL of individual *i*.

The regression coefficients on the SNP genotypes that would optimize the prediction accuracy of *H* can be calculated as (Kempthorne and Nordskog 1959):

$$\mathbf{b}_A = \mathbf{P}_A^{-1}\mathbf{G}_A\mathbf{v} \, , \tag{4.3}$$

in which $\mathbf{P}_A$ is the covariance matrix (based on LD) between all SNPs in population *A* and $\mathbf{G}_A$ is the covariance matrix between SNPs and QTL in population *A*. Then the prediction accuracy of predicting the QTL genotype in another population, i.e., population *B*, using $\mathbf{b}_A$ can be calculated as (Lin 1978):

$$r_{IH} = \frac{\mathbf{b'}_A \, \mathbf{G}_B\mathbf{v}}{\sqrt{\mathbf{b'}_A \, \mathbf{P}_B\mathbf{b}_A\mathbf{v'C}_B\mathbf{v}}} \, , \tag{4.4}$$

in which $\mathbf{G}_B$ is the covariance matrix between SNPs and QTL in population *B*, $\mathbf{P}_B$ is the covariance matrix of SNPs in population *B* and $\mathbf{C}_B$ is the covariance matrix of QTL in population *B*.

4.2.1.2 Using a genomic best linear unbiased prediction model to estimate breeding values

To investigate the relationship between the prediction accuracies of the QTL genotypes and the accuracies of predicting genomic breeding values, the following Genomic-relationship-matrix Residual Maximum Likelihood (GREML) model was used:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zg} + \mathbf{e} \, , \tag{4.5}$$

in which **y** is a vector containing phenotypes, **b** is a vector containing fixed effects, **X** is an incidence matrix that allocates the fixed effects to the individuals, **g** is a vector containing the predicted genomic breeding values ~$N(0, \mathbf{GRM}\,\sigma_g^2)$, **GRM** is a genomic relationship matrix based on SNPs (calculation of **GRM** is explained later),

Z is an incidence matrix that allocates the genomic breeding values to the individuals and **e** is a vector containing the residuals ~$N(0,\mathbf{I}\sigma_e^2)$. The GREML model is equivalent to the commonly known genomic best linear unbiased prediction (GBLUP) model, except that it estimates the variances using residual maximum likelihood (REML) instead of assuming that the variances are known.

4.2.2 Simulations to investigate the prediction accuracies

4.2.2.1 Genotypes

Genotypes of 1285 dairy cows from the Netherlands were used, originating from three different breeds (1033 Holstein Friesians (HF), 105 Groningen White Headed (GWH), and 147 Meuse-Rhine-Yssel (MRY)). The genotypes of MRY and GWH animals were obtained by isolating DNA from whole blood samples of the animals. Blood samples were collected in accordance with the guidelines for the care and use of animals as approved by the ethical committee on animal experiments of ID-LELYSTAD (protocol: 2011062). No approval was obtained for the HF genotypes, because these genotypes were obtained from an existing database.

All animals originated for at least 87.5% from one of the three breeds, so were considered to be pure-bred animals. The HF animals were genotyped with the Illumina BovineSNP50 Beadchip (50k, Illumina, San Diego, CA), and genotypes were imputed to high-density (777k) using 3150 HF animals in the reference population as described in Pryce *et al.* (2014). The GWH and MRY animals were genotyped with the Illumina BovineHD Beadchip (777k, Illumina, San Diego, CA). The quality checks and the criteria for including the SNP genotypes in the combined dataset of the three breeds are described in Wientjes *et al*. (2015b). For each of the individuals, both genotype (coded as 0, 1 and 2) and phased allele information (coded as 0 and 1) was available. Phasing of the allele genotypes was done using the software package Beagle (Browning and Browning 2009). From those high-density genotypes, arbitrarily the SNP genotypes of three chromosomes (*Bos Taurus* chromosome 13, 23 and 28) were selected to reduce computation time and to increase the power of the study to estimate breeding values. The three selected chromosomes contained 31,503 SNPs, which was about 10% of the SNPs from the entire combined dataset. The characteristics of the 31,503 SNPs used in this study are shown in Table 4.1.

From all 31,503 SNPs, randomly 5000 SNPs were selected to become candidate QTL from which the actual QTL were sampled. The other 26,503 SNPs were used as SNP markers in this study. With this approach, it was possible to randomly sample QTL from the candidate QTL in each of the replicates, while keeping the set of SNP markers constant across the replicates to reduce the computational demands. To limit the number of possible singularities in the matrices needed for the selection index calculations, SNPs with a correlation above 0.85 or below -0.85 with another SNP on the same chromosome were deleted, irrespective of their allele frequency. Moreover, SNPs that were not segregating in one of the breeds were deleted as well. Deleting those SNPs reduced the total number of SNPs from 26,503 to 4541, of which 1655 SNPs were located on BTA 13, 1515 on BTA 23, and 1371 on BTA 28.

Table 4.1 Characteristics of the SNPs in each of the different breeds.

Characteristics of the SNPs	HF	GWH	MRY
Number of segregating SNPs	31,483	30,449	31,262
Number of breed-specific SNPs	14	6	3
Average MAF[a] of all SNPs	0.279	0.251	0.266
Average MAF[a] of segregating SNPs	0.279	0.260	0.268
Number of SNPs with MAF[a] ≤ 0.1	4266	6530	5308
Number of SNPs with 0.1 < MAF[a] ≤ 0.2	5587	5803	5609
Number of SNPs with 0.2 < MAF[a] ≤ 0.3	6558	5745	6623
Number of SNPs with 0.3 < MAF[a] ≤ 0.4	7430	6718	6657
Number of SNPs with 0.4 < MAF[a] ≤ 0.5	7662	6707	7306

[a]MAF = Minor allele frequency;
HF = Holstein Friesian; GWH = Groningen White Headed; MRY = Meuse-Rhine-Yssel.

4.2.2.2 Phenotypes

Phenotypes were simulated for each individual by randomly sampling 3000, 300, 30, or 3 QTL from the group of 5000 candidate QTL and by sampling their allele substitution effects from $N(0,1)$, using the same effects for each of the breeds. An additive model, without considering epistatic interactions or dominance effects, was assumed. The simulated allele substitution effects were multiplied with the QTL genotypes, coded as 0, 1 and 2, to calculate a true breeding value (TBV) for each of the individuals. Those TBVs were rescaled to a mean of 0 and a variance of 1 across breeds for all of the scenarios. Thus, when the number of QTL underlying the trait was lower, each QTL explained a larger part of the genetic variance. For each individual, an environmental effect was sampled from N(0, $\left(\frac{1}{h^2}-1\right)$*variance of TBV corrected for mean TBV within breed), in which h^2 is the heritability of the simulated trait. This approach enables to sample the environmental term from the same distribution for each individual, independent of the breed, and to keep the heritability more or less constant across the breeds (Wientjes *et al.* 2015b). The phenotype for each individual was calculated as the sum of its TBV and its randomly sampled environmental effect. Please note that the TBVs were only corrected for the mean TBV to calculate the environmental variance, the TBVs and the phenotypes still contained the breed effect.

Two different heritabilities were used to simulate phenotypes, namely 0.3 and 0.95. The same subsets of QTL were used to simulate phenotypes for the two heritabilities, but allele substitution effects and environmental effects were different. For all scenarios, simulations were replicated 100 times for each scenario. A more detailed description of the simulations of phenotypes can be found in Wientjes *et al.* (2015b).

In general, QTL underlying complex traits are expected to have a lower minor allele frequency (MAF) than the SNPs, due to ascertainment bias of the SNPs on the chip (Matukumalli *et al.* 2009; Kemper and Goddard 2012). To investigate if selecting QTL randomly from the SNPs could affect our results, phenotypes were also simulated by selecting QTL from the 5000 candidate QTL with an average MAF across the breeds below 0.1. The average MAF across the breeds was calculated by giving an equal weight to each of the three breeds, indicating that the allele frequency in each of the breeds ranged between 0 and 0.3, resulting in sampling QTL from 480 candidate QTL. Simulating phenotypes by selecting QTL with a low MAF was only done using 3 QTL underlying the trait and a heritability of 0.95 using 100 replicates.

4.2.2.3 Scenarios

The consistency of multi-locus LD and accuracy of genomic prediction were evaluated in five different scenarios (Table 4.2). In the base scenario, within population genomic prediction was applied, using HF individuals both in the reference population and as selection candidates. The other four scenarios used across-population genomic prediction, indicating that the population of the selection candidates (GWH or MRY) was not included in the reference population, and that all individuals of the predicted population were used for the validation. To perform validation in the within-population scenario, 10-fold cross validation was used in which the individuals were randomly divided in 10 equally sized groups using each group once as selection candidates and the other groups as reference population. In each replicate, the division of the individuals over the groups was the same.

Table 4.2 Overview of the breeds used in the different reference populations and as selection candidates.

	Reference population		Predicted individuals	
Scenario	**Breed(s)**	**Nb of individuals**	**Breed**	**Nb of individuals**
Base	HF	928-929	HF	103-104
1	HF	1033	GWH	105
2	HF + MRY	1180	GWH	105
3	HF	1033	MRY	147
4	HF + GWH	1138	MRY	147

HF = Holstein Friesian; GWH = Groningen White Headed; MRY = Meuse-Rhine-Yssel.

4.2.2.4 Selection index calculations

The selection index calculations were performed for each scenario by defining a selection index to predict QTL genotypes in the reference population (Equation 4.3) and to calculate the prediction accuracy of this selection index in the selection candidates (Equation 4.4). In the **P-**, **G-**, and **C-**matrices (Equation 4.3 and 4.4), we used the correlations between SNPs and QTL that were calculated based on the phased alleles of SNPs and QTL of all individuals in either the reference population or the group of selection candidates. By using correlations instead of covariances, each SNP explains an equal amount of the genetic variance, similar to the commonly used assumption in GREML. Moreover, the square of the correlation between phased alleles at two loci, r^2, is commonly used as a measure for LD between loci (Hill and Robertson 1968).

Across the different replicates, the subset of SNPs was constant, as indicated previously. This indicates that the **P-**matrices within both the reference population and the selection candidates were constant across the replicates. The set of QTL differed for each replicate, so both the **G-** and **C-**matrices were specific for each of the replicates. Correlations among SNPs and QTL and between SNPs and QTL on different chromosomes were taken into account as well to make the analyses consistent with the GREML analyses that did not differentiate between the chromosomes. To prevent problems due to non-positive definiteness of the final matrices, the **P-** and **C-**matrices were bended following the unweighted bending procedure described by Jorjani *et al.* (2003) by setting the eigenvalues of the matrix lower than $10e^{-6}$ to $10e^{-6}$.

Two different weightings of the QTL in the overall breeding goal, vector **v** in Equation 4.2, 4.3 and 4.4, were used; either QTL were weighted equally (**v** is a vector of ones), or each QTL was weighted based on its simulated allele substitution effect to take into account that it is more important to accurately predict the QTL genotype of QTL with large effects than for QTL with small effects. Weighting the QTL based on their allele substitution effects was only performed for the phenotypes simulated using a heritability of 0.95, both when QTL were randomly selected and when QTL were selected with a low MAF.

In the analyses described above, all SNPs across the whole genome were taken into account to explain the QTL genotypes. The SNPs more closely located to a QTL are supposed to have a higher and more consistent LD with the QTL across populations (e.g., Andreescu *et al.* 2007; De Roos *et al.* 2008; Zhou *et al.* 2013). To investigate if the accuracy of predicting QTL genotypes would be increased when focusing only on the SNPs surrounding a QTL, the analyses with 3 randomly selected QTL underlying the trait were repeated using only the four surrounding SNPs (two at either side) of each QTL. When the number of SNPs from one side of the QTL was insufficient, i.e., when the QTL was located at the end of a chromosome, more SNPs from the other side of the QTL were added to obtain four SNPs per QTL. Those analyses were only performed by using an equal weight of the QTL in the overall breeding goal.

4.2.2.5 Estimating breeding values using GREML

To estimate breeding values for the individuals, the GREML model (Equation 4.5) was run in ASReml (Gilmour *et al.* 2009), including breed as the only fixed effect. The **GRM** matrix that was used in the model was calculated as $\mathbf{GRM} = \frac{\mathbf{XX'}}{n}$ (VanRaden 2008; Yang *et al.* 2010), in which n represents the number of SNP markers (n = 4541) and the **X**-matrix contains standardized genotypes, calculated as $x_{ij} = \frac{g_{ij} - 2p_j}{\sqrt{2p_j(1-p_j)}}$, in which g_{ij} codes the genotype for individual i at marker locus j as 0, 1 and 2, and p_j is the allele frequency at marker locus j for the second allele (for which the homozygote genotype is coded 2) averaged over the three breeds. After adjusting the inbreeding level in **GRM** to the inbreeding level in the pedigree based relationship matrix **A**, the **GRM** matrix was regressed back to the **A** matrix to reduce the effect of sampling the SNPs on the chip. For each of the scenarios, a different **GRM** matrix was calculated, containing only the individuals included in that scenario. For a more detailed description of calculating **GRM**, see Wientjes *et al.* (2015b).

For each population, the accuracy of genomic prediction was calculated as the correlation between the estimated breeding values and the simulated TBVs.

Averages and standard errors of the accuracies of genomic prediction were calculated across replicates.

4.3 Results

4.3.1 Regression coefficients

The regression coefficients on the SNP genotypes to predict the QTL genotypes derived in the Holstein Friesian reference population using selection index calculations (Equation 4.3; $\mathbf{b}_{RP}$) are presented in Figure 4.1 for one of the replicates with 3 randomly selected QTL underlying the trait. This figure clearly shows that the SNPs surrounding a QTL were given a higher weight to predict the QTL genotypes, due to the greater correlations between those SNPs and the QTL. When QTL were weighted based on their different allele substitution effects, mainly the SNPs surrounding the QTL with a large effect were given a higher weight. The same patterns were also seen when the number of QTL was higher, although the pattern was less clear due to the higher number of QTL (see Appendix Figure A4.1, Figure A4.2, and Figure A4.3), and when the MAF of QTL was lower (see Appendix Figure A4.4).

4.3.2 Accuracy of predicting QTL genotypes using selection index theory

Accuracies of predicting the QTL genotypes for the selection candidates, using a selection index derived in the reference population based on all SNPs, are shown in Figure 4.2 when QTL were randomly sampled. Since this prediction accuracy is a measure of the consistency of multi-locus LD (MLLD) between the selection candidates and the reference population, hereafter this accuracy will be referred to as acc_MLLD. In the within-population scenarios, average acc_MLLD was around 0.94. As expected, average acc_MLLD was much lower for the across population scenarios due to differences in LD across populations with an average acc_MLLD of ~0.37 for GWH and ~0.34 for MRY using HF as reference population. Adding another population to the HF reference population did not affect the prediction accuracy.

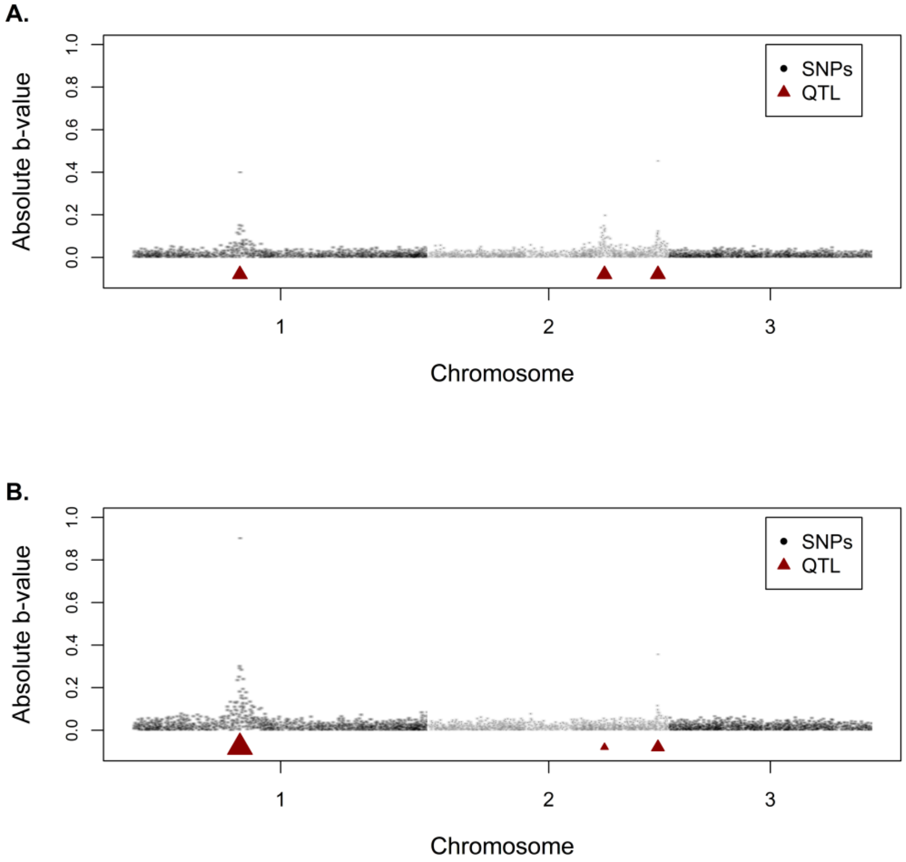


Figure 4.1 Absolute estimated regression coefficients (b-values) for each SNP to predict the QTL genotypes of 3 randomly selected QTL. Absolute regression coefficients for each of the SNPs estimated in a Holstein Friesian reference population (b_{RP}) to predict the QTL genotypes of 3 randomly selected QTL with (A) equal weight for each of the QTL, or (B) QTL weighted differently, based on their allele substitution effects, in the overall breeding goal. The size of the triangle represents the weight of the QTL in the overall breeding goal of the selection index calculations, i.e., the allele substitution effect in (B).

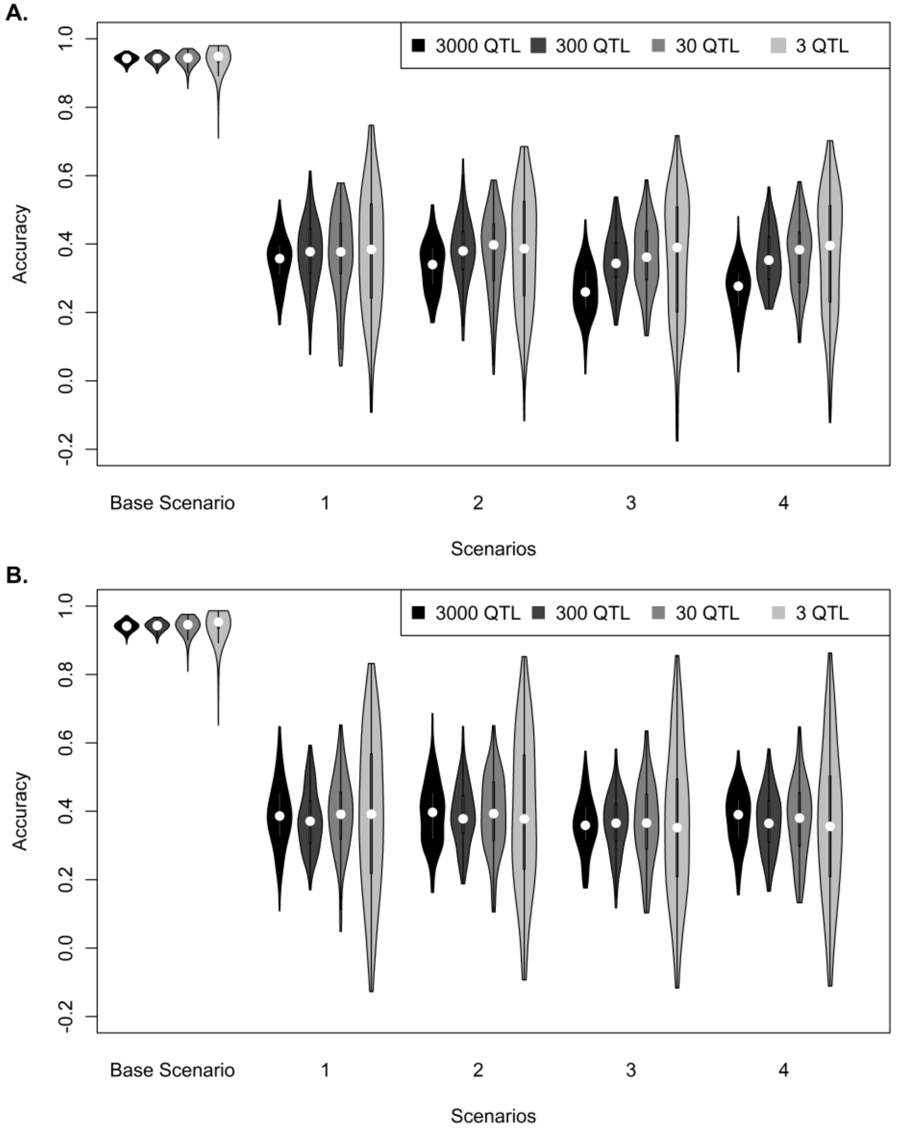


Figure 4.2 Accuracies of predicing genotypes of randomly sampled QTL using selection index theory. Violin plot depicting the accuracies of selection index theory to predict the QTL genotypes of randomly sampled QTL using (A) equal weight for each of the QTL, or (B) QTL weighted differently, based on their allele substitution effects, in the overall breeding goal for five different scenarios. Base = reference population Holstein Friesian (HF), selection candidates HF; 1 = reference population HF, selection candidates Groningen White Headed (GWH); 2 = reference population HF and Meuse-Rhine-Yssel (MRY), selection candidates GWH; 3 = reference population HF, selection candidates MRY; 4 = reference population HF and GWH, selection candidates MRY.

The average acc_MLLD seems to be independent from the number of QTL underlying the trait for the within- as well as for the across-population scenarios, both when QTL had an equal weight and when QTL were weighted based on their allele substitution effects. Only when 3000 QTL were underlying the trait and QTL had an equal weight in the breeding goal, acc_MLLD was slightly lower compared to the across-population scenarios with fewer QTL. Standard errors were in general very small, but tended to be slightly larger for the scenarios with a lower number of QTL.

Weighting the QTL equally or based on their allele substitution effects resulted in similar values for acc_MLLD, both for the within- and across-population scenarios. This was also expected beforehand, since the consistency of multi-locus LD across populations was supposed to be a characteristic of the investigated populations. Giving different weights to the QTL only resulted in giving more emphasis on predicting QTL with a large effect, but it had no effect on the LD structure of that QTL with the surrounding SNPs. The only exception to this pattern was again the across-population scenario with 3000 QTL underlying the trait, where acc_MLLD was higher when QTL were weighted differently compared to weighting the QTL equally.

By focusing only on the four SNPs surrounding a QTL, the accuracy of predicting the QTL genotypes of the selection candidates decreased by 19% for the within-population scenario (Table 4.3). For the across-population scenarios, however, the prediction accuracy increased by approximately 53% (Table 4.3). As a consequence, the difference in prediction accuracy of the QTL genotypes between the within- and across-population scenarios was substantially reduced compared to the analyses using all SNPs.

In Figure 4.3, the values for acc_MLLD are shown when 3 QTL were underlying the trait and when QTL were sampled with a low MAF. The results show that acc_MLLD was lower for all scenarios when the MAF of the QTL was lower, confirming the expectation that the strength of LD is reduced when the MAF of the QTL is lower. The decrease in acc_MLLD was, however, much lower for the within-population scenario where acc_MLLD was around 95% of the acc_MLLD with QTL randomly sampled, than for the across-population scenarios where acc_MLLD was around 60 – 70% of the acc_MLLD with QTL randomly sampled.

Table 4.3 Average prediction accuracies of QTL genotypes using all SNPs or only the neighboring SNPs of the QTL. The results are for different within- and across-population scenarios with 3 QTL underlying the trait and with an equal weight of the QTL in the overall breeding goal.

Scenario	Reference population	Selection candidates	Average prediction accuracy (s.e.)			
			All SNPs		Four surrounding SNPs	
Base	**HF**	**HF**	0.942	(0.003)	0.766	(0.011)
1	**HF**	**GWH**	0.378	(0.018)	0.569	(0.020)
2	**HF + MRY**	**GWH**	0.377	(0.017)	0.579	(0.020)
3	**HF**	**MRY**	0.362	(0.018)	0.562	(0.020)
4	**HF + GWH**	**MRY**	0.373	(0.018)	0.567	(0.021)

HF = Holstein Friesian; GWH = Groningen White Headed; MRY = Meuse-Rhine-Yssel.

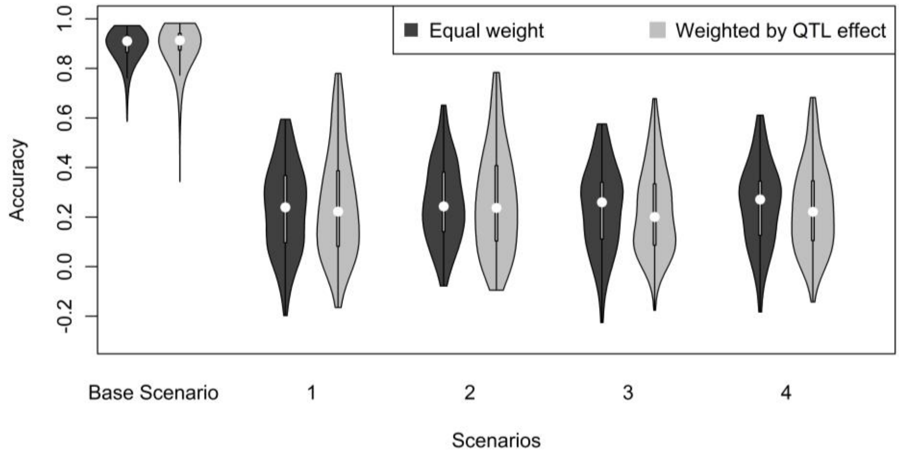


Figure 4.3 Accuracies of predicing genotypes of QTL with low MAF using selection index theory. Violin plot depicting the accuracies of selection index theory to predict the QTL genotypes of three QTL with low MAF using an equal weight for each of the QTL, or different weights for each QTL, based on their allele substitution effects, in the overall breeding goal for five different scenarios. Base = reference population Holstein Friesian (HF), selection candidates HF; 1 = reference population HF, selection candidates Groningen White Headed (GWH); 2 = reference population HF and Meuse-Rhine-Yssel (MRY), selection candidates GWH; 3 = reference population HF, selection candidates MRY; 4 = reference population HF and GWH, selection candidates MRY.

4.3.3 Accuracy of genomic prediction

Accuracies of predicting genomic estimated breeding values, hereafter denoted as acc_GEBV, achieved with a GBLUP type of model are shown in Figure 4.4, for a heritability of 0.95 (A) and a heritability of 0.3 (B). At a heritability of 0.95, the average acc_GEBV for the within-population scenario was around 0.95, and was much lower and in the range of 0.3 – 0.4 across populations. At a heritability of 0.3, average acc_GEBV was lower for all scenarios, with values around 0.75 for the within-population scenario and values around 0.2 for the across-population scenarios. For all scenarios, acc_GEBV was independent from the number of QTL underlying the trait and standard errors were reasonably small, although slightly larger for the across-population scenarios compared to the within-population scenarios.

The acc_GEBV for GWH individuals were somewhat higher (~0.04 at a heritability of 0.95; and ~0.005 at a heritability of 0.3) than predicting MRY individuals using a HF reference population. When the reference population was extended with the other population, acc_GEBV increased slightly, although not significantly, for both populations (~0.015).

Table 4.4 shows the average acc_GEBV when 3 QTL were underlying the trait with QTL randomly selected and QTL selected to have a low MAF for a heritability of 0.95. Those results show that average acc_GEBV was in all scenarios lower when QTL had a low MAF compared to randomly selected QTL. The accuracies achieved for QTL with a low MAF were 98% and 65% of the accuracies for randomly selected QTL for respectively the within- and across-population scenarios, indicating that the decrease in accuracy was smaller for the within-population scenario compared to the across-population scenarios.

Table 4.4 Average accuracies (s.e.) of genomic prediction using QTL randomly sampled or QTL with low minor allele frequency (MAF). The results are for different within- and across-population scenarios with 3 QTL underlying the trait and a heritability of 0.95.

Scenario	Reference population	Selection candidates	Average prediction accuracy (s.e.)	
			QTL randomly sampled	QTL with low MAF
Base	HF	HF	0.949 (0.001)	0.932 (0.002)
1	HF	GWH	0.341 (0.021)	0.233 (0.022)
2	HF + MRY	GWH	0.361 (0.022)	0.246 (0.022)
3	HF	MRY	0.304 (0.020)	0.186 (0.018)
4	HF + GWH	MRY	0.310 (0.021)	0.189 (0.019)

HF = Holstein Friesian; GWH = Groningen White Headed; MRY = Meuse-Rhine-Yssel.

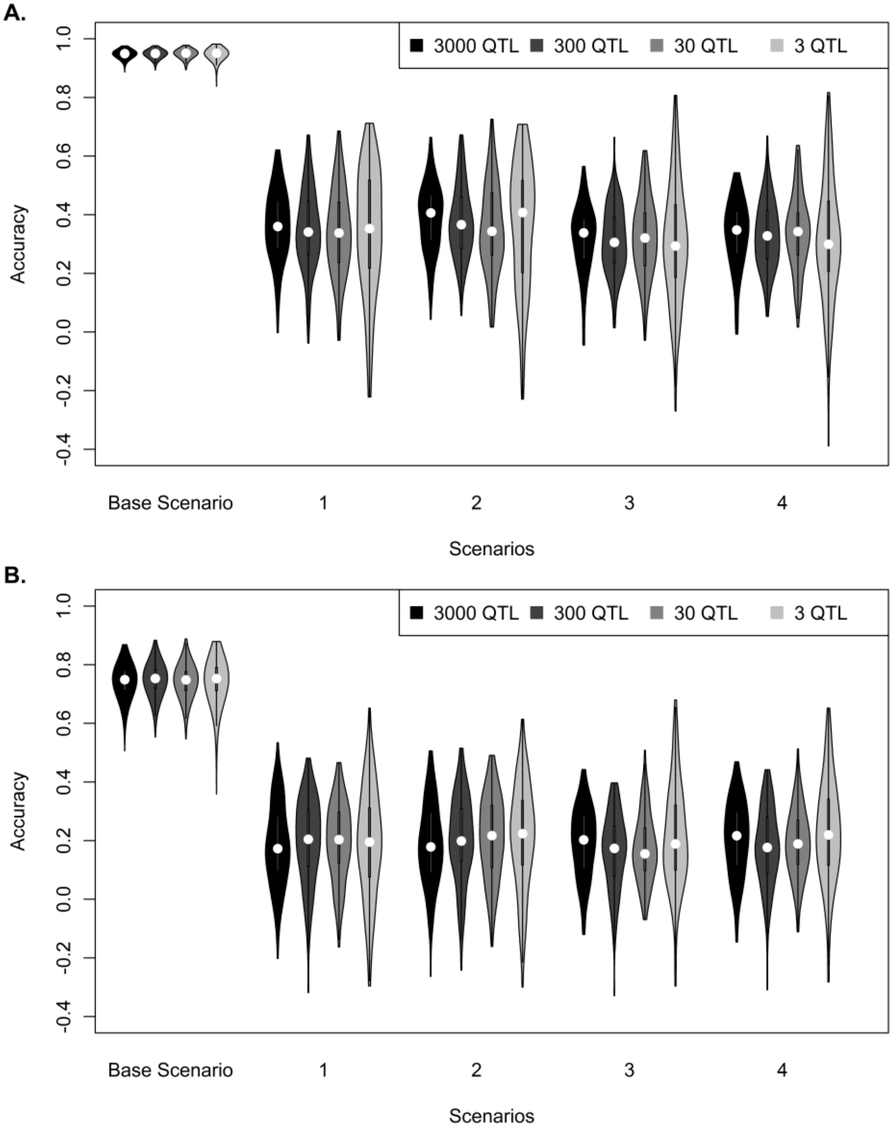


Figure 4.4 Accuracies of predicting genomic breeding values using GREML for different scenarios using multiple populations. Violin plot depicting the accuracies of genomic prediction using GREML and a (A) heritability of 0.95, or (B) heritability of 0.3 for five different scenarios. Base = reference population Holstein Friesian (HF), selection candidates HF; 1 = reference population HF, selection candidates Groningen White Headed (GWH); 2 = reference population HF and Meuse-Rhine-Yssel (MRY), selection candidates GWH; 3 = reference population HF, selection candidates MRY; 4 = reference population HF and GWH, selection candidates MRY.

4.3.4 Accuracy of predicting genomic breeding values (acc_GEBV) versus accuracy of predicting QTL genotypes (acc_MLLD)

To investigate the relationship between acc_MLLD and acc_GEBV across different across-population genomic prediction scenarios, the average acc_GEBV are plotted against the average acc_MLLD in Figure 4.5 for the four across-population scenarios with 3 QTL underlying the trait. As expected, the average acc_MLLD was for most scenarios equal or higher than the average acc_GEBV. When the heritability was 0.95 and QTL were randomly sampled, the average acc_MLLD was ~0.03 higher than acc_GEBV in the across-population scenarios, and the average acc_MLLD and acc_GEBV were similar in the within-population scenarios. The differences were larger when the heritability was 0.3 (~0.17 in the across-population scenarios, and ~0.20 in the within-population scenarios). When QTL were sampled with a low MAF, the differences were comparable to the differences with QTL randomly sampled at a heritability of 0.95 for the across-population scenarios. In the within-population scenarios, however, the average acc_GEBV was ~0.04 higher than acc_MLLD.

The correlation between acc_GEBV and acc_MLLD was expected to be high and positive, since a high consistency of multi-locus LD across reference individuals and selection candidates is supposed to be very important in getting a high accuracy of genomic prediction. Across the four different across-population scenarios and at the same number of randomly sampled QTL underlying the trait and a heritability of 0.95, the average correlation between acc_GEBV and acc_MLLD was 0.91 (range 0.76 to 1.00) when each QTL had an equal weight in the breeding goal, and on average 0.94 (range 0.86 to 1.00) when each QTL had a different weight, based on their different allele substitution effects. When the heritability was only 0.3, the average correlation was lower (0.79). At a heritability of 0.95 and 3 QTL sampled with a low MAF, the correlations were 0.33 and 0.95 when QTL were respectively equally weighted or weighted based on their different allele substitution effects. Altogether, those results show that the measure for consistency of multi-locus LD, acc_MLLD, as calculated in this study using selection index theory, is highly related to the accuracy of genomic prediction obtained with GBLUP.

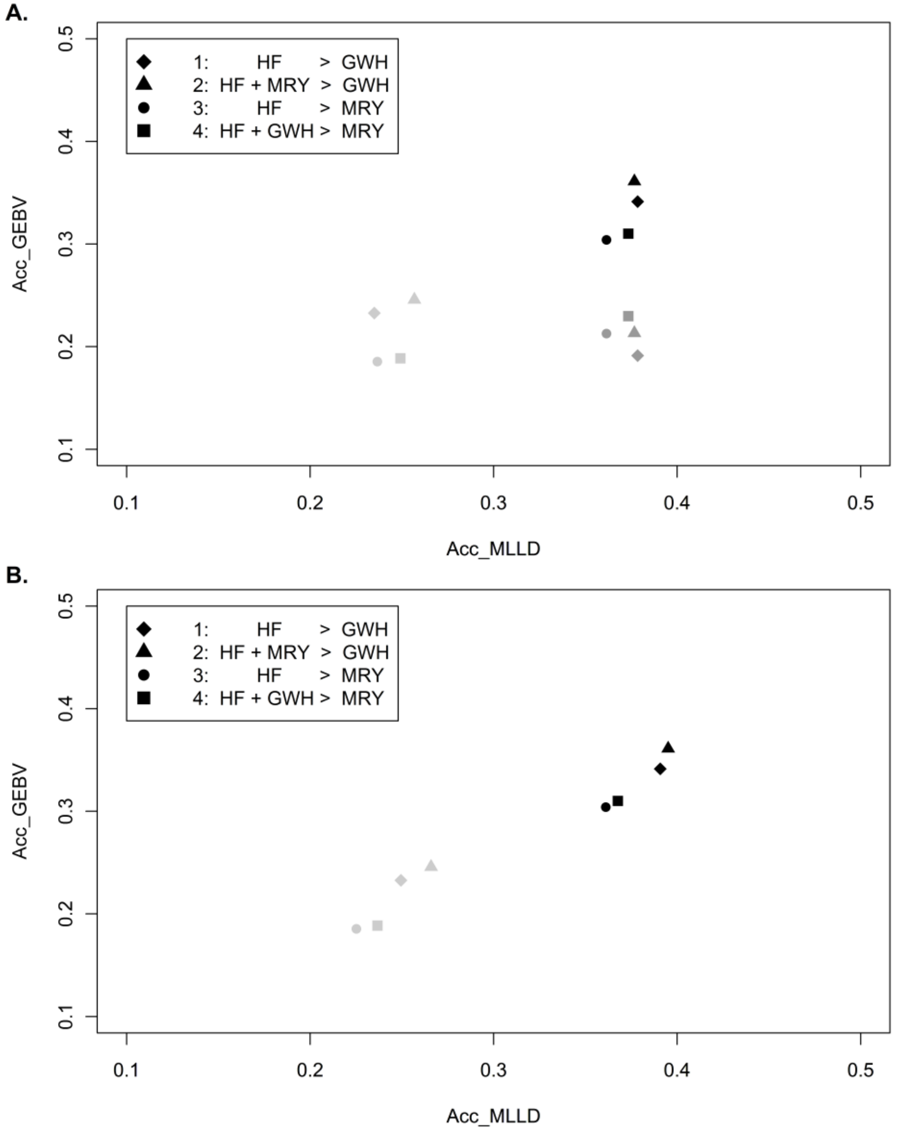


Figure 4.5 Average accuracies of genomic prediction (Acc_GEBV) versus average accuracies of predicting QTL genotypes (Acc_MLLD) with 3 QTL. Average accuracies of genomic prediction (Acc_GEBV) versus average accuracies of selection index theory to predict the QTL genotypes (Acc_MLLD) with (A) equal weight for each of the QTL, or (B) QTL weighted based on their allele substitution effects in the overall breeding goal and with 3 QTL underlying the trait randomly sampled using a heritability of 0.95 (black) or 0.3 (dark grey), or QTL selected with a low MAF and a heritability of 0.95 (light grey) for four different scenarios; HF = Holstein Friesian; MRY = Meuse-Rhine-Yssel; GWH = Groningen White Headed.

4.4 Discussion

4.4.1 Using selection index theory to investigate the consistency of multi-locus LD

The first objective of this study was to investigate the consistency of multi-locus LD across different populations using selection index theory. Our results indicate that the strength of LD reduces when the MAF of the QTL reduces and that LD between QTL and SNPs is at least partly different across populations, especially for loci with a low MAF, resulting in a lower accuracy of predicting the QTL genotypes of selection candidates from another population. When focusing in genomic prediction models only on the SNPs closely located to a QTL, the accuracy of predicting the QTL genotypes of individuals from another population increased, indicating that consistency of LD across populations is higher at shorter distances on the genome. Those findings are in agreement with other studies investigating the consistency of linkage phase between pairs of markers across populations (Gautier *et al.* 2007; De Roos *et al.* 2008), but provide a more complete picture as it considers multi-locus LD. Moreover, the measure for the consistency of multi-locus LD seems to be independent from the number of QTL underlying the trait and the weighting of the QTL in the overall breeding goal of the selection index calculations, but it is depending on the properties of the QTL like allele frequency pattern. Therefore, the consistency of multi-locus LD, as calculated with selection index theory using all SNPs, can be seen as a characteristic of the properties of the QTL for the investigated populations.

4.4.2 Consistency of multi-locus LD and accuracy of genomic prediction

The second objective of this paper was to investigate the relationship between consistency of multi-locus LD and accuracy of genomic prediction across different within- and across-population genomic prediction scenarios. As expected, the correlation between average consistency of multi-locus LD and average accuracy of genomic prediction across the different across-population scenarios was positive and strong, both at a heritability of 0.95 and 0.3, and when QTL were randomly selected or selected to have a low MAF. The correlations were slightly stronger when QTL were weighted based on their allele substitution effects in the overall breeding goal, since it is more important that the linkage phases between SNPs and QTL with a high effect are consistent across reference and selection individuals compared to QTL with a small effect.

At a heritability of 0.95 and with QTL randomly selected, the correlations between consistency of multi-locus LD and accuracy of genomic prediction were around 0.9. This indicates that around 81% of the variance in accuracy of genomic prediction could be explained by differences in consistency of multi-locus LD. The remaining part of the variance might be explained by the accuracy of estimating SNP effects, which influenced the accuracy of genomic prediction, but not the consistency of multi-locus LD. The accuracy of estimating SNP effects in the reference population depends on the allele frequency of the QTL, the number of QTL underlying the trait, the heritability of the trait and the size of the reference population (Meuwissen *et al.* 2001; Daetwyler *et al.* 2008; Goddard 2009). In general, estimated SNP effects are less accurate for traits with a low heritability and for SNPs linked to QTL with a low frequency. This is confirmed by the lower correlations between consistency of multi-locus LD and accuracy of genomic prediction found in this study when the heritability was only 0.3 and when QTL were selected to have a low MAF. The difference in accuracy obtained when QTL were randomly selected compared to selecting QTL with a low MAF was higher for the across-population scenarios compared to the within-population scenarios. This can be explained by the fact that QTL with a low MAF in the reference population explain only a small part of the genetic variance within the selection candidates when they are from the same population (Daetwyler *et al.* 2008). Due to differences in allele frequencies across populations, the penalty of incorrectly estimating the effects of SNPs linked to QTL with a low MAF might be much higher when selection candidates are from a different population (Daetwyler *et al.* 2008). Combining two or more populations in the reference population might increase the probability that the QTL explaining a large part of the genetic variance in the selection candidates are segregating at reasonable allele frequencies in the reference population. This could explain the slight increase in accuracy of across-population genomic prediction when another population was added to the reference population, as seen in this study as well as in other studies (Hayes *et al.* 2009; Pryce *et al.* 2011; Wientjes *et al.* 2015b). Another explanation for the slight increase in accuracy when combining multiple populations in the reference population could be the assigning of the effect of QTL to SNPs that are more closely located to the QTL (Hayes *et al.* 2009), for which the consistency of LD across populations is higher (Andreescu *et al.* 2007; De Roos *et al.* 2008; Zhou *et al.* 2013). This latter explanation is, however, not confirmed by the values for the consistency of multi-locus LD calculated in this study.

Both the accuracy of predicting the QTL genotype and accuracy of genomic prediction were very high in the single population scenario. Those high values might indicate a strong level of LD within the population, but might also be caused by a high level of family relationships within the population, since family relationships and level of LD are entangled (Falconer and Mackay 1996). Both population level LD and LD due to family relationships are helpful in predicting the QTL genotype, resulting in higher accuracies of genomic prediction when the level of family relationships between reference and selection candidates is higher, as was already shown in other studies (Habier *et al.* 2007; Wientjes *et al.* 2013). Across populations, close family relationships are in general absent, so across-population genomic prediction is only depending on the level of LD across the populations, resulting in lower accuracies of genomic prediction. Both the accuracy of predicting the QTL genotype and accuracy of genomic prediction decreased when the MAF of QTL was lower, with a much smaller decrease in the within-population scenario compared to the across-population scenarios. This might be a result of the possibility to tag QTL with low MAF by the SNPs within a population due to the high level of family relationships. Across populations, it is much more difficult to tag those QTL by the SNPs, since only the level of LD across the populations can be used. This indicates that the effect of the MAF of QTL might be much larger for across-population genomic prediction compared to within-population genomic prediction.

By focusing only on the four neighboring SNPs of a QTL, the accuracy of predicting the QTL genotype of the selection candidates substantially decreased within a population, but substantially increased in the across-population scenarios. This indicates that SNPs further away from the QTL on the genome can be helpful in predicting the QTL genotype within a population, but can be detrimental for across-population settings, due to the lower consistency of LD across populations (Andreescu *et al.* 2007; De Roos *et al.* 2008; Zhou *et al.* 2013). The potential of combining populations using the current methods of genomic prediction based on all SNPs would therefore be overestimated by only considering the consistency of LD across populations at short distances on the genome. On the other hand, the results do show that the accuracy of across- and multi-population genomic prediction could potentially be increased by focusing only on the neighboring SNPs of a QTL, for which the consistency of LD is higher across populations.

Within this study, different numbers of QTL were selected and allele substitution effects were drawn from a normal distribution. The actual distribution of allele substitution effects may perhaps be closer to a gamma distribution (Hayes and Goddard 2001), showing few QTL with large effects and many QTL with small effects. In such case, the achieved accuracy mainly depends on the ability to tag those few QTL (Calus *et al.* 2008), so effectively is rather similar to our simulations with only 3 QTL underlying the trait. Since the number of QTL underlying the trait had no effect on the consistency of multi-locus LD and the accuracy of genomic prediction in the GBLUP model, we expect that the results of our study are also valid when QTL effects follow a gamma distribution.

Altogether, the results of this study show that consistency of multi-locus LD can be used to get more insight in possible underlying reasons and potential ways to increase the low empirical accuracies of across-population genomic prediction described in literature (e.g., Pryce *et al.* 2011; Erbe *et al.* 2012; Calus *et al.* 2014), as follows. When a low accuracy of across-population genomic prediction is accompanied by a low consistency of multi-locus LD, a higher marker density might be used to increase the accuracy of genomic prediction. When a low accuracy is not accompanied by a low consistency of multi-locus LD, it indicates that the accuracy of estimating SNP effects is low. This might be caused by differences in allele substitution effects across populations, due to the presence of non-additive effects and differences in allele frequencies across populations (Falconer and Mackay 1996). In genetic analyses, those differences can be taken into account by estimating the genetic correlation across the populations (Karoui *et al.* 2012; Wientjes *et al.* 2015b). Another reason for the low accuracy of estimating SNP effects might be that the allele frequency of the QTL explaining a large part of the genetic variance in the selection candidates is too low in the reference population, the effect of this might be reduced by including another population to the reference population.

4.4.3 Potential applications

Our results showed that consistency of multi-locus LD across populations was not influenced by the number of QTL nor by the weighting of QTL in the overall breeding goal. This indicates that the consistency of multi-locus LD is not trait-dependent and that, even when the actual QTL are unknown, reliable estimates of the consistency of multi-locus LD can be obtained by sampling loci from the SNPs. The characteristics of the QTL, such as allele frequency, however, influenced the consistency of multi-locus LD and accuracy of genomic prediction. The effect of MAF of QTL on accuracy was already shown in other studies (Daetwyler *et al.* 2013; Wientjes *et al.* 2015a), but the results of this study confirm the hypothesis that this effect was due to a reduction in the strength of LD between SNPs and QTL. Therefore, it is highly recommended, assuming that the knowledge about the distribution of allele frequencies of QTL increases in the next decade, to select loci that have comparable allele frequencies as the actual QTL underlying the trait of interest in future applications. Since the main conclusions of this study remain valid when the characteristics of the QTL are taken into account, we expect that those conclusions are also valid for traits with other characteristics, for other breeds and even for other species.

The computational demands for the selection index calculations would be high when including all SNPs on the genome. For practical applications, it might therefore be beneficial to only include a subset of the chromosomes in the analyses which have a representative LD pattern for the whole genome. Computational demands can also be reduced by decreasing the number of QTL, which also reduces the number of potential singularities in the correlation matrices between QTL, since the number of QTL did not have a large impact on the accuracy of predicting the QTL genotype. The number of QTL did, however, influence the variance across the replicates. Therefore, multiple replicates would be necessary when a rather small number of QTL is selected.

4.5 Conclusions

In this paper, selection index theory was used to obtain a measure for the consistency of multi-locus LD across the reference and selection populations. As expected, the consistency of multi-locus LD across populations, when reference and selection candidates were from different populations, was much lower compared to the consistency of multi-locus LD within a population, when reference and selection individuals belonged to the same population. Moreover, the consistency of multi-locus LD was much lower for QTL with a low MAF compared to randomly selected QTL. The average consistency of multi-locus LD is shown to be independent from the number of QTL and the weighting of the QTL in the overall breeding goal of the selection index. Therefore, consistency of multi-locus LD can be seen as a characteristic of the properties of the QTL for the investigated populations. Across different across-population scenarios, consistency of multi-locus LD was highly correlated with the achieved accuracy of genomic prediction using a GBLUP type of model, confirming that consistency of LD is an import factor determining the accuracy of across-population genomic prediction. Therefore, the consistency of multi-locus LD can provide more insight in underlying reasons for a low empirical accuracy of across-population genomic prediction. By focusing only on the SNPs closely located to a QTL, the accuracy of predicting the QTL genotypes of individuals from another population increased. This shows that accuracy of across- and multi-population genomic prediction could be increased by focusing only on the neighboring SNPs of a QTL, for which the consistency of LD is higher across populations.

4.6 Acknowledgements

This study was financially supported by Breed4Food (KB-12-006.03-005-ASG-LR), a public-private partnership in the domain of animal breeding and genomics, and CRV (Arnhem, The Netherlands). The RobustMilk project and the National Institute of Food and Agriculture (NIFA) are acknowledged for providing the 50k genotypes of the HF cows, and the gDMI consortium is acknowledged for imputing those to 777k genotypes. The Dutch Milk Genomics Initiative and the project 'Melk op Maat', funded by Wageningen University (the Netherlands), the Dutch Dairy Association (NZO, Zoetermeer, the Netherlands), the cooperative cattle improvement organization CRV BV (Arnhem, the Netherlands), the Dutch Technology Foundation (STW, Utrecht, the Netherlands), the Dutch Ministry of Economic Affairs (The Hague, the Netherlands) and the Provinces of Gelderland and Overijssel (Arnhem, the Netherlands), are thanked for providing the 777k genotypes of the GWH and MRY cows. The authors acknowledge Myrthe Maurice - Van Eijndhoven for collecting the data of the GWH and MRY cows, and the herd owners for their help in collecting the data.

4.7 Appendix

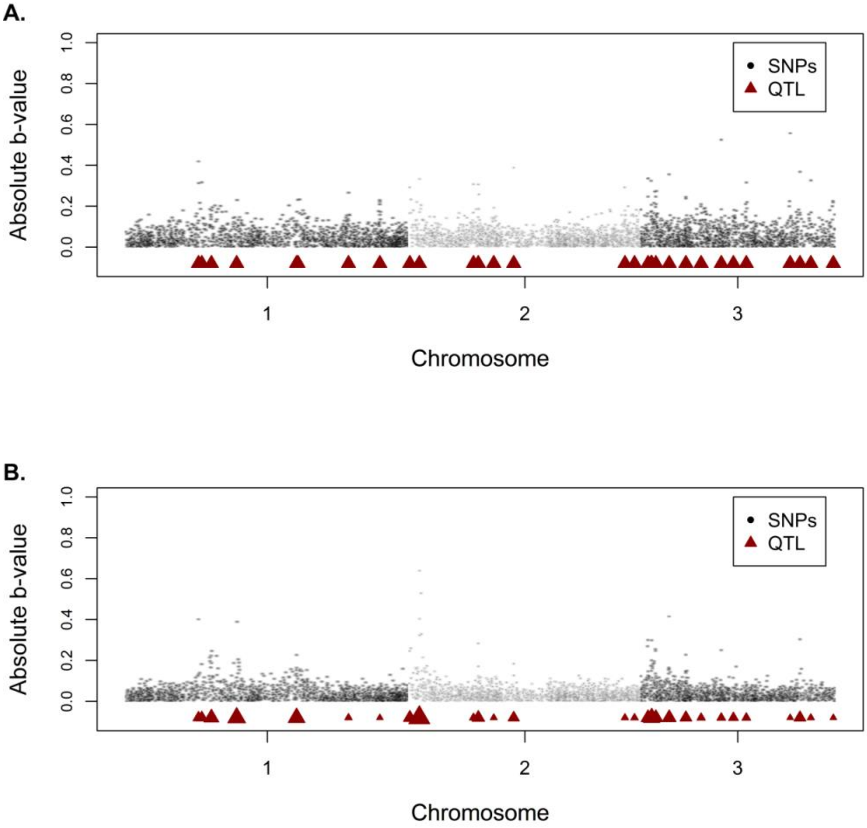


Figure A4.1 Absolute estimated regression coefficients (b-values) for each SNP to predict the QTL genotypes of 30 randomly selected QTL. Absolute regression coefficients for each of the SNPs estimated in a Holstein Friesian reference population (b_{RP}) to predict the QTL genotypes of 30 randomly selected QTL with (A) equal weight for each of the QTL, or (B) QTL weighted differently, based on their allele substitution effects, in the overall breeding goal. The size of the triangle represents the weight of the QTL in the overall breeding goal of the selection index calculations, i.e., the allele substitution effect in (B).

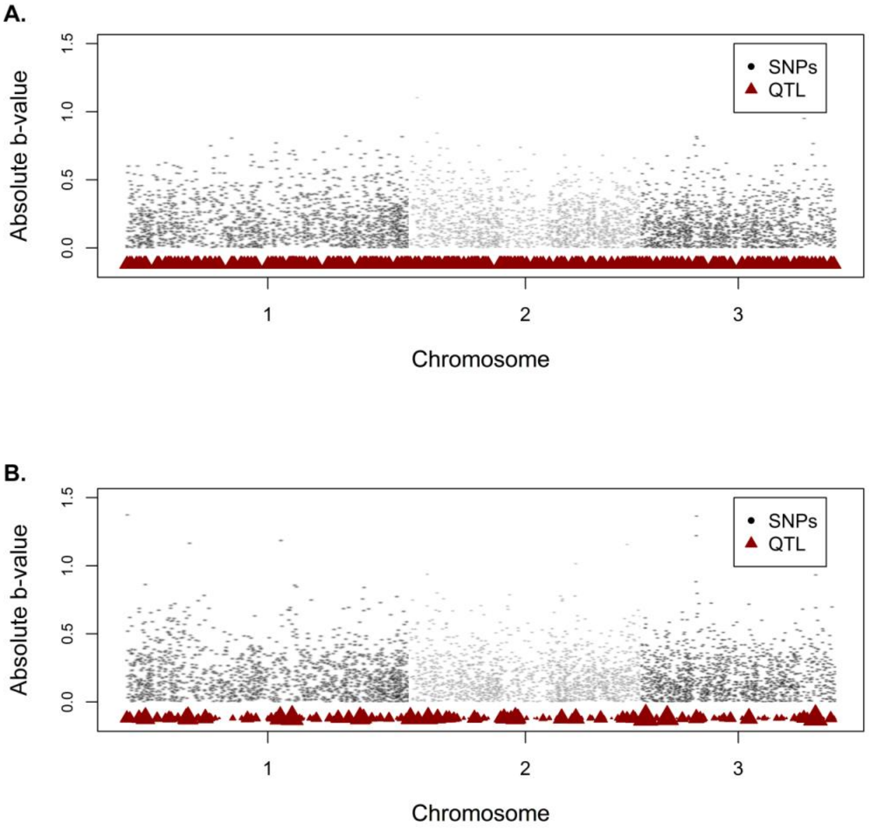


Figure A4.2 Absolute estimated regression coefficients (b-values) for each SNP to predict the QTL genotypes of 300 randomly selected QTL. Absolute regression coefficients for each of the SNPs estimated in a Holstein Friesian reference population (b_{RP}) to predict the QTL genotypes of 300 randomly selected QTL with (A) equal weight for each of the QTL, or (B) QTL weighted differently, based on their allele substitution effects, in the overall breeding goal. The size of the triangle represents the weight of the QTL in the overall breeding goal of the selection index calculations, i.e., the allele substitution effect in (B).

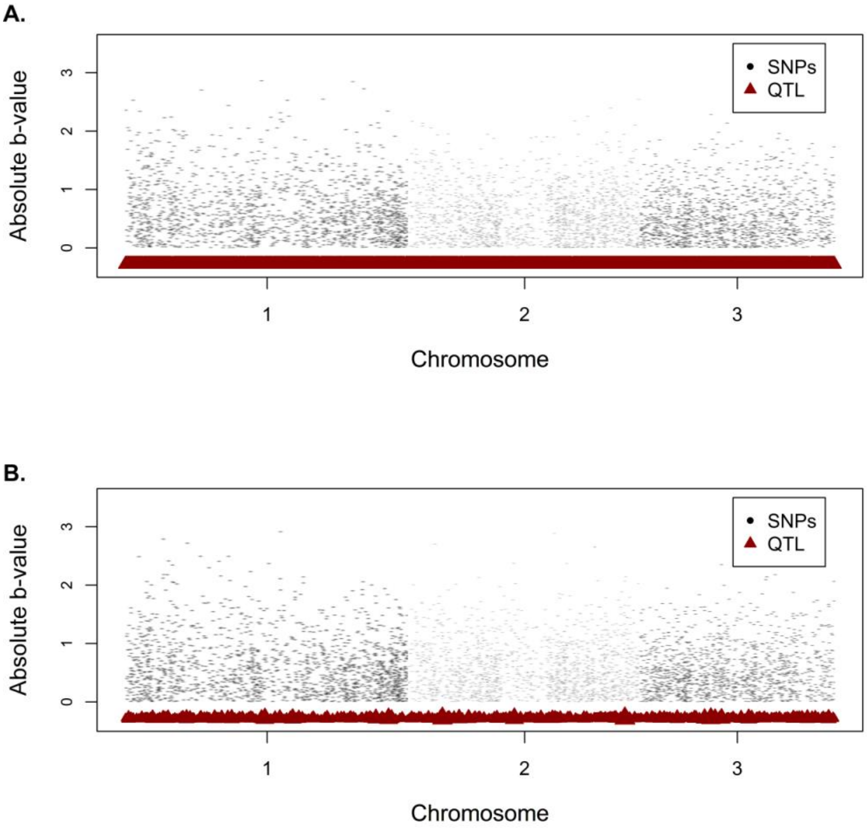


Figure A4.3 Absolute estimated regression coefficients (b-values) for each SNP to predict the QTL genotypes of 3000 randomly selected QTL.Absolute regression coefficients for each of the SNPs estimated in a Holstein Friesian reference population (b_{RP}) to predict the QTL genotypes of 3000 randomly selected QTL with (A) equal weight for each of the QTL, or (B) QTL weighted differently, based on their allele substitution effects, in the overall breeding goal. The size of the triangle represents the weight of the QTL in the overall breeding goal of the selection index calculations, i.e., the allele substitution effect in (B).

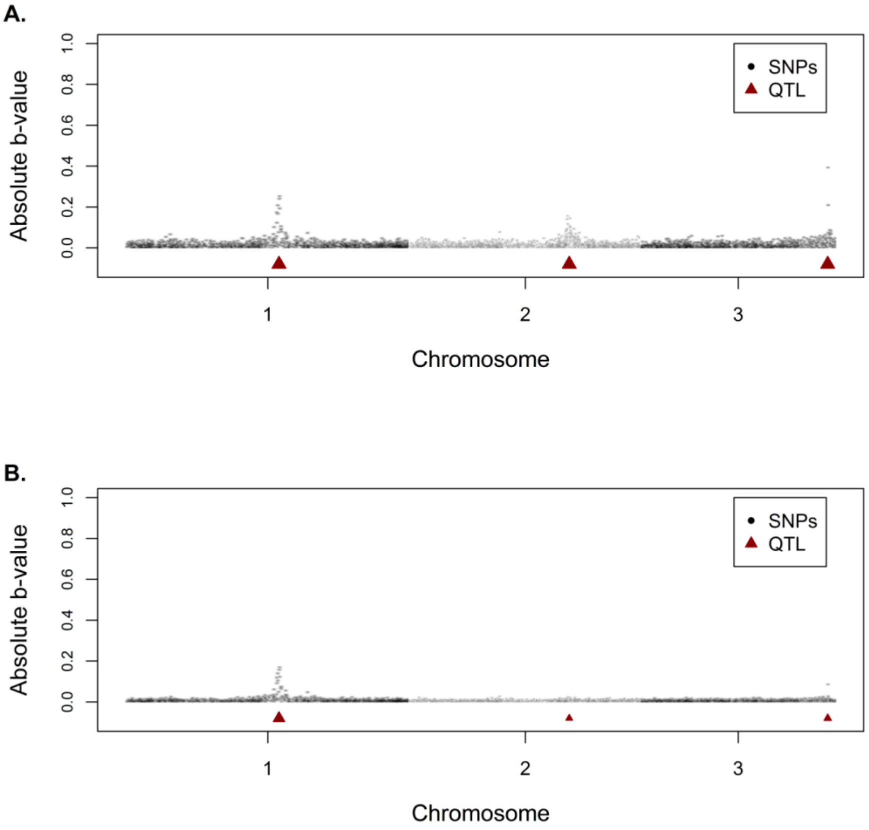


Figure A4.4 Absolute estimated regression coefficients (b-values) for each SNP to predict the QTL genotypes of 3 QTL with a low MAF. Absolute regression coefficients for each of the SNPs estimated in a Holstein Friesian reference population (b_{RP}) to predict the QTL genotypes of 3 QTL with a low MAF with (A) equal weight for each of the QTL, or (B) QTL weighted differently, based on their allele substitution effects, in the overall breeding goal. The size of the triangle represents the weight of the QTL in the overall breeding goal of the selection index calculations, i.e., the allele substitution effect in (B).

4.8 References

Abasht, B., E. Sandford, J. Arango, P. Settar, J. E. Fulton*, et al.*, 2009 Extent and consistency of linkage disequilibrium and identification of DNA markers for production and egg quality traits in commercial layer chicken populations. BMC Genom. 10: S2.

Andreescu, C., S. Avendano, S. R. Brown, A. Hassen, S. J. Lamont*, et al.*, 2007 Linkage disequilibrium in related breeding lines of chickens. Genetics 177: 2161-2169.

Browning, B. L. and S. R. Browning, 2009 A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am. J. Hum. Genet. 84: 210-223.

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

Calus, M. P. L., H. Huang, A. Vereijken, J. Visscher, J. Ten Napel*, et al.*, 2014 Genomic prediction based on data from three layer lines: a comparison between linear methods. Genet. Sel. Evol. 46: 57.

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., M. P. L. Calus, R. Pong-Wong, G. De los Campos and J. M. Hickey, 2013 Genomic prediction in animals and plants: Simulation of data, validation, reporting, and benchmarking. Genetics 193: 347-365.

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: 1545-1553.

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

Falconer, D. S. and T. F. C. Mackay, 1996 *Introduction to quantitative genetics*. Pearson Education Limited, Harlow.

Gautier, M., T. Faraut, K. Moazami-Goudarzi, V. Navratil, M. Foglio*, et al.*, 2007 Genetic and haplotypic structure in 14 European and African cattle breeds. Genetics 177: 1059-1070.

Gilmour, A. R., B. Gogel, B. Cullis, R. Thompson, D. Butler*, et al.*, 2009 *ASReml user guide release 3.0*. VSN International Ltd, Hemel Hempstead.

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

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., 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.

Hayes, B. J. and M. E. Goddard, 2001 The distribution of the effects of genes affecting quantitative traits in livestock. Genet. Sel. Evol. 33: 209-229.

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

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

Hazel, L. N. and J. L. Lush, 1942 The efficiency of three methods of selection. J. Hered. 33: 393-399.

Hazel, L. N., 1943 The genetic basis for constructing selection indexes. Genetics 28: 476-490.

Heifetz, E. M., J. E. Fulton, N. O'Sullivan, H. Zhao, J. C. M. Dekkers*, et al.*, 2005 Extent and consistency across generations of linkage disequilibrium in commercial layer chicken breeding populations. Genetics 171: 1173-1181.

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

Jorjani, H., L. Klei and U. Emanuelson, 2003 A simple method for weighted bending of genetic (co)variance matrices. J. Dairy Sci. 86: 677-679.

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

Kemper, K. E. and M. E. Goddard, 2012 Understanding and predicting complex traits: Knowledge from cattle. Hum. Mol. Genet. 21: R45-R51.

Kempthorne, O. and A. W. Nordskog, 1959 Restricted selection indices. Biometrics 15: 10-19.

Lin, C. Y., 1978 Index selection for genetic improvement of quantitative characters. Theor. Appl. Genet. 52: 49-56.

Matukumalli, L. K., C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan*, et al.*, 2009 Development and characterization of a high density SNP genotyping assay for cattle. PLoS ONE 4: e5350.

Meuwissen, T. H. E. and M. E. Goddard, 2000 Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. Genetics 155: 421-430.

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.

Pryce, J. E., B. Gredler, S. Bolormaa, P. J. Bowman, C. Egger-Danner*, et al.*, 2011 Short communication: Genomic selection using a multi-breed, across-country reference population. J. Dairy Sci. 94: 2625-2630.

Pryce, J. E., J. Johnston, B. J. Hayes, G. Sahana, K. A. Weigel*, et al.*, 2014 Imputation of genotypes from low density (50,000 markers) to high density (700,000 markers) of cows from research herds in Europe, North America, and Australasia using 2 reference populations. J. Dairy Sci. 97: 1799-1811.

Sawyer, S. L., N. Mukherjee, A. J. Pakstis, L. Feuk, J. R. Kidd*, et al.*, 2005 Linkage disequilibrium patterns vary substantially among populations. Europ. J. Hum. Genet. 13: 677-686.

Smith, H. F., 1936 A discriminant function for plant selection. Ann. Eugen. 7: 240-250.

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

Veroneze, R., P. S. Lopes, S. E. F. Guimarães, F. F. Silva, M. S. Lopes*, et al.*, 2013 Linkage disequilibrium and haplotype block structure in six commercial pig lines. J. Anim. Sci. 91: 3493-3501.

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.

Wientjes, Y. C. J., M. P. L. Calus, M. E. Goddard and B. J. Hayes, 2015a Impact of QTL properties on the accuracy of multi-breed genomic prediction. Genet. Sel. Evol. 47: 42.

Wientjes, Y. C. J., R. F. Veerkamp, P. Bijma, H. Bovenhuis, C. Schrooten*, et al.*, 2015b Empirical and deterministic accuracies of across-population genomic prediction. Genet. Sel. Evol. 47: 5.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders*, et al.*, 2010 Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42: 565-569.

Zhou, L., X. Ding, Q. Zhang, Y. Wang, M. S. Lund*, et al.*, 2013 Consistency of linkage disequilibrium between Chinese and Nordic Holsteins and genomic prediction for Chinese Holsteins using a joint reference population. Genet. Sel. Evol. 45: 7.

CHAPTER 5

IMPACT OF QTL PROPERTIES ON THE ACCURACY OF MULTI-BREED GENOMIC PREDICTION

Y.C.J. WIENTJES[1,2]

M.P.L. CALUS[1]

M.E. GODDARD[3,4,5]

B.J. HAYES[3,5,6]

[1] ANIMAL BREEDING AND GENOMICS CENTRE,
WAGENINGEN UR LIVESTOCK RESEARCH,
6700 AH WAGENINGEN, THE NETHERLANDS
[2] ANIMAL BREEDING AND GENOMICS CENTRE,
WAGENINGEN UNIVERSITY,
6700 AH WAGENINGEN, THE NETHERLANDS
[3] DEPARTMENT OF ENVIRONMENT AND PRIMARY INDUSTRIES,
AGRIBIO, LA TROBE UNIVERSITY, VICTORIA 3083, AUSTRALIA
[4] FACULTY OF LAND AND ENVIRONMENT,
UNIVERSITY OF MELBOURNE, VICTORIA 3010, AUSTRALIA
[5] DAIRY FUTURES COOPERATIVE RESEARCH CENTRE,
LA TROBE UNIVERSITY, VICTORIA 3083, AUSTRALIA;
[6] LA TROBE UNIVERSITY, VICTORIA 3083, AUSTRALIA

GENETIC SELECTION EVOLUTION (2015) 47:42

Abstract

Background: Although simulation studies show that combining multiple breeds in one reference population increases accuracy of genomic prediction, this is not always confirmed in empirical studies. This discrepancy might be due to the assumptions on quantitative trait loci (QTL) properties applied in simulation studies, including number of QTL, spectrum of QTL allele frequencies across breeds, and distribution of allele substitution effects. We investigated the effects of QTL properties and of including a random across- and within-breed animal effect in a genomic best linear unbiased prediction (GBLUP) model on accuracy of multi-breed genomic prediction using genotypes of Holstein Friesian and Jersey cows.

Methods: Genotypes of three classes of variants obtained from whole-genome sequence data, with moderately low, very low or extremely low average minor allele frequencies (MAF), were imputed in 3000 Holstein Friesian and 3000 Jersey cows that had real high-density genotypes. Phenotypes of traits controlled by QTL with different properties were simulated by sampling 100 or 1000 QTL from one class of variants and their allele substitution effects either randomly from a gamma distribution, or computed such that each QTL explained the same variance, i.e., rare alleles had a large effect. Genomic breeding values for 1000 selection candidates per breed were estimated using GBLUP models including a random across- and a within-breed animal effect.

Results: For all three classes of QTL allele frequency spectra, accuracies of genomic prediction were not affected by the addition of 2000 individuals of the other breed to a reference population of the same breed as the selection candidates. Accuracies of both single- and multi-breed genomic prediction decreased as MAF of QTL decreased, especially when rare alleles had a large effect. Accuracies of genomic prediction were similar for the models with and without a random within-breed animal effect, probably because of insufficient power to separate across- and within-breed animal effects.

Conclusions: Accuracy of both single- and multi-breed genomic prediction depends on the properties of the QTL that underlie the trait. As QTL MAF decreased, accuracy decreased, especially when rare alleles had a large effect. This demonstrates that QTL properties are key parameters that determine the accuracy of genomic prediction.

Key words: accuracy, multi-breed genomic prediction, allele frequency, QTL property, allele substitution effect

5.1 Background

In livestock breeding programs, genomic information is widely used to estimate genomic breeding values for selection candidates. Genomic estimated breeding values (GEBV) are calculated from marker effects estimated in a reference population that consists of animals with phenotypes and marker genotypes. Accuracy of GEBV for selection candidates, that typically have no phenotypes of their own, depends on the size of the reference population i.e., the larger the size of the reference population, the more accurately breeding values can be predicted (e.g., Meuwissen *et al.* 2001; Daetwyler *et al.* 2008; VanRaden *et al.* 2009). For numerically small breeds, assembling such a large reference population is challenging, therefore, an attractive approach would be to combine purebred reference populations from different breeds or lines to establish large reference populations (De Roos *et al.* 2009; Zhong *et al.* 2009; Erbe *et al.* 2012; Simeone *et al.* 2012). However, the benefit of adding another breed or line to the reference population may be reduced by the inconsistency in allele substitution effects across breeds (Spelman *et al.* 2002; Thaller *et al.* 2003; Wientjes *et al.* 2015), by between-breed differences in linkage disequilibrium (LD) between single-nucleotide polymorphisms (SNPs) and quantitative trait loci (QTL) that influence a trait across breeds or lines (e.g., De Roos *et al.* 2008; Zhong *et al.* 2009; Pryce *et al.* 2011), as well as by the absence of close family relationships between breeds or lines (Wientjes *et al.* 2013). In addition, the accuracy of prediction using both single-breed and multi-breed reference populations may be affected by the properties of the QTL that control a trait, i.e., number of QTL for the trait, joint distribution of QTL allele frequencies across breeds, and distribution of QTL effects.

In *Bos taurus* cattle populations, LD phase is conserved across breeds among SNP alleles at short distances (5 to 30 kb) (De Roos *et al.* 2008). Therefore, a high marker density might overcome the problem of differences in LD between SNPs and QTL across breeds or lines (De Roos *et al.* 2008). Indeed, simulation studies using high-density markers showed that prediction accuracy increased when reference populations were combined across breeds compared to single-breed reference populations (De Roos *et al.* 2009; Ibánẽz-Escriche *et al.* 2009). However, in empirical studies, the increase in prediction accuracy was smaller and sometimes absent (Hayes *et al.* 2009; Pryce *et al.* 2011; Calus *et al.* 2014), even when more than 600,000 SNPs were used (Harris *et al.* 2011; Erbe *et al.* 2012; Bolormaa *et al.* 2013). Part of this difference between accuracies obtained from simulation and empirical studies could be explained by the assumptions made in simulation studies on the properties of the QTL that underlie a trait, which may not completely reflect the reality. One of these QTL properties that could affect prediction accuracy is the pattern of QTL allele frequencies. For most complex traits, the QTL that underlie a trait have a low minor allele frequency (MAF) (Goddard and Hayes 2009; Yang *et al.* 2010; Kemper and Goddard 2012). Due to ascertainment bias of SNP chips (Matukumalli *et al.* 2009), SNPs tend to have higher MAF than QTL, which reduces the LD between QTL and SNPs and therefore the accuracy of genomic prediction, particularly across breeds and lines. Besides differences in allele frequencies between SNPs and QTL, differences in allele frequencies of QTL across breeds may also influence prediction accuracy. In extreme cases, QTL may even only segregate in one of the breeds. When the SNPs that flank a breed-specific QTL are segregating across breeds, the apparent effect of SNPs may vary across breeds. The above examples show that the properties of QTL that underlie a trait are likely to affect the accuracy of multi-breed or line genomic prediction.

In spite of potential differences in QTL properties across breeds, most studies on multi-breed genomic prediction estimate only one effect for each SNP across all breeds, (e.g., Hayes *et al.* 2009; Brøndum *et al.* 2011; Erbe *et al.* 2012). Olson *et al.* (2012) and Makgahlela *et al.* (2013) accounted for differences in SNP effects across breeds by fitting a multi-trait model in which the same trait in different breeds was treated as a different trait and both studies showed a minor increase in prediction accuracy using ~40,000 SNPs. Another way to account for breed-specific SNP effects and at the same time benefit from increasing the size of the reference population by adding another breed could be to estimate an across-breed SNP effect and a within-breed SNP effect. Khansefid *et al.* (2014) showed that this can be done by including a random across-breed animal effect and a within-breed animal effect in a genomic best linear unbiased prediction (GBLUP) model.

The first objective of this study was to investigate the effect of the properties of the QTL that underlie the trait on the accuracy of multi-breed genomic prediction. The second objective was to investigate the effect of a GBLUP model with a random across-breed animal and a within-breed animal effect on the accuracy of multi-breed genomic prediction. In this study, real genotypes of Holstein Friesian and Jersey dairy cows were used. Phenotypes were simulated using different properties of QTL by sampling 100 or 1000 QTL from three different classes of markers with average MAF that ranged from moderately low (representing allele frequencies expected under a neutral model) to extremely low values, and by simulating allele substitution effects using two different models.

5.2 Methods

For this study, two different datasets were used. For the first dataset, including genotypes of Australian cows, samples were collected for DNA extraction as approved by the Department of Primary Industries Victoria Animals Ethics Committee (protocol: 2010-19). For the second dataset, sequence information from the 1000 bull genomes project was used, for which DNA for most animals was extracted from semen. Only for Angus animals, samples were collected for DNA extraction as approved by the New South Wales Department of Primary Industries Animals Ethics Committee.

5.2.1 Genotypes

Genotypes were available for 3000 Holstein Friesian cows and 3000 Jersey cows from Australia. Individuals were genotyped with the Illumina BovineHD Beadchip (777k, Illumina, San Diego, CA) or the Illumina BovineSNP50 Beadchip (50k, Illumina, San Diego, CA). Animals genotyped at the lower density (50k) were imputed to high-density (777k) using the software package Beagle 3.0 (Browning and Browning 2009) and a reference population of 1072 animals (Holstein Friesian and Jersey) that were genotyped with the 777k chip. Quality was checked using a larger dataset that included those 6000 individuals. SNPs of low quality based on the same criteria as described in Erbe *et al.* (2012) were removed, leaving 606,384 SNPs for the analyses.

In order to obtain plausible QTL allele frequencies that ranged from frequencies of loci that are effectively neutral to frequencies of loci that are expected to have large pleiotropic effects on fitness, sequence data of variants in annotated classes from the 1000 bull genomes project (Daetwyler *et al.* 2014) were used. This included sequence information of 129 Holstein Friesian, 15 Jersey, 47 Angus and 43 Simmental animals. Variants in this dataset were annotated as either synonymous mutations (80,515 mutations), missense mutations (97,296 mutations), and premature stop codon mutations (4064 mutations), with about the same number of variants in each class as presented in Daetwyler *et al.* (2014). More information about the samples, alignment, variant calling and filtering, and annotation of the sequenced animal genomes is in Daetwyler *et al.* (2014).

Table 5.1 Characteristics of different classes of variants used to simulate QTL.

Characteristic per class	Holstein Friesian	Jersey	Total
Moderately low average MAF[a]			
Segregating variants	63,119	55,363	65,920
Number of breed-specific variants	10,557	2801	13,358
Percentage of breed-specific variants	16.0	4.2	20.3
Average MAF of the 65,920 segregating variants (± standard deviation)	0.130 ± 0.169	0.115 ± 0.168	0.122 ± 0.146
Very low average MAF[b]			
Segregating variants	61,302	49,473	67,097
Number of breed-specific variants	17,624	5795	23,419
Percentage of breed-specific variants	26.3	8.6	34.9
Average MAF of the 67,097 segregating variants (± standard deviation)	0.082 ± 0.146	0.072 ± 0.142	0.077 ± 0.127
Extremely low average MAF[c]			
Segregating variants	1804	1245	2142
Number of breed-specific variants	897	338	1235
Percentage of breed-specific variants	41.9	15.8	57.7
Average MAF of the 2142 segregating variants (± standard deviation)	0.017 ± 0.067	0.015 ± 0.066	0.016 ± 0.059

[a] annotated as synonymous mutations;
[b] annotated as missense mutations;
[c] annotated as premature stop codon mutations;
MAF = minor allele frequency.

Our aim was to simulate different groups of QTL that had decreasing MAF and that were increasingly more difficult to tag with SNPs on the SNP chip and were equally distributed across the whole genome. Therefore, the three classes of annotated variants that varied in average MAF (Table 5.1) and MAF pattern (see Appendix Figure A5.1 and Figure A5.2), were used to represent different patterns of QTL MAF; the synonymous mutations represented QTL with on average a moderately low MAF (average MAF of 0.122), the missense mutations represented QTL with on average a very low MAF (average MAF of 0.077), and the premature stop codon mutations represented QTL with on average an extremely low MAF (average MAF of 0.016). It should be noted that these classes of variants were only used to represent differences in patterns of QTL MAF and not differences in biological functions of the QTL.

Genotypes for the three classes of variants were imputed in 3000 Holstein Friesian and 3000 Jersey animals with real high-density SNP genotypes (Browning and Browning 2009). Imputation was done using all sequenced animals from the reference population, which included the Angus and Simmental animals, since it has been shown that using animals from other breeds improves imputation accuracy (Brøndum *et al.* 2012; Daetwyler *et al.* 2014). Allele frequency patterns of the imputed variants were similar to the allele frequency patterns in sequenced animals (see Appendix Figure A5.1 and Figure A5.2). Other characteristics of the three classes of imputed variants are shown in Table 5.1. For imputed and real sequence data, the number of segregating variants was much smaller for the Jersey population than for the Holstein Friesian population. This is probably due to the small number of Jersey sequenced genomes in the dataset, since more polymorphic SNPs are detected when the group of genotyped individuals is larger (Li and Leal 2009; The International HapMap 3 Consortium 2010; Jansen *et al.* 2013). Reliabilities (i.e., R^2 values) of imputation were low (average reliabilities estimated by Beagle were equal to 0.67 for variants with on average a moderately low MAF, 0.51 for variants with on average a very low MAF, and 0.32 for variants with on average an extremely low MAF), which probably results from the relatively small number of animals with sequence data in combination with the low MAF of the variants. This decrease in reliabilities of imputation as average MAF of variants decreases confirms the assumption that LD between variants with a low MAF and neighboring SNPs on the commercial SNP chip decreases, i.e., that tagging the variants with SNPs on the chip was increasingly more difficult.

5.2.2 Simulation of phenotypes

Traits that were controlled by QTL with different properties were simulated by varying: 1) the average MAF of the QTL that underlie the trait, by sampling QTL from one of the three classes described above, 2) the number of QTL that underlie the trait, and 3) the distribution of allele substitution effects. In each simulation, 100 or 1000 QTL were sampled assuming that they followed one of the three QTL MAF patterns i.e., moderately low average MAF, very low average MAF, or extremely low average MAF. All variants that segregated in the entire dataset, consisting of 3000 Holstein Friesian and 3000 Jersey individuals, were considered as potential QTL, which resulted in 65,920 potential QTL with a moderately low average MAF, 67,097 with a very low average MAF, and 2142 with an extremely low average MAF. It should be noted that the percentage of breed-specific variants increased as the MAF of the variants decreased (Table 5.1).

Allele substitution effects were sampled using two different models: 1) a pseudo-infinitesimal model, where small allele substitution effects were randomly assigned to QTL independently of allele frequency (RANDOM model), and 2) a ‘rare allele, large effect’ model, where larger allele substitution effects were assigned to QTL with a lower MAF such that each QTL explained an equal amount of the total genetic variance (VAR model). Under the RANDOM model, allele substitution effects were randomly sampled from a gamma distribution with a shape parameter of 0.4 and a scale parameter of 1.66, following Meuwissen *et al.* (2001). Under the VAR model, the variance explained by each QTL was kept constant across all QTL by computing allele substitution effects as $a = \sqrt{\frac{Var(QTL)}{2p(1-p)}}$, where a is the allele substitution effect assuming a purely additive model, $Var(QTL)$ is the variance of the QTL which is constant across the QTL and was set to 1, and p is the allele frequency of the QTL across all 6000 individuals (3000 Holstein Friesian and 3000 Jersey cows). Under the two models, both alleles at a given QTL had an equal chance to have a positive or a negative effect on the simulated trait and the effect was the same in both breeds. The simulated allele substitution effects were multiplied by the genotype codes (0, 1, or 2) to calculate a true breeding value (TBV) for each individual. Over all individuals and across the breeds, TBV were rescaled to a mean of 0 and a variance of 1.

Allele frequencies for the loci selected as QTL differed between the two breeds (see Appendix Figure A5.3). These differences in allele frequencies resulted in differences in average TBV between breeds. To calculate the genetic variance as the variance across TBV, breed effects were first subtracted from all TBV to avoid breed effects influencing the simulated heritability. Thereafter, the environmental effect per individual was sampled from a normal distribution with a mean of 0 and variance $\left(\frac{1}{h^2}-1\right)$*(variance of TBV corrected for breed effect). For each individual, the phenotype was calculated as the sum of its TBV, including its breed effect and the randomly sampled environmental effect.

In this study, a rather simple situation was simulated to be able to investigate the effect of QTL properties on the accuracy of both single- and multi-breed genomic prediction. Heritabilities and allele substitution effects were assumed to be the same across breeds, such that phenotypic differences between breeds were only due to differences in QTL allele frequencies. Phenotypes were simulated using a heritability of 0.8, which is similar to the heritability of daughter yield deviation of a bull for milk yield if the bull has approximately 100 daughters. We chose this rather high heritability value to achieve high accuracies of genomic prediction, which resulted in more pronounced differences in accuracies between the different scenarios for the small reference population size used in the simulations. According to the formula of Daetwyler *et al.* (2008; 2010), a trait with a heritability of 0.8 is expected to yield the same accuracy as a trait with a heritability of 0.25 but using a reference population that includes 3.2 times more animals.

To decide on the number of replicates, the variance of the squared accuracy (r^2) was calculated from the sampling variance of a correlation coefficient as (Fisher 1954):

$$Var(r^2) = \frac{(1-r^2)^2}{N-1}, \tag{5.1}$$

where *N* is the number of selection candidates. Thereafter, the required number of replicates (*n*) was calculated as (Ott and Longnecker 2001):

$$n > \frac{\left(1.96^2 * Var(r^2)\right)}{0.02^2}, \tag{5.2}$$

where 1.96 refers to the z-value on the standard normal distribution relating to a confidence interval of 95%, and 0.02 is the maximum allowable difference between the estimated and true mean. This resulted in a maximum required number of replicates of 9.62 with an actual accuracy of 0, and a minimum required number of replicates of 0.004 with an actual accuracy of -0.99 or 0.99. Thus, 10 replicates are sufficient to cover the whole spectrum of possible accuracies.

5.2.3 Investigating the accuracy of genomic prediction

For each replicate, the accuracy of genomic prediction was empirically calculated for a fixed group of 1000 Holstein Friesian and 1000 Jersey selection candidates that were selected from the 3000 animals per breed that were used in this study. Due to the presence of overlapping generations and the use of cow data with small progeny groups, selection candidates were randomly sampled from the full dataset. The other 2000 Holstein Friesian and 2000 Jersey cows were used as reference animals in seven reference populations (Table 5.2), with different numbers of Holstein Friesian and Jersey individuals that ranged from a single-breed reference population to a multi-breed reference population with equal numbers of animals of both breeds. Each of the smaller reference populations was a random subset from the larger reference populations.

Table 5.2 Overview of the different reference populations.

	Reference population	
Scenarios	**Number of Holstein Friesian**	**Number of Jersey**
1	2000	2000
2	2000	500
3	2000	100
4	2000	0
5	500	2000
6	100	2000
7	0	2000

Since LD pattern between QTL and SNPs differed across breeds and some QTL segregated only in one of the breeds, SNP effects were expected to differ across breeds. To account for these differences in SNP effects, a Genomic-relatedness-matrix Residual Maximum Likelihood model (GREML) including both a random across-breed animal effect and a within-breed animal effect was run in ASReml (Gilmour *et al.* 2009). A GREML model has the same features as the commonly known GBLUP model (assuming a normal distribution of SNP effects), but it estimates the variances and the breeding values simultaneously using REML. This was done using the following model, hereafter called the base model:

$$\mathbf{y} = \mathbf{1}_n \mu + \mathbf{Z}\mathbf{g}_a + \mathbf{Z}\mathbf{g}_w + \mathbf{e}, \qquad (5.3)$$

where **y** is a vector containing the simulated phenotypes, $\mathbf{1}_n$ is a vector of ones, μ is the overall mean across breeds, $\mathbf{g}_a$ and $\mathbf{g}_w$ are vectors of the genomic breeding values predicted either across-breeds or within-breeds ($\mathbf{g}_a \sim N(0,\mathbf{G}_a\sigma^2_{g_a})$ and $\mathbf{g}_w \sim N(0,\mathbf{G}_w\sigma^2_{g_w})$), **Z** is an incidence matrix that allocates genomic breeding values (both $\mathbf{g}_a$ and $\mathbf{g}_w$) to the individuals and **e** is a vector containing the residuals $\sim N(0,\mathbf{I}\sigma^2_e)$. Note that only one $\sigma^2_{g_a}$ and one $\sigma^2_{g_w}$ was estimated, which reflect the variances in the base population of the genomic relationship matrices ($\mathbf{G}_a$ and $\mathbf{G}_w$), which was set to be the population immediately before Holstein Friesian and Jersey breeds diverged by using the method of Erbe *et al.* (2012). As a first step to calculate $\mathbf{G}_a$ and $\mathbf{G}_w$, the **G** matrix was calculated as (Erbe *et al.* 2012):

$$\mathbf{G} = \frac{\mathbf{WW}'}{2\sum_{j=1}^{n} p_j(1-p_j)}, \qquad (5.4)$$

where *n* is the number of loci, **W** is a matrix of standardized genotypes for individual *i* at locus *j* calculated as $w_{ij} = g_{ij} - 2p_j$, where g_{ij} codes the genotype as 0, 1 and 2, and p_j is the allele frequency for the second allele (for which the homozygote genotype is coded 2) calculated as $p_j = \alpha p_{j,HF} + (1-\alpha)p_{j,Jer}$. In this last equation, $p_{j,HF}$ is the allele frequency in the Holstein Friesian population, $p_{j,Jer}$ is the allele frequency in the Jersey population and α is calculated as $\alpha = \frac{F_{Jer}}{F_{Jer} + F_{HF}}$, and represents the proportion of Holstein Friesian haplotypes in the ancestral population. The inbreeding coefficient for the Jersey population was calculated as:

$$F_{Jer} = 1 - \frac{\sum_{j=1}^{n} 2p_{j,Jer}(1-p_{j,Jer})}{\sum_{j=1}^{n} \left[p_{j,HF}(1-p_{j,Jer}) + p_{j,Jer}(1-p_{j,HF})\right]} . \qquad (5.5)$$

The inbreeding coefficient for the Holstein Friesian population was calculated in the same way by substituting the two breeds accordingly. As described by Erbe *et al.* (2012), inbreeding in **G** can be adjusted for the inbreeding that occurred relative to the base set at the time of divergence of the two breeds as $\mathbf{G}^* = \mathbf{G}(1-F) + 2F$. In this equation, *F* is the inbreeding relative to an F1 base population calculated as $F = \frac{F_{Jer}F_{HF}}{F_{Jer} + F_{HF}}$. The relationship matrix based on the pedigree, **A**, was rescaled to the same base by rescaling the within-Holstein Friesian block as $\mathbf{A}^* = \mathbf{A}[1-(F_{HF} - f_{HF})] + 2(F_{HF} - f_{HF})$, in which f_{HF} is the amount of inbreeding in the Holstein Friesian population since the base of the pedigree. The within-Jersey block was rescaled in the same way and the across-breed block was set to 0. Thereafter, the rescaled $\mathbf{G}^*$ matrix was regressed back to the rescaled $\mathbf{A}^*$ matrix following Yang *et al.* (2010) and Goddard *et al.* (2011) to calculate $\mathbf{G}_a$. The regression was done separately across- and within-breed as well as per bin of pedigree relationship (< 0.10, 0.10-0.25, 0.25-0.50, >0.5), because the sampling error on elements of $\mathbf{G}^*$ depends on the level of family relationships. Across these bins of relationships, the different regression coefficients ranged from 0.994 to 0.999 when all 606,384 SNPs were used to calculate $\mathbf{G}_a$. The $\mathbf{G}_w$ matrix was formed from the $\mathbf{G}_a$ matrix by setting the elements between individuals of different breeds to zero, while the within-breed elements of $\mathbf{G}_w$ were equal to the corresponding elements in $\mathbf{G}_a$.

In this base model, genomic breeding values were predicted across breeds as well as within breeds. For each selection candidate, the genomic breeding values across- and within-breed were summed to calculate the total genomic breeding value. The accuracy of genomic prediction was calculated per breed as the correlation between the total genomic breeding values and the simulated true breeding values of all selection candidates of that breed.

Analyses were performed using different numbers of SNPs to set-up $\mathbf{G}_a$ and $\mathbf{G}_w$, namely: 1) 606,384 SNPs, 2) 60,000 SNPs, 3) 606,384 SNPs plus the genotypes of all imputed variants representing QTL, and 4) 60,000 SNPs plus the genotypes of all imputed variants representing QTL. The 60,000 SNPs were randomly selected from the 606,384 SNPs to study the accuracy that could be achieved with a lower marker density. When genotypes for the imputed variants representing QTL were included in the dataset used to calculate $\mathbf{G}_a$ and $\mathbf{G}_w$, genotypes of all imputed variants in the three classes were used i.e., 80,515 variants with a moderately low average MAF, 97,296 with a very low average MAF and 4064 with an extremely low average MAF. In this way, the potential accuracy of genomic prediction was studied when the causal mutations, i.e., the QTL, were included in the marker dataset.

The power of the base model to separate across- and within-breed animal effects was investigated for one of the scenarios, namely the RANDOM scenario with 1000 QTL and 2000 Holstein Friesian and 2000 Jersey animals in the reference population. Due to computational reasons, only one of the scenarios was investigated. The base model that included a random across-breed animal effect and a within-breed animal effect, was run once for each specific replicate in this scenario and the total genetic variance was calculated. Thereafter, the model was run again by fixing the within-breed variance to 1, 10, 20, 30, 40, 50, 60, 70, 80, 90, 99% of the total genetic variance and assigning the remaining part to the across-breed variance.

To test for significance, twice the difference in log-likelihood between the model with fixed variance components and the model with estimated variance components was compared with the 5% significance threshold (2.71) taken from a mixed Chi-square distribution with 0 and 1 degrees of freedom.

To investigate the advantage in terms of prediction accuracy of using a GBLUP type of model with a random across-breed animal effect and a within-breed animal effect over a model with only a random across-breed animal effect, the analyses were repeated using a model where $\mathbf{Zg}_w$ was removed. The effect of a fixed breed effect on accuracy of multi-breed genomic prediction was also studied by running the base model including breed as a fixed effect. Both alternative models were run for the RANDOM and VAR scenario using all reference populations when 100 QTL controlled the trait.

5.3 Results

The results presented in this section are the averages across the 10 replicates, with standard errors computed across the replicates. In general, the standard errors across replicates were small. To further investigate if 10 replicates were sufficient for this study, the impact of the number of replicates was analyzed by comparing the averages after 10 replicates with the averages after the first five replicates. In general, the absolute difference in accuracy was only ~0.01 between the averages after five and 10 replicates for all scenarios using the base model and QTL with a moderately low average MAF, very low average MAF or extremely low average MAF. Standard errors were, as expected, slightly higher with five replicates. The low standard errors and the small differences in averages after five and 10 replicates indicate that using only 10 replicates did not affect the conclusions of our study.

5.3.1 QTL properties

Average accuracies for the base model using all 606,384 SNPs for the different reference populations are shown in Figure 5.1 when 100 QTL controlled the simulated trait, both for the RANDOM (A) and VAR (B) scenarios. For all reference populations, accuracies were greater for the RANDOM scenario than for the VAR scenario, regardless of the average MAF of QTL. Moreover, accuracies were slightly greater for Jersey selection candidates than for Holstein Friesian selection candidates when the number of individuals in the reference population from the evaluated breed was the same, which reflects the smaller effective population size of this breed.

As the number of reference individuals of a breed decreased, the achieved prediction accuracies for the selection candidates from the same breed decreased as expected for all scenarios. For the RANDOM scenario, prediction accuracy decreased by ~0.51 for the Jersey and ~0.01 for the Holstein Friesian selection candidates when the number of Jersey individuals changed from 2000 to 0 in the reference population, and it decreased by ~0.01 for the Jersey and ~0.50 for the Holstein Friesian selection candidates when the number of Holstein Friesian individuals changed from 2000 to 0 in the reference population. For the VAR scenario, the decrease in accuracy due to a decreasing number of animals from the breed itself was also large, although this decrease was less pronounced due to smaller accuracies, and the decrease in accuracy due a decreasing number of animals from the other breed was negligible. Thus, the effect of including another breed in the reference population on prediction accuracy was small for both scenarios.

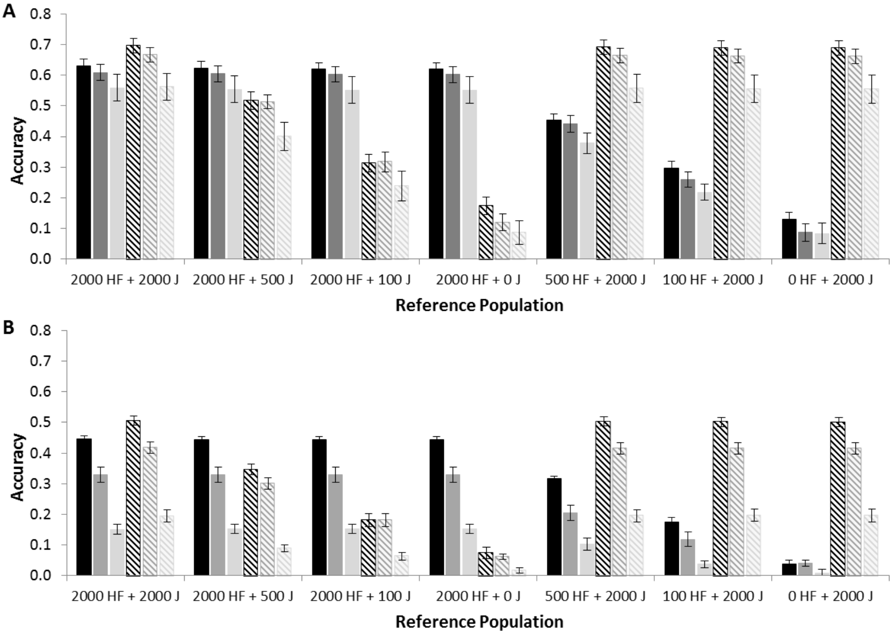


Figure 5.1 Accuracies of genomic prediction for traits that are controlled by QTL with different properties when 100 QTL underlie the trait. Average accuracies of genomic prediction (± standard errors) for Holstein Friesian (HF, solid fill) and Jersey (J, diagonal fill) animals using a model that included a random across-breed animal effect and a within-breed animal effect, 606,384 SNPs, seven different reference populations and using simulated allele substitution effects (A) randomly sampled from a gamma distribution or (B) with each QTL explaining an equal proportion of the genetic variance, when 100 QTL underlying the trait were sampled from variants with on average a moderately low allele frequency (black), very low minor allele frequency (dark grey) or extremely low minor allele frequency (light grey).

In general, accuracies were greatest for QTL with a moderately low average MAF and smallest for QTL with an extremely low average MAF. The differences in accuracies between classes of QTL with different average MAF were more pronounced for the VAR scenario than for the RANDOM scenario, mainly as a result of a smaller accuracy for QTL with a very low average MAF and a much smaller accuracy for QTL with an extremely low average MAF. These results are consistent with the estimated heritabilities for each scenario (Table 5.3); estimated heritabilities decreased when the average MAF of QTL decreased and the differences were more pronounced for the VAR scenario than for the RANDOM scenario. For all scenarios, the estimated heritability was below the simulated heritability, but for the RANDOM scenario, the differences were small. This indicates that it was difficult for the GBLUP model to capture all the genetic variance when the QTL that underlie the simulated trait had on average a low MAF, especially when rare alleles had a large effect.

Table 5.3 Average estimated heritabilities of QTL with different properties. Average heritabilities (standard errors across replicates) estimated with a model including a random across-breed animal effect and a within-breed animal effect and using 606,384 SNPs to calculate the genomic relationship matrix using different reference populations, different average minor allele frequencies (MAF) of the 100 QTL that underlie the trait and using simulated allele substitution effects randomly sampled from a gamma distribution (RANDOM) or with each QTL explaining an equal proportion of the genetic variance (VAR).

			RANDOM			VAR		
Scen.	**Nb HF**	**Nb J**	**Moderately low MAF**	**Very low MAF**	**Extremely low MAF**	**Moderately low MAF**	**Very low MAF**	**Extremely low MAF**
1	**2000**	**2000**	0.78 (0.003)	0.77 (0.002)	0.72 (0.011)	0.60 (0.001)	0.44 (0.002)	0.21 (0.002)
2	**2000**	**500**	0.76 (0.004)	0.75 (0.006)	0.70 (0.023)	0.54 (0.002)	0.38 (0.004)	0.18 (0.001)
3	**2000**	**100**	0.75 (0.005)	0.75 (0.007)	0.70 (0.027)	0.54 (0.002)	0.36 (0.004)	0.18 (0.002)
4	**2000**	**0**	0.75 (0.005)	0.75 (0.007)	0.70 (0.029)	0.54 (0.002)	0.37 (0.004)	0.18 (0.002)
5	**500**	**2000**	0.79 (0.004)	0.78 (0.002)	0.70 (0.008)	0.64 (0.001)	0.47 (0.002)	0.22 (0.006)
6	**100**	**2000**	0.78 (0.007)	0.76 (0.005)	0.62 (0.017)	0.62 (0.001)	0.44 (0.002)	0.19 (0.004)
7	**0**	**2000**	0.78 (0.008)	0.76 (0.006)	0.58 (0.025)	0.61 (0.002)	0.42 (0.002)	0.17 (0.004)

Scen. = scenarios; Nb HF = number of Holstein Friesian animals; Nb J = number of Jersey animals; MAF = minor allele frequency.

For the RANDOM scenario, the number of QTL underlying a trait had a limited effect on prediction accuracies (Figures 5.1 and 5.2); accuracies were slightly greater for QTL with a very low average MAF (~0.03) or extremely low average MAF (~0.07) when 1000 QTL instead of 100 controlled the trait. This reduced the effect of the average MAF of QTL on accuracy with 1000 QTL compared to 100 QTL. For the VAR scenario, the effect of the number of QTL on accuracy was very small for all situations (Figures 5.1 and 5.2). Estimated heritabilities with 1000 QTL underlying the trait were similar to those with 100 QTL underlying the trait, both for the RANDOM and VAR scenarios (see Appendix Table A5.1).

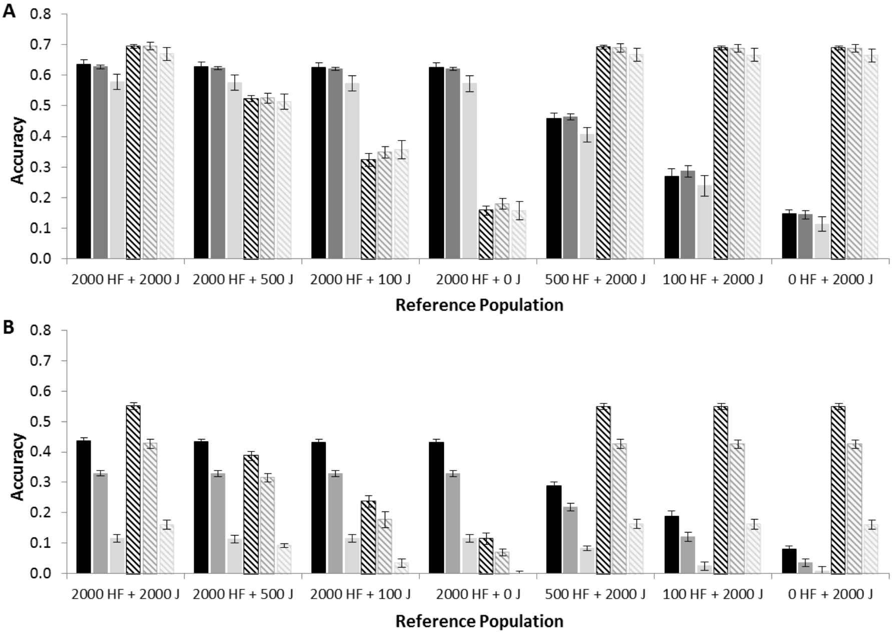


Figure 5.2 Accuracies of genomic prediction for traits that are controlled by QTL with different properties when 1000 QTL underlie the trait. Average accuracies of genomic prediction (± standard errors) for Holstein Friesian (HF, solid fill) and Jersey (J, diagonal fill) animals using a model that included a random across-breed animal effect and a within-breed animal effect, 606,384 SNPs, seven different reference populations and using simulated allele substitution effects (A) randomly sampled from a gamma distribution or (B) with each QTL explaining an equal proportion of the genetic variance, when 1000 QTL underlying the trait were sampled from variants with on average a moderately low allele frequency (black), very low minor allele frequency (dark grey) or extremely low minor allele frequency (light grey).

5.3.2 Marker densities and mutations

With 100 QTL underlying the trait, average accuracies achieved with the base model that used genomic relationship matrices based on different marker densities, with or without the simulated QTL, are shown in Figure 5.3 for the RANDOM scenario (A) and VAR scenario (B). For both scenarios, a decrease in the number of SNPs used to calculate the genomic relationship matrices from 606,384 to 60,000 resulted in similar accuracies of genomic prediction, although values were slightly, but consistently, lower (~0.007) with 60,000 SNPs than with 606,384 SNPs. Estimated heritabilities using 60,000 or 606,384 SNPs were also similar (Table 5.4).

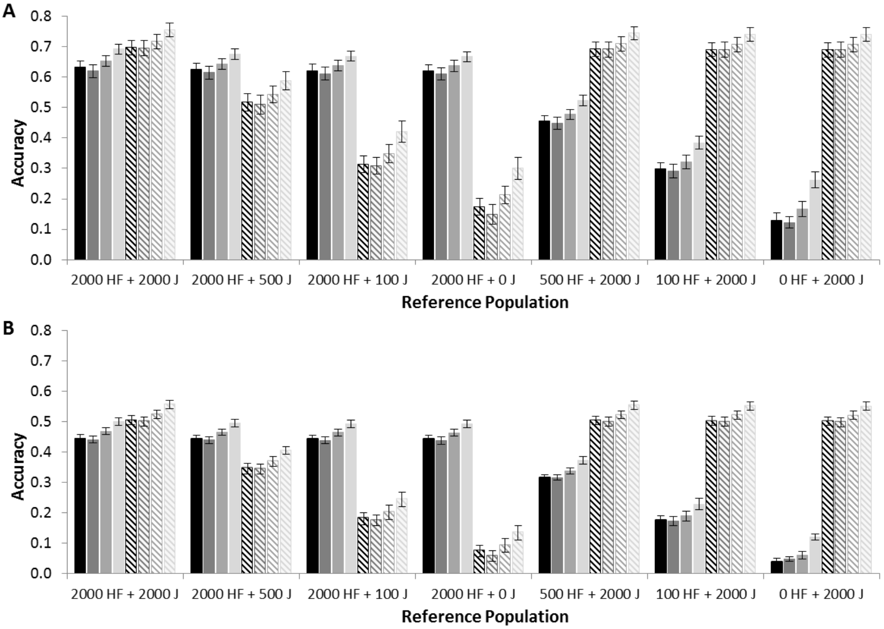


Figure 5.3 Accuracies of genomic prediction using different marker densities to calculate the genomic relationship matrix. Average accuracies of genomic prediction (± standard errors) for Holstein Friesian (HF, solid fill) and Jersey (J, diagonal fill) animals using a model that included a random across-breed animal effect and a within-breed animal effect, seven different reference populations and using simulated allele substitution effects (A) randomly sampled from a gamma distribution or (B) with each QTL explaining an equal proportion of the genetic variance, when 100 QTL underlying the trait were sampled from variants with on average a moderately low minor allele frequency. The genomic relationship matrices were calculated using 606,384 SNPs (black), 60,000 SNPs (dark grey), 606,384 SNPs plus all sampled QTL (grey), or 60,000 SNPs plus all sampled QTL (light grey).

Adding the genotypes of the simulated QTL to the SNPs used to calculate the genomic relationship matrices increased prediction accuracy (Figure 5.3), and the percentage of increase was higher when the average MAF of QTL was lower (see Appendix Figure A5.4 and Figure A5.5). This increase in accuracy was greater when 60,000 SNPs were used (increase in accuracy of ~0.08 and ~0.06 for the RANDOM and VAR scenarios, respectively) than when 606,384 SNPs were used (increase in accuracy of ~0.02 for both scenarios). Thus, prediction accuracies were greatest when 60,000 SNPs plus the genotypes of the simulated QTL were used to calculate the genomic relationship matrices. As expected, the same pattern was observed with estimated heritabilities (Table 5.4). This indicates that including the simulated QTL in the marker set to calculate genomic relationship matrices improved the ability of the model to capture all the genetic variance present in the reference population, probably because the QTL can capture the effects without depending on LD between marker and QTL.

Table 5.4 Average estimated heritabilities using different marker densities to calculate the genomic relationship matrix. Average heritabilities (standard errors across replicates) estimated with a model including a random across-breed animal effect and a within-breed animal effect using different reference populations, 100 QTL underlying the trait with on average a moderately low minor allele frequency and using simulated allele substitution effects randomly sampled from a gamma distribution (RANDOM) or with each QTL explaining an equal proportion of the genetic variance (VAR). The genomic relationship matrix was calculated using 606,384 SNPs (600k), 60,000 SNPs (60k), 606,384 SNPs plus all sampled QTL (600k + QTL), or 60,000 SNPs plus all sampled QTL (60k + QTL).

Scen.	Nb HF	Nb J	RANDOM				VAR			
			600k	60k	600k+QTL	60k+QTL	600k	60k	600k+QTL	60k+QTL
1	2000	2000	0.78 (0.003)	0.77 (0.003)	0.80 (0.002)	0.84 (0.001)	0.59 (0.001)	0.58 (0.001)	0.61 (0.001)	0.64 (0.001)
2	2000	500	0.76 (0.004)	0.74 (0.004)	0.78 (0.003)	0.82 (0.003)	0.54 (0.003)	0.53 (0.003)	0.57 (0.003)	0.59 (0.003)
3	2000	100	0.75 (0.005)	0.73 (0.005)	0.77 (0.004)	0.80 (0.005)	0.54 (0.004)	0.53 (0.004)	0.57 (0.004)	0.59 (0.004)
4	2000	0	0.75 (0.005)	0.73 (0.005)	0.77 (0.004)	0.81 (0.005)	0.55 (0.004)	0.54 (0.004)	0.58 (0.004)	0.60 (0.004)
5	500	2000	0.79 (0.004)	0.78 (0.005)	0.80 (0.003)	0.83 (0.002)	0.61 (0.006)	0.60 (0.005)	0.63 (0.005)	0.66 (0.005)
6	100	2000	0.78 (0.007)	0.77 (0.007)	0.80 (0.006)	0.82 (0.005)	0.58 (0.006)	0.58 (0.006)	0.60 (0.006)	0.63 (0.005)
7	0	2000	0.78 (0.008)	0.77 (0.008)	0.79 (0.007)	0.82 (0.006)	0.56 (0.006)	0.55 (0.006)	0.57 (0.005)	0.60 (0.005)

Scen. = scenarios; Nb HF = number of Holstein Friesian animals; Nb J = number of Jersey animals.

5.3.3 Different models

The base model of this study contained a random across-breed animal effect and a within-breed animal effect to account for differences in SNP effects across breeds. For the multi-breed reference populations, the proportion of variance explained by the within-breed animal component was equal to ~27% and ~52% for the RANDOM and VAR scenarios, respectively, when QTL had a moderately low average MAF, ~33% and ~53% when QTL had a very low average MAF, and ~40% and ~63% when QTL had an extremely low average MAF.

The power to separate across-breed animal and within-breed animal effects was investigated in Figure 5.4. This figure shows that for the three classes of QTL with different average MAF and for most of the replicates, the model that estimated across- and within-breed animal variances was not significantly better than a model without a random within-breed animal effect ($P < 0.05$). This is because the log-likelihood is rather flat. Moreover, prediction accuracies and heritabilities estimated with the base model that included a random across-breed animal effect and a within-breed animal effect were very similar to those estimated with a model without a random within-breed animal effect for all scenarios (results not shown). These results indicate that the power to separate across- and within-breed animal effects was limited in our simulated data. Similar prediction accuracies were achieved with a model that included a fixed breed effect (results not shown). Thus, for all scenarios for which a random within-breed animal effect and/or fixed breed effect is included in the model, accuracies of genomic prediction were not affected, and therefore, they are not shown.

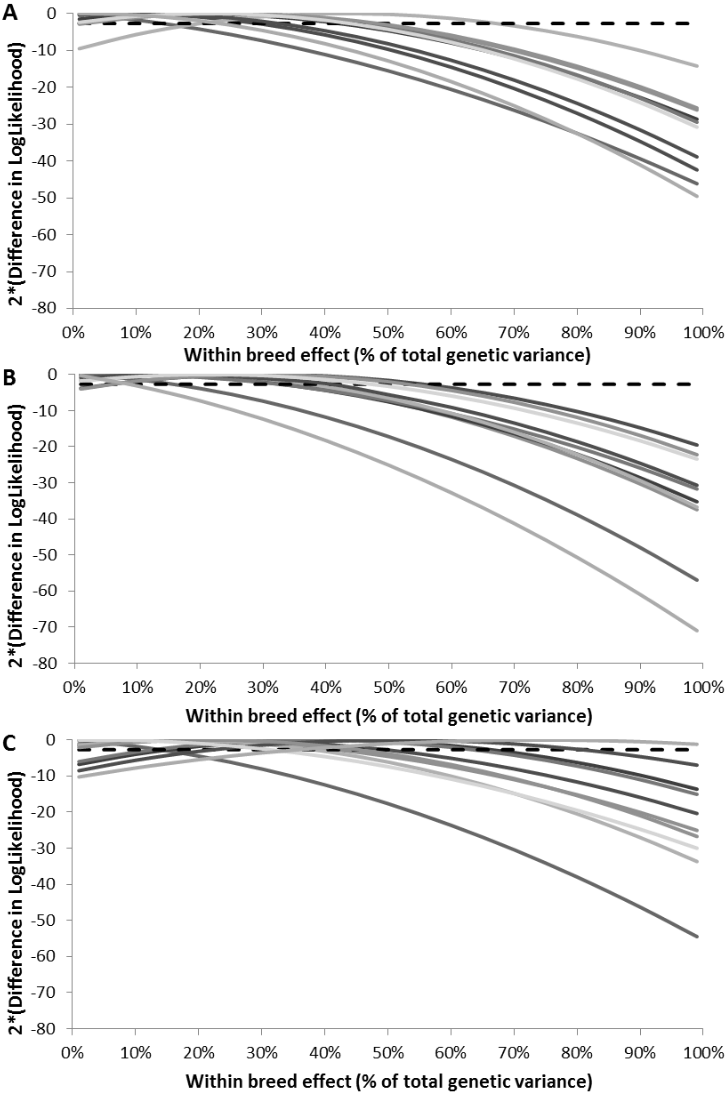


Figure 5.4 Log-likelihood comparison of models with fixed or estimated random across-breed and within-breed animal effects. Twice the difference in log-likelihood for each of the 10 replicates and 5% significance threshold (black dotted line) using models with fixed variance components for the random across-breed animal effect and a within-breed animal effect compared to a model that estimated both variance components. The genomic relationship matrix was calculated based on 606,384 SNPs, the reference population consisted of 2000 Holstein Friesian and 2000 Jersey animals, allele substitution effects were sampled from a gamma distribution, when 1000 QTL underlying the trait were sampled from variants with on average a (A) moderately low allele frequency, (B) very low minor allele frequency or (C) extremely low minor allele frequency.

5.4 Discussion

5.4.1 Accuracy of multi-breed genomic prediction

For an accurate prediction of genomic breeding values, a large group of animals with both genotypes and phenotypes is required (e.g., Meuwissen *et al.* 2001; Daetwyler *et al.* 2008; VanRaden *et al.* 2009). Therefore, an attractive approach is to enlarge small reference populations of a particular breed by using information from other breeds. This might be especially interesting for traits that are difficult to measure, such as feed efficiency and dry matter intake in dairy cattle (De Haas *et al.* 2012; Pryce *et al.* 2014), and for numerically small breeds. In this study, the effect of adding another breed to the reference population on prediction accuracy was investigated in different scenarios using Holstein Friesian and Jersey animals. Accuracy of genomic prediction was not significantly increased by adding 2000 individuals of the other breed to a reference population of animals from the same breed as the selection candidates regardless of marker density. The accuracy of across-breed genomic prediction, i.e., using a reference population consisting only of individuals from the other breed, ranged from 0.01 to 0.19. The positive accuracies of across-breed genomic prediction indicated that useful information was present in the other breed, although adding animals from the other breed to the reference population did not increase prediction accuracy. This suggests that the number of reference individuals from the other breed compared to the number of reference individuals from the breed of the selection candidates was relatively too small to see an increase in accuracy, as suggested by Hozé *et al.* (2014). The benefit of using a multi-breed reference population might also depend on the model used to analyze the data, Bayesian models, for example, might gain more from multiple breeds (Kemper *et al.* 2015).

5.4.2 Effect of QTL properties on the accuracy of genomic prediction

The first objective of this study was to investigate the effect of properties of QTL that underlie the trait on the accuracy of multi-breed genomic prediction using Holstein Friesian and Jersey animals. Phenotypes of traits that are controlled by QTL with different properties were simulated by sampling 100 or 1000 QTL from three different classes of variants that had an average MAF ranging from moderately low to extremely low, and by sampling allele substitution effects either based on a model where effect size was independent of allele frequency (RANDOM) or based on a 'rare allele, large effect' model (VAR). The three different classes of variants were imputed using sequenced animal genomes, such that the QTL displayed characteristics that were present on the actual bovine genome. Our results showed that the accuracy of both single-breed and multi-breed genomic prediction was influenced by the properties of the QTL that control the trait. A lower QTL MAF decreased prediction accuracy and this effect was more pronounced when QTL with the lowest MAF had the largest effect, which is consistent with the results from other studies that showed that the prediction model could better capture the genetic variance and provided a greater accuracy of genomic prediction when a small group of QTL explained a large part of the genetic variance (Goddard 2009; Hayes *et al.* 2010).

A decrease in QTL MAF was expected to decrease accuracy of multi-breed genomic prediction, since the percentage of breed-specific variants increased when the MAF of the variants decreased, thereby reducing the potential benefit of adding another breed. Moreover, LD between SNPs and QTL decreases as the allele frequency of QTL becomes more extreme, due to ascertainment bias of the SNPs on the chip (Matukumalli *et al.* 2009). The existence of ascertainment bias was confirmed by the fact that imputation reliabilities decreased when QTL MAF decreased and that the prediction accuracies increased most when QTL had the lowest MAF and QTL genotypes were added to the markers. Moreover, the low LD between SNPs and QTL is reflected in the increasing underestimation of the heritability as the average QTL MAF decreased. This is in agreement with other studies, that showed that simulating QTL with a low MAF resulted in underestimated heritability estimates (Yang *et al.* 2010; De los Campos *et al.* 2013) and lower accuracy of genomic prediction (Daetwyler *et al.* 2013; De los Campos *et al.* 2013). QTL for many complex traits have a low MAF (Goddard and Hayes 2009; Yang *et al.* 2010; Kemper and Goddard 2012), which indicates that the probability of underestimating the heritability for those traits is high. Heritability may also be underestimated because only a subset of the animals from a population is used in the analyses. When QTL MAF are low and the size of the reference population is small, the probability that all these QTL are segregating in the reference population is reduced. Therefore, the increase in accuracy of genomic prediction achieved by enlarging the reference population, as shown by (e.g., Meuwissen *et al.* 2001; Daetwyler *et al.* 2008; VanRaden *et al.* 2009), might not only result from a more accurate prediction of SNP effects, but also from capturing a larger proportion of the alleles that segregate in the complete population.

Many previous simulation studies have simulated QTL based on SNP characteristics (De Roos *et al.* 2009; Ibánẽz-Escriche *et al.* 2009; Wientjes *et al.* 2015). However, the SNPs that are commonly used on chips are selected to have a reasonably high MAF and to segregate in multiple breeds. In our data, the average MAF of the SNPs across breeds was 0.27, which is much higher than the average MAF of the other variants (Table 5.1). As shown in Figure 5.5, prediction accuracies increase as the average QTL MAF increases; therefore, it is clear that using the MAF pattern of SNPs to simulate QTL will result in a substantially larger expected accuracy of both across-breed and multi-breed genomic prediction. This can explain why the benefits of using information from another breed are much larger in other simulation studies compared to our simulation study (De Roos *et al.* 2009; Ibánẽz-Escriche *et al.* 2009; Wientjes *et al.* 2015) and compared to empirical studies (e.g., Hayes *et al.* 2009; Pryce *et al.* 2011; Erbe *et al.* 2012).

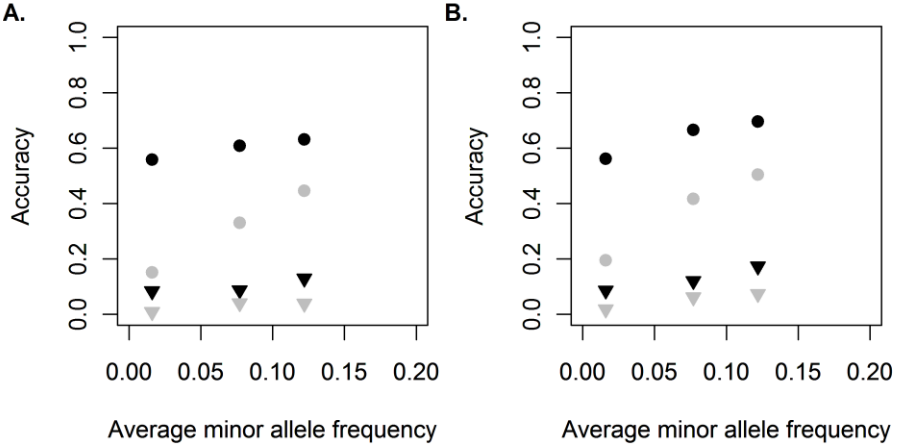


Figure 5.5 Accuracy of across- and multi-breed genomic prediction versus average minor allele frequency of QTL. The average accuracy of across- and multi-breed genomic prediction for (A) Holstein Friesian and (B) Jersey selection candidates versus the average minor allele frequency of the 100 simulated QTL. Black points represent the scenarios with allele substitution effects randomly sampled from a gamma distribution and grey points represent the scenario with each QTL explaining an equal proportion of the genetic variance. The circles represent the accuracy for the multi-breed reference population with 2000 Holstein Friesian and 2000 Jersey animals, the triangles represent the accuracy of across-breed genomic prediction with a reference population of 2000 animals from the other breed.

It should be noted that there are two caveats regarding our results, but we consider that they do not affect the overall conclusions greatly. First, the effect of low MAF on accuracy and heritability may be somewhat exaggerated by the imperfect imputation of causal variants. This means that the QTL are not as well tracked by the SNPs as they should be. Second, the formula used to calculate the **G** matrix might be more appropriate for the scenario with allele substitution effects that are sampled independently of allele frequencies than for the scenario using the 'rare allele, large effect' model, which might be better analyzed by the **G** matrix described by Yang *et al.* (2010). However, for a fair comparison of the scenarios, we decided to use the same **G** matrix for both scenarios.

5.4.3 Marker densities and mutations

In this study, the data was analyzed with a GBLUP type of model using genomic relationship matrices based on 606,384 or 60,000 SNPs. Reducing the number of SNPs from 606,384 to 60,000 resulted in similar accuracies of genomic prediction. This is in agreement with empirical studies using dairy cattle data that showed that increasing the number of SNPs from 50k to high-density (777 k) had almost no effect on the accuracy of multi-breed genomic prediction (e.g., Harris *et al.* 2011; Erbe *et al.* 2012), in contrast to earlier expectations (De Roos *et al.* 2008).

For all scenarios, accuracy of genomic prediction was slightly greater when the simulated QTL were added to the subset of markers used to calculate the genomic relationship matrices. This indicates that the model could better capture QTL effects with the markers, which led to higher estimated heritabilities and accuracies, when the simulated QTL were used as markers, which was also shown in other studies (Kizilkaya *et al.* 2010; Meuwissen and Goddard 2010). The increase in prediction accuracy due to adding the simulated QTL was larger when 60,000 SNPs were used than when 606,384 SNPs were used. This is likely an artifact of the GBLUP model for which all markers are assumed to explain the same amount of variance. This means that as the number of markers increases, each marker effect is *a priori* smaller. Thus, with a larger number of markers, the effects of true markers in the dataset are diluted to a greater degree. By using sequence data in the analyses, the causal variants or QTL are supposed to be included in the data, as well as a large number of other variants. Therefore, on the one hand, the expected benefit of sequence data achieved with a GBLUP model is small, and smaller than that with Bayesian models, which allow some marker effects to be zero, as demonstrated by Meuwissen and Goddard (Meuwissen and Goddard 2010). On the other hand, our result does demonstrate that if the marker set can be enriched with real causative mutations from the sequence data, as we did here by including the QTL in the SNP dataset, accuracies can be increased. The larger increase in prediction accuracy achieved with a smaller number of other variants in the dataset highlights the importance to filter sequence variants that are included in genomic prediction, for example by using biological information (MacLeod *et al.* 2014).

Both in the single-breed and multi-breed scenarios using Holstein Friesian and Jersey animals, the percentage of increase in accuracies due to adding the QTL genotypes to the markers was higher when the average MAF of QTL was lower. This can be explained by the fact that the QTL with a lower MAF were in lower LD with the SNPs on the chip, particularly across breeds. Besides differences in LD across breeds, the accuracy of multi-breed genomic prediction might also be influenced by other factors, such as the absence of family relationships or differences in allele frequencies (e.g., Daetwyler *et al.* 2008; Habier *et al.* 2010; Wientjes *et al.* 2013). As explained by Daetwyler *et al.* (2008), a QTL with a large effect and a low allele frequency in one breed can be imprecisely estimated within that breed. Since that QTL only explains a small proportion of the genetic variance in that breed, the negative effect on the accuracy of single-breed genomic prediction might be small. If the estimated effect was used to predict breeding values for another breed, the effect on accuracy would be more detrimental when the allele frequency of that QTL is higher in that breed. This indicates that it is important that the QTL and SNPs that segregate in the selection candidate population are also segregating with a reasonable allele frequency in the reference population to be able to estimate the effects accurately. When the relationships between selection candidates and reference individuals are larger, the probability that SNPs and QTL segregating in the selection candidate population are segregating in the reference population becomes higher as well. Overall, these results indicate that the accuracy of across-breed genomic prediction is small because of differences in LD (e.g., De Roos *et al.* 2008; Zhong *et al.* 2009), absence of family relationships (e.g., Habier *et al.* 2010; Wientjes *et al.* 2013), and differences in allele frequency across breeds (e.g., Daetwyler *et al.* 2008); in addition, all these factors are probably entangled with each other.

5.4.4 Effect of random within-breed animal effect on the accuracy of genomic prediction

The second objective of this study was to investigate the effect of including a random across-breed animal effect and a within-breed animal effect in a GBLUP model on the accuracy of multi-breed genomic prediction. Our results showed that, in contrast to our expectations, adding a random within-breed animal effect did not influence prediction accuracy. In particular, if the QTL were breed-specific and if the SNPs segregated in both breeds, which was to a high extent the case when the average MAF of QTL was extremely low, an increase in accuracy due to the inclusion of a random within-breed animal effect was expected because of differences in apparent SNP effects across breeds. The power of this approach to separate across- and within-breed animal effects was limited when allele substitution effects were randomly assigned to QTL, which may explain why adding a within-breed animal effect was not beneficial. For the scenarios for which each QTL explained the same variance, the power to separate both effects might differ, but adding a within-breed animal effect was still not beneficial in terms of accuracy. Using a larger reference population with more animals of each breed may enable to properly separate across-breed animal and within-breed animal effects in a better way, but enlarging reference populations for numerically small breeds is challenging. Thus, to give a conclusive answer about this objective, more data is needed to investigate if it is possible to separate random across-breed and within-breed animal effects, and if this is case, then it is necessary to investigate whether it is beneficial for multi-breed genomic prediction.

5.5 Conclusions

The results of this study show that the accuracy of both single- and multi-breed genomic prediction depends on the properties of the QTL that control the trait. A decrease in average QTL MAF decreased accuracy of genomic prediction, especially when rare alleles had a large effect. Therefore, we demonstrated that the properties of the QTL that control traits (i.e., allele frequency spectra of QTL, distribution of QTL effects) are key parameters that determine the accuracy of both single- and multi-breed genomic predictions. Based on these results, the properties of QTL that underlie a trait can explain the limited benefit or the absence of benefit of combining information from multiple breeds that is described in empirical studies as opposed to the substantial benefit that is achieved in simulation studies. Accuracy of single-, but especially multi-breed genomic prediction, could be increased by using sequence data, since the causative mutations are probably included in the dataset. The results show that the increase in accuracy was consistently, although not significantly, larger when the number of other variants included in the dataset was smaller. Finally, adding a random within-breed animal effect to a GBLUP type of model had no effect on the accuracy of genomic prediction, most likely because the power to separate random across-breed and within-breed animal effects was low.

5.6 Acknowledgements

The authors thank Roel Veerkamp, Chris Schrooten and Henk Bovenhuis for their useful comments. This study was financially supported by Breed4Food (KB-12-006.03-005-ASG-LR), a public-private partnership in the domain of animal breeding and genomics, the cooperative cattle improvement organization CRV BV (Arnhem, The Netherlands), and EU FP7 IRSES SEQSEL (Grant no. 317697) by funding the research visit of Yvonne Wientjes at the Department of Environmental and Primary Industries in Victoria, Australia. The 1000 bulls genomes consortium is acknowledged for providing the sequence data, and in particular Paul Stothard for providing the annotations.

5.7 Appendix

Table A5.1 Average estimated heritabilities for QTL with different properties when 1000 QTL underlie the trait. Average heritabilities (standard errors across replicates) estimated with a model including a random across-breed animal effect and a within-breed animal effect and using 606,384 SNPs to calculate the genomic relationship matrix using different reference populations, different average minor allele frequencies (MAF) of the 1000 QTL that underlie the trait and using simulated allele substitution effects randomly sampled from a gamma distribution (RANDOM) or with each QTL explaining an equal proportion of the genetic variance (VAR).

			RANDOM						**VAR**					
Scenarios	**Nb HF**	**Nb J**	**Moderately low MAF**		**Very low MAF**		**Extremely low MAF**		**Moderately low MAF**		**Very low MAF**		**Extremely low MAF**	
1	2000	2000	0.79	(0.020)	0.77	(0.021)	0.72	(0.023)	0.59	(0.028)	0.45	(0.031)	0.21	(0.031)
2	2000	500	0.80	(0.030)	0.76	(0.032)	0.72	(0.033)	0.54	(0.040)	0.39	(0.042)	0.16	(0.037)
3	2000	100	0.80	(0.032)	0.76	(0.035)	0.71	(0.036)	0.54	(0.043)	0.39	(0.045)	0.15	(0.039)
4	2000	0	0.80	(0.033)	0.75	(0.035)	0.70	(0.037)	0.55	(0.044)	0.39	(0.046)	0.15	(0.036)
5	500	2000	0.78	(0.025)	0.77	(0.025)	0.73	(0.028)	0.61	(0.034)	0.49	(0.039)	0.26	(0.041)
6	100	2000	0.78	(0.026)	0.78	(0.026)	0.72	(0.030)	0.58	(0.037)	0.46	(0.041)	0.23	(0.042)
7	0	2000	0.78	(0.027)	0.78	(0.027)	0.73	(0.030)	0.56	(0.038)	0.45	(0.042)	0.23	(0.043)

Nb HF = Number of Holstein Friesian animals; Nb J = Number of Jersey animals; MAF = minor allele frequency

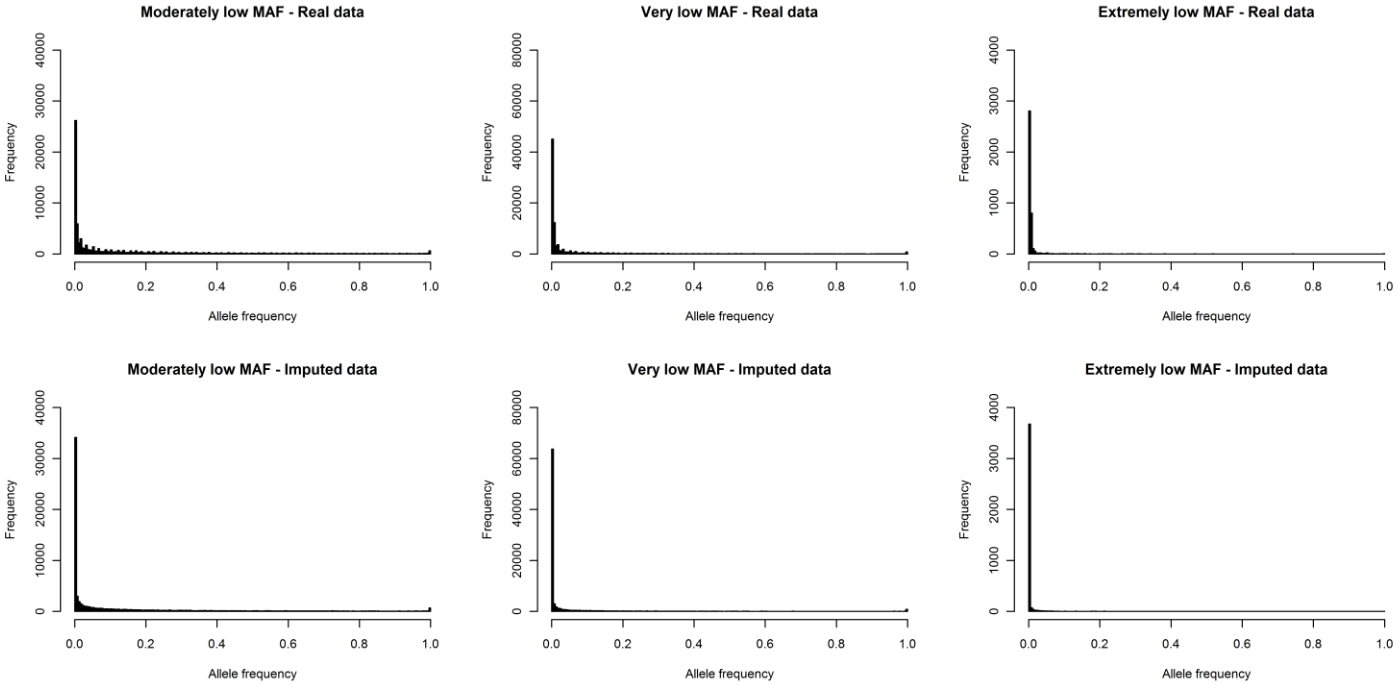


Figure A5.1 Allele frequency distribution of imputed and genotyped variants in Holstein Friesian animals. Distribution of allele frequencies of variants with on average a moderately low minor allele frequency (MAF), very low MAF or extremely low MAF in real data and imputed data for Holstein Friesian animals.

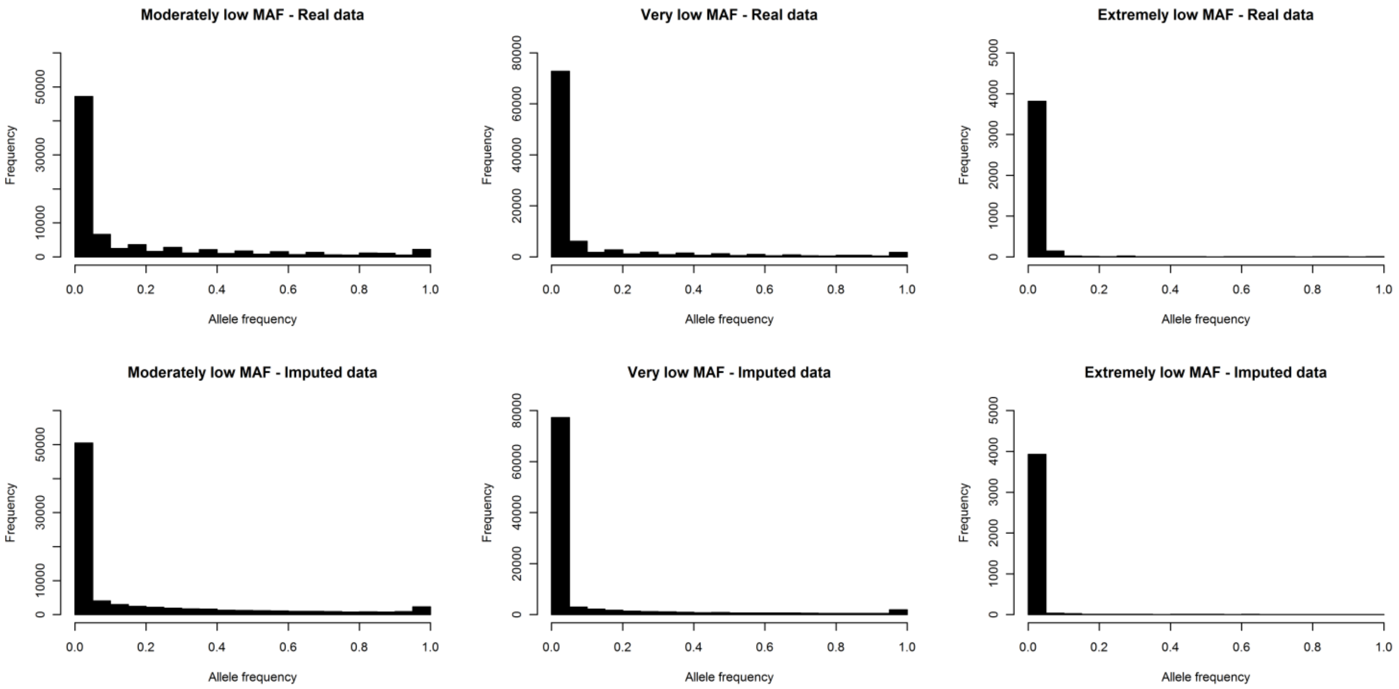


Figure A5.2 Allele frequency distribution of imputed and genotyped variants in Jersey animals. Distribution of allele frequencies of variants with on average a moderately low minor allele frequency (MAF), very low MAF or extremely low MAF in real data and imputed data for Jersey animals.

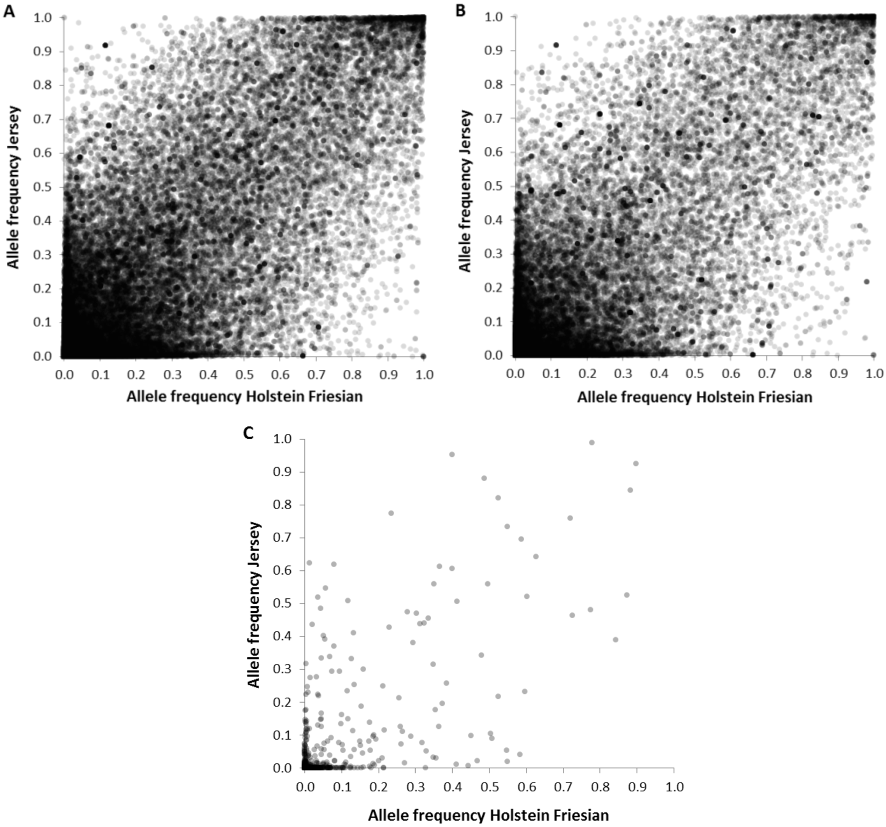


Figure A5.3 Allele frequencies of Holstein Friesian versus Jersey animals. Patterns of allele frequencies for Holstein Friesian versus Jersey animals. (A) Variants with on average a moderately low minor allele frequency; (B) Variants with on average a very low minor allele frequency; (C) Variants with on average an extremely low minor allele frequency.

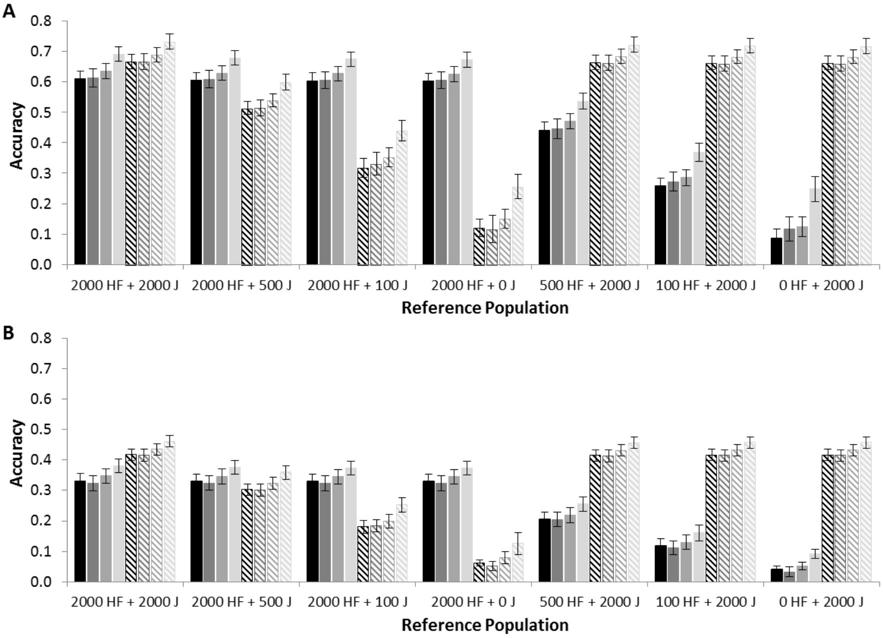


Figure A5.4 Accuracies of genomic prediction using different marker densities to calculate the genomic relationship matrix and QTL with very low minor allele frequency. Average accuracies of genomic prediction (± standard errors) for Holstein Friesian (HF, solid fill) and Jersey (J, diagonal fill) animals using a model that included a random across-breed animal effect and a within-breed animal effect, seven different reference populations and using simulated allele substitution effects (A) randomly sampled from a gamma distribution or (B) with each QTL explaining an equal proportion of the genetic variance, when 100 QTL underlying the trait were sampled from variants with on average a very low minor allele frequency. The genomic relationship matrices were calculated using 606,384 SNPs (black), 60,000 SNPs (dark grey), 606,384 SNPs plus all sampled QTL (grey), or 60,000 SNPs plus all sampled QTL (light grey).

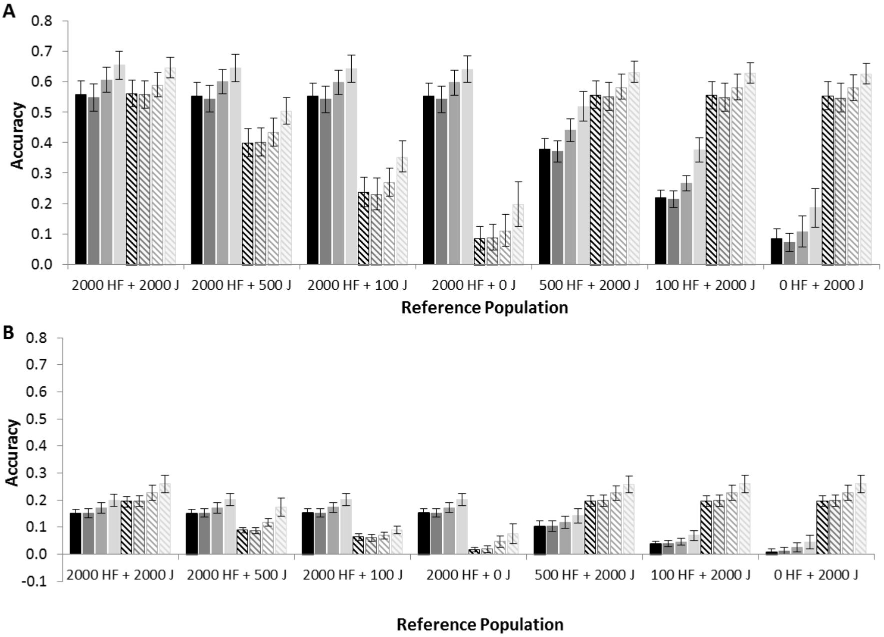


Figure A5.5 Accuracies of genomic prediction using different marker densities to calculate the genomic relationship matrix and QTL with extremely low minor allele frequency. Average accuracies of genomic prediction (± standard errors) for Holstein Friesian (HF, solid fill) and Jersey (J, diagonal fill) animals using a model that included a random across-breed animal effect and a within-breed animal effect, seven different reference populations and using simulated allele substitution effects (A) randomly sampled from a gamma distribution or (B) with each QTL explaining an equal proportion of the genetic variance, when 100 QTL underlying the trait were sampled from variants with on average an extremely low minor allele frequency. The genomic relationship matrices were calculated using 606,384 SNPs (black), 60,000 SNPs (dark grey), 606,384 SNPs plus all sampled QTL (grey), or 60,000 SNPs plus all sampled QTL (light grey).

5.8 References

Bolormaa, S., J. E. Pryce, K. Kemper, K. Savin, B. J. Hayes*, et al.*, 2013 Accuracy of prediction of genomic breeding values for residual feed intake and carcass and meat quality traits in Bos taurus, Bos indicus, and composite beef cattle. J. Anim. Sci. 91: 3088-3104.

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

Brøndum, R. F., P. Ma, M. S. Lund and G. Su, 2012 Short communication: Genotype imputation within and across Nordic cattle breeds. J. Dairy Sci. 95: 6795-6800.

Browning, B. L. and S. R. Browning, 2009 A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am. J. Hum. Genet. 84: 210-223.

Calus, M. P. L., H. Huang, A. Vereijken, J. Visscher, J. Ten Napel*, et al.*, 2014 Genomic prediction based on data from three layer lines: a comparison between linear methods. Genet. Sel. Evol. 46: 57.

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

Daetwyler, H. D., M. P. L. Calus, R. Pong-Wong, G. De los Campos and J. M. Hickey, 2013 Genomic prediction in animals and plants: Simulation of data, validation, reporting, and benchmarking. Genetics 193: 347-365.

Daetwyler, H. D., A. Capitan, H. Pausch, P. Stothard, R. Van Binsbergen*, et al.*, 2014 Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat. Genet. 46: 858-865.

De Haas, Y., M. P. L. Calus, R. F. Veerkamp, E. Wall, M. P. Coffey*, et al.*, 2012 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 los Campos, G., A. I. Vazquez, R. Fernando, Y. C. Klimentidis and D. Sorensen, 2013 Prediction of complex human traits using the genomic best linear unbiased predictor. PLoS Genet. 9: e1003608.

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: 1545-1553.

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

Fisher, R. A., 1954 *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh.

Gilmour, A. R., B. Gogel, B. Cullis, R. Thompson, D. Butler*, et al.*, 2009 *ASReml user guide release 3.0*. VSN International Ltd, Hemel Hempstead.

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

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

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.

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.

Harris, B. L., F. E. Creagh, A. M. Winkelman and D. L. Johnson, 2011 Experiences with the Illumina high density Bovine BeadChip. Interbull Bull. 44: 3-7.

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

Hayes, B. J., J. E. Pryce, A. J. Chamberlain, P. J. Bowman and M. E. Goddard, 2010 Genetic architecture of complex traits and accuracy of genomic prediction: Coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. PLoS Genet. 6: e1001139.

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

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

Jansen, S., B. Aigner, H. Pausch, M. Wysocki, S. Eck*, et al.*, 2013 Assessment of the genomic variation in a cattle population by re-sequencing of key animals at low to medium coverage. BMC Genom. 14: 446.

Kemper, K. E. and M. E. Goddard, 2012 Understanding and predicting complex traits: Knowledge from cattle. Hum. Mol. Genet. 21: R45-R51.

Kemper, K. E., C. M. Reich, P. J. Bowman, C. J. Vander Jagt, A. J. Chamberlain*, et al.*, 2015 Improved precision of QTL mapping using a nonlinear Bayesian method in a multi-breed population leads to greater accuracy for across-breed genomic predictions. Genet. Sel. Evol. 47: 29.

Khansefid, M., J. E. Pryce, S. Bolormaa, S. P. Miller, Z. Wang*, et al.*, 2014 Estimation of genomic breeding values for residual feed intake in a multibreed cattle population. J. Anim. Sci. 92: 3270-3283.

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

Li, B. and S. M. Leal, 2009 Discovery of rare variants via sequencing: Implications for the design of complex trait association studies. PLoS Genet. 5: e1000481.

MacLeod, I. M., B. J. Hayes, C. J. Vander Jagt, K. E. Kemper, M. Haile-Mariam*, et al.*, 2014 A Bayesian analysis to exploit imputed sequence variants for QTL discovery. Proc. 10th World Congr. Genet. Appl. Livest. Prod., ASAS, Vancouver.

Makgahlela, M. L., E. A. Mäntysaari, I. Strandén, M. Koivula, U. S. Nielsen*, et al.*, 2013 Across breed multi-trait random regression genomic predictions in the Nordic Red dairy cattle. J. Anim. Breed. Genet. 130: 10-19.

Matukumalli, L. K., C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan*, et al.*, 2009 Development and characterization of a high density SNP genotyping assay for cattle. PLoS ONE 4: e5350.

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

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

Ott, R. L. and M. Longnecker, 2001 *An introduction to statistical methods and data analysis (fifth edition)*. Duxbury, Pacific Grove.

Pryce, J., W. Wales, Y. De Haas, R. Veerkamp and B. Hayes, 2014 Genomic selection for feed efficiency in dairy cattle. Animal 8: 1-10.

Pryce, J. E., B. Gredler, S. Bolormaa, P. J. Bowman, C. Egger-Danner*, et al.*, 2011 Short communication: Genomic selection using a multi-breed, across-country reference population. J. Dairy Sci. 94: 2625-2630.

Simeone, R., I. Misztal, I. Aguilar and Z. G. Vitezica, 2012 Evaluation of a multi-line broiler chicken population using a single-step genomic evaluation procedure. J. Anim. Breed. Genet. 129: 3-10.

Spelman, R. J., C. A. Ford, P. McElhinney, G. C. Gregory and R. G. Snell, 2002 Characterization of the DGAT1 gene in the New Zealand dairy population. J. Dairy Sci. 85: 3514-3517.

Thaller, G., W. Krämer, A. Winter, B. Kaupe, G. Erhardt*, et al.*, 2003 Effects of DGAT1 variants on milk production traits in German cattle breeds. J. Anim. Sci. 81: 1911-1918.

The International HapMap 3 Consortium, 2010 Integrating common and rare genetic variation in diverse human populations. Nature 467: 52-58.

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel*, et al.*, 2009 Invited review: Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92: 16-24.

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.

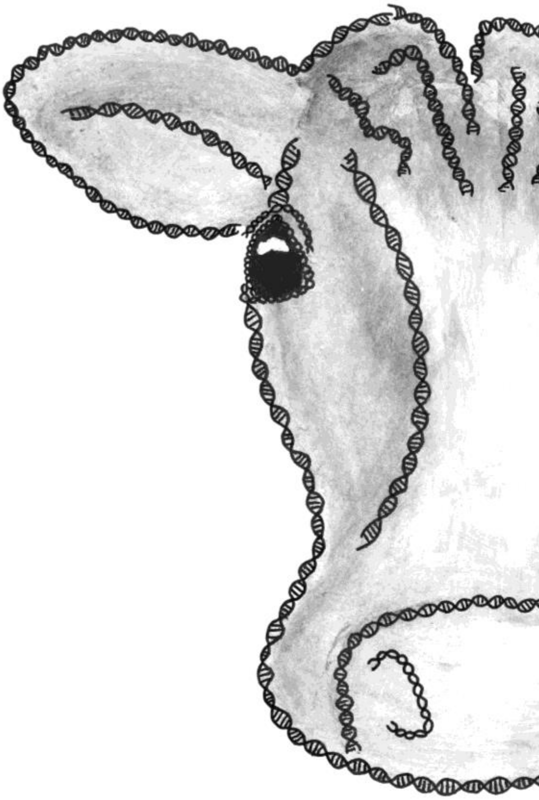
Wientjes, Y. C. J., R. F. Veerkamp, P. Bijma, H. Bovenhuis, C. Schrooten*, et al.*, 2015 Empirical and deterministic accuracies of across-population genomic prediction. Genet. Sel. Evol. 47: 5.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders*, et al.*, 2010 Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42: 565-569.

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.

CHAPTER 6

AN EQUATION TO PREDICT THE ACCURACY OF GENOMIC VALUES BY COMBINING DATA FROM MULTIPLE TRAITS, BREEDS, LINES, OR ENVIRONMENTS



Y.C.J. WIENTJES[1,2]

R.F. VEERKAMP[1,2]

P. BIJMA[1]

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,
6700 AH WAGENINGEN, THE NETHERLANDS

ACCEPTED FOR PUBLICATION IN GENETICS

Abstract

Predicting the accuracy of estimated genomic values using genome-wide marker information is an important step in designing training populations. Currently, different deterministic equations are available to predict accuracy within populations, but not for multi-population scenarios where data from multiple breeds, lines or environments is combined. Therefore, our objective was to develop and validate a deterministic equation to predict the accuracy of genomic values when different populations are combined in one training population. The input parameters of the derived prediction equation are the number of individuals and the heritability from each of the populations in the training population, the genetic correlations between the populations, i.e., the correlation between allele substitution effects of quantitative trait loci, the effective number of chromosome segments across predicted and training populations, and the proportion of the genetic variance in the predicted population captured by the markers in each of the training populations. Validation was performed based on real genotype information of 1033 Holstein Friesian cows that were divided in three different populations by combining half-sib families in the same population. Phenotypes were simulated for multiple scenarios, differing in heritability within populations and in genetic correlations between the populations. Results showed that the derived equation can accurately predict the accuracy of estimating genomic values for different scenarios of multi-population genomic prediction. Therefore, the derived equation can be used to investigate the potential accuracy of different multi-population genomic prediction scenarios and to decide on the most optimal design of training populations.

Key words: genomic prediction, multi-population, accuracy, prediction equation

6.1 Introduction

Genomic markers can be used to estimate genomic values of individuals, also known as additive genetic values or breeding values, that are used to select animals (e.g., Dekkers 2007; De Roos *et al.* 2011) and plants for breeding (e.g., Heffner *et al.* 2009; Jannink *et al.* 2010), and in humans to predict the genetic risk of diseases (e.g., Wray *et al.* 2007; De Los Campos *et al.* 2010). In genomic prediction, genome-wide single-nucleotide polymorphism (SNP) marker information is used to predict genomic values based on SNP effects estimated in a training population consisting of individuals with known SNP genotypes and phenotypes (Meuwissen *et al.* 2001). The accuracy of estimating genomic values is in general higher when the size of the training population is larger, when the level of linkage disequilibrium (LD) between the SNPs and the quantitative trait loci (QTL) underlying the trait is higher, and when the predicted individuals are more related to the individuals in the training population (e.g., Daetwyler *et al.* 2008; Zhong *et al.* 2009; De los Campos *et al.* 2013; Wientjes *et al.* 2013).

For numerically small populations, the size of the training population is limited which restricts the accuracy of genomic prediction. Therefore, combining different populations in one training population for estimating SNP effects is an appealing approach to increase the size of the training population and, thereby, the accuracy of predicting genomic values. The potential accuracy of combing different populations in one training population has been investigated by combining populations from different breeds (e.g., Hayes *et al.* 2009a; Harris and Johnson 2010), lines (e.g., Zhong *et al.* 2009; Calus *et al.* 2014; Lehermeier *et al.* 2014), subpopulations (e.g., De los Campos *et al.* 2013), or countries (e.g., Lund *et al.* 2011; Haile-Mariam *et al.* 2015). The increase in accuracy by adding individuals from another population to the training population is in most cases much lower than the increase in accuracy obtained by adding an equal number of individuals from the same population. This is a result of differences that exist between populations, like differences in allele frequencies, LD patterns (De Roos *et al.* 2008; Zhong *et al.* 2009; De los Campos *et al.* 2012), allele substitution effects of QTL (Spelman *et al.* 2002; Thaller *et al.* 2003; Wientjes *et al.* 2015b), environments in combination with genotype by environment interactions (Lund *et al.* 2011; Haile-Mariam *et al.* 2015), the presence of QTL that are only segregating in one population (Kemper *et al.* 2015), and the absence of close family relationships across populations.

Different deterministic equations are available to calculate the accuracy of genomic prediction when the training population is a subset from the same population as the predicted individuals (Daetwyler *et al.* 2008; VanRaden 2008; Goddard 2009). One type of deterministic equation is based on prediction error variance of the mixed model equation and is using the genomic relationships within the training population and between training and predicted individuals (VanRaden 2008). This equation has been extended to enable the calculation of the accuracy when different populations are combined in one training population (Wientjes *et al.* 2015b). A disadvantage of this equation is, however, that individuals have to be genotyped before the accuracy can be calculated. Therefore, this equation cannot be used to decide on the most optimal design of training populations. Another type of deterministic equation is able to predict the accuracy before genotype information is available and is based on population parameters, such as the size of the training population, the heritability of the trait and the effective number of chromosome segments (Daetwyler *et al.* 2008; Daetwyler *et al.* 2010). This equation can be used to investigate the accuracy of different training population designs, however, the equation is not applicable for situations with more than one population in the training population.

The first objective of this study is to develop a deterministic equation using population parameters to predict the accuracy of genomic values when different populations are combined in one training population. The different combined populations might for example be populations from different breeds, lines or environments, or populations measured for different traits. The second objective is to validate the derived equation. For the validation, different scenarios of multi-population genomic prediction were considered by dividing 1033 Holstein Friesian cows with real genotypes and simulated phenotypes in three populations, assuming different heritabilities within populations and different genetic correlations between populations. Moreover, the equation was used to investigate the potential accuracy for one specific dairy cattle scenario and one specific human scenario.

6.2 Materials and methods

6.2.1 Theory

The accuracy of estimated genomic values (r_{EGV}) is defined as the correlation between estimated and true genomic values. The overall accuracy depends on the square root of the proportion of genetic variance captured by the SNPs (r_{LD}) and on the accuracy of estimating SNP effects (r_{effect}) (Daetwyler 2009; Goddard 2009). The r_{LD} depends on the strength of LD between QTL and SNPs; the stronger the LD, the higher the proportion of the genetic variance that is captured by the SNPs. The

r_{effect} depends on the characteristics of the trait, the population in which the effects are estimated and the population in which the effects are used to predict genomic values. First, we will derive r_{effect} for a training population consisting of two distinct populations, based on the same assumptions as underlying a commonly used prediction equation for single-population genomic prediction. Thereafter, r_{effect} is combined with r_{LD} to account for the proportion of the genetic variance captured by the SNPs to derive the accuracy of multi-population genomic prediction.

Using the assumptions that n_G independent loci are underlying the trait and that each locus is explaining an equal amount of the genetic variance, Daetwyler *et al.* (2008) derived the following prediction equation for r_{effect} when considering single-population genomic prediction:

$$r_{effect} = \sqrt{\frac{h^2 n_P}{h^2 n_P + n_G}}\ , \tag{6.1}$$

in which h^2 is the heritability of the trait and n_P is the number of individuals with phenotypes and genotypes included in the training population. The original derivation of this equation is rather complex and difficult to extend to multi-population genomic prediction. As shown by Wientjes *et al.* (2015b), the same equation can also be derived by partitioning the variance of the average phenotype of n_P individuals into a part explained by one locus $\left(\sigma_a^2 / n_G\right)$ and a part not explained by that locus $\left(\frac{\sigma_p^2 \text{-} (\sigma_a^2 / n_G)}{n_P}\right)$, in which σ_a^2 is the total genetic variance and σ_p^2 is the phenotypic variance. In general, the accuracy of predicting an effect is equal to the square root of the proportion of the total variance explained by that effect. So, the accuracy of predicting the effect of one locus equals:

$$r_{locus} = \sqrt{\frac{\left(\sigma_a^2 / n_G\right)}{\left(\sigma_a^2 / n_G\right) + \left(\frac{\sigma_p^2 \text{-} (\sigma_a^2 / n_G)}{n_P}\right)}}\ . \tag{6.2}$$

Since each locus is assumed to explain only very little variance, $\sigma_p^2 \text{-} (\sigma_a^2 / n_G) \approx \sigma_p^2$. Due to the assumption that each locus is explaining an equal amount of the genetic variance, the accuracy of estimating the effect of one locus is the same for each of the loci, and represents the overall accuracy of estimating SNP effects:

$$r_{effect} = \sqrt{\frac{\left(\sigma_a^2 / n_G\right)}{\left(\sigma_a^2 / n_G\right) + \left(\sigma_p^2 / n_P\right)}} = \sqrt{\frac{h^2 n_P}{h^2 n_P + n_G}} \ . \tag{6.3}$$

Thus, this approach results in the same equation to predict the accuracy as derived by Daetwyler *et al.* (2008). The derivation described in Equations 6.2 and 6.3 is, however, much simpler, and an analogy of this derivation will be used to derive the accuracy of multi-population genomic prediction.

Similar to Daetwyler *et al*. (2008), we assume that n_G independent loci are underlying the trait and that each locus explains an equal amount of the genetic variance. The effects of the loci might be different in each population, which is measured by the genetic correlation between populations. Furthermore, we will assume that $n_{P,A}$ individuals from population *A* and $n_{P,B}$ individuals from population *B* with phenotype and genotype information are combined into one training population to estimate SNP effects. Those estimated SNP effects are then used to predict genomic values of individuals from population *C*, that could be a sample from one of the training populations or from a different population. The information from populations *A* and *B*, used to estimate SNP effects, is combined in a selection index approach (Hazel 1943), using the average phenotype of $n_{P,A}$ individuals from population *A* (x_A) and the average phenotype of $n_{P,B}$ individuals from population *B* (x_B) as records, and the genomic values of individuals from population *C* as breeding goal traits:

$$I_i = \hat{g}_{C_i} = b_A x_A + b_B x_B \, , \tag{6.4}$$

in which b_A and b_B are the regression coefficients on the average phenotype of individuals from population *A* (x_A) and *B* (x_B) to predict genomic values for individual *i* from population *C* ($\hat{g}_{C_i}$).

The regression coefficients of genomic values of individuals from population *C* on the average phenotype of population *A* and *B* can be calculated as:

$$\mathbf{b} = \begin{bmatrix} b_A \\ b_B \end{bmatrix} = \mathbf{P}^{-1}\mathbf{g} \, , \tag{6.5}$$

in which **P** is the (co)variance-matrix of x_A and x_B and **g** is a vector with covariances between x_A and x_B and the true genomic value of individual *i* from population *C* (g_{C_i}):

$$\mathbf{P} = \begin{bmatrix} Var(x_A) & Cov(x_A, x_B) \\ Cov(x_A, x_B) & Var(x_B) \end{bmatrix}, \tag{6.6}$$

and:

$$\mathbf{g} = \begin{bmatrix} Cov(x_A, g_{C_i}) \\ Cov(x_B, g_{C_i}) \end{bmatrix}. \tag{6.7}$$

In analogy with Wientjes *et al.* (2015b), the variance of the average phenotype of $n_{P,A}$ individuals can be partitioned into a part explained by one locus $\left(\sigma^2_{a_A} / n_G\right)$ and a part not explained by that locus $\left(\frac{\sigma^2_{p_A} - \left(\sigma^2_{a_A} / n_G\right)}{n_{P,A}} \approx \frac{\sigma^2_{p_A}}{n_{P,A}}\right)$, in which $\sigma^2_{a_A}$ is the total genetic variance in population *A* and $\sigma^2_{p_A}$ is the total phenotypic variance in population *A*. So, the total variance of x_A can be written as:

$$Var(x_A) = \frac{\sigma^2_{a_A}}{n_G} + \frac{\sigma^2_{p_A}}{n_{P,A}}. \tag{6.8}$$

Note that $\sigma^2_{p_A} / n_{P,A}$ represents the part of the phenotypic variance not explained by that locus, i.e., the residual variance ($\sigma^2_{e_{A,j}}$) for one locus *j*.

The covariance between the average phenotypes in the two populations can be partitioned in a part explained by one locus, a part not explained by that locus and twice the covariance between the two parts. In an additive model, $Cov(a,e) = 0$ and the parts not explained by a locus, i.e., the residual variances, are expected to be independent across populations, indicating that only the covariance between the populations of the part explained by one locus is assumed to differ from zero. Therefore, the covariance can be written as:

$$Cov(x_A, x_B) = r_{G_{A,B}} \frac{\sigma_{a_A}\sigma_{a_B}}{n_G}, \tag{6.9}$$

in which σ_{a_A} and σ_{a_B} are the genetic standard deviations in respectively population *A* and *B* and $r_{G_{A,B}}$ is the genetic correlation between population *A* and *B*. Hence:

$$\mathbf{P} = \begin{bmatrix} \frac{\sigma^2_{a_A}}{n_G} + \frac{\sigma^2_{p_A}}{n_{P,A}} & r_{G_{A,B}} \frac{\sigma_{a_A}\sigma_{a_B}}{n_G} \\ r_{G_{A,B}} \frac{\sigma_{a_A}\sigma_{a_B}}{n_G} & \frac{\sigma^2_{a_B}}{n_G} + \frac{\sigma^2_{p_B}}{n_{P,B}} \end{bmatrix}, \tag{6.10}$$

in which $\sigma^2_{a_B}$ is the total genetic variance in population *B* and $\sigma^2_{p_B}$ is the total phenotypic variance in population *B*.

Since an additive model is assumed, the covariance between the average phenotype of population *A* and the true genomic value of individual *i* from population *C* is also equal to the covariance between the populations of the part explained by one locus:

$$Cov(x_A, g_{C_i}) = r_{G_{A,C}} \frac{\sigma_{a_A}\sigma_{a_C}}{n_G}, \tag{6.11}$$

in which σ_{a_C} is the genetic standard deviation in population *C* and $r_{G_{A,C}}$ is the genetic correlation between population *A* and *C*. Hence:

$$\mathbf{g} = \begin{bmatrix} r_{G_{A,C}} \dfrac{\sigma_{a_A}\sigma_{a_C}}{n_G} \\ r_{G_{B,C}} \dfrac{\sigma_{a_B}\sigma_{a_C}}{n_G} \end{bmatrix}, \tag{6.12}$$

in which $r_{G_{B,C}}$ is the genetic correlation between population *B* and *C*. Substituting Equations 6.10 and 6.12 in Equation 6.5 results in:

$$\mathbf{b} = \mathbf{P}^{-1}\mathbf{g} = \begin{bmatrix} \dfrac{\sigma^2_{a_A}}{n_G} + \dfrac{\sigma^2_{p_A}}{n_{P,A}} & r_{G_{A,B}} \dfrac{\sigma_{a_A}\sigma_{a_B}}{n_G} \\ r_{G_{A,B}} \dfrac{\sigma_{a_A}\sigma_{a_B}}{n_G} & \dfrac{\sigma^2_{a_B}}{n_G} + \dfrac{\sigma^2_{p_B}}{n_{P,B}} \end{bmatrix}^{-1} \begin{bmatrix} r_{G_{A,C}} \dfrac{\sigma_{a_A}\sigma_{a_C}}{n_G} \\ r_{G_{B,C}} \dfrac{\sigma_{a_B}\sigma_{a_C}}{n_G} \end{bmatrix}. \tag{6.13}$$

With some algebra (see Appendix A), it can be shown that the accuracy of this selection index, representing the accuracy of estimating SNP effects, equals:

$$r_{HI} = r_{effect} = \sqrt{\frac{\mathbf{b'g}}{\sigma^2_H}} = \sqrt{\frac{\mathbf{g'P}^{-1}\mathbf{g}}{\left(\sigma^2_{a_C}/n_G\right)}}$$

$$= \sqrt{\begin{bmatrix} r_{G_{A,C}}\sqrt{\dfrac{h^2_A}{n_G}} & r_{G_{B,C}}\sqrt{\dfrac{h^2_B}{n_G}} \end{bmatrix} \begin{bmatrix} \dfrac{h^2_A}{n_G} + \dfrac{1}{n_{P,A}} & r_{G_{A,B}} \dfrac{\sqrt{h^2_A h^2_B}}{n_G} \\ r_{G_{A,B}} \dfrac{\sqrt{h^2_A h^2_B}}{n_G} & \dfrac{h^2_B}{n_G} + \dfrac{1}{n_{P,B}} \end{bmatrix}^{-1} \begin{bmatrix} r_{G_{A,C}}\sqrt{\dfrac{h^2_A}{n_G}} \\ r_{G_{B,C}}\sqrt{\dfrac{h^2_B}{n_G}} \end{bmatrix}}. \tag{6.14}$$

When only one population is included in the training population, Equation 6.14 reduces to:

$$r_{effect} = \sqrt{\left[r_{G_{A,C}}\sqrt{\frac{h^2_A}{n_G}}\right]\left[\frac{h^2_A}{n_G} + \frac{1}{n_{P,A}}\right]^{-1}\left[r_{G_{A,C}}\sqrt{\frac{h^2_A}{n_G}}\right]} = r_{G_{A,C}}\sqrt{\frac{h^2_A n_{P,A}}{h^2_A n_{P,A} + n_G}}. \tag{6.15}$$

This equation is equivalent to the equation of Wientjes *et al.* (2015b) for across-population genomic prediction. When estimated SNP effects are applied in another subset of the same population as the training population, i.e., $r_{G_{A,C}}$ is 1, Equation 6.15 becomes equivalent to the equation derived by Daetwyler *et al.* (2008) to predict the accuracy of estimating SNP effects within a population (Equation 6.1).

As explained before, the accuracy of genomic prediction depends on r_{effect} as well as on r_{LD}, accounting for the proportion of the genetic variance captured by the SNPs. It might for example be that the SNP effects are accurately estimated (r_{effect}=1), but when LD between QTL and SNPs is not complete, not all genetic variance can be captured by the SNPs and the accuracy of genomic prediction is still not 1. Moreover, when a number of QTL is segregating in the predicted population and not in the training population, part of the genetic variance in the predicted population can never be captured by the SNPs in the training population. Altogether, this indicates that the proportion of the genetic variance in the predicted population that can be captured by the SNPs in the training population is specific for a combination of training and predicted population. Therefore, r_{LD} affects the covariance between the phenotypes in the training population and the aggregated genotype of the predicted individuals (Equation 6.12), which results in:

$$\mathbf{g}=\begin{bmatrix} r_{LD_{A,C}}\left(r_{G_{A,C}}\dfrac{\sigma_{a_A}\sigma_{a_C}}{n_G}\right) \\ r_{LD_{B,C}}\left(r_{G_{B,C}}\dfrac{\sigma_{a_B}\sigma_{a_C}}{n_G}\right)\end{bmatrix}, \qquad (6.16)$$

in which $r_{LD_{A,C}}$ is the square root of the proportion of the genetic variance in predicted population *C* captured by the SNPs in training population *A*, and $r_{LD_{B,C}}$ is the square root of the proportion of the genetic variance in predicted population *C* captured by the SNPs in training population *B*. Using Equation 6.16 instead of Equation 6.12 in the remaining part of the derivation results in the following equation to predict the accuracy of genomic prediction:

$$r_{EGV}=\sqrt{\begin{bmatrix} r_{LD_{A,C}}r_{G_{A,C}}\sqrt{\dfrac{h_A^2}{n_G}} & r_{LD_{B,C}}r_{G_{B,C}}\sqrt{\dfrac{h_B^2}{n_G}}\end{bmatrix}\begin{bmatrix}\dfrac{h_A^2}{n_G}+\dfrac{1}{n_{P,A}} & r_{G_{A,B}}\dfrac{\sqrt{h_A^2h_B^2}}{n_G} \\ r_{G_{A,B}}\dfrac{\sqrt{h_A^2h_B^2}}{n_G} & \dfrac{h_B^2}{n_G}+\dfrac{1}{n_{P,B}}\end{bmatrix}^{-1}\begin{bmatrix} r_{LD_{A,C}}r_{G_{A,C}}\sqrt{\dfrac{h_A^2}{n_G}} \\ r_{LD_{B,C}}r_{G_{B,C}}\sqrt{\dfrac{h_B^2}{n_G}}\end{bmatrix}}.$$

(6.17)

In this study, $r_{LD_{A,C}}$ and $r_{LD_{B,C}}$ were assumed to be characteristics of the training and predicted populations, and depending on the SNP density and the properties of the QTL underlying the trait. Therefore, an empirical approach was needed to estimate values for $r_{LD_{A,C}}$ and $r_{LD_{B,C}}$. The values were estimated in the scenarios when only one population (*A* or *B*) was used as training population, by calculating r_{LD} as $r_{LD} = {}^{r_{EGV}}\!/_{r_{effect}}$, in which r_{EGV} was the empirical accuracy and r_{effect} the predicted accuracy assuming all genetic variance in the predicted population was captured by the SNPs. The empirically estimated values for $r_{LD_{A,C}}$ and $r_{LD_{B,C}}$ were used to predict the accuracy when population *A* and *B* were combined in the training population to predict genomic values for individuals from population *C*.

6.2.2 Derivation of M_e to replace n_G

An important assumption underlying the derived equation is that n_G independent loci are underlying the trait. In a finite population, loci do not segregate independently due to the existence of LD between loci. The equation predicting the accuracy of SNP effects using a single population (Equation 6.1), derived by Daetwyler *et al.* (2008), accounts for that by replacing n_G by the effective number of chromosome segments, M_e, in the population (Daetwyler *et al.* 2010). The M_e within a population is a statistical concept, and can be interpreted as the effective number of chromosome segments that are independently segregating in that population. Or in other words, it represents the effective number of effects that has to be estimated to predict genomic values for individuals from that population. In the derived equation for multi-population genomic prediction, different populations are combined in the training population, each with different values for M_e. For predicting genomic values for individuals from population *C*, using estimated SNP effects in population *A*, the effective number of estimated effects is equal to the effective number of chromosome segments shared between population *A* and *C* ($M_{e_{A,C}}$). Equivalently, when estimated SNP effects in population *B* are used, the effective number of estimated effects is equal to $M_{e_{B,C}}$. In analogy of M_e within a population, the M_e across populations can be interpreted as the effective number of segments that are segregating in a combined population, when considering the differences in LD between the populations. Therefore, we propose the following adjustment to Equation 6.17:

$$r_{EGV} = \sqrt{\begin{bmatrix} r_{LD_{A,C}} r_{G_{A,C}} \sqrt{\dfrac{h_A^2}{M_{e_{A,C}}}} & r_{LD_{B,C}} r_{G_{B,C}} \sqrt{\dfrac{h_B^2}{M_{e_{B,C}}}} \end{bmatrix} \begin{bmatrix} \dfrac{h_A^2}{M_{e_{A,C}}} + \dfrac{1}{n_{P,A}} & r_{G_{A,B}} \dfrac{\sqrt{h_A^2 h_B^2}}{\sqrt{M_{e_{A,C}} M_{e_{B,C}}}} \\ r_{G_{A,B}} \dfrac{\sqrt{h_A^2 h_B^2}}{\sqrt{M_{e_{A,C}} M_{e_{B,C}}}} & \dfrac{h_B^2}{M_{e_{B,C}}} + \dfrac{1}{n_{P,B}} \end{bmatrix}^{-1} \begin{bmatrix} r_{LD_{A,C}} r_{G_{A,C}} \sqrt{\dfrac{h_A^2}{M_{e_{A,C}}}} \\ r_{LD_{B,C}} r_{G_{B,C}} \sqrt{\dfrac{h_B^2}{M_{e_{B,C}}}} \end{bmatrix}}, \tag{6.18}$$

in which $M_{e_{A,C}}$ is the effective number of segments across population *A* and *C*, and $M_{e_{B,C}}$ the effective number of segments across population *B* and *C*. The same equation can also be derived when a selection index is used combining estimated genomic values for individuals from population *C* based on training populations of respectively population *A* or *B*, as is shown in Appendix B.

The M_e within a population can be calculated as (Goddard *et al.* 2011):

$$M_e = \frac{1}{Var(\mathbf{G}_{ij} - E(\mathbf{G}_{ij}))}, \tag{6.19}$$

in which $\mathbf{G}_{ij}$ contains the genomic relationship and $E(\mathbf{G}_{ij})$ the expected values for the genomic relationships between all individuals *i* and *j* from that population, with the variance taken over all pair-wise relationships between individuals *i* and *j*. In analogy to Equation 6.19, the values for M_e across populations can be calculated using (Wientjes *et al.* 2015b):

$$M_{e_{1,2}} = \frac{1}{Var(\mathbf{G}_{Pop.1_i,Pop.2_j} - E(\mathbf{G}_{Pop.1_i,Pop.2_j}))}, \tag{6.20}$$

in which $\mathbf{G}_{Pop.1_i,Pop.2_j}$ contains the genomic relationships and $E(\mathbf{G}_{Pop.1_i,Pop.2_j})$ contains the expected values for the genomic relationships between all individuals *i* from population 1 and individuals *j* from population 2, again the variance is taken over all pair-wise relationships between individuals *i* and *j*. The genomic relationships can be calculated following Yang *et al.* (2010), by calculating the genomic relationships between individual *i* from population *y* and individual *j* from population *z* as $G_{y_i,z_j} = \frac{1}{n}\sum_k G_{(y_i,z_j)_k} = \frac{1}{n}\sum_k \frac{(x_{y_ik} - 2p_{yk})(x_{z_jk} - 2p_{zk})}{\sqrt{2p_{yk}(1-p_{yk})}\sqrt{2p_{zk}(1-p_{zk})}}$, and the genomic relationship of individual *i* from population *y* with itself as $G_{y_{ii}} = \frac{1}{n}\sum_k G_{(y_{ii})_k} = 1 + \frac{1}{n}\sum_k \frac{x_{y_ik}^2 - (1+2p_{yk})x_{y_ik} + 2p_{yk}^2}{2p_{yk}(1-p_{yk})}$, in which *n* is the number of SNPs, x_{y_ik} and x_{z_jk} are the genotypes at locus *k* coded as 0, 1, and 2, and p_{yk} and

p_{zk} are the allele frequencies for the second allele (with homozygote genotype coded as 2) at locus *k* for respectively population *y* and *z*. The genomic relationships used to calculate *M_e* are based on population-specific allele frequencies to ensure that unrelated individuals have an expected genomic relationship of 0, which is an underlying assumption of the equation to calculate M_e (Goddard *et al.* 2011).

In most human studies, individuals included in the data are unrelated (e.g., Yang *et al.* 2010; Lee *et al.* 2012; Maier *et al.* 2015). This indicates that the expected values for all genomic relationships (*E*(**G**)) would approximately be zero, so Equation 6.20 simplifies to $M_{e_{1,2}} = \frac{1}{Var(\mathbf{G}_{Pop.1_i,Pop.2_j})}$. In most livestock populations, individuals are related, so *E*(**G**) would not be zero and *E*(**G**) could be approximated by the pedigree relationship matrix **A**, i.e., $M_{e_{1,2}} = \frac{1}{Var(\mathbf{G}_{Pop.1_i,Pop.2_j} - \mathbf{A}_{Pop.1_i,Pop.2_j})}$. When both the **G** and **A** matrix are used to calculate M_e, both matrices should be scaled to the same base population. This can be achieved by rescaling the inbreeding level in **G** to the inbreeding in **A**, for example by using the following adjustment separately for each of the within-population and across-population blocks (Powell *et al.* 2010):

$$\mathbf{G}^* = \left(1 - \overline{F_b}\right)\mathbf{G} + 2\overline{F_b}\,\mathbf{J}, \tag{6.21}$$

in which $\overline{F_b}$ is the average pedigree inbreeding level of individuals in population *b* and **J** is a matrix filled with ones.

The **G**-*E*(**G**) values are expected to follow a normal distribution around zero for each value of *E*(**G**). The pedigree relationships between individuals in **A**, however, depend on the depth of the pedigree for both individuals. In general, the pedigree relationships will more closely resemble *E*(**G**) when the pedigree for both individuals is deeper. When the pedigree is not deep or complete enough for all or a subset of the individuals, extra variation in **G**-**A** is introduced, resulting in an underestimation of M_e when **A** is used to represent *E*(**G**). Since the depth of the pedigree can differ across individuals, the impact of an insufficient pedigree depth on the calculated M_e can be reduced by only taking the relationships of individuals with the most complete pedigree into account to calculate M_e. To check if selecting those individuals indeed minimized the impact of an insufficient pedigree depth, values of **G**-**A** can be plotted versus values of **A**. When the values for **G**-**A** are lower for higher **A** values, as is shown in Figure 6.1, an insufficient pedigree depth is still influencing the calculation of M_e. To account for this particular pattern, an exponential function was fitted through the data. For all values of **A** in the data, the parameters of the function were estimated in R (R Development Core Team 2011) and the fitted values of the function were subtracted from the values of **G**-**A** before calculating M_e.

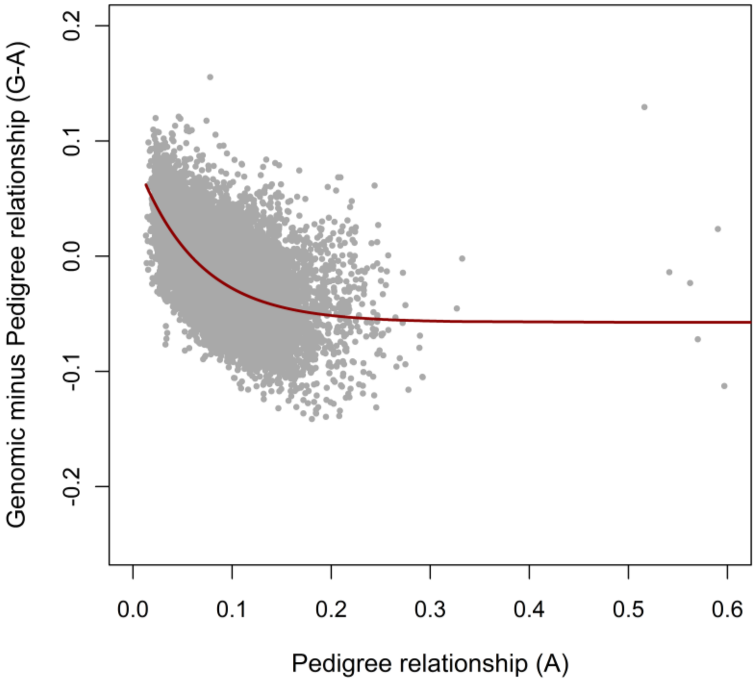


Figure 6.1 The genomic minus pedigree relationships (**G**-**A**) versus the pedigree relationships (**A**) for across population elements between individuals of two populations. The red line is the fitted exponential function ($f = a + 1/e^{bx+c}$) used to correct G-A values to reduce the impact of an insufficient pedigree depth.

6.2.3 Validation

After deriving the equation, the aim was to validate it for a broad range of scenarios, differing in heritabilities within populations and genetic correlations between populations. Those scenarios resemble combining populations from different environments or measured for different traits. For the validation, real genotypes and simulated phenotypes were used. In each of the scenarios, an empirical accuracy was calculated and compared with the predicted accuracy using the derived equation to investigate how accurate the accuracy was predicted. In this part, the used genotype information, the simulated phenotypes and the estimation of the empirical accuracy is explained. The genotype and pedigree information from all individuals, as well as the simulated phenotypes are available upon request.

6.2.3.1 Genotypes

Genotypes were available for 1033 dairy cows from the Netherlands, each originating for at least 87.5% from the Holstein Friesian breed, i.e., all animals were pure-bred Holstein Friesians. Genotyping was done using the Illumina BovineSNP50 Beadchip (50k, Illumina, San Diego, CA), after which genotypes were imputed to higher density (777k) using 3150 Holstein Friesian animals as reference population (Pryce *et al.* 2014). The accuracy of imputation across imputed loci, as reflected by the Beagle R^2 value, was on average 0.96, indicating high imputation accuracy. As quality control, SNPs with a call rate smaller than 95%, an unknown mapping position, located on the sex chromosomes, a minor allele frequency (MAF) <0.005, for which only two genotypes were observed, and in complete linkage disequilibrium with a neighboring SNP were deleted. This quality control step reduced the number of SNPs for this study to 422,405.

A total of 50,000 candidate QTL were selected from the 422,405 SNPs, and in each replicate QTL were randomly sampled from the candidate QTL to simulate phenotypes for each individual. The candidate QTL were selected from the SNPs using two different approaches: 1) Candidate QTL were randomly selected (RANDOM), and 2) Candidate QTL were selected from the SNPs with a MAF below 0.2 (LOW MAF), since the MAF of QTL underlying complex traits is expected to be lower than the MAF of SNPs (Goddard and Hayes 2009; Yang *et al.* 2010; Kemper and Goddard 2012) due to ascertainment bias of the SNPs on the SNP chips (Matukumalli *et al.* 2009). For each of the two approaches, the remaining 372,405 SNPs were used as markers. In this way, the QTL underlying a trait could be randomly sampled from the candidate QTL in each of the replicates, while the subset of SNP markers was constant across replicates for both RANDOM and LOW MAF.

6.2.3.2 Phenotypes

The 1033 individuals were divided into three groups to represent different populations. The first two groups (population 1 and 2) contained 450 individuals and represented the different training populations (population *A* and *B* in the derived equation). The last group (population 3) contained 133 individuals and represented the group of predicted individuals for which genomic values were estimated (population *C* in the derived equation). The division over the groups was performed using pedigree information, by allocating paternal and maternal half-sib families to the same population. In this way, relationships within a population were higher than between populations, as usually would be expected for distinct populations.

For both the RANDOM and LOW MAF approach of selecting candidate QTL, phenotypes were simulated by randomly sampling 4000 QTL from the group of 50,000 candidate QTL. The QTL underlying the trait were the same in each of the populations. For each QTL, allele substitution effects were sampled from a multivariate normal distribution, with a mean of 0 and standard deviation of 1, using different genetic correlations between the populations. Only additive effects and no dominance or epistatic interactions were assumed. True genomic values (TGVs) were calculated by multiplying the QTL genotypes, coded as 0, 1 and 2, by the simulated allele substitution effects of the population to which the individual belonged. Across populations, the TGVs were rescaled to a mean of 0 and variance of 1. In each of the populations, the genetic variance was calculated as the variance of the TGVs for the individuals from that population. For all individuals, the environmental effect was sampled from $N(0, \left(\frac{1}{h^2}-1\right) * \ Var(TGV_i))$, in which $Var(TGV_i)$ is the variance of TGV in population *i* to which the individual belonged. For each individual, the simulated TGV and environmental effect was summed to calculate the phenotype.

6.2.3.3 Scenarios

Seven different scenarios of multi-population genomic prediction were investigated, differing in heritabilities and genetic correlations between the populations (Table 6.1). The first four scenarios represent multi-environment genomic prediction, where populations in different environments were combined in one training population in which SNP effects were estimated. In those scenarios, the heritability was assumed to be the same in each population (0.95), but genetic correlations between populations varied from 0.4 to 1. The last three scenarios represent multi-trait genomic prediction, where populations measured for different traits are combined in one training population. In those scenarios, each population had a different heritability of 0.3 or 0.95 and genetic correlations between populations were 0.6 or 1. The values for the heritabilities of 0.3 and 0.95 were chosen to have a clear contrast between the populations.

Table 6.1 Overview of the different scenarios to simulate phenotypes.

	Heritability		**Genetic correlation**		
Scenarios[a]	**Pop. 1**	**Pop. 2**	**Pop. 1 - 2**	**Pop. 1 - 3**	**Pop. 2 - 3**
Same heritability					
SH1.0-0.6	0.95	0.95	0.60	1.00	0.60
SH0.8-0.6	0.95	0.95	0.60	0.80	0.60
SH0.8-0.4	0.95	0.95	0.60	0.80	0.40
SH0.4-0.4	0.95	0.95	0.60	0.40	0.40
Different heritabilities					
DH1.0-1.0	0.95	0.30	1.00	1.00	1.00
DH1.0-0.6	0.95	0.30	0.60	1.00	0.60
DH0.6-1.0	0.95	0.30	0.60	0.60	1.00

[a] Scenarios are labeled as follows: The names of the scenarios with the same heritability in each population start with **SH**, followed by the genetic correlation between population 1 and 3, and the genetic correlation between population 2 and 3. The names of scenarios with different heritabilities in each population start with **DH**, followed by the genetic correlation between population 1 and 3, and the genetic correlation between population 2 and 3.

In each scenario, population 1, population 2, or population 1 and 2 were used as training population and population 3 contained the predicted individuals. Each scenario was analyzed using both approaches of selecting QTL; RANDOM and LOW MAF. Simulations were replicated 100 times in each scenario.

6.2.3.4 Calculating M_e

Values for M_e across the different populations were calculated based on the difference between the genomic and pedigree relationship matrix. Since the subset of SNPs slightly differed between the two approaches of selecting candidate QTL, RANDOM and LOW MAF, values for M_e were calculated for each of the approaches. To reduce the impact of incompleteness of the pedigree, only individuals with at least 3 generations of complete pedigree were taken into account, resulting in 329 individuals in population 1, 270 individuals in population 2, and 90 individuals in population 3. Thereafter, an exponential function was fitted through the data to further reduce the impact of an insufficient pedigree depth, as explained before. The **G** matrix was the same for all replicates, since the subset of 372,405 SNPs was constant for all replicates while QTL were re-sampled every replicate, resulting in the same M_e for all replicates. Therefore, only one accuracy could be predicted for all replicates of the same approach of selecting candidate QTL, representing the expected average accuracy of estimating SNP effects.

6.2.3.5 Empirical accuracy of genomic prediction

The empirical accuracies of genomic prediction were obtained both with a single-trait and a multi-trait GBLUP type of model run in ASReml (Gilmour *et al.* 2009) using the simulated phenotypes. In both models, population was included as fixed effect to account for differences in mean phenotype between populations. Genomic values for the predicted individuals were estimated using a genomic relationship matrix, **G**, containing all training and predicted individuals, and simulated phenotypes of the training individuals. The **G** matrix included in the models was calculated using the allele frequencies across all individuals without taking the population into account. The other steps in calculating **G** were the same as explained above.

In the single-trait model, variances were estimated using REML. Therefore, the model used was termed GREML instead of GBLUP, where variances are assumed to be known. In the single-trait model, the phenotypes of the different populations were pooled in one population, without taking the genetic correlations between the populations into account. The differences in heritability were, however, taken into account by weighting the phenotypes differently and in this way acknowledging that the phenotypes in one population were more accurately representing the genomic values of the individuals compared to the phenotypes in the other population. It was assumed that the heritability of the phenotypes from the population with the lowest heritability, i.e., a heritability of 0.3, represented the trait heritability based on one measurement. The phenotypes of the individuals from this population were given a weight of 1. The heritability of the other population, i.e., a heritability of 0.95, represented the heritability based on multiple measurements of the same trait. Or in other words, it represented the reliability of the phenotype based on more than one record. This indicates that the genetic variance can be assumed to be the same in both populations. The weight for the phenotypes of individuals from the population with the highest reliability (r^2) was equal to the ratio of the residual variances in both populations, which can be calculated as:

$$w = \frac{1-h^2}{h^2/_{r^2} - h^2} . \tag{6.22}$$

Following Equation 6.22, a weight of 44.33 was given to the phenotypes from the population with a heritability of 0.95. One possible scenario where phenotypes could be weighted differently is in dairy cattle populations, where phenotypes of cows are generally based on one single measurement and phenotypes of bulls are based on different numbers of progeny, for which the same weights can be obtained following Garrick *et al.* (2009).

The multi-trait model considered the phenotypes for the same trait in the different populations as different traits with a genetic correlation between the traits. Estimating all genetic correlations in the multi-trait model was not possible, since phenotypes of the predicted individuals were not included in the model. Therefore, genetic correlations and variance components were assumed to be known and fixed to the simulated values, and the multi-trait model was termed GBLUP.

For each of the models, the accuracy of genomic prediction was calculated as the correlation between the simulated TGVs and predicted genomic values. Note that the single- and multi-trait GBLUP models use both SNP information and simulated phenotypes, that differed across the replicates. Therefore, averages and standard errors across the replicates were calculated and compared to the predicted accuracies.

6.2.4 Evaluating the potential accuracies of two scenarios

The derived equation can be used to investigate the accuracy of different scenarios of multi-population genomic prediction. To show the potential of the equation for this aim, we used Equation 6.18 to evaluate the potential accuracy for two specific scenarios, assuming that all genetic variance in the predicted population was captured by the SNPs in the training population ($r_{LD_{A,C}} = r_{LD_{B,C}} = 1$).

The first scenario is relevant for dairy cattle breeding, where bulls with deregressed estimated genetic values based on daughter information are in general used in the training population, with a heritability equal to the reliability of the estimated genetic values. Different studies have investigated the potential to increase the accuracy of genomic prediction by adding cows to the training population with their own phenotypes, that are in general less reliable than estimated genetic values (e.g., Calus *et al.* 2013; Cooper *et al.* 2015). This approach was studied using the prediction equation (Equation 6.18) when different numbers of cows (range 0 to 50,000) were added to a training population of 10,000 bulls, assuming a heritability of 0.05 for the phenotypes of cows which is representing the heritability of a fertility trait in dairy cattle (e.g., Karoui *et al.* 2012), different reliabilities (range 0 to 1) for the estimated genetic values of bulls, and a genetic correlation of 1 between the estimated genetic values of bulls and own phenotypes of cows. The values for M_e were set to the values derived from the cattle genotype data used in this study.

The second scenario is representing a scenario in human studies, in which it was assumed that different numbers of individuals from a population from African descent (range 0 to 100,000) were added to a training population of 5000 individuals from European descent to increase the accuracy of predicting genetic risk for the European population. As an example, parameters for the trait schizophrenia were used, with a heritability of 0.28 in the European population, a heritability of 0.24 in the African population and a genetic correlation of 0.66 between the populations (De Candia *et al.* 2013). The M_e in the European population ($M_{e_{A,C}}$ in Equation 6.18) was set to 43,000, based on the equation $M_e = \frac{2N_eL}{\ln(4N_eL)}$ (Goddard 2009), an effective population size (N_e) of 10,000 (McEvoy *et al.* 2011), and a genome length (L) of 30 Morgan (Venter *et al.* 2001). The M_e across the populations ($M_{e_{B,C}}$ in Equation 6.18) was varied (range 43,000 to 2,000,000).

6.3 Results

In this section, the results of the prediction equation are first presented assuming that all genetic variance in the predicted population (population 3) is captured by the SNPs in the training population. Those predicted accuracies were used to calculate $r_{LD_{1,3}}$ and $r_{LD_{2,3}}$ based on the ratio between the empirical and predicted accuracy of genomic prediction when only one of the populations, population 1 or population 2, was used as training population. As a next step, the calculated values for $r_{LD_{1,3}}$ and $r_{LD_{2,3}}$ were used to predict the accuracy of genomic prediction when population 1 and 2 were combined in the training population.

6.3.1 Calculating M_e

In Table 6.2, the different estimated M_e values across populations are shown. Due to only small differences in the subset of SNPs used to calculate **G**, estimated M_e values were very similar for the scenarios with QTL randomly sampled (RANDOM) and QTL sampled with a low MAF (LOW MAF). Using population-specific allele frequencies or allele frequencies across populations only had a very small effect on the estimated values for M_e, as well as on the predicted accuracies (range -0.9%– +1.3%). This indicates that, for this study, the use of population-specific allele frequencies or the allele frequency across populations did not influence the results, due to the very similar allele frequencies across the three populations. Therefore, the predicted accuracies are only shown for the M_e values calculated based on a **G** matrix using the allele frequencies across the populations.

Table 6.2 Estimated M_e values across populations using population-specific allele frequencies or the allele frequency across populations to set-up **G**.

Scenario	Population-specific allele frequency	Allele frequency across populations
QTL with low MAF		
Population 1 - 3	1541	1515
Population 2 - 3	1616	1652
QTL randomly sampled		
Population 1 - 3	1620	1585
Population 2 - 3	1694	1741

6.3.2 Scenarios with QTL randomly sampled (RANDOM)

In this part, results are presented for the RANDOM scenarios of simulating phenotypes. For those scenarios, the predicted accuracies and average empirical accuracies of genomic prediction obtained with a single-trait model using either a single or combined training population and different scenarios of simulated phenotypes, are shown in Figure 6.2. The first four scenarios show the accuracies when different genetic correlations between the populations were simulated, with the same heritability in each of the populations. Those scenarios show that when only one population was used as training population, predicted and empirical accuracies were, as expected, higher when the genetic correlation between training and predicted individuals was higher. There was only a small difference between the accuracies obtained using population 1 or 2 as training population when the genetic correlation with the predicted individuals was the same, because both populations were about equally related to the predicted individuals. Combining the two populations in one training population always resulted in an increase in both predicted and empirical accuracy. The magnitude of the increase in accuracy depended on the genetic correlation between the predicted individuals and the added population; the higher the genetic correlation, the higher the increase in accuracy.

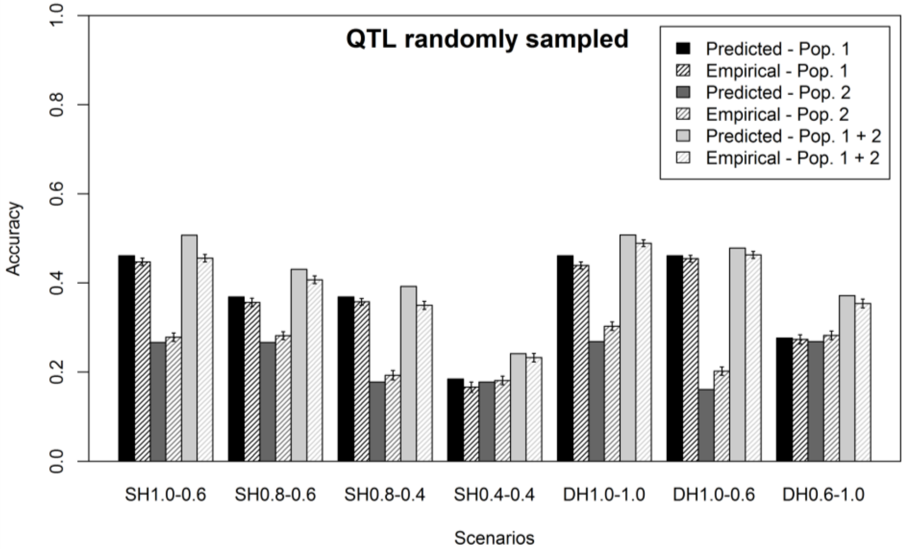


Figure 6.2 Predicted and empirical accuracies of genomic prediction (± standard errors) using a single-trait model, one or two populations in the training population, QTL randomly sampled from the SNPs, and assuming in the prediction equation that all genetic variance in the predicted population was captured by the SNPs in the training population. The different scenarios represent the different genetic correlations and heritabilities used to simulate phenotypes. The scenarios starting with **SH** have the same heritability in the two training populations, the scenarios starting with **DH** have a different heritability. For each scenario, **SH** or **DH** is followed by the genetic correlation between population 1 and 3, and the genetic correlation between population 2 and 3.

The last three scenarios show the predicted and empirical accuracies using different heritabilities in each of the populations and genetic correlations of 1 or 0.6 between populations. Those scenarios show that when only one population was used as training population, predicted and empirical accuracies were, as expected, higher when the heritability in the training population was higher. For this study, a heritability of 0.3 resulted in approximately 60% of the accuracy obtained with a heritability of 0.95. Adding 450 individuals from the population with a low heritability to a training population of 450 individuals from the population with a high heritability, however, still resulted in an increase in accuracy. The increase in both predicted and empirical accuracy was again lower when the genetic correlation was lower, similar to the scenarios with the same heritability in each population.

For each of the scenarios, the predicted accuracy of genomic prediction shown in Figure 6.2 is assuming that $r_{LD_{1,3}} = r_{LD_{2,3}}$ =1. In general, predicted accuracies were very slightly overestimating the empirical accuracies of genomic prediction (±1%), both when the heritability was the same in each population and when the heritability was different. When population 1 was used as training population, the overestimation was on average 4% (range 1% – 11%). When population 2 was used as training population, the empirical accuracy was slightly underestimated by the predicted accuracy with on average 8% (range -20% – -2%). When both populations were combined in the training population, the overestimation was on average 6% (range 3% – 12%). Those results indicate that when QTL were randomly sampled from the SNPs, most of the genetic variance in the predicted individuals was tagged by the SNPs in the training population, especially when population 2 was used as training population, and the estimated value for $r_{LD_{1,3}}$ was 0.96 and for $r_{LD_{2,3}}$ 1. Using those calculated values to predict the accuracy of genomic prediction for the combined training population reduced the overestimation of the empirical accuracy to 3%.

6.3.3 Scenarios sampling QTL with low MAF (LOW MAF)

In this part, results are presented for the LOW MAF scenarios of simulating phenotypes. For those scenarios, the predicted and average empirical accuracies of genomic prediction obtained with a single-trait model using either a single or combined training population are shown in Figure 6.3, assuming $r_{LD_{1,3}} = r_{LD_{2,3}}$ =1. All empirical accuracies for the LOW MAF scenarios were lower than the accuracies obtained for the RANDOM scenarios. The predicted accuracies, however, were similar to the predicted accuracies for the RANDOM scenarios. So, the predicted accuracies for the LOW MAF scenarios overestimated the empirical accuracies to a greater extent. On average, the overestimation was ±15%, and again higher when population 1 was used as training population, compared to using population 2 as training population (population 1: 20%; population 2: 7%; combined training population: 20%). Those results indicate that, as expected, a smaller proportion of the genetic variance in the predicted individuals was tagged by the SNPs in the training population when QTL were sampled with a low MAF and the estimated value for $r_{LD_{1,3}}$ was 0.84 and for $r_{LD_{2,3}}$ 0.94. Using those calculated values to predict the accuracy of genomic prediction for the combined training population, reduced the overestimation of the empirical accuracy to 5%.

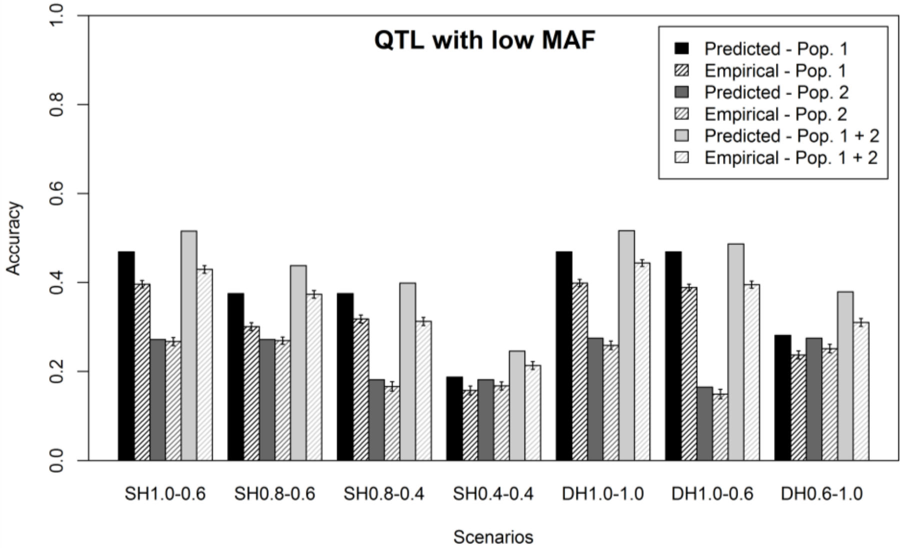


Figure 6.3 Predicted and empirical accuracies of genomic prediction (± standard errors) using a single-trait model, one or multiple populations in the training population, QTL sampled with a low minor allele frequency (MAF), and assuming in the prediction equation that all genetic variance in the predicted population was captured by the SNPs in the training population. The different scenarios represent the different genetic correlations and heritabilities used to simulate phenotypes. The scenarios starting with **SH** have the same heritability in the two training populations, the scenarios starting with **DH** have a different heritability. For each scenario, **SH** or **DH** is followed by the genetic correlation between population 1 and 3, and the genetic correlation between population 2 and 3.

6.3.4 Single-trait versus multi-trait model

The analyses using a combined training population were performed using both a single-trait model as well as a multi-trait model, where the same trait in the different populations was modelled as a different correlated trait. The accuracies from both models are shown in Figure 6.4, for the (**A**) RANDOM, as well as for the (**B**) LOW MAF scenarios. In this figure, the predicted accuracies for the combined training populations use the estimated values of $r_{LD_{1,3}}$ and $r_{LD_{2,3}}$, estimated when only population 1 or 2 was included in the training population. In general, the accuracies obtained with the multi-trait model were equal to or higher than the accuracies obtained with the single-trait model, depending on the genetic correlations. When the genetic correlations between both training populations and the predicted population were the same, accuracies obtained with the single- and multi-trait model were similar. When the genetic correlations were different,

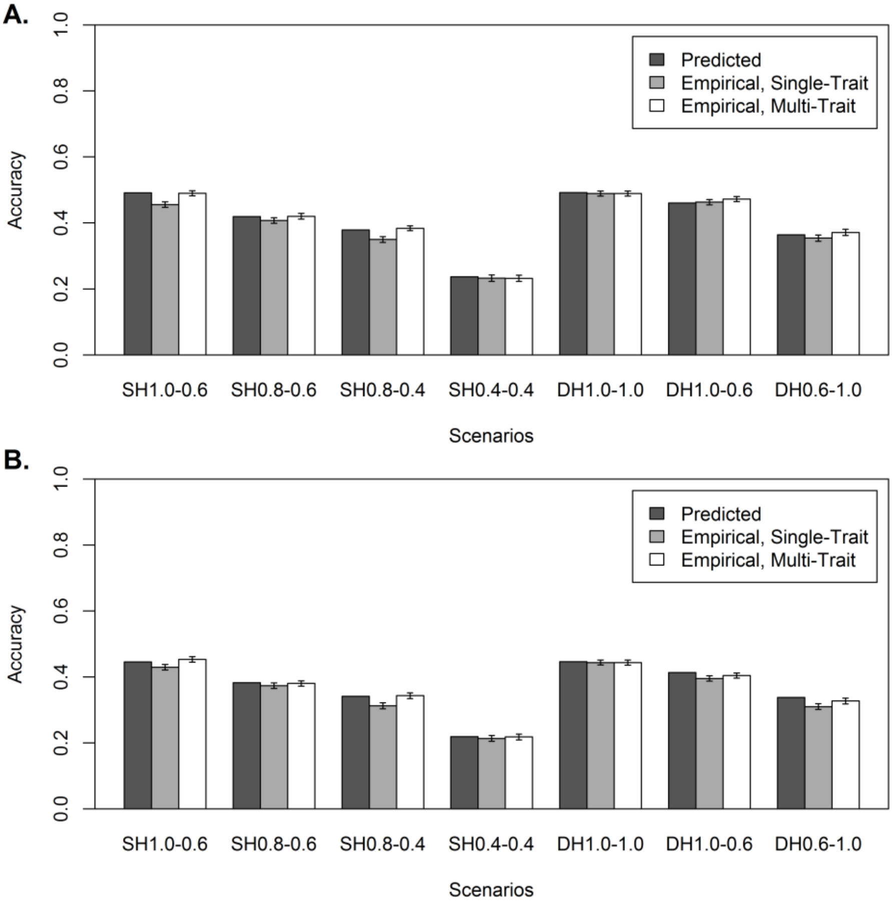


Figure 6.4 Predicted and empirical accuracies of genomic prediction (± standard errors) using a training population consisting of two populations and QTL (A) randomly sampled, or (B) with a low minor allele frequency, and accounting for the proportion of genetic variance in the predicted population captured by the SNPs in the training population in the prediction equation. Empirical accuracies were either obtained with a single-trait model or a multi-trait model. The different scenarios represent the different genetic correlations and heritabilities used to simulate phenotypes. The scenarios starting with **SH** have the same heritability in the two training populations, the scenarios starting with **DH** have a different heritability. For each scenario, **SH** or **DH** is followed by the genetic correlation between population 1 and 3, and the genetic correlation between population 2 and 3.

the predicted accuracy of genomic prediction using the estimated values of $r_{LD_{1,3}}$ and $r_{LD_{2,3}}$ reduced on average across replicates to 0% (range -2% to +2%) for the RANDOM scenarios and to 1% (range -2% to +3%) for the LOW MAF scenarios. This indicates that the equation can accurately predict the accuracy of genomic prediction when the proportion of the genetic variance in the predicted population not captured by the SNPs in the training population is known and taken into account.

6.3.5 The potential accuracies of two scenarios

The potential accuracies when cows with own phenotypes are added to a training population of 10,000 bulls with deregressed estimated genetic values, is shown in Figure 6.5, for different numbers of cows added to the training population and different reliabilities for the estimated genetic values. This figure shows that when the reliability of the estimated genetic values of the bulls was low, relatively a small amount of cows had to be added to the training population to see a substantial increase in accuracy. When the reliability of the estimated genetic values was high (above 0.7), a high accuracy was already obtained with 10,000 bulls in the training population (accuracies were above 0.9), and enlarging the training population by adding cows with own phenotypes only resulted in a minor increase in accuracy.

The potential accuracies for the human scenario where a population from African descent is added to a training population of European descent to predict the genetic risk of individuals from the European population is shown in Figure 6.6, with different numbers of individuals from the African population added to the training population and different values for M_e across the populations. This figure shows that when M_e across the two populations was low, adding individuals from another population could substantially improve the accuracy of predicting genetic risk. When the M_e across the two populations was large (>20 times the M_e within the European population), adding individuals from the other population only resulted in a minor increase in accuracy. This indicates that to improve the accuracy of predicting genomic values, using training individuals from populations that are more closely related and have a more consistent LD pattern, resulting in lower values for M_e across populations, is more beneficial than using training individuals from populations that are only distantly related.

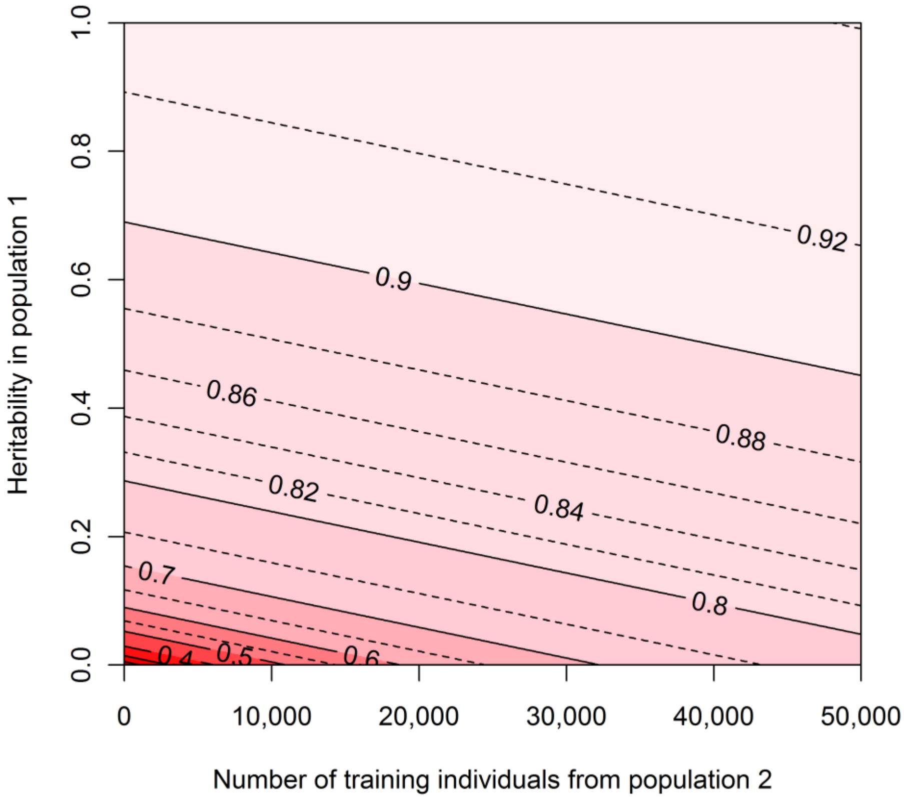


Figure 6.5 Predicted accuracies with different numbers of individuals from population 2 added to a training population consisting of 10,000 individuals from population 1 with different heritabilities for the trait. The input parameters represent a scenario in dairy cattle were a cow population with own phenotypes (population 2) was added to a bull population with estimated genetic values based on daughter information (population 1). Due to different numbers of daughters used to estimate genetic values for the bulls, the heritability or reliability of the phenotype in population 1 ranged between 0 and 1. The heritability for the trait in population 2 was 0.05, and genetic correlations between the training populations and between both training populations and the predicted population were 1. The values for M_e were equal to the values in the simulations ($M_{e_{1,3}}$ = 1620, $M_{e_{2,3}}$ = 1694).

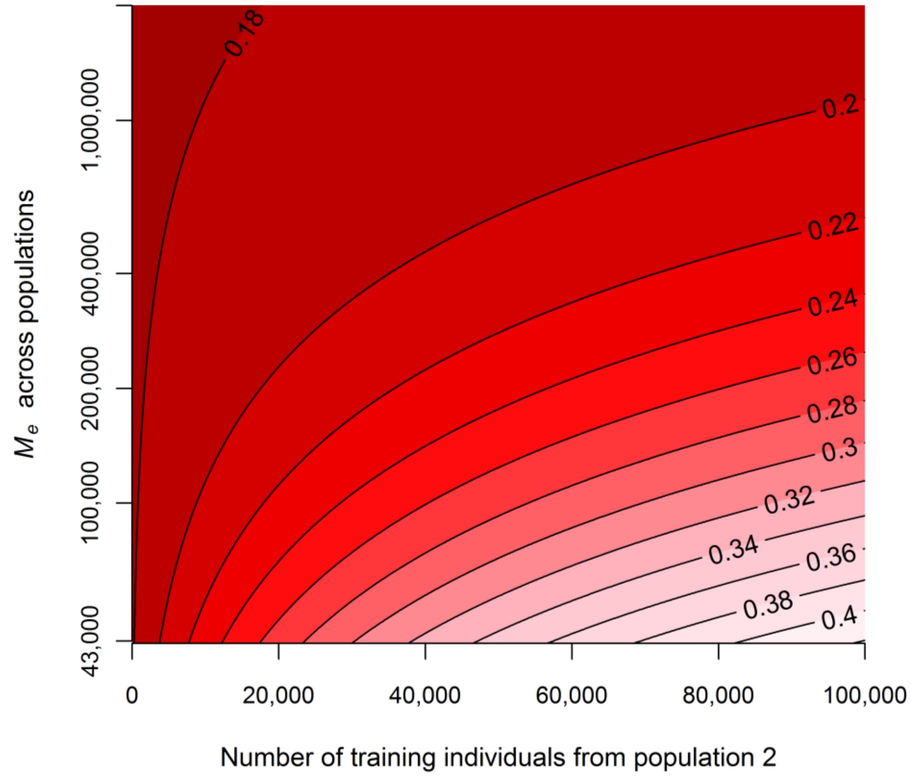


Figure 6.6 Predicted accuracies with different numbers of individuals from population 2 added to a training population consisting of individuals from population 1 with different values for the effective number of chromosome segments, $M_{e,}$ across population 1 and 2. The input parameters represent a human scenario where a population from African descent (population 2) was added to a population from European descent (population 1) to predict the genetic risk for Schizophrenia in the European population (population 3 = population 1), with heritabilities of 0.28 in population 1 and 0.24 in population 2 and a genetic correlation of 0.66 between populations 1 and 2 (De Candia *et al.* 2013). The M_e in population 1 was set to 43,000, based on the equation $M_e = \frac{2N_eL}{\ln(4N_eL)}$ (Goddard 2009) and an effective population size of 10,000 (McEvoy *et al.* 2011).

6.4 Discussion

In this paper, a deterministic equation was derived using population parameters to predict the accuracy of genomic values when different populations are combined in the training population. The equation was validated in this study using simulations to resemble the combining of populations from different environments and measured for different correlated traits, i.e., multi-environment and multi-trait genomic prediction, with different heritabilities in each population and genetic correlations between populations different from 1. In all simulated scenarios, the equation was able to accurately predict the accuracy of genomic prediction when the proportion of the genetic variance in the predicted population captured by the SNPs in the training population was known and taken into account.

For the validation of the derived equation, real cattle genotypes from Dutch Holstein Friesian cows, divided in three populations based on the pedigree, and simulated phenotypes were used. The simulations showed that the equation is able to handle heterogeneous data in different populations, such as differences in heritability in each population and genetic correlations between populations different from 1. In principle, the equation can handle data from more divergent populations as well, such as populations from different environments, breeds or lines. The proportion of the genetic variance captured by the SNPs can, however, be expected to be lower across more divergent populations, as will be discussed later. To confirm that the equation indeed gives accurate predictions for those other scenarios when the proportion of the genetic variance captured by the SNPs is known, further validation of the equation is required using a broader range of populations, preferably with real genotype and phenotype information.

6.4.1 Potential of the derived equation

The equation gives insight in important parameters for multi-population genomic prediction and can be used to compare different scenarios. The equation for example shows that when the M_e across populations is two times higher than M_e within a population, two times more individuals from the other population have to be added to obtain the same increase in accuracy when the heritabilities are the same, the genetic correlations between populations is 1, and all genetic variance can be captured. When those last criteria are not met, even more individuals from the other population have to be added to obtain the same increase in accuracy.

Another way in which the equation can be used is to investigate the potential accuracy of different scenarios, as was done in Figure 6.5 and 6.6. In Figure 6.6, the equation was applied to a scenario where human populations from European and African descent were combined in one training population to predict Schizophrenia risk for the European population, a scenario that was suggested by de Candia *et al.* (2013). The results show that when the LD pattern is very different across populations, resulting in a high M_e across populations, it is very unlikely to see an increase in prediction accuracy, even when a lot of individuals from the other population are added. Moreover, it shows that the sensitivity of the accuracy for M_e is much smaller at larger values of M_e across populations compared to small values of M_e, which is in agreement with the results found within a population (Brard and Ricard 2015). The equation can be used in the same way to investigate other scenarios of multi-population genomic prediction. Estimates for the input parameters, such as the M_e across predicted and training populations, the heritability of the trait in each of the training populations, the genetic correlations between the populations (r_G), and the part of the genetic variance in the predicted population captured by the SNPs in the training population (r_{LD}) should, however, be known. Apart from the heritability, for which estimates are straightforward to calculate, each of the input parameters and how to estimate values for those parameters will be discussed in more detail in the following paragraphs.

6.4.2 Effective number of chromosome segments (M_e)

In the derived prediction equation, M_e across populations is an important parameter. This parameter can be interpreted as a statistical concept and represents the effective number of segments that are segregating in a combined population, which is a measure for the effective number of effects that has to be estimated in one population to predict genomic values for individuals from another population. It depends on the consistency in LD between the populations; when the LD pattern is completely different between the populations, each of the segments has to be very small to segregate in both populations, resulting in a large M_e across the populations. The importance of this parameter indicates that the predicted population influences the accuracy of estimating SNP effects in the training population. Consider for example a situation where one population is included in the training population to predict a trait that is influenced by one QTL having two SNPs in complete LD in that training population. For predicting genomic values within a subset of the same population, it does not matter to which of the SNPs the effect of the QTL is allocated. When the estimated SNP effects are used in another population, for which only one of the SNPs is in complete LD and the other is completely independent from the QTL, it is important to which of the SNPs the effect of the QTL is allocated. When the effect is equally distributed across the two SNPs, only half of the effect of the QTL is captured for that population, which reduces the accuracy. This indicates that the accuracy of estimating SNP effects in the training population is indeed depending on the predicted population, which is reflected in the M_e across populations.

It is good to note that M_e represents the number of effects that have to be estimated in a GBLUP type of model, basically assuming an infinitesimal model. When a Bayesian variable selection model is used, the number of estimated effects is only equal to M_e when the effective number of QTL underlying the trait is larger than M_e, otherwise the number of estimated effects is equal to the effective number of QTL (Daetwyler *et al.* 2010; Van den Berg *et al.* 2015). This indicates that when the number of QTL is substantially lower than M_e and a Bayesian variable selection model is used, the number of estimated effects is equal to the effective number of QTL, which is the value that should be used in the equation to predict the accuracy of genomic values.

As shown in this study as well as in other studies (Goddard *et al.* 2011; Wientjes *et al.* 2013; Wientjes *et al.* 2015b), the value for M_e can be calculated using the relationship matrices based on genomic information and pedigree information. This indicates that when a small subset, for example 100 individuals with pedigree information, from each population is genotyped, estimates for M_e can be obtained. The value for M_e within population can also be obtained based on the effective population size and genome size, for which different equations exist (Goddard 2009; Hayes *et al.* 2009b; Goddard *et al.* 2011). The different equations, however, result in quite different estimates for M_e (Wientjes *et al.* 2013; Brard and Ricard 2015). Moreover, it is not possible to use the equations based on effective population size to estimate the value for M_e across populations. In general, the value for M_e across populations can be expected to be higher than within populations (Wientjes *et al.* 2013; Wientjes *et al.* 2015b), since M_e is depending on the strength of LD between loci (Goddard *et al.* 2011), and LD is at least partly different across populations (Sawyer *et al.* 2005; De Roos *et al.* 2008; Veroneze *et al.* 2013; Wientjes *et al.* 2015c). In this study, the estimated M_e within a population was around 1350 for all three populations. The values for M_e across populations were approximately 20% higher and around 1600. In a study using different breeds, the M_e values across populations were reported to be around 10 times larger than M_e within a population (Wientjes *et al.* 2015b), which is a result of the lower variation in relationships across breeds than across populations of the same breed.

6.4.3 Genetic correlation between populations (r_G)

Another input parameter is the genetic correlation between the populations, which is the correlation between the allele substitution effects of the QTL. In a simulation study with at least 100 individuals in each of the populations, it was shown that this parameter can accurately be estimated using a genomic multi-trait model, where the same trait in different populations was treated as a different trait (Wientjes *et al.* 2015b). For closely related populations with an overlapping pedigree, such as populations in different countries that have some common co-ancestry, the genetic correlation can also be estimated using a pedigree relationship matrix (Schaeffer 1994). For more distantly related populations, such as different breeds or lines, the pedigree would probably not be deep enough to capture the relationships across populations and a relationship matrix based on genomic information is required (Karoui *et al.* 2012; Huang *et al.* 2014).

6.4.4 Genetic variance captured by the SNPs (r_{LD})

Results of this study show that the empirical accuracy of genomic prediction was depending on the MAF of the QTL underlying the simulated trait; when QTL had on average a lower MAF than the SNPs, the accuracy reduced. This is in agreement with results of other studies using single- or multi-population genomic prediction (Daetwyler *et al.* 2013; Wientjes *et al.* 2015a). The reason for this is a decrease in the strength of LD between QTL and SNPs when the MAF of QTL is lower than the MAF of SNPs (Khatkar *et al.* 2008; Yan *et al.* 2009; Wientjes *et al.* 2015c), reducing the proportion of the genetic variance captured by the SNPs. As stated before, the MAF of QTL underlying complex traits is expected to be lower than the MAF of SNPs (Goddard and Hayes 2009; Yang *et al.* 2010; Kemper and Goddard 2012), indicating that it is highly likely that not all the genetic variance can be captured by the SNPs in real data.

The square root of the proportion of the genetic variance captured by the SNPs is represented in the prediction equation as r_{LD}, and is depending on the density of the SNP chip, the characteristics of the QTL underlying the trait, and the investigated populations (Daetwyler 2009; Erbe *et al.* 2013). This parameter can only be estimated based on empirical data, by comparing the predicted and empirical accuracy. Using this approach, r_{LD} was estimated to be around 1 when QTL were randomly sampled from the SNPs and around 0.85 when QTL had a low MAF in this study. In other studies using real data, the square of r_{LD}, i.e., r_{LD}^2, was estimated to be around 0.8 using a 50k chip in Holstein Friesian dairy populations for Net Merit (Daetwyler 2009) and production traits (Erbe *et al.* 2013), and slightly lower in Brown Swiss dairy populations for production traits (Erbe *et al.* 2013; Román-Ponce *et al.* 2014). The studies estimating r_{LD}^2 only focused on one population. Across populations, the value for r_{LD} is supposed to be lower and depending on the number of generations since the separation of the populations; the higher the number of generations, the lower the consistency in LD (e.g., Andreescu *et al.* 2007; De Roos *et al.* 2008) and the higher the chance on QTL segregating in only one population (Kemper *et al.* 2015). Therefore, the values of $\sqrt{0.8}$ =0.89 for r_{LD} found in the empirical studies can probably be seen as the upper limit of r_{LD}, which can only be obtained when the predicted and training population are subsets from the same population. The more divergent the predicted and training population are, the lower the value of r_{LD} and the further away the value is from the upper limit of r_{LD} within a population.

6.4.5 Single-trait versus multi-trait model

Empirical accuracies were obtained using both a single-trait model as well as a multi-trait model. The results showed that the use of a multi-trait model was beneficial when the genetic correlation between the two training populations and the predicted population was different. In an empirical study with three different chicken lines with different genetic correlations between populations, a multi-trait model resulted in more or less similar accuracies than a single-trait model (Huang *et al.* 2014). In an empirical study with three dairy cattle breeds, a multi-trait model using estimated genetic correlations resulted in more or less similar accuracies than a multi-trait model with genetic correlations fixed at 0.95 (Karoui *et al.* 2012). The combining of dairy cattle populations from three different countries, however, showed a higher accuracy for a multi-trait model compared to a single-trait model (De Haas *et al.* 2012). So, empirical studies have shown that multi-trait models yield similar or slightly higher accuracies than single-trait models, however, genetic correlations were generally estimated with large standard errors.

The observed increase in accuracy of using a multi-trait model when genetic correlations between the two training populations and the predicted population were different can be explained as follows. When the genetic correlations are different, it is beneficial to take into account that estimated SNP effects from one training population are more related to SNP effects in the predicted population than estimated SNP effects from the other training population. When the genetic correlation was the same, the use of a multi-trait model was not beneficial, even not when the genetic correlation among the training populations was different from 1. This can be explained by the fact that estimated SNP effects in each of the training populations are equally related to SNP effects in the predicted population. In the single-trait model, averages of the SNP effects in both training populations are estimated, which have the same correlation with the SNP effects in the predicted population as the SNP effects in each of the training populations. Therefore, taking the genetic correlation between the training populations into account had no effect on the obtained accuracy for those scenarios.

6.5 Conclusion

A deterministic equation is derived to predict the accuracy of genomic values when the training population comprises individuals of different populations, such as different breeds, lines or environments, or populations measured for different traits. In this study, the equation was validated for different multi-environment and multi-trait scenarios. Results showed that the accuracy of estimating genomic values can be accurately predicted for those scenarios, provided that the effective number of chromosome segments across predicted and training populations, the heritability of the trait in each of the training populations, the genetic correlations between the populations, and the proportion of the genetic variance in the predicted population captured by the SNPs in the training population are known. Therefore, the derived equation can be used to investigate the potential accuracy of different multi-population genomic prediction scenarios and to decide on the most optimal design of training populations.

6.6 Acknowledgements

The authors are thankful for useful comments from Chris Schrooten and Henk Bovenhuis. This study was financially supported by Breed4Food (KB-12-006.03-005-ASG-LR), a public-private partnership in the domain of animal breeding and genomics, and CRV (Arnhem, The Netherlands). The RobustMilk project and the National Institute of Food and Agriculture (NIFA) are acknowledged for providing the 50k genotypes of the HF cows, and the gDMI consortium is acknowledged for imputing those to 777k genotypes.

6.7 Appendix

6.7.1 Appendix A: Deriving the accuracy of estimating SNP effects in a combined training population

The accuracy of the selection index, representing the accuracy of estimating the effect of one locus, can be calculated as:

$$r_{HI} = r_{effect} = \sqrt{\frac{\mathbf{b'g}}{\sigma_H^2}} = \sqrt{\frac{\mathbf{g'P^{-1}g}}{\left(\sigma_{a_C}^2 / n_G\right)}}$$

$$= \sqrt{\begin{bmatrix} r_{G_{A,C}} \dfrac{\sigma_{a_A}}{n_G} & r_{G_{B,C}} \dfrac{\sigma_{a_B}}{n_G} \end{bmatrix} \begin{bmatrix} \dfrac{\sigma_{a_A}^2}{n_G} + \dfrac{\sigma_{p_A}^2}{n_{P,A}} & r_{G_{A,B}} \dfrac{\sigma_{a_A}\sigma_{a_B}}{n_G} \\ r_{G_{A,B}} \dfrac{\sigma_{a_A}\sigma_{a_B}}{n_G} & \dfrac{\sigma_{a_B}^2}{n_G} + \dfrac{\sigma_{p_B}^2}{n_{P,B}} \end{bmatrix}^{-1} \begin{bmatrix} r_{G_{A,C}} \dfrac{\sigma_{a_A}}{n_G} \\ r_{G_{B,C}} \dfrac{\sigma_{a_B}}{n_G} \end{bmatrix} n_G}$$

$$= \sqrt{\begin{bmatrix} r_{G_{A,C}} \dfrac{\sigma_{a_A}}{\sqrt{n_G}} & r_{G_{B,C}} \dfrac{\sigma_{a_B}}{\sqrt{n_G}} \end{bmatrix} \begin{bmatrix} \dfrac{\sigma_{a_A}^2}{n_G} + \dfrac{\sigma_{p_A}^2}{n_{P,A}} & r_{G_{A,B}} \dfrac{\sigma_{a_A}\sigma_{a_B}}{n_G} \\ r_{G_{A,B}} \dfrac{\sigma_{a_A}\sigma_{a_B}}{n_G} & \dfrac{\sigma_{a_B}^2}{n_G} + \dfrac{\sigma_{p_B}^2}{n_{P,B}} \end{bmatrix}^{-1} \begin{bmatrix} r_{G_{A,C}} \dfrac{\sigma_{a_A}}{\sqrt{n_G}} \\ r_{G_{B,C}} \dfrac{\sigma_{a_B}}{\sqrt{n_G}} \end{bmatrix}} \,. \qquad \text{(A6.1)}$$

For simplicity, we will start by referring to the first element of this inversed **P** matrix as A, to the off-diagonal elements as B and to the last element as C. Hence, Equation A6.1 can be written as:

$$r_{effect} = \sqrt{\begin{bmatrix} r_{G_{A,C}} \dfrac{\sigma_{a_A}}{\sqrt{n_G}} & r_{G_{B,C}} \dfrac{\sigma_{a_B}}{\sqrt{n_G}} \end{bmatrix} \begin{bmatrix} A & B \\ B & C \end{bmatrix} \begin{bmatrix} r_{G_{A,C}} \dfrac{\sigma_{a_A}}{\sqrt{n_G}} \\ r_{G_{B,C}} \dfrac{\sigma_{a_B}}{\sqrt{n_G}} \end{bmatrix}}$$

$$= \sqrt{\left[\left(r_{G_{A,C}} \frac{\sigma_{a_A}}{\sqrt{n_G}} A + r_{G_{B,C}} \frac{\sigma_{a_B}}{\sqrt{n_G}} B\right) r_{G_{A,C}} \frac{\sigma_{a_A}}{\sqrt{n_G}} + \left(r_{G_{A,C}} \frac{\sigma_{a_A}}{\sqrt{n_G}} B + r_{G_{B,C}} \frac{\sigma_{a_B}}{\sqrt{n_G}} C\right) r_{G_{B,C}} \frac{\sigma_{a_B}}{\sqrt{n_G}}\right]}$$

$$= \sqrt{\left[r_{G_{A,C}}^2 \frac{\sigma_{a_A}^2}{n_G} A + 2 r_{G_{B,C}} \frac{\sigma_{a_B}}{\sqrt{n_G}} r_{G_{A,C}} \frac{\sigma_{a_A}}{\sqrt{n_G}} B + r_{G_{B,C}}^2 \frac{\sigma_{a_B}^2}{n_G} C\right]} \,. \qquad \text{(A6.2)}$$

The inverse of the **P** matrix can be written as:

$$\begin{bmatrix} \frac{\sigma^2_{a_A}}{n_G}+\frac{\sigma^2_{p_A}}{n_{P,A}} & r_{G_{A,B}}\frac{\sigma_{a_A}\sigma_{a_B}}{n_G} \\ r_{G_{A,B}}\frac{\sigma_{a_A}\sigma_{a_B}}{n_G} & \frac{\sigma^2_{a_B}}{n_G}+\frac{\sigma^2_{p_B}}{n_{P,B}} \end{bmatrix}^{-1}$$

$$=\frac{1}{\left(\frac{\sigma^2_{a_A}}{n_G}+\frac{\sigma^2_{p_A}}{n_{P,A}}\right)\left(\frac{\sigma^2_{a_B}}{n_G}+\frac{\sigma^2_{p_B}}{n_{P,B}}\right)-\left(r_{G_{A,B}}\frac{\sigma_{a_A}\sigma_{a_B}}{n_G}\right)^2}\begin{bmatrix} \frac{\sigma^2_{a_B}}{n_G}+\frac{\sigma^2_{p_B}}{n_{P,B}} & -r_{G_{A,B}}\frac{\sigma_{a_A}\sigma_{a_B}}{n_G} \\ -r_{G_{A,B}}\frac{\sigma_{a_A}\sigma_{a_B}}{n_G} & \frac{\sigma^2_{a_A}}{n_G}+\frac{\sigma^2_{p_A}}{n_{P,A}} \end{bmatrix}$$

$$=\begin{bmatrix} \frac{\frac{\sigma^2_{a_B}}{n_G}+\frac{\sigma^2_{p_B}}{n_{P,B}}}{\left(\frac{\sigma^2_{a_A}}{n_G}+\frac{\sigma^2_{p_A}}{n_{P,A}}\right)\left(\frac{\sigma^2_{a_B}}{n_G}+\frac{\sigma^2_{p_B}}{n_{P,B}}\right)-\left(r_{G_{A,B}}\frac{\sigma_{a_A}\sigma_{a_B}}{n_G}\right)^2} & \frac{-r_{G_{A,B}}\frac{\sigma_{a_A}\sigma_{a_B}}{n_G}}{\left(\frac{\sigma^2_{a_A}}{n_G}+\frac{\sigma^2_{p_A}}{n_{P,A}}\right)\left(\frac{\sigma^2_{a_B}}{n_G}+\frac{\sigma^2_{p_B}}{n_{P,B}}\right)-\left(r_{G_{A,B}}\frac{\sigma_{a_A}\sigma_{a_B}}{n_G}\right)^2} \\ \frac{-r_{G_{A,B}}\frac{\sigma_{a_A}\sigma_{a_B}}{n_G}}{\left(\frac{\sigma^2_{a_A}}{n_G}+\frac{\sigma^2_{p_A}}{n_{p,A}}\right)\left(\frac{\sigma^2_{a_B}}{n_G}+\frac{\sigma^2_{p_B}}{n_{P,B}}\right)-\left(r_{G_{A,B}}\frac{\sigma_{a_A}\sigma_{a_B}}{n_G}\right)^2} & \frac{\frac{\sigma^2_{a_A}}{n_G}+\frac{\sigma^2_{p_A}}{n_{P,A}}}{\left(\frac{\sigma^2_{a_A}}{n_G}+\frac{\sigma^2_{p_A}}{n_{P,A}}\right)\left(\frac{\sigma^2_{a_B}}{n_G}+\frac{\sigma^2_{p_B}}{n_{P,B}}\right)-\left(r_{G_{A,B}}\frac{\sigma_{a_A}\sigma_{a_B}}{n_G}\right)^2} \end{bmatrix}.$$

(A6.3)

Hence, Equation A6.2 can be written as:

$$r_{effect}=\sqrt{\frac{r^2_{G_{A,C}}\frac{\sigma^2_{a_A}}{n_G}\left(\frac{\sigma^2_{a_B}}{n_G}+\frac{\sigma^2_{p_B}}{n_{P,B}}\right)-2r_{G_{B,C}}\frac{\sigma_{a_B}}{\sqrt{n_G}}r_{G_{A,C}}\frac{\sigma_{a_A}}{\sqrt{n_G}}r_{G_{A,B}}\frac{\sigma_{a_A}\sigma_{a_B}}{n_G}+r^2_{G_{B,C}}\frac{\sigma^2_{a_B}}{n_G}\left(\frac{\sigma^2_{a_A}}{n_G}+\frac{\sigma^2_{p_A}}{n_{P,A}}\right)}{\left(\frac{\sigma^2_{a_A}}{n_G}+\frac{\sigma^2_{p_A}}{n_{P,A}}\right)\left(\frac{\sigma^2_{a_B}}{n_G}+\frac{\sigma^2_{p_B}}{n_{P,B}}\right)-\left(r_{G_{A,B}}\frac{\sigma_{a_A}\sigma_{a_B}}{n_G}\right)^2}}.$$

(A6.4)

Dividing both the numerator and the denominator by $\sigma^2_{p_A}$ and $\sigma^2_{p_B}$, results in:

$$r_{effect} = \sqrt{\frac{r^2_{G_{A,C}}\frac{h_A^2}{n_G}\left(\frac{h_B^2}{n_G}+\frac{1}{n_{P,B}}\right) - 2r_{G_{B,C}}\frac{\sqrt{h_B^2}}{\sqrt{n_G}}r_{G_{A,C}}\frac{\sqrt{h_A^2}}{\sqrt{n_G}}r_{G_{A,B}}\frac{\sqrt{h_A^2}\sqrt{h_B^2}}{n_G} + r^2_{G_{B,C}}\frac{h_B^2}{n_G}\left(\frac{h_A^2}{n_G}+\frac{1}{n_{P,A}}\right)}{\left(\frac{h_A^2}{n_G}+\frac{1}{n_{P,A}}\right)\left(\frac{h_B^2}{n_G}+\frac{1}{n_{P,B}}\right) - \left(r_{G_{A,B}}\frac{\sqrt{h_A^2}\sqrt{h_B^2}}{n_G}\right)^2}}$$

$$= \sqrt{\begin{bmatrix} r_{G_{A,C}}\sqrt{\frac{h_A^2}{n_G}} & r_{G_{B,C}}\sqrt{\frac{h_B^2}{n_G}} \end{bmatrix} \begin{bmatrix} \frac{h_A^2}{n_G}+\frac{1}{n_{P,A}} & r_{G_{A,B}}\frac{\sqrt{h_A^2h_B^2}}{n_G} \\ r_{G_{A,B}}\frac{\sqrt{h_A^2h_B^2}}{n_G} & \frac{h_B^2}{n_G}+\frac{1}{n_{P,B}} \end{bmatrix}^{-1} \begin{bmatrix} r_{G_{A,C}}\sqrt{\frac{h_A^2}{n_G}} \\ r_{G_{B,C}}\sqrt{\frac{h_B^2}{n_G}} \end{bmatrix}} . \qquad \text{(A6.5)}$$

Since each locus is assumed to explain the same amount of the genetic variance, the accuracy of estimating the effect of one SNP is the same for each of the SNPs, and represents the overall accuracy of estimating SNP effects (r_{effect}).

6.7.2 Appendix B: Alternative way of deriving the prediction equation

In this section, an alternative derivation of the prediction equation is presented. In this derivation, the estimated genomic values for population *C* based on two different training populations (population *A* and population *B*), are combined in a selection index to calculate the estimated genomic values for population *C* when the two populations would be combined in one training population. The estimated genomic value for individual *i* from population *C* (EGV_{A,C_i}) can be calculated using the estimated marker effects in a training population of population *A*, following:

$$EGV_{A,C_i} = r_{G_{A,C}} \sum_j X_{C_{i,j}} \hat{\beta}_{A_j} , \qquad \text{(B6.1)}$$

in which $r_{G_{A,C}}$ is the genetic correlation between population *A* and *C*, $X_{C_{i,j}}$ is the genotype of individual *i* from population *C* for marker *j*, and $\hat{\beta}_{A_j}$ is the estimated effect of marker *j* in population *A*. In an equivalent way, the estimated genomic value for individual *i* from population *C* can be calculated using the estimated markers effects in a training population of population B, i.e., EGV_{B,C_i} .

Both estimated genomic values, EGV_{A,C_i} and EGV_{B,C_i} , can be combined in a selection index to estimate the genomic value for individual *i* from population *C* when both population *A* and *B* would be combined in the training population (EGV_{A+B,C_i}), following:

$$EGV_{A+B,C_i} = b_A EGV_{A,C_i} + b_B EGV_{B,C_i}\,, \tag{B6.2}$$

in which b_A and b_B are the regression coefficients on EGV_{A,C_i} and EGV_{B,C_i} to predict the estimated genomic value for individual *i* from population *C* for the combined training population (EGV_{A+B,C_i}).

The regression coefficients on EGV_{A,C_i} and EGV_{B,C_i} that would maximize the estimation of the genomic value for individual *i* from population *C* can be calculated as:

$$\mathbf{b} = \begin{bmatrix} b_A \\ b_B \end{bmatrix} = \mathbf{P}^{-1}\mathbf{g}\,, \tag{B6.3}$$

in which **P** is the (co)variance-matrix between the information sources EGV_{A,C_i} and EGV_{B,C_i} , and **g** is a vector with covariances between the information sources, EGV_{A,C_i} and EGV_{B,C_i} , and the true genomic value for individual *i* from population C (TGV_{C_i}):

$$\mathbf{P} = \begin{bmatrix} Var(EGV_{A,C_i}) & Cov(EGV_{A,C_i}, EGV_{B,C_i}) \\ Cov(EGV_{A,C_i}, EGV_{B,C_i}) & Var(EGV_{B,C_i}) \end{bmatrix}, \tag{B6.4}$$

and:

$$\mathbf{g} = \begin{bmatrix} Cov(EGV_{A,C_i}, TGV_{C_i}) \\ Cov(EGV_{B,C_i}, TGV_{C_i}) \end{bmatrix}. \tag{B6.5}$$

In the following part, we will assume that the variances of the estimated and true genomic values are scaled, such that the true genomic values in population *C* have a variance of 1. The variance of the estimated genomic values for population *C* using population *A* in the training population is then equal to the reliability of predicting genomic values for population *C*:

$$Var(EGV_{A,C_i}) = r^2_{EGV_{A,C}}\,. \tag{B6.6}$$

The covariance between EGV_{A,C_i} and EGV_{B,C_i} can be written as:

$$Cov\left(EGV_{A,C_i}, EGV_{B,C_i}\right) = Cov\left(r_{G_{A,C}} \sum_j x_{C_{i,j}} \hat{\beta}_{A_j}, r_{G_{B,C}} \sum_j x_{C_{i,j}} \hat{\beta}_{B_j} \right)$$

$$= r_{G_{A,C}} r_{G_{B,C}} Cov\left(\sum_j x_{C_{i,j}} \hat{\beta}_{A_j}, \sum_j x_{C_{i,j}} \hat{\beta}_{B_j} \right) = r_{G_{A,C}} r_{G_{B,C}} Cov\left(\sum_j \hat{\beta}_{A_j}, \sum_j \hat{\beta}_{B_j} \right). \tag{B6.7}$$

The covariance between the effects marker estimated in population *A* and *B* can be written as:

$$Cov\left(\sum_j \hat{\beta}_{A_j}, \sum_j \hat{\beta}_{B_j}\right) = r_{\hat{\beta}_{A_j},\hat{\beta}_{B_j}} \sqrt{Var(\hat{\beta}_{A_j})Var(\hat{\beta}_{B_j})} \; . \qquad \text{(B6.8)}$$

Using the path coefficient method as described by Dekkers (2007), it can be shown that the correlation between the estimated marker effects is equal to:

$$r_{\hat{\beta}_{A_j},\hat{\beta}_{B_j}} = r_{G_{A,B}} r_{effect\,A} r_{effect\,B} \; , \qquad \text{(B6.9)}$$

in which $r_{G_{A,B}}$ is the genetic correlation between population *A* and *B*, and $r_{effect\,A}$ and $r_{effect\,B}$ are the accuracies of estimating the marker effects in respectively population *A* and *B*. The square root of the variance of the estimated marker effects in each of the populations is equal to the accuracy of the estimated marker effects, i.e., $\sqrt{Var(\hat{\beta}_{A_j})} = r_{effect\,A}$, therefore:

$$Cov\left(\sum_j \hat{\beta}_{A_j}, \sum_j \hat{\beta}_{B_j}\right) = r_{G_{A,B}} r_{effect\,A} r_{effect\,B} r_{effect\,A} r_{effect\,B} = r_{G_{A,B}} r^2_{effect\,A} r^2_{effect\,B} \; . \qquad \text{(B6.10)}$$

And:

$$Cov\left(EGV_{A,C_i}, EGV_{B,C_i}\right) = r_{G_{A,C}} r_{G_{B,C}} r_{G_{A,B}} r^2_{effect\,A} r^2_{effect\,B} \; . \qquad \text{(B6.11)}$$

The accuracy of estimating marker effects in population *A* multiplied by the genetic correlation between population *A* and *C* equals the accuracy of the estimated genomic values, i.e., $r_{EGV_{A,C}} = r_{G_{A,C}} r_{Effect\,A}$, under the assumption that all genetic variance of the predicted population is captured by the training populations. Hence, the covariance can be written as:

$$Cov\left(EGV_{A,C_i}, EGV_{B,C_i}\right) = r_{G_{A,B}} \frac{r^2_{EGV_{A,C}} r^2_{EGV_{B,C}}}{r_{G_{A,C}} r_{G_{B,C}}} \; . \qquad \text{(B6.12)}$$

Hence, **P** can be written as:

$$\mathbf{P} = \begin{bmatrix} r^2_{EGV_{A,C}} & r_{G_{A,B}} \dfrac{r^2_{EGV_{A,C}} r^2_{EGV_{B,C}}}{r_{G_{A,C}} r_{G_{B,C}}} \\ r_{G_{A,B}} \dfrac{r^2_{EGV_{A,C}} r^2_{EGV_{B,C}}}{r_{G_{A,C}} r_{G_{B,C}}} & r^2_{EGV_{B,C}} \end{bmatrix} . \qquad \text{(B6.13)}$$

The covariance between the estimated genomic values for individual *i* from population *C* using population *A* as training population is also equal to the reliability of predicting genomic values for population *C,* i.e., $Cov\left(EGV_{A,C_i}, TGV_{C_i}\right) = r^2_{EGV_{A,C}}$. Hence, **g** can be written as:

$$\mathbf{g} = \begin{bmatrix} r^2_{EGV_{A,C}} \\ r^2_{EGV_{B,C}} \end{bmatrix}. \tag{B6.14}$$

Since it is assumed that the variance of the true genomic values in population *C* is scaled to 1, the accuracy of this selection index, representing the accuracy of estimating genomic values for population *C* based on a training population of population *A* and *B*, can be calculated as:

$$r_{EGV_{A+B,C}} = \sqrt{\frac{\mathbf{g}'\mathbf{P}^{-1}\mathbf{g}}{\left(\sigma^2_{a_C}\right)}} = \sqrt{\mathbf{g}'\mathbf{P}^{-1}\mathbf{g}}$$

$$= \sqrt{\begin{bmatrix} r^2_{EGV_{A,C}} & r^2_{EGV_{B,C}} \end{bmatrix} \begin{bmatrix} r^2_{EGV_{A,C}} & r_{G_{A,B}} \dfrac{r^2_{EGV_{A,C}} r^2_{EGV_{B,C}}}{r_{G_{A,C}} r_{G_{B,C}}} \\ r_{G_{A,B}} \dfrac{r^2_{EGV_{A,C}} r^2_{EGV_{B,C}}}{r_{G_{A,C}} r_{G_{B,C}}} & r^2_{EGV_{B,C}} \end{bmatrix}^{-1} \begin{bmatrix} r^2_{EGV_{A,C}} \\ r^2_{EGV_{B,C}} \end{bmatrix}}. \tag{B6.15}$$

For simplicity, we will start by referring to the first element of matrix $\mathbf{P}^{-1}$ as A, to the off-diagonal elements as B and to the last element as C. Hence, Equation B6.15 can be written as:

$$r_{EGV_{A+B,C}} = \sqrt{\begin{bmatrix} r^2_{EGV_{A,C}} & r^2_{EGV_{B,C}} \end{bmatrix} \begin{bmatrix} A & B \\ B & C \end{bmatrix} \begin{bmatrix} r^2_{EGV_{A,C}} \\ r^2_{EGV_{B,C}} \end{bmatrix}}$$

$$= \sqrt{(r^2_{EGV_{A,C}} A + r^2_{EGV_{B,C}} B) r^2_{EGV_{A,C}} + (r^2_{EGV_{A,C}} B + r^2_{EGV_{B,C}} C) r^2_{EGV_{B,C}}}$$

$$= \sqrt{r^4_{EGV_{A,C}} A + 2 r^2_{EGV_{B,C}} B r^2_{EGV_{A,C}} + r^4_{EGV_{B,C}} C}\ . \tag{B6.16}$$

The matrix $\mathbf{P}^{-1}$ can be written as:

$$\begin{bmatrix} r^2_{EGV_{A,C}} & r_{G_{A,B}} \dfrac{r^2_{EGV_{A,C}} r^2_{EGV_{B,C}}}{r_{G_{A,C}} r_{G_{B,C}}} \\ r_{G_{A,B}} \dfrac{r^2_{EGV_{A,C}} r^2_{EGV_{B,C}}}{r_{G_{A,C}} r_{G_{B,C}}} & r^2_{EGV_{B,C}} \end{bmatrix}^{-1}$$

$$= \frac{1}{r^2_{EGV_{A,C}} r^2_{EGV_{B,C}} - \left(r_{G_{A,B}} \dfrac{r^2_{EGV_{A,C}} r^2_{EGV_{B,C}}}{r_{G_{A,C}} r_{G_{B,C}}} \right)^2} \begin{bmatrix} r^2_{EGV_{B,C}} & -r_{G_{A,B}} \dfrac{r^2_{EGV_{A,C}} r^2_{EGV_{B,C}}}{r_{G_{A,C}} r_{G_{B,C}}} \\ -r_{G_{A,B}} \dfrac{r^2_{EGV_{A,C}} r^2_{EGV_{B,C}}}{r_{G_{A,C}} r_{G_{B,C}}} & r^2_{EGV_{A,C}} \end{bmatrix}$$

$$= \begin{bmatrix} \dfrac{r^2_{EGV_{B,C}}}{r^2_{EGV_{A,C}} r^2_{EGV_{B,C}} - \left(r_{G_{A,B}} \dfrac{r^2_{EGV_{A,C}} r^2_{EGV_{B,C}}}{r_{G_{A,C}} r_{G_{B,C}}} \right)^2} & \dfrac{-r_{G_{A,B}} \dfrac{r^2_{EGV_{A,C}} r^2_{EGV_{B,C}}}{r_{G_{A,C}} r_{G_{B,C}}}}{r^2_{EGV_{A,C}} r^2_{EGV_{B,C}} - \left(r_{G_{A,B}} \dfrac{r^2_{EGV_{A,C}} r^2_{EGV_{B,C}}}{r_{G_{A,C}} r_{G_{B,C}}} \right)^2} \\ \dfrac{-r_{G_{A,B}} \dfrac{r^2_{EGV_{A,C}} r^2_{EGV_{B,C}}}{r_{G_{A,C}} r_{G_{B,C}}}}{r^2_{EGV_{A,C}} r^2_{EGV_{B,C}} - \left(r_{G_{A,B}} \dfrac{r^2_{EGV_{A,C}} r^2_{EGV_{B,C}}}{r_{G_{A,C}} r_{G_{B,C}}} \right)^2} & \dfrac{r^2_{EGV_{A,C}}}{r^2_{EGV_{A,C}} r^2_{EGV_{B,C}} - \left(r_{G_{A,B}} \dfrac{r^2_{EGV_{A,C}} r^2_{EGV_{B,C}}}{r_{G_{A,C}} r_{G_{B,C}}} \right)^2} \end{bmatrix}.$$

(B6.17)

Hence, Equation B6.16 can be written as:

$$r_{EGV_{A+B,C}}$$

$$= \sqrt{\frac{r^4_{EGV_{A,C}} r^2_{EGV_{B,C}} - 2r^2_{EGV_{B,C}} r_{G_{A,B}} \frac{r^2_{EGV_{A,C}} r^2_{EGV_{B,C}}}{r_{G_{A,C}} r_{G_{B,C}}} r^2_{EGV_{A,C}} + r^4_{EGV_{B,C}} r^2_{EGV_{A,C}}}{r^2_{EGV_{A,C}} r^2_{EGV_{B,C}} - \left(r_{G_{A,B}} \frac{r^2_{EGV_{A,C}} r^2_{EGV_{B,C}}}{r_{G_{A,C}} r_{G_{B,C}}} \right)^2}}$$

$$= \sqrt{\frac{r^2_{EGV_{A,C}} - 2r_{G_{A,B}} \frac{r^2_{EGV_{A,C}} r^2_{EGV_{B,C}}}{r_{G_{A,C}} r_{G_{B,C}}} + r^2_{EGV_{B,C}}}{1 - r^2_{G_{A,B}} \frac{r^2_{EGV_{A,C}} r^2_{EGV_{B,C}}}{r^2_{G_{A,C}} r^2_{G_{B,C}}}}} \; . \qquad \text{(B6.18)}$$

If we assume that all genetic variance in population *C* can be captured by the SNPs in the training population, the accuracies for each of the populations can be replaced by the corresponding equation to predict the accuracy of genomic prediction (Daetwyler *et al.* 2008; Daetwyler *et al.* 2010; Wientjes *et al.* 2015b):

$$r_{EGV\ A,C} = \sqrt{r^2_{G_{A,C}} \frac{h_A^2 n_{P,A}}{h_A^2 n_{P,A} + M_{e_{A,C}}}} = \sqrt{r^2_{G_{A,C}} \frac{\frac{h_A^2}{M_{e_{A,C}}}}{\frac{h_A^2}{M_{e_{A,C}}} + \frac{1}{n_{P,A}}}} \; . \qquad \text{(B6.19)}$$

And:

$$r_{EGV\ B,C} = \sqrt{r^2_{G_{B,C}} \frac{\frac{h_B^2}{M_{e_{B,C}}}}{\frac{h_B^2}{M_{e_{B,C}}} + \frac{1}{n_{P,B}}}} \; . \qquad \text{(B6.20)}$$

Using this in Equation B6.18 results in:

$$r_{EGV_{A+B,C}}$$

$$= \sqrt{\frac{r_{G_{A,C}}^{2}\left(\frac{\frac{h_A^2}{M_{e_{A,C}}}}{\frac{h_A^2}{M_{e_{A,C}}}+\frac{1}{n_{P,A}}}\right)-2r_{G_{A,B}}\frac{r_{G_{A,C}}^{2}\left(\frac{\frac{h_A^2}{M_{e_{A,C}}}}{\frac{h_A^2}{M_{e_{A,C}}}+\frac{1}{n_{P,A}}}\right)r_{G_{B,C}}^{2}\left(\frac{\frac{h_B^2}{M_{e_{B,C}}}}{\frac{h_B^2}{M_{e_{B,C}}}+\frac{1}{n_{P,B}}}\right)}{r_{G_{A,C}}r_{G_{B,C}}}+r_{G_{B,C}}^{2}\left(\frac{\frac{h_B^2}{M_{e_{B,C}}}}{\frac{h_B^2}{M_{e_{B,C}}}+\frac{1}{n_{P,B}}}\right)}{1-r_{G_{A,B}}^{2}\frac{r_{G_{A,C}}^{2}\left(\frac{\frac{h_A^2}{M_{e_{A,C}}}}{\frac{h_A^2}{M_{e_{A,C}}}+\frac{1}{n_{P,A}}}\right)r_{G_{B,C}}^{2}\left(\frac{\frac{h_B^2}{M_{e_{B,C}}}}{\frac{h_B^2}{M_{e_{B,C}}}+\frac{1}{n_{P,B}}}\right)}{r_{G_{A,C}}^{2}r_{G_{B,C}}^{2}}}}$$

$$= \sqrt{\frac{r_{G_{A,C}}^{2}\left(\frac{\frac{h_A^2}{M_{e_{A,C}}}}{\frac{h_A^2}{M_{e_{A,C}}}+\frac{1}{n_{P,A}}}\right)-2r_{G_{A,B}}r_{G_{A,C}}\left(\frac{\frac{h_A^2}{M_{e_{A,C}}}}{\frac{h_A^2}{M_{e_{A,C}}}+\frac{1}{n_{P,A}}}\right)r_{G_{B,C}}\left(\frac{\frac{h_B^2}{M_{e_{B,C}}}}{\frac{h_B^2}{M_{e_{B,C}}}+\frac{1}{n_{P,B}}}\right)+r_{G_{B,C}}^{2}\left(\frac{\frac{h_B^2}{M_{e_{B,C}}}}{\frac{h_B^2}{M_{e_{B,C}}}+\frac{1}{n_{P,B}}}\right)}{1-r_{G_{A,B}}^{2}\left(\frac{\frac{h_A^2}{M_{e_{A,C}}}}{\frac{h_A^2}{M_{e_{A,C}}}+\frac{1}{n_{P,A}}}\right)\left(\frac{\frac{h_B^2}{M_{e_{B,C}}}}{\frac{h_B^2}{M_{e_{B,C}}}+\frac{1}{n_{P,B}}}\right)}}\ .$$

(B6.21)

Multiplying both the numerator and the denominator by $\left(\frac{h_A^2}{M_{e_{A,C}}}+\frac{1}{n_{P,A}}\right)$ and $\left(\frac{h_B^2}{M_{e_{B,C}}}+\frac{1}{n_{P,B}}\right)$, results in:

$$r_{EGV_{A+B,C}}$$

$$=\sqrt{\frac{r_{G_{A,C}}^2\left(\frac{h_A^2}{M_{e_{A,C}}}\right)\left(\frac{h_B^2}{M_{e_{B,C}}}+\frac{1}{n_{P,B}}\right)-2r_{G_{A,B}}r_{G_{A,C}}\left(\frac{h_A^2}{M_{e_{A,C}}}\right)r_{G_{B,C}}\left(\frac{h_B^2}{M_{e_{B,C}}}\right)+r_{G_{B,C}}^2\left(\frac{h_B^2}{M_{e_{B,C}}}\right)\left(\frac{h_A^2}{M_{e_{A,C}}}+\frac{1}{n_{P,A}}\right)}{\left(\frac{h_A^2}{M_{e_{A,C}}}+\frac{1}{n_{P,A}}\right)\left(\frac{h_B^2}{M_{e_{B,C}}}+\frac{1}{n_{P,B}}\right)-r_{G_{A,B}}^2\left(\frac{h_A^2}{M_{e_{A,C}}}\right)\left(\frac{h_B^2}{M_{e_{B,C}}}\right)}}$$

$$=\sqrt{\begin{bmatrix} r_{G_{A,C}}\frac{\sqrt{h_A^2}}{\sqrt{M_{e_{A,C}}}} & r_{G_{B,C}}\frac{\sqrt{h_B^2}}{\sqrt{M_{e_{B,C}}}}\end{bmatrix}\begin{bmatrix}\frac{h_A^2}{M_{e_{A,C}}}+\frac{1}{n_{P,A}} & r_{G_{A,B}}\frac{\sqrt{h_A^2h_B^2}}{\sqrt{M_{e_{A,C}}M_{e_{B,C}}}}\\ r_{G_{A,B}}\frac{\sqrt{h_A^2h_B^2}}{\sqrt{M_{e_{A,C}}M_{e_{B,C}}}} & \frac{h_B^2}{M_{e_{B,C}}}+\frac{1}{n_{P,B}}\end{bmatrix}^{-1}\begin{bmatrix} r_{G_{A,C}}\frac{\sqrt{h_A^2}}{\sqrt{M_{e_{A,C}}}}\\ r_{G_{B,C}}\frac{\sqrt{h_B^2}}{\sqrt{M_{e_{B,C}}}}\end{bmatrix}}.$$

(B6.22)

This last equation is equivalent to the equation derived before, using the same assumption that all genetic variance of the predicted population is captured by the SNPs in the training populations.

6.8 References

Andreescu, C., S. Avendano, S. R. Brown, A. Hassen, S. J. Lamont*, et al.*, 2007 Linkage disequilibrium in related breeding lines of chickens. Genetics 177: 2161-2169.

Brard, S. and A. Ricard, 2015 Is the use of formulae a reliable way to predict the accuracy of genomic selection? J. Anim. Breed. Genet. 132: 207-217.

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

Calus, M. P. L., H. Huang, A. Vereijken, J. Visscher, J. Ten Napel*, et al.*, 2014 Genomic prediction based on data from three layer lines: a comparison between linear methods. Genet. Sel. Evol. 46: 57.

Cooper, T. A., G. R. Wiggans and P. M. VanRaden, 2015 Short communication: Analysis of genomic predictor population for Holstein dairy cattle in the United States—Effects of sex and age. J. Dairy Sci. 98: 2785-2788.

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., 2009, *Genome-wide evaluation of populations*. PhD thesis: Animal Breeding and Genomics Centre, Wageningen, Wageningen University, Wageningen, NL

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

Daetwyler, H. D., M. P. L. Calus, R. Pong-Wong, G. De los Campos and J. M. Hickey, 2013 Genomic prediction in animals and plants: Simulation of data, validation, reporting, and benchmarking. Genetics 193: 347-365.

De Candia, T. R., S. H. Lee, J. Yang, B. L. Browning, P. V. Gejman*, et al.*, 2013 Additive genetic variation in schizophrenia risk is shared by populations of African and European descent. Am. J. Hum. Genet. 93: 463-470.

De Haas, Y., M. P. L. Calus, R. F. Veerkamp, E. Wall, M. P. Coffey*, et al.*, 2012 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 Los Campos, G., D. Gianola and D. B. Allison, 2010 Predicting genetic predisposition in humans: the promise of whole-genome markers. Nat. Rev. Genet. 11: 880-886.

De los Campos, G., Y. C. Klimentidis, A. I. Vazquez and D. B. Allison, 2012 Prediction of expected years of life using whole-genome markers. PLoS ONE 7: e40964.

De los Campos, G., A. I. Vazquez, R. Fernando, Y. C. Klimentidis and D. Sorensen, 2013 Prediction of complex human traits using the genomic best linear unbiased predictor. PLoS Genet. 9: e1003608.

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., C. Schrooten, R. F. Veerkamp and J. A. M. Van Arendonk, 2011 Effects of genomic selection on genetic improvement, inbreeding, and merit of young versus proven bulls. J. Dairy Sci. 94: 1559-1567.

Dekkers, J. C. M., 2007 Prediction of response to marker-assisted and genomic selection using selection index theory. J. Anim. Breed. Genet. 124: 331-341.

Erbe, M., B. Gredler, F. R. Seefried, B. Bapst and H. Simianer, 2013 A function accounting for training set size and marker density to model the average accuracy of genomic prediction. PLoS ONE 8: e81046.

Garrick, D. J., J. F. Taylor and R. L. Fernando, 2009 Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet. Sel. Evol. 41: 1.

Gilmour, A. R., B. Gogel, B. Cullis, R. Thompson, D. Butler*, et al.*, 2009 *ASReml user guide release 3.0*. VSN International Ltd, Hemel Hempstead.

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

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

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.

Haile-Mariam, M., J. E. Pryce, C. Schrooten and B. J. Hayes, 2015 Including overseas performance information in genomic evaluations of Australian dairy cattle. J. Dairy Sci. 98: 3443–3459.

Harris, B. L. and D. L. Johnson, 2010 Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. J. Dairy Sci. 93: 1243-1252.

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

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

Hazel, L. N., 1943 The genetic basis for constructing selection indexes. Genetics 28: 476-490.

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

Huang, H., J. J. Windig, A. Vereijken and M. P. Calus, 2014 Genomic prediction based on data from three layer lines using non-linear regression models. Genet. Sel. Evol. 46: 75.

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.

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

Kemper, K. E. and M. E. Goddard, 2012 Understanding and predicting complex traits: Knowledge from cattle. Hum. Mol. Genet. 21: R45-R51.

Kemper, K. E., B. J. Hayes, H. D. Daetwyler and M. E. Goddard, 2015 How old are quantitative trait loci and how widely do they segregate? J. Anim. Breed. Genet. 132: 121-134.

Khatkar, M. S., F. W. Nicholas, A. R. Collins, K. R. Zenger, J. A. L. Cavanagh*, et al.*, 2008 Extent of genome-wide linkage disequilibrium in Australian Holstein-Friesian cattle based on a high-density SNP panel. BMC Genom. 9: 187.

Lee, S. H., T. R. DeCandia, S. Ripke, J. Yang, P. F. Sullivan*, et al.*, 2012 Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. Nat. Genet. 44: 247-250.

Lehermeier, C., N. Krämer, E. Bauer, C. Bauland, C. Camisan*, et al.*, 2014 Usefulness of multiparental populations of maize (*Zea mays* L.) for genome-based prediction. Genetics 198: 3-16.

Lund, M. S., S. P. W. De Roos, A. G. De Vries, T. Druet, V. Ducrocq*, et al.*, 2011 A common reference population from four European Holstein populations increases reliability of genomic predictions. Genet. Sel. Evol. 43: 43.

Maier, R., G. Moser, G.-B. Chen, S. Ripke, W. Coryell*, et al.*, 2015 Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. Am. J. Hum. Genet. 96: 283–294.

Matukumalli, L. K., C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan*, et al.*, 2009 Development and characterization of a high density SNP genotyping assay for cattle. PLoS ONE 4: e5350.

McEvoy, B. P., J. E. Powell, M. E. Goddard and P. M. Visscher, 2011 Human population dispersal "Out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs. Genome Res. 21: 821-829.

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.

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

Pryce, J. E., J. Johnston, B. J. Hayes, G. Sahana, K. A. Weigel*, et al.*, 2014 Imputation of genotypes from low density (50,000 markers) to high density (700,000 markers) of cows from research herds in Europe, North America, and Australasia using 2 reference populations. J. Dairy Sci. 97: 1799-1811.

R Development Core Team, 2011 R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.

Román-Ponce, S. I., A. B. Samoré, M. A. Dolezal, A. Bagnato and T. H. E. Meuwissen, 2014 Estimates of missing heritability for complex traits in Brown Swiss cattle. Genet. Sel. Evol. 46: 36.

Sawyer, S. L., N. Mukherjee, A. J. Pakstis, L. Feuk, J. R. Kidd*, et al.*, 2005 Linkage disequilibrium patterns vary substantially among populations. Europ. J. Hum. Genet. 13: 677-686.

Schaeffer, L. R., 1994 Multiple-country comparison of dairy sires. J. Dairy Sci. 77: 2671-2678.

Spelman, R. J., C. A. Ford, P. McElhinney, G. C. Gregory and R. G. Snell, 2002 Characterization of the DGAT1 gene in the New Zealand dairy population. J. Dairy Sci. 85: 3514-3517.

Thaller, G., W. Krämer, A. Winter, B. Kaupe, G. Erhardt*, et al.*, 2003 Effects of DGAT1 variants on milk production traits in German cattle breeds. J. Anim. Sci. 81: 1911-1918.

Van den Berg, S., M. P. L. Calus, T. H. E. Meuwissen and Y. C. J. Wientjes, 2015 Across population genomic prediction scenarios in which Bayesian variable selection outperforms GBLUP. Submitted to BMC Genet.

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

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural*, et al.*, 2001 The sequence of the human genome. Science 291: 1304-1351.

Veroneze, R., P. S. Lopes, S. E. F. Guimarães, F. F. Silva, M. S. Lopes*, et al.*, 2013 Linkage disequilibrium and haplotype block structure in six commercial pig lines. J. Anim. Sci. 91: 3493-3501.

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.

Wientjes, Y. C. J., M. P. L. Calus, M. E. Goddard and B. J. Hayes, 2015a Impact of QTL properties on the accuracy of multi-breed genomic prediction. Genet. Sel. Evol. 47: 42.

Wientjes, Y. C. J., R. F. Veerkamp, P. Bijma, H. Bovenhuis, C. Schrooten*, et al.*, 2015b Empirical and deterministic accuracies of across-population genomic prediction. Genet. Sel. Evol. 47: 5.

Wientjes, Y. C. J., R. F. Veerkamp and M. P. L. Calus, 2015c Using selection index theory to estimate consistency of multi-locus linkage disequilibrium across populations. BMC Genet. 16: 87.

Wray, N. R., M. E. Goddard and P. M. Visscher, 2007 Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res. 17: 1520-1528.

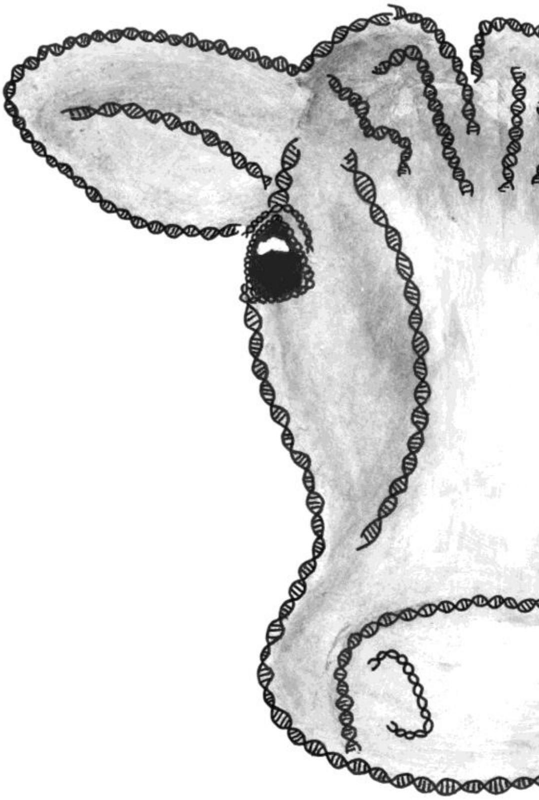
Yan, J., T. Shah, M. L. Warburton, E. S. Buckler, M. D. McMullen*, et al.*, 2009 Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. PLoS ONE 4: e8451.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders*, et al.*, 2010 Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42: 565-569.

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.

CHAPTER 7

GENERAL DISCUSSION



7.1 Introduction

In livestock breeding programs, genotype information of thousands of single-nucleotide polymorphism (SNP) markers spread across the whole genome is widely used to select the genetically best animals to produce the next generation. In this approach, known as genomic selection, animals are selected based on genomic estimated breeding values (GEBVs), predicted using SNP genotype information of those animals as well as SNP effects estimated in a reference population containing animals with known phenotypes and SNP genotypes (Meuwissen *et al.* 2001). The accuracy of predicting GEBVs, i.e., the accuracy of genomic prediction, determines the response to selection (Falconer and Mackay 1996). One of the factors limiting the accuracy of genomic prediction in numerically small populations is the size of the reference population. This might result in an increasing genetic gap between populations of numerically small breeds compared to populations of the more commonly used breeds, e.g., the Holstein Friesian breed in dairy cattle. One way to increase the size of the reference population for numerically small populations is to add individuals from other populations to the reference population, for example individuals from different countries, breeds, or lines. The suitability of individuals from another population to increase the accuracy of genomic prediction is, however, reduced by differences between the populations, such as differences in linkage disequilibrium (LD) between the SNPs and the quantitative trait loci (QTL) underlying the trait, differences in allele frequencies of SNPs and QTL, differences in allele substitution effects of QTL, and the absence of close family relationships between populations.

The overall objective of this thesis was to investigate the accuracy of multi-population genomic prediction in dairy cattle. This overall objective was divided in two sub-objectives. The first sub-objective was to investigate the effect of different factors on the accuracy of multi-population genomic prediction. The factors that were studied were the effect of absence of close family relationships (Chapter 2), and the effect of differences across populations in allele substitution effects (Chapter 3 + 6), linkage disequilibrium patterns (Chapter 2 + 3 + 4 + 6), and allele frequencies (Chapter 5). The second sub-objective was to derive deterministic equations to calculate or predict the accuracy of multi-population genomic prediction. In this thesis, two different equations were derived, one using genomic relationships between individuals (Chapter 3), and one using population parameters (Chapter 6).

This general discussion is divided in five parts. In the first part, the potential of multi-population genomic prediction is discussed for different scenarios, combining either populations from the same breed from different countries, closely related breeds, or distantly related breeds. In the second part, the impact of the model used to estimate GEBVs on the accuracy of multi-population genomic prediction is discussed. In the third part, the possibility to estimate the genetic correlation, an important parameter determining the potential of multi-population genomic prediction, using SNP information is discussed. In the fourth part, the relation between the effective number of chromosome segments and the consistency of multi-locus LD across populations is discussed. Both measures were used in the previous Chapters and reflect on the consistency of LD across populations, which is influencing the accuracy of multi-population genomic prediction. Finally, in the fifth part, possible research directions to improve the accuracy of multi-population genomic prediction are discussed.

7.2 Potential of multi-population genomic prediction

Combining two or more populations in one reference population was expected to result in an increase in accuracy of genomic prediction by increasing the size of the reference population. Some studies have indeed confirmed this and showed an average increase in accuracy of approximately 10% by combining populations (e.g., Brøndum *et al.* 2011; Lund *et al.* 2011; Zhou *et al.* 2013). However, other studies showed no increase or even a decrease in accuracy by combining populations (e.g., Erbe *et al.* 2012; Karoui *et al.* 2012; Olson *et al.* 2012). Therefore, an interesting question is: Under which conditions will combining populations in one reference population result in an increase in accuracy of genomic prediction?

To answer this question, the potential of different scenarios of multi-population genomic prediction is discussed, differing in the relatedness between the populations that are combined. For clarity, it is first assumed that the consistency of LD across populations is the only difference between closely related and distantly related populations (7.2.1). Later on, the effect of differences between closely and distantly related populations in allele substitution effects across populations, reflected by the genetic correlation between populations (7.2.2), and differences in the proportion of the genetic variance captured by the SNPs across populations (7.2.3) are investigated as well. Subsequently, the results are compared with the results from empirical studies (7.2.4), followed by some concluding remarks regarding the potential of multi-population genomic prediction (7.2.5).

7.2.1 Differences in consistency of LD across populations

In this part, the potential to combine information from two populations that are either closely or distantly related populations is discussed, assuming that the consistency of LD across populations is the only difference between closely and distantly related populations. The relatedness between populations is generally lower for populations that split more generations ago, for which the effect of drift and the number of recombination and mutation events in each population since the separation of the populations is higher (Falconer and Mackay 1996). Due to the higher number of recombination and mutations events, the consistency of LD between the population becomes lower (e.g., Andreescu *et al.* 2007; Gautier *et al.* 2007; De Roos *et al.* 2008), resulting in a larger effective number of chromosome segments (M_e) across the populations (Chapter 3; Goddard *et al.* 2011). The M_e across populations is used here to reflect the consistency of LD across populations.

The potential accuracy for predicting GEBVs for selection candidates from population 1, using different numbers of individuals from population 1 and 2 in the reference population, is investigated using the prediction equation derived in Chapter 6. First, this prediction equation is used to investigate how valuable individuals from another population are compared to the value of an individual from the population of the selection candidates. For this purpose, it is assumed that only one population, either population 1 or 2, is included in the reference population. When only population 1 is included in the reference population, the prediction equation reduces to the prediction equation for within-population genomic prediction, as derived by Daetwyler *et al.* (2008; 2010). When only population 2 is included in the reference population, the prediction equation reduces to the prediction equation for across-population genomic prediction, as derived in Chapter 3. To calculate the number of individuals from population 2 ($n_{P,2}$) that can obtain the same accuracy as $n_{P,1}$ individuals from population 1, the predicted accuracy for within- and across- population genomic prediction were equalized:

$$\sqrt{\frac{n_{P,1}h_1^2}{n_{P,1}h_1^2+M_{e_1}}}=r_{G_{1,2}}\sqrt{\frac{n_{P,2}h_2^2}{n_{P,2}h_2^2+M_{e_{1,2}}}}\ , \qquad (7.1)$$

in which h_1^2 and h_2^2 are the heritabilities in population 1 and 2 respectively, M_{e_1} is the effective number of chromosome segments in population 1, $M_{e_{1,2}}$ is the effective number of chromosome segments across population 1 and 2, and $r_{G_{1,2}}$ is the genetic correlation between population 1 and 2. By solving this equation for

$n_{P,2}$ assuming a genetic correlation of 1, it can be shown that the number of individuals from population 2 should equal:

$$n_{P,2} = \frac{h_1^2}{h_2^2}\frac{M_{e_{1,2}}}{M_{e_1}}n_{P,1} \tag{7.2}$$

to obtain the same increase in accuracy as with $n_{P,1}$ individuals from population 1. This indicates that when the heritabilities are the same in both populations, the value of individuals from another population compared to the value of individuals from the population of the selection candidates linearly depends on the ratio between M_e across the populations and M_e within the population of the selection candidates.

Second, the prediction equation derived in Chapter 6 is used to investigate the potential accuracy of combining population 1 and 2 for predicting GEBVs for individuals from population 1. For this purpose, it is assumed that the heritability in both populations was 0.3. For both populations, the number of individuals in the reference populations was varied between 0 and 10,000. Moreover, the effective number of chromosome segments in population 1 was set to 1000, which is about equal to the M_e in a Holstein Friesian population (Chapter 2; Chapter 6; Brard and Ricard 2015). The M_e between the populations was varied between the scenarios, representing either populations from the same breed from different countries (7.2.1.1), closely related breeds (7.2.1.2), or distantly related breeds (7.2.1.3).

7.2.1.1 Same breed from different countries

In this paragraph, the accuracy of genomic prediction is described when two populations from the same breed from different countries are combined in the reference population. This, for example, represents combining Holstein Friesian populations from two or more countries (e.g., Lund *et al.* 2011; Haile-Mariam *et al.* 2015). Even though the populations are from the same breed, the relationships between individuals from the same population are likely to be slightly higher than between individuals of different populations. Therefore, the M_e between the populations was set to be twice the M_e within population 1 (M_{e_1} =1000 and $M_{e_{1,2}}$ =2000), indicating that one individual from population 1 is just as informative as two individuals from population 2 (Equation 7.2). The predicted accuracies for this scenario are shown in Figure 7.1.

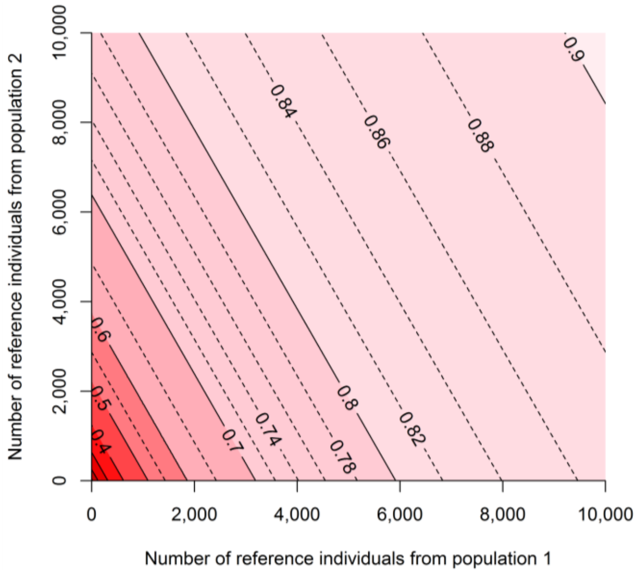


Figure 7.1 Predicted accuracy of GEBVs for population 1 using a reference population with different numbers of individuals from population 1 and 2. The heritability of the trait is 0.3 for both populations and the genetic correlation between the populations is 1. The M_e within population 1 is set to 1000 and the M_e between the populations to 2000, representing populations from the same breed from different countries.

The results show that when the reference population from population 1 is small, e.g., 1000 individuals, a substantial increase in accuracy can be obtained by adding 10,000 individuals from population 2 (from 0.48 to 0.80). When the reference population from population 1 is large, e.g., 10,000 individuals, the accuracy obtained with only population 1 in the reference population is already high, resulting in a much smaller increase in accuracy by adding 10,000 individuals from population 2 (from 0.87 to 0.90).

7.2.1.2 Closely related breed

Another possibility to enlarge the reference population is by adding individuals from a closely related breed. This, for example, represents the scenario where Holstein Friesian and Meuse-Rhine-Yssel or Groningen White Headed individuals are combined in the reference population, as described in Chapter 3 and 4. The M_e between Holstein Friesian and Meuse-Rhine-Yssel or Groningen White Headed individuals was found to be 10 times the M_e within the Holstein Friesian population (M_{e_1} =1000 and $M_{e_{1,2}}$ =10,000). Those values for M_e indicate that, in this scenario, one individual from population 1 is just as informative as 10 individuals from population 2 (Equation 7.2). The predicted accuracies for this scenario are shown in Figure 7.2.

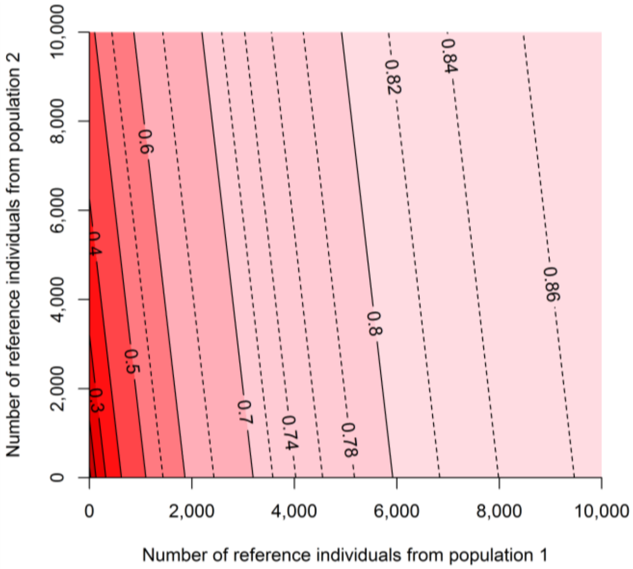


Figure 7.2 Predicted accuracy of GEBVs for population 1 using a reference population with different numbers of individuals from population 1 and 2. The heritability of the trait is 0.3 for both populations and the genetic correlation between the populations is 1. The M_e within population 1 is set to 1000 and the M_e between the populations to 10,000, representing closely related breeds.

As can be expected, the increase in accuracy is much lower when individuals from a closely related breed are added to the reference population compared to individuals from a population from the same breed from a different country. The increase in accuracy by adding 10,000 individuals from a closely related breed is still reasonably large when only 1000 individuals from population 1 are included in the reference population (from 0.48 to 0.61). The increase in accuracy is almost negligible when the reference population already consisted of 10,000 individuals from population 1 (from 0.87 to 0.88). So, the addition of a closely related breed is only helpful when the reference population of the breed itself is small, which can be the case for numerically small breeds as well as for traits that are difficult or expensive to measure in numerically large populations.

7.2.1.3 Distantly related breed

The last option to enlarge the reference population is by adding individuals from a distantly related breed. This, for example, represents combining Holstein Friesian and Jersey or Angus individuals in one reference population, as described in different studies (Chapter 5; Hayes *et al.* 2009; Harris and Johnson 2010; Khansefid *et al.* 2014). Using high-density SNP genotypes obtained from sequence data of 58 Angus and 30 Holstein Friesian individuals from the United States (1000 bull genomes consortium), I estimated an M_e of approximately 20,000 between Holstein Friesian and Angus, which is about 20 times the M_e within the Holstein Friesian population. Therefore, an M_e of 20,000 across populations was used to represent very distantly related populations. This indicates that for this scenario one individual from population 1 is just as informative as 20 individuals from population 2 (Equation 7.2). The accuracies for this scenario are shown in Figure 7.3.

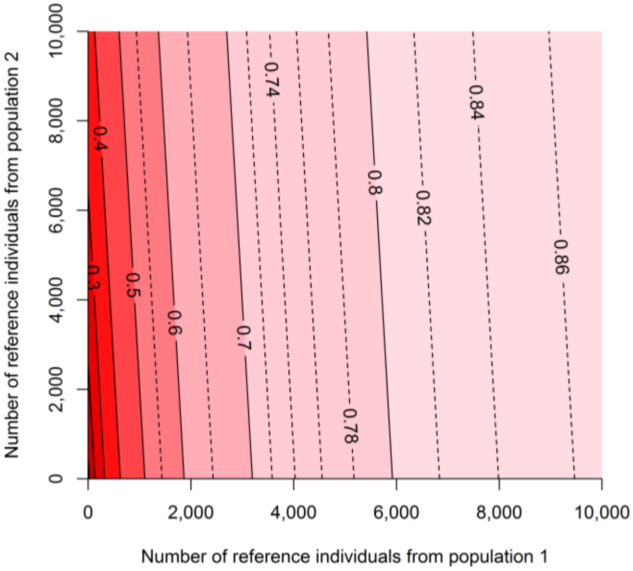


Figure 7.3 Predicted accuracy of GEBVs for population 1 using a reference population with different numbers of individuals from population 1 and 2. The heritability of the trait is 0.3 for both populations and the genetic correlation between the populations is 1. The M_e within population 1 is set to 1000 and the M_e between the populations to 20,000, representing distantly related breeds.

The increase in accuracy when distantly related breeds are combined in the reference population is much lower than when closely related breeds are combined. The increase in accuracy by adding 10,000 individuals from population 2 to a small reference population of population 1, e.g., 1000 individuals, is, however, still substantial (from 0.48 to 0.56). When the reference population of population 1 is large, e.g., 10,000 individuals, adding 10,000 individuals from population 2 does not increase the accuracy (0.87). This indicates that only when the reference population of the breed itself is small, an increase in accuracy can be expected by adding individuals from a distantly related breed.

7.2.2 Differences in genetic correlations between populations

In the previous paragraph, it was assumed that the M_e across populations was the only factor representing the distance between populations and, thereby, influencing the accuracy of multi-population genomic prediction. In real data, this is unlikely to be the case. The genetic correlation between populations can, for example, be expected to be lower than 1, i.e., the allele substitution effects are likely to differ between populations, due to genotype by environment interactions (Falconer 1952; Schaeffer 1994; Lillehammer *et al.* 2007), and due to differences in genetic background of the populations in combination with non-additive effects (Falconer and Mackay 1996; Huang *et al.* 2012). Therefore, the accuracies reported before are likely to be overestimated.

For populations that separated more generations ago, the differences in allele frequencies are generally higher, due to selection and random genetic drift occurring separately in each of the populations (Falconer and Mackay 1996). In combination with non-additive effects, those differences in allele frequencies can result in differences in allele substitution effects of the QTL underlying the trait (Falconer and Mackay 1996; Huang *et al.* 2012) and can, thereby, reduce the correlation between allele substitution effects of QTL from different populations. So, distantly related populations, with a reasonably large M_e across the populations, generally have a lower genetic correlation between populations compared to closely related populations, with a reasonably small M_e across the populations (Lehermeier *et al.* 2015).

Genetic correlations lower than 1 between populations reduce the value of adding individuals from another population to the reference population. When only individuals from another population are used, the maximum accuracy that can be obtained is equal to the genetic correlation between the populations. By solving

Equation 7.1 including the genetic correlation as a variable, it can be shown that the number of individuals from population 2 should equal:

$$n_{P,2} = \frac{\frac{h_1^2}{h_2^2} M_{e_{1,2}} n_{P,1}}{h_1^2 n_{P,1}(r_{G_{1,2}}^2 - 1) + r_{G_{1,2}}^2 M_{e_1}}\ , \tag{7.3}$$

to obtain the same increase in accuracy as with $n_{P,1}$ individuals from population 1. Equation 7.3 indicates that when the absolute genetic correlation between the populations is lower than 1, the value of individuals from another population is reduced and the relationship between $n_{P,1}$ and $n_{P,2}$ is no longer linear.

In Figure 7.4, the estimates of $n_{P,2}$ are plotted assuming different values for $n_{P,1}$ and different genetic correlations, for each of the three scenarios of combining populations. Those results show that the value of individuals from population 2 compared to the value of individuals from population 1 decreases when the genetic correlation between population 1 and 2 deviates more from 1, and thereby becomes closer to 0. For example, when the genetic correlation is 0.8 instead of 1, 9434 instead of 4000 individuals from population 2 are needed to obtain the same increase in accuracy as with 2000 individuals from population 1 when population 1 and 2 are from the same breed from different countries ($M_{e_{1,2}}$ =2 M_{e_1}), and 94,340 instead of 40,000 individuals from population 2 when population 1 and 2 are distantly related breeds ($M_{e_{1,2}}$ =20 M_{e_1}). It is, however, unrealistic to expect a genetic correlation of 0.8 between distantly related breeds. Karoui *et al*. (2012), for example, estimated genetic correlations of around 0.53 for production traits between three distantly related French dairy cattle breeds; Normande, Holstein Friesian and Montbéliarde. The distance between those three breeds is comparable to the distance between Holstein Friesian and Jersey, and slightly larger than the distance between Holstein Friesian and Angus (Gautier *et al.* 2010; Decker *et al.* 2014). When the genetic correlation between very distantly related populations would be 0.53, it is not even possible to obtain the same accuracy as obtained with 2000 individuals from population 1 by using only individuals from population 2, since an accuracy above 0.53 is already obtained with 2000 individuals from population 1 at a heritability of 0.3. Therefore, it can be expected that improving the accuracy of genomic prediction by combining populations from distantly related breeds is impossible. Combining closely related breeds or populations from the same breed from different countries can help to increase the accuracy, but only when the population of the selection candidates in the reference population is small and a large number of individuals from the other population is added.

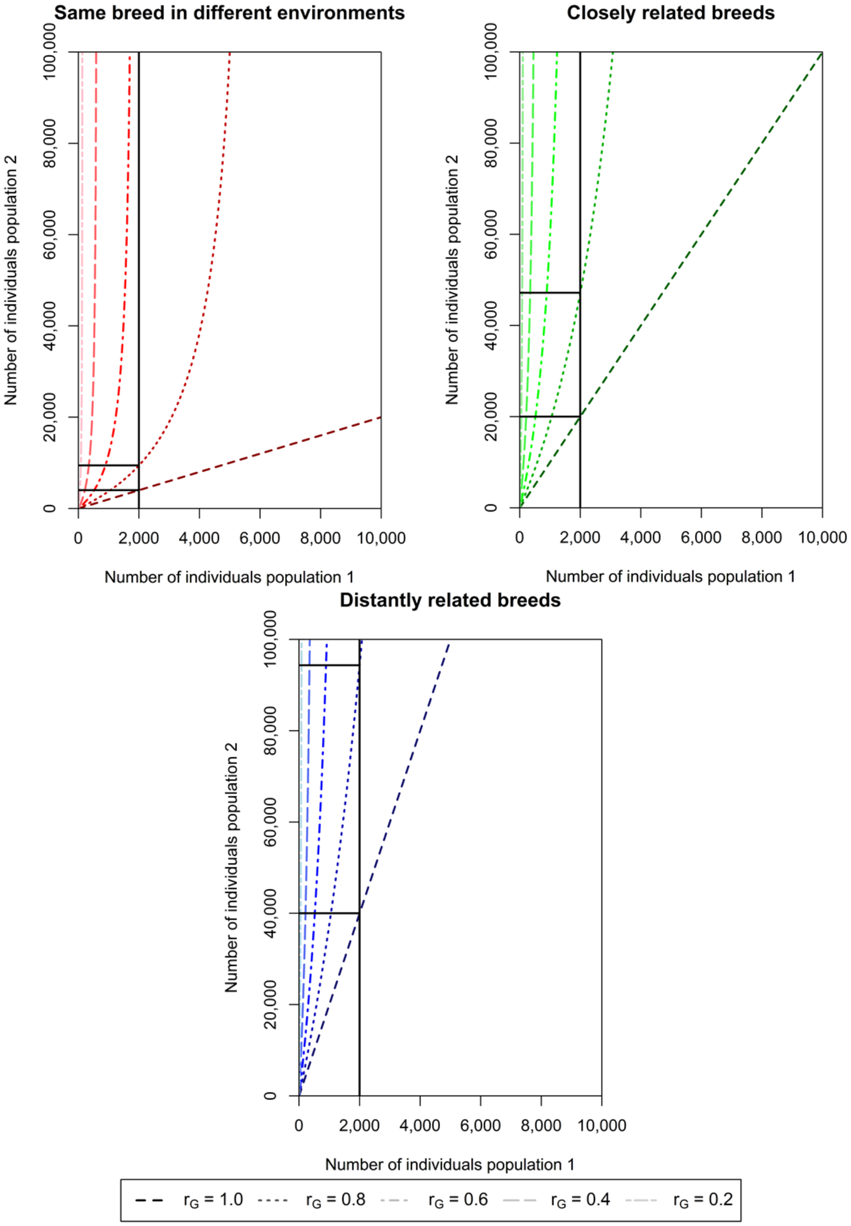


Figure 7.4 The number of individuals in population 2 that have to be used to obtain the same accuracy as different numbers of individuals in population 1 using different genetic correlations. The heritability of the trait is 0.3 for both populations, and the different populations either represent the same breed from different countries, closely related breeds or distantly related breeds.

7.2.3 Differences in the proportion of the genetic variance captured by SNPs across populations

So far, it was assumed that all of the genetic variance in the predicted population can be captured by the SNPs in all reference populations. Due to differences in allele frequencies between QTL and SNPs, this assumption will not hold within a population, as is shown in Chapter 5 and 6. Moreover, the proportion of the genetic variance captured by the SNPs in another population is likely to be higher for closely related populations than for distantly related populations. This indicates that the proportion of the genetic variance in one population that can be captured by the SNPs in another population can also influence the accuracy of multi-population genomic prediction.

As explained before, allele frequencies are likely to differ between populations, with generally larger differences in allele frequencies between populations that were separated a longer time ago. Those differences in allele frequencies can result in differences in the part of the genetic variance for a trait explained by a specific QTL, even though that QTL is segregating in both populations. This indicates that a QTL that is segregating at a high allele frequency and explaining a large part of the genetic variance in one population, might only explain a very small part in another population, as was the case for DGAT1 in a Holstein Friesian population compared to a Meuse-Rhine-Yssel population (Maurice-Van Eijndhoven *et al.* 2015).

When the number of generations since the separation of the population increases, the number of population-specific mutations will also increase, resulting in a higher number of QTL segregating in only one population (Kemper *et al.* 2015a). Those QTL can never be explained by SNPs in another population, indicating that when population-specific QTL explain a larger part of the genetic variance, the potential of increasing the accuracy by adding another population is lower. It might also be that the QTL is fixed in one population, thereby reducing the potential benefit of using this population as a reference population for another population. To maximize the number of QTL segregating in the reference population, it might help to combine multiple populations in the reference population, preferably multiple closely related breeds.

In summary, distantly related populations have a larger value for M_e across populations, a lower genetic correlation between the populations, and a smaller part of the genetic variances that can be captured in the other population compared to closely related populations. All those three factors reduce the potential to use information across populations, especially across distantly related populations.

7.2.4 Theoretical versus empirical potential of multi-population genomic prediction

Based on paragraphs 7.2.1 till 7.2.3, it can be concluded that combining populations in one reference population is expected to be beneficial when; 1) the combined populations are closely related, 2) the population of the selection candidates in the reference population is small, and 3) the number of individuals added from the other population is very large. This indicates that the design of the multi-population reference population has a large impact on the potential benefit of combining populations. So far, this conclusion is only based on theory. Therefore, in this part, this theoretical conclusion is compared to empirical results to prove that the design of the reference population can also explain the differences in obtained benefits of combining populations described in literature.

In dairy cattle, different studies have investigated the potential to combine a bull and a cow reference population (Calus *et al.* 2013; Cooper *et al.* 2015), a scenario that was also studied in Chapter 6. Those populations generally have a high level of family relationships between the populations, indicating that those populations are very closely related to each other. The heritability of the phenotypes might, however, differ between the populations, since phenotypes of bulls are normally based on performance records of many daughters with a high reliability and phenotypes of cows are only based on own performance records. Therefore, the increase in accuracy by adding cows to the reference population was lower than what can be expected by adding the same number of bulls, but still an increase in accuracy was observed (Calus *et al.* 2013; Cooper *et al.* 2015). This indicates that combining those closely related populations was indeed beneficial, which is in agreement with the theoretical expectation.

Different studies investigated the obtained accuracy of multi-population genomic prediction by combining populations from the same breed from different countries, for example by combining different Holstein Friesian populations (e.g., Lund *et al.* 2011; De Haas *et al.* 2015; Haile-Mariam *et al.* 2015), Jersey populations (Haile-Mariam *et al.* 2015; Wiggans *et al.* 2015), or Brown Swiss populations (Zumbach *et al.* 2010; Jorjani *et al.* 2011). Since the relationships between those populations might be high, due to the use of partly the same sires, those different populations are generally closely related. Based on the theoretical expectation, those scenarios are expected to result in a substantial increase in accuracy, which was generally observed as well. The benefit of combining those different populations might, however, be slightly lower than expected from the relatedness between the populations, due to genotype by environment interactions which reduce the genetic correlation between the populations (Falconer and Mackay 1996). Combining populations from the same breeds from different countries is especially attractive for breeds with small populations in different countries, as is, for example, the case for Jersey or Brown Swiss breed (Wiggans *et al.* 2011; VanRaden *et al.* 2012; Wiggans *et al.* 2015), or for traits that are difficult or expensive to measure and, therefore, only measured at a small scale, such as dry matter intake and feed efficiency (De Haas *et al.* 2012; Pryce *et al.* 2014).

Combining more distantly related populations, for example populations from different breeds, resulted in a lower increase in accuracy, which is in agreement with the theoretical expectation. Combining the Nordic Red breeds, however, showed an average increase in accuracy of more than 10% (e.g., Brøndum *et al.* 2011; Zhou *et al.* 2014), which can be explained by the reasonably high relatedness between those populations. When more distantly related breeds, like different French dairy cattle breeds, were combined, an increase in accuracy was only observed when the reference population of the selection candidates was small and a large number of individuals from another population was added (Karoui *et al.* 2012; Hozé *et al.* 2014b). The studies combining the very distantly related Holstein Friesian and Jersey breeds in general showed no increase, or even a decrease in accuracy (e.g., Erbe *et al.* 2012; Olson *et al.* 2012). Those finding confirms that combining populations is only beneficial when populations are closely related and when a large number of individuals is added compared to the size of the reference population from the population itself. For selecting closely related breeds, phylogenetic trees, like described by Gautier *et al.* (2010) and Decker *et al*. (2014), can be helpful, since they provide insight in the relationships between breeds.

7.2.5 Concluding remarks regarding multi-population genomic prediction

Overall, it can be concluded that the potential to improve the accuracy by combining populations in one reference population is depending on the design of the reference population. The most optimal design of multi-population genomic prediction is to combine individuals from the same breed from different countries. When this is not possible, for example because the breed is only kept in one country, adding individuals from a closely related breed might help to increase the accuracy. The value of individuals from another breed is, however, lower than the value of individuals from the same breed and depending on the relatedness between the breeds. It is difficult to estimate the maximum value for M_e or the minimum value of the genetic correlation to be able to see an increase in accuracy, since the increase in accuracy is influenced by the heritability. At a high heritability, a high accuracy can already be obtained with a small number of individuals for the same population, reducing the impact of further enlarging the reference population by the breed itself or by another breed. In my opinion, however, it is clear that populations with an M_e across populations ≥20 times the M_e within the population and a genetic correlation ≤ 0.5 are too divergent to be combined in a reference population.

7.3 Genomic prediction model

The accuracy of both single- and multi-population genomic prediction varies with the model used to estimate GEBVs. At the moment, the commonly used models can roughly be divided in two different types; genomic best linear unbiased prediction (GBLUP) models and Bayesian variable selection models. The original GBLUP model, as described by Meuwissen *et al.* (2001), assumes that all SNPs explain an equal amount of the genetic variance, so basically assumes an infinitesimal model. All SNPs or independent segments were also assumed to explain an equal amount of the genetic variance in the derivation of the prediction equation derived in Chapter 6, used to investigate different scenarios of multi-population genomic prediction in paragraph 7.2, as well as in the derivation of other prediction equations (Daetwyler *et al.* 2008; VanRaden 2008; Goddard 2009; Daetwyler *et al.* 2010). Therefore, the prediction equations are predicting the accuracies that can be obtained with GBLUP. In contrast to GBLUP, the Bayesian variable selection model accommodates for some SNPs or segments explaining a larger part of the genetic variance compared to other SNPs or segments (Meuwissen *et al.* 2001). Due to this difference, the accuracy of a Bayesian variable selection model might deviate from the predicted accuracy.

For within-population genomic prediction, it is shown that the accuracy of GBLUP can accurately be predicted using a prediction equation when all genetic variance is captured by the SNPs (Daetwyler *et al.* 2008; Daetwyler *et al.* 2010). The accuracy of a Bayesian variable selection model, however, was larger than the predicted accuracy when the effective number of QTL underlying the trait was smaller than M_e, and about equal to the predicted accuracy when the effective number of QTL was equal to M_e (Daetwyler *et al.* 2010). The same principle was shown to be valid for across-population genomic prediction (Van den Berg *et al.* 2015). Since the value of M_e is much higher across populations than within populations, it is much more likely to find an effective number of QTL smaller than M_e for across-population scenarios than for within-population scenarios, given that the total number of QTL underlying the trait is more or less the same across populations. This indicates that the use of Bayesian variable selection models might increase the benefit of combining information from multiple populations. It is important to stress here that for distantly related populations, a low genetic correlation (Lehermeier *et al.* 2015) and a high number of QTL segregating in only one population can be expected (Kemper *et al.* 2015a). The negative impact of those two factors cannot be reduced by using a Bayesian variable selection model, which is only reducing the impact of a large M_e across populations. Therefore, I expect that even when a Bayesian variable selection model is used, the benefit of combining information from distantly related breeds is low and negligible. For closely related breeds, with a reasonably high genetic correlation and a relatively low number of breed-specific QTL, using a Bayesian variable selection model might help to use information across populations, especially for traits influenced by a low number of QTL or by a few QTL with large effect.

In a Bayesian variable selection model, a subset of SNPs is selected to explain the genetic variance. When the effective number of QTL underlying the trait is substantially smaller than M_e, the selection of a subset of SNPs has a clear advantage, since it reduces the number of effects that has to be estimated. This indicates that when the effective number of QTL is smaller than M_e, the number of effects that has to be estimated in a Bayesian variable selection model is lower than M_e, resulting in a higher accuracy of genomic prediction. This can be taken into account in the prediction equation by replacing the M_e by the effective number of QTL underlying the trait. At the moment, however, it is very difficult to get accurate estimates for the effective number of QTL underlying a trait. Therefore, it remains difficult to predict the exact benefit of using a Bayesian variable selection model. In general, I expect that it is very likely that a Bayesian variable selection model can slightly increase the accuracy when closely related breeds are combined in the reference population.

Besides M_e, the proportion of the genetic variance captured by the SNPs might also be influenced by the model used to analyze the data. In Chapter 4, it was shown that the SNPs very close to a QTL have a higher consistency of multi-locus LD across populations, and therefore, those SNPs were better able to predict the QTL genotype of individuals from another population. Within a population, however, the ability to predict the QTL genotypes using only a subset of the SNPs was lower than when all SNPs were used. So, for genomic prediction within a population, the selection of a subset of SNPs will probably result in a decrease in the number of effects that have to be estimated, but also in a decrease in the proportion of the genetic variance captured by the SNPs. The same process of selecting SNPs is also taking place in a Bayesian variable selection model, indicating that such a model is expected to explain a smaller part of the genetic variance compared to a GBLUP model, which was indeed seen before (Kemper *et al.* 2015b). The smaller the subset of selected SNPs, the lower the number of effects that has to be estimated and the lower the proportion of the genetic variance explained by the SNPs. For across-population genomic prediction, selecting SNPs surrounding the QTL would result in a lower number of estimated effects as well as in a higher proportion of the genetic variance captured by the SNPs, since the consistency of LD is higher at shorter distances on the genome (De Roos *et al.* 2008; Zhou *et al.* 2013). Therefore, selecting SNPs in a Bayesian variable selection model is expected to increase the proportion of the genetic variance captured by the SNPs in another population when the number of QTL underlying the trait is low. This can result in an even higher expected accuracy for a Bayesian variable selection model compared to GBLUP based on the difference in the number of effects that have to be estimated.

Altogether, I expect to see a larger advantage of using Bayesian variable selection models for across- and multi-population genomic prediction than for within-population genomic prediction. Most of the production traits in dairy cattle are suggested to be influenced by a large number of QTL. For those traits, I do not expect to see an increase in accuracy by using a Bayesian variable selection model compared to a GBLUP model for within-population genomic prediction. For across- and multi-population genomic prediction, however, I expect to see an increase in accuracy, although this increase is probably reasonably low, in the range of 0-10%. For traits known to be mainly influenced by only a small number of QTL, such as fat percentage in milk in dairy cattle, I expect to see a small increase in accuracy by using a Bayesian variable selection model compared to a GBLUP model for within-population genomic prediction. For across- and multi-population genomic prediction, the increase in accuracy is probably much larger and might be up to 30-50%.

7.4 Estimating the genetic correlation

As discussed, an important parameter determining the potential to combine populations in one reference population is the genetic correlation between the populations. The genetic correlation represents the correlation between allele substitution effects of the true QTL underlying the trait (Falconer and Mackay 1996). The true QTL and their effects are generally unknown, which makes it impossible to calculate the true genetic correlation. Therefore, an important question is how to estimate the genetic correlation in real data.

A multi-trait model, where the same trait in the different populations is modelled as a different trait, can be used to estimate the genetic correlation between populations. For populations from the same breed from different countries, it is often possible to estimate the genetic correlation by using only pedigree information, since partly the same sires might be used in both populations (Schaeffer 1994). For more distantly related populations, pedigree information is often not able to accurately describe the relationships between individuals from different populations. Therefore, a relationship matrix based on genomic information is essential for estimating the genetic correlation between more distantly related populations. At the moment, different studies have estimated the genetic correlation using a multi-trait GBLUP model (Chapter 3; Karoui *et al.* 2012; Carillier *et al.* 2014; Huang *et al.* 2014; Legarra *et al.* 2014; Lehermeier *et al.* 2015). In the multi-trait GBLUP model, the (co)variance structure between the GEBVs on the scale of both populations (*A* and *B*) is assumed to follow (Karoui *et al.* 2012):

$$\begin{bmatrix} \mathbf{GEBV}_A \\ \mathbf{GEBV}_B \end{bmatrix} \sim N(\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \sigma_A^2\mathbf{G} & \sigma_{AB}\mathbf{G} \\ \sigma_{AB}\mathbf{G} & \sigma_B^2\mathbf{G} \end{bmatrix} \right), \tag{7.4}$$

in which **GEBV**$_A$ (**GEBV**$_B$) is a vector with GEBVs for all individuals on the scale of population *A* (*B*), σ_A^2 (σ_B^2) is the variance of estimated SNP effects in population *A* (*B*), σ_{AB} is the covariance between the estimated SNP effects in population *A* and *B*, and **G** is the genomic relationship matrix computed from the SNPs containing all genotyped individuals. The genetic correlation estimated with the multi-trait model is the correlation between **GEBV**$_A$ and **GEBV**$_B$. This correlation is equivalent to the correlation between estimated SNP effects in each population, since **GEBV**$_A$ and **GEBV**$_B$ are calculated using the same genotypes for the same subset of individuals and different SNP effects, that are specific for each population. The estimated SNP effects might differ between the populations due to differences in QTL effects, but also due to differences in the LD between QTL and SNPs and the accuracy of estimating the effects. This indicates that the estimated genetic correlation at the SNPs is likely to be lower than the true genetic correlation at the QTL, when the LD pattern is not consistent for the two populations (Gianola *et al.* 2015) and when SNP effects are not estimated with 100% accuracy (Calo *et al.* 1973). Therefore, it can be expected that the estimated genetic correlation based on SNP information is underestimating the true genetic correlation between the populations.

Besides differences in LD, the estimated genetic correlation might also be influenced by the genomic relationship matrix (**G**) used in the multi-trait GBLUP model. In Chapter 3, the genetic correlation between different populations was calculated using a **G** matrix based on high-density genotypes, by using the average allele frequencies across the populations to set-up **G**. By using those average allele frequencies, the base population of **G** is a kind of admixed population and not the current population. This results in an increase in the relationships within a population and relationships below zero between populations (Karoui *et al.* 2012; Makgahlela *et al.* 2013). In general, negative relationships are just as informative as positive relationships. So, when the negative relationships between populations are absolutely higher than the relationships between unrelated individuals within a population, which is for at least part of the relationships the case in Chapter 3 and in Karoui *et al.* (2012), more information can be shared between individuals from different populations than between unrelated individuals from the same population, which is counterintuitive. Moreover, the higher relationships within a population indicate that the inbreeding level is higher than expected when the current population was used as base population, resulting in higher estimated genetic variances. Altogether, it can be concluded that a **G** matrix based on average allele frequencies might not be the most appropriate **G** matrix for estimating the genetic correlation.

A more appropriate way to calculate the **G** matrix for estimating the genetic correlation might be the approach described by Erbe *et al*. (2012). This approach of calculating **G** sets the base population at the time when the populations split, indicating that the relationships within a population include a high inbreeding level and relationships between populations are on average zero, assuming unrelated individuals. The genetic correlation estimated with this approach would, however, refer to the genetic correlation in the base population, before the two populations separated. Another approach of calculating **G** would be to use population-specific allele frequencies, resulting in relationships between populations of on average zero and relationships between unrelated individuals within a population of on average zero as well, since the current population was used as base population. This would indicate that the average relationships between unrelated individuals within a population and relationships between individuals from different populations are on the same level, which is counterintuitive as well.

To check which of the **G** matrices should be used to obtain the most accurate estimate of the genetic correlation, the analyses of Chapter 3 were repeated using the different **G** matrices based on high-density genotypes. In Table 7.1, the estimated genetic correlations between the Holstein Friesian and Groningen White Headed populations are shown for three different simulated genetic correlations. Those results show that the estimated genetic correlations were generally close to the simulated genetic correlation, with the most accurate estimate when the **G** matrix described by Erbe *et al*. (2012) was used. Differences in LD were present in those populations, as is shown in Chapter 4, indicating that the effect of differences in LD between populations on the estimated genetic correlation was very minimal. Therefore, it can be hypothesized that genetic correlations can be estimated in an accurate way by using a multi-trait GBLUP model, at least when high-density genotypes were used to set-up the genomic relationship matrix. When the density of the SNPs is lower, the impact of differences in LD on the estimated genetic correlation might be higher.

Table 7.1 Estimated genetic correlations (standard errors across replicates) between the populations using different approaches.

Set-up G matrix	Estimated genetic correlations (s.e.)					
	True r_G = 1		True r_G = 0.6		True r_G = 0.2	
Average AFreq[a]	0.89	(0.01)	0.56	(0.02)	0.16	(0.02)
Method of Erbe[b]	0.91	(0.01)	0.58	(0.02)	0.16	(0.02)
Pop.-specific AFreq[c]	0.86	(0.01)	0.51	(0.02)	0.15	(0.02)

[a] **G** matrix is calculated using average allele frequencies across the populations;
[b] **G** matrix is calculated using the method described by Erbe *et al.* (2012);
[c] **G** matrix is calculated using population-specific allele frequencies;
r_G = genetic correlation.

7.5 Consistency of LD between populations

Besides the genetic correlation, the consistency of LD across populations is another important factor influencing the accuracy of multi-population genomic prediction. In this thesis, two different values referring to the consistency of LD across populations are described, namely the consistency of multi-locus LD ($r_{MLLD_{RP,SK}}$; Chapter 4) and the effective number of chromosome segments ($M_{e_{RP,SK}}$; Chapter 3 and 6) between the reference population and selection candidates. Both $r_{MLLD_{RP,SK}}$ and $M_{e_{RP,SK}}$ are affected by the relatedness between reference and selection candidates. When reference and selection individuals are highly related to each other, for example due to a high level of family relationships, the LD pattern can be expected to be highly consistent between reference and selection individuals with the same allele of a SNP in high LD with a QTL allele. This indicates that $r_{MLLD_{RP,SK}}$ can be approximately 1 and $M_{e_{RP,SK}}$ is more or less similar to M_e within the reference population ($M_{e_{RP}}$), resulting in a high accuracy of genomic prediction (Chapter 2). When the level of family relationship between reference and selection individuals is low, the LD pattern might be less consistent and a different SNP or a different SNP allele can be in high LD with a QTL allele across populations. This indicates that $r_{MLLD_{RP,SK}}$ can be much lower than 1 and $M_{e_{RP,SK}}$ much larger than $M_{e_{RP}}$, resulting in a lower accuracy of genomic prediction (Chapter 2). Since the two measures both refer to the consistency of LD across populations, an interesting question is how the relation between those two measures can be described?

The value for M_e can be calculated as the inverse of the average LD between all pairs of loci on the same chromosome (Goddard *et al.* 2011), i.e., $\frac{1}{\bar{r}_{LD}^2}$. The consistency of multi-locus LD indicates how related the LD pattern of the selection candidates is to the LD pattern in the reference population, i.e., a consistency of multi-locus LD of 0.5 ($r_{MLLD_{RP,SK}}$ =0.5) indicates that the average LD between selection candidates and reference individuals is equal to $0.5^2 * \bar{r}_{LD}^2$, in which $\bar{r}_{LD}^2$ is the average LD in the reference population. So, for this case, the value for $M_{e_{RP,SK}}$ is $\frac{1}{0.5^2 * \bar{r}_{LD}^2} = \frac{1}{0.5^2} M_{e_{RP}}$. Or in general terms; $M_{e_{RP,SK}}$ can be calculated following:

$$M_{e_{RP,SK}} = \frac{1}{r_{MLLD_{RP,SK}}^2} M_{e_{RP}} . \tag{7.5}$$

This shows that the M_e between reference and selection individuals ($M_{e_{RP,SK}}$) is directly related to the consistency of multi-locus LD between the same individuals ($r_{MLLD_{RP,SK}}$). By knowing the M_e within the reference population and either $M_{e_{RP,SK}}$ or $r_{MLLD_{RP,SK}}$, the value for the other parameter can be calculated directly.

In Chapter 3, values for $M_{e_{RP}}$ and $M_{e_{RP,SK}}$ were obtained using Holstein Friesian individuals as reference population ($M_{e_{HF}}$ =185) and either Groningen White Headed ($M_{e_{HF,GWH}}$ =1809) or Meuse-Rhine-Yssel ($M_{e_{HF,MRY}}$ =2435) individuals as selection candidates. In Chapter 4, values for $r_{MLLD_{RP,SK}}$ were obtained based on the same data ($r_{MLLD_{HF,GWH}}$ =0.37; $r_{MLLD_{HF,MRY}}$ =0.33). Applying the estimates for $M_{e_{HF}}$, and respectively $r_{MLLD_{HF,GWH}}$ and $r_{MLLD_{HF,MRY}}$ in Equation 7.5, results in an estimate for $M_{e_{HF,GWH}}$ of 1351, and for $M_{e_{HF,MRY}}$ of 1699. Those estimates are not exactly the same as the estimates from Chapter 3, however, they are in good agreement with each other, especially when considering that estimating the consistency of multi-locus LD and M_e is prone to sampling variance.

7.6 Research directions for multi-population genomic prediction

As discussed in this general discussion, the currently used models do not show a large potential for using genomic information across populations, especially not across different breeds. This does not necessarily mean that there is no information that can be shared between different populations and might be related to the used models. By using other approaches, it might still be possible to use at least some information from one population to predict GEBVs for individuals from another population. In this part, the potential of the following research directions for multi-population genomic prediction is discussed: using sequence data in genomic prediction (7.6.1), using information of significant regions across populations (7.6.2), and including non-additive effects in the prediction model (7.6.3).

7.6.1 Sequence data in genomic prediction

In the last decade, the availability of whole-genome sequence data increased rapidly due to decreasing costs of this technology. Whole-genome sequence data is assumed to contain all variants, including the causal mutations or causal QTL underlying the traits of interest. Therefore, by using sequence data in genomic prediction models, the dependency on LD between QTL and SNPs is removed. This might especially be of interest for reference populations combining multiple populations, since the LD between QTL and SNPs is different across populations (e.g., Chapter 4; Andreescu *et al.* 2007; De Roos *et al.* 2008).

Simulation studies indeed showed an increase in accuracy of within-population genomic prediction by using sequence data compared to low- or high-density SNP data (Meuwissen and Goddard 2010; Clark *et al.* 2011; Druet *et al.* 2013; MacLeod *et al.* 2014a). Unfortunately, this increase is not completely confirmed in studies using real data, both for GBLUP and Bayesian variable selection models in Drosophila (Ober *et al.* 2012) and dairy cattle (Van Binsbergen *et al.* 2015). In another simulation study, the increase in accuracy by using sequence data was found to be even higher for multi-population genomic prediction (~16.5%) compared to single-population genomic prediction (~4.7%) (Iheshiulor *et al.* 2014), but, disappointingly, the increase in accuracy by using sequence data was only 2% for across-population genomic prediction in a study using real dairy cattle data across different traits (Hayes *et al.* 2014). A plausible explanation for those unexpected findings is that even though the causal QTL are included in the data, the genomic prediction models are not able to use this information properly and are still distributing the effect of the QTL across multiple variants.

Altogether, those results indicate that with the current prediction models, no or only a very small increase in accuracy can be expected when sequence data is used compared to high-density SNP information, even for multi-breed genomic prediction. In Chapter 5, it is shown that adding causal QTL to the SNP data resulted in an increase in accuracy, with a much larger increase in accuracy when the initial number of SNPs was lower. The lower initial number of SNPs reduced the dilution of the causal QTL effect over the SNPs, which resulted in a higher accuracy of estimating the effect. This shows that it is essential to reduce the number of variants in sequence data, without deleting the QTL, to be able to see an increase in accuracy. One approach of selecting variants is by using biological information. In a simulation study, it was shown that only including causal QTL in the model resulted in accuracies approaching 1 (Pérez-Enciso *et al.* 2015). Only including all SNPs in the genes affecting the trait also resulted in an increase in accuracy, however, when the genes were not selected with 100% accuracy, the accuracy dropped drastically (Pérez-Enciso *et al.* 2015). In studies using real data, variants of sequence data have been weighted differently based on the annotation of the variant, by giving a higher weight for coding versus non-coding variants (MacLeod *et al.* 2014b), or based on available information of significant SNPs, by giving a higher weight to SNPs shown to be significantly related to the trait in previous studies versus SNPs not shown to be related (Hayes *et al.* 2014; MacLeod *et al.* 2014b). Surprisingly, the accuracies of genomic prediction were not largely affected by the different weighting of the variants in the prediction model (Hayes *et al.* 2014). Based on those studies, it can be concluded that at the moment, the available knowledge about the genetic architecture of the different traits is insufficient to benefit from including biological information in the model.

Another approach to reduce the number of estimated effects is by using a principal component analysis on a genotype matrix (*n* x *p*), including for all *n* individuals the genotype for all *p* SNPs, and fitting the most important principal components as a variable in a regression model (e.g., Solberg *et al.* 2009; Macciotta *et al.* 2010; Dadousis *et al.* 2014). Although a study using real cattle data with 50,000 SNPs did not show an increase in accuracy by using principal components compared to GBLUP (Dadousis *et al.* 2014), this approach might increase the accuracy when sequence data is used. The reason for this expectation is that for sequence data, the number of estimated effects can decrease more drastically using principal components due to the higher dependencies between the variants.

Since principal components analyses are able to recover the structure of the data, using principal components might especially be attractive for structured data, such as multi-population and multi-breed reference populations.

Another approach to increase the potential of sequence information is to reduce the long-range LD present in the reference population as a result of a high level of relatedness in the population. This can, for example, be done by reducing the average relatedness within the reference population, which was already shown to be beneficial for within-population genomic prediction using SNP data (Pszczola *et al.* 2012). For sequence data, the LD between QTL and multiple SNPs can still be expected to be too high across unrelated individuals from the same population, indicating that finding the causal QTL using information from only one population is almost impossible. Combining individuals from distantly related populations and generations in the reference population, thereby reducing the consistency of LD in the reference population, might be necessary. By combining those distantly related populations, the level of family relationships in the reference population is reduced and the genomic prediction models are forced to focus more on short-range LD compared to long-range LD. This is supposed to improve the prediction performance across generations as well as across populations. In Chapter 4, for example, it is shown that short-range LD is more consistent across breeds and by focusing on the SNPs closely located to a QTL, the prediction performance across breeds can be improved. An important assumption underlying this approach is that the same QTL are underlying the trait and that the QTL have the same effect, which is unlikely to be the case for distantly related populations. This would greatly reduce the potential of this approach and, therefore, I do not expect to see a large increase in accuracy by using this approach.

7.6.2 Information of significant regions

Another approach to increase the accuracy of multi-population genomic prediction is to focus on sharing information about significant regions between populations. Even though the causal QTL might partly be different across populations, the QTL underlying the trait might still be located in the same regions on the genome. Therefore, it can be expected that the regions containing SNPs with a large effect show an overlap across populations. This is supported by the findings in literature, showing that pre-selecting SNPs with a large effect in one French cattle breed can help to increase the accuracy for some traits in another French cattle breed (Hozé *et al.* 2014a). Including information of regions with a large effect in one population in the prior of a Bayesian model for another population also helped to increase the prediction accuracy (Brøndum *et al.* 2012).

Moreover, it was shown that the accuracy of predicting GEBVs for residual feed intake in Holstein Friesian animals can be increased by giving a larger weight to SNPs significantly associated with residual feed intake in beef cattle (Khansefid *et al.* 2014). This indicates that there is a large potential to increase the amount of information that can be shared across populations by shifting the emphasis from combining populations to increase the accuracy of estimating SNP effects to combining populations to find regions associated with a trait. Another advantage is that in this approach, it is not necessary to account for differences in apparent SNP effects between breeds that might exist due to differences in causal QTL, LD, and allele substitution effects of QTL, since the effects are estimated separately in each population.

The studies mentioned before have all used a two-step approach, in which first significant SNPs or regions are localized in one population and information of those SNPs or regions is used later on as input for the model in another population (Brøndum *et al.* 2012; Hozé *et al.* 2014a; Khansefid *et al.* 2014). For practical applications, combining the localization of significant regions and estimating the effects might be attractive. This can, for example, be done by combining the information from both populations for defining which SNPs to include in a Bayesian variable selection model with a large effect, and consecutively estimate the SNP effects separately in each population. This suggestion is comparable to the multi-task Bayesian learning model described by Chen *et al.* (2014), which was shown to be able to increase the accuracy for a population with a low number of individuals in the reference population and keeping the accuracy of the population with a high number of individuals in the reference population at the same level. The chance of missing QTL that are only segregating in the population with a low number of individuals in the reference population is, however, reasonably high. Moreover, QTL with a large effect that have an opposite linkage phase with the surrounding SNPs might be missed as well. Therefore, the ideal model would be able to use the information from both populations to decide on which SNPs to assign a large effect, but would still be flexible enough to be able to assign a large effect to other SNPs when there is convincing evidence for that in one of the populations.

7.6.3 Non-additive effects

The estimated SNP effects might not only be different across populations due to differences in the causal QTL or LD with the QTL, but might also be different due to the existence of non-additive effects at the QTL in combination with differences in allele frequencies. This indicates that estimating non-additive effects in the model can help to find more consistent effects across populations, which can help to increase the prediction accuracy when populations are combined. Although different studies using genomic information have shown that non-additive effects, such as dominance and epistasis, exist in livestock populations (Carlborg *et al.* 2004; Große-Brinkhaus *et al.* 2010; Lopes *et al.* 2014), including non-additive effects in estimating GEBVs within one population has not been able to greatly improve the accuracy of GEBVs (Huang *et al.* 2012; Su *et al.* 2012). This is suggested to be a result of capturing a substantial part of the non-additive effects by the average allele substitution effects estimated in an additive model (Falconer and Mackay 1996). This is especially the case when QTL have a low minor allele frequency (Hill *et al.* 2008), since at a low minor allele frequency, the genotypes in a population are mainly at one side of the spectrum being either homozygous for one of the alleles or heterozygous, with only a very small number of individuals homozygous for the minor allele. This indicates that, for QTL with a low minor allele frequency, the number of individuals homozygous for the minor allele might be too low to accurately disentangle the additive and dominance effect.

Due to differences in allele frequencies across populations, a reasonably high number of homozygous, heterozygous, and opposing homozygous individuals might be present when two populations are combined, which is beneficial for estimating non-additive effects. By estimating dominance effects for each of the SNPs in a linear model, the number of estimated effects is doubled compared to fitting only additive effects. When also first-order epistatic effects between the SNPs are fitted in a linear model, the number of estimated effects is an exponential function of the number of SNPs. Therefore, the number of effects that have to be estimated can become that high, that it is impossible to estimate all of them in a genomic prediction model. So, the accuracy of genomic prediction when multiple populations are combined in the reference population might be increased by including dominance effects in a linear model, but probably not by including epistatic effects, even though epistatic interactions might explain a large part of the phenotypic variance. Other models would be needed to efficiently estimate the first- or even higher-order epistatic interactions. Non-parametric models, such as kernel regressions, have the potential to fit epistatic interactions, without explicitly modelling all pair-wise interactions (Gianola and Van Kaam 2008; González-Recio *et al.* 2008; De los Campos *et al.* 2010). Although those models have the potential to exploit the non-additive genetic effects in a more efficient way, there is no empirical evidence for this yet (Gianola *et al.* 2014).

7.6.4 Concluding remarks regarding the different research directions

In total, three different research directions to improve the accuracy of multi-population genomic prediction were discussed in this paragraph, namely using sequence data in genomic prediction, using information of significant regions across populations, and including non-additive effects in the prediction model. Overall, I conclude that using information of significant regions across populations has the highest potential to increase the accuracy of multi-population genomic prediction in the coming 10 years. One of the main advantages of this approach is that it uses information from other populations to identify regions related to a trait, but that it is still able to estimate the effects separately within each population. Using this approach, it was shown to be possible to share information between very distantly related populations, for instance between populations from different beef cattle breeds and a Holstein Friesian population. For more closely related populations, I expect an even higher potential to share information, due to the smaller differences between the populations. For a practical application of this approach, the current models should be slightly modified. The first steps to implement those changes are already taken by different research groups, indicating that a practical application of this approach should be possible in the near future.

Estimating non-additive effects in the prediction has the potential to improve the consistency of the estimated effects across the populations, and, therefore, can improve the accuracy of predicting GEBVs when multiple populations are combined in the reference population. A large disadvantage of this approach is the enormous increase in the number of effects that has to be estimated when both dominance and epistatic interactions are explicitly modeled. This might even be more pronounced for multi-population genomic prediction, since the number of estimated effects is already larger in those scenarios, due to the larger effective number of chromosome segments across populations than within populations. Non-parametric models can potentially help to efficiently estimate non-additive effects, however, those models still need to be optimized. In my opinion, the modelling of non-additive effects can also be improved by increasing our knowledge about the genetic architecture of traits, since it can provide information about genomic regions influencing a trait and about genomic regions that are likely to contain large non-additive effects.

The main advantage of using sequence data in genomic prediction is that it probably contains the causal variants. To make optimal use of those causal variants, the number of other variants should be as low as possible. In my opinion, the best way to obtain this is by including biological information in the model, which requires to have a good understanding of the genetic architecture. At the moment, this information is not yet available, which reduces the potential of using sequence information in genomic prediction. Therefore, I strongly recommend to more thoroughly study the genetic architecture of the most important traits in animal breeding (e.g., milk production traits in dairy cattle). When it is possible to use information about the genetic architecture in genomic prediction, the dependency on a consistent LD phase between SNPs and QTL is reduced. Moreover, the number of effects that has to be estimated can be reduced to the number of QTL underlying the trait. This also increases the potential to model non-additive effects, which can further improve the accuracy of both single- and multi-population genomic prediction in the long-term.

7.7 References

1000 bull genomes consortium. "1000 bull genomes project." Retrieved July 28, 2015, from www.1000bullgenomes.com.

Andreescu, C., S. Avendano, S. R. Brown, A. Hassen, S. J. Lamont, *et al.*, 2007 Linkage disequilibrium in related breeding lines of chickens. Genetics 177: 2161-2169.

Brard, S. and A. Ricard, 2015 Is the use of formulae a reliable way to predict the accuracy of genomic selection? J. Anim. Breed. Genet. 132: 207-217.

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

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

Calo, L. L., R. E. McDowell, L. D. Van Vleck and P. D. Miller, 1973 Genetic aspects of beef production among Holstein-Friesians pedigree selected for milk production. J. Anim. Sci. 37: 676-682.

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

Carillier, C., H. Larroque and C. Robert-Granié, 2014 Comparison of joint versus purebred genomic evaluation in the French multi-breed dairy goat population. Genet. Sel. Evol. 46: 67.

Carlborg, Ö., P. M. Hocking, D. W. Burt and C. S. Haley, 2004 Simultaneous mapping of epistatic QTL in chickens reveals clusters of QTL pairs with similar genetic effects on growth. Genet. Res. 83: 197-209.

Chen, L., C. Li, S. Miller and F. Schenkel, 2014 Multi-population genomic prediction using a multi-task Bayesian learning model. BMC Genet. 15: 53.

Clark, S. A., J. M. Hickey and J. H. J. Van Der Werf, 2011 Different models of genetic variation and their effect on genomic evaluation. Genet. Sel. Evol. 43: 18.

Cooper, T. A., G. R. Wiggans and P. M. VanRaden, 2015 Short communication: Analysis of genomic predictor population for Holstein dairy cattle in the United States—Effects of sex and age. J. Dairy Sci. 98: 2785-2788.

Dadousis, C., R. F. Veerkamp, B. Heringstad, M. Pszczola and M. P. Calus, 2014 A comparison of principal component regression and genomic REML for genomic prediction across populations. Genet. Sel. Evol. 46: 60.

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

De Haas, Y., M. P. L. Calus, R. F. Veerkamp, E. Wall, M. P. Coffey, *et al.*, 2012 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 Haas, Y., J. E. Pryce, M. P. L. Calus, E. Wall, D. P. Berry, *et al.*, 2015 Genomic prediction of dry matter intake in dairy cattle from an international data set consisting of research herds in Europe, North America, and Australasia. J. Dairy Sci. 98: 6522-6534.

De los Campos, G., D. Gianola, G. J. M. Rosa, K. A. Weigel and J. Crossa, 2010 Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet. Res. 92: 295-308.

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.

Decker, J. E., S. D. McKay, M. M. Rolf, J. Kim, A. Molina Alcalá*, et al.*, 2014 Worldwide patterns of ancestry, divergence, and admixture in domesticated cattle. PLoS Genet. 10: e1004254.

Druet, T., I. M. Macleod and B. J. Hayes, 2013 Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. Heredity 112: 39-47.

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

Falconer, D. S., 1952 The problem of environment and selection. Amer. Nat. 86: 293-298.

Falconer, D. S. and T. F. C. Mackay, 1996 *Introduction to quantitative genetics*. Pearson Education Limited, Harlow.

Gautier, M., T. Faraut, K. Moazami-Goudarzi, V. Navratil, M. Foglio*, et al.*, 2007 Genetic and haplotypic structure in 14 European and African cattle breeds. Genetics 177: 1059-1070.

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

Gianola, D. and J. B. C. H. M. Van Kaam, 2008 Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178: 2289-2303.

Gianola, D., G. Morota and J. Crossa, 2014 Genome-enabled prediction of complex traits with kernel methods: What have we learned? Proc. 10th World Congr. Genet. Appl. Livest. Prod., ASAS, Vancouver.

Gianola, D., G. De los Campos, M. A. Toro, H. Naya, C.-C. Schön*, et al.*, 2015 Do molecular markers inform about pleiotropy? Genetics 201: 23-29.

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

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*, et al.*, 2008 Nonparametric methods for incorporating genomic information into genetic evaluations: An application to mortality in broilers. Genetics 178: 2305-2313.

Große-Brinkhaus, C., E. Jonas, H. Buschbell, C. Phatsara, D. Tesfaye*, et al.*, 2010 Epistatic QTL pairs associated with meat quality and carcass composition traits in a porcine Duroc× Pietrain population. Genet. Sel. Evol. 42: 39.

Haile-Mariam, M., J. E. Pryce, C. Schrooten and B. J. Hayes, 2015 Including overseas performance information in genomic evaluations of Australian dairy cattle. J. Dairy Sci. 98: 3443–3459.

Harris, B. L. and D. L. Johnson, 2010 Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. J. Dairy Sci. 93: 1243-1252.

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

Hayes, B. J., I. M. MacLeod, H. D. Daetwyler, P. J. Bowman, A. C. Chamberlain*, et al.*, 2014 Genomic prediction from whole genome sequence in livestock: the 1000 bull genomes project. Proc. 10th World Congr. Genet. Appl. Livest. Prod., ASAS, Vancouver.

Hill, W. G., M. E. Goddard and P. M. Visscher, 2008 Data and theory point to mainly additive genetic variance for complex traits. PLoS Genet. 4: e1000008.

Hozé, C., S. Fritz, F. Phocas, D. Boichard, V. Ducrocq*, et al.*, 2014a Genomic evaluation using combined reference populations from Montbéliarde and French Simmental breeds. Proc. 10th World Congr. Genet. Appl. Livest. Prod., ASAS, Vancouver.

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

Huang, H., J. J. Windig, A. Vereijken and M. P. Calus, 2014 Genomic prediction based on data from three layer lines using non-linear regression models. Genet. Sel. Evol. 46: 75.

Huang, W., S. Richards, M. A. Carbone, D. Zhu, R. R. H. Anholt*, et al.*, 2012 Epistasis dominates the genetic architecture of Drosophila quantitative traits. Proc. Nat. Acad. Sci. U. S. A. 109: 15553-15559.

Iheshiulor, O. O. M., J. A. Woolliams, X. Yu, R. Wellmann and T. H. E. Meuwissen, 2014 Genomic predictions using whole genome sequence data and multi-breed reference populations. Proc. 10th World Congr. Genet. Appl. Livest. Prod., ASAS, Vancouver.

Jorjani, H., J. Jakobsen, M. A. Nilforooshan, E. Hjerpe, B. Zumbach*, et al.*, 2011 Genomic evaluation of BSW populations, InterGenomics: Results and Deliverables. Interbull Bull. 43: 5-8.

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

Kemper, K. E., B. J. Hayes, H. D. Daetwyler and M. E. Goddard, 2015a How old are quantitative trait loci and how widely do they segregate? J. Anim. Breed. Genet. 132: 121-134.

Kemper, K. E., C. M. Reich, P. J. Bowman, C. J. Vander Jagt, A. J. Chamberlain*, et al.*, 2015b Improved precision of QTL mapping using a nonlinear Bayesian method in a multi-breed population leads to greater accuracy for across-breed genomic predictions. Genet. Sel. Evol. 47: 29.

Khansefid, M., J. E. Pryce, S. Bolormaa, S. P. Miller, Z. Wang*, et al.*, 2014 Estimation of genomic breeding values for residual feed intake in a multibreed cattle population. J. Anim. Sci. 92: 3270-3283.

Legarra, A., G. Baloche, F. Barillet, J. Astruc, C. Soulas*, et al.*, 2014 Within-and across-breed genomic predictions and genomic relationships for Western Pyrenees dairy sheep breeds Latxa, Manech, and Basco-Béarnaise. J. Dairy Sci. 97: 3200-3212.

Lehermeier, C., C.-C. Schön and G. De los Campos, 2015 Assessment of genetic heterogeneity in structured plant populations using multivariate whole-genome regression models. Genetics 201: 323-337.

Lillehammer, M., M. Árnyasi, S. Lien, H. G. Olsen, E. Sehested*, et al.*, 2007 A genome scan for quantitative trait locus by environment interactions for production traits. J. Dairy Sci. 90: 3482-3489.

Lopes, M. S., J. W. M. Bastiaansen, B. Harlizius, E. F. Knol and H. Bovenhuis, 2014 A genome-wide association study reveals dominance effects on number of teats in pigs. PLoS ONE 9: e105867.

Lund, M. S., S. P. W. De Roos, A. G. De Vries, T. Druet, V. Ducrocq*, et al.*, 2011 A common reference population from four European Holstein populations increases reliability of genomic predictions. Genet. Sel. Evol. 43: 43.

Macciotta, N. P. P., G. Gaspa, R. Steri, E. L. Nicolazzi, C. Dimauro*, et al.*, 2010 Using eigenvalues as variance priors in the prediction of genomic breeding values by principal component analysis. J. Dairy Sci. 93: 2765-2774.

MacLeod, I. M., B. J. Hayes and M. E. Goddard, 2014a The effects of demography and long term selection on the accuracy of genomic prediction with sequence data. Genetics 198: 1671–1684.

MacLeod, I. M., B. J. Hayes, C. J. Vander Jagt, K. E. Kemper, M. Haile-Mariam*, et al.*, 2014b A Bayesian analysis to exploit imputed sequence variants for QTL discovery. Proc. 10th World Congr. Genet. Appl. Livest. Prod., ASAS, Vancouver.

Makgahlela, M. L., I. Strandén, U. S. Nielsen, M. J. Sillanpää and E. A. Mäntysaari, 2013 The estimation of genomic relationships using breedwise allele frequencies among animals in multibreed populations. J. Dairy Sci. 96: 5364-5375.

Maurice-Van Eijndhoven, M. H. T., H. Bovenhuis, R. F. Veerkamp and M. P. L. Calus, 2015 Overlap in genomic variation associated with milk fat composition in Holstein Friesian and Dutch native dual-purpose breeds. J. Dairy Sci. 98: 6510-6521.

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

Ober, U., J. F. Ayroles, E. A. Stone, S. Richards, D. Zhu*, et al.*, 2012 Using whole-genome sequence data to predict quantitative trait phenotypes in Drosophila melanogaster. PLoS Genet. 8: e1002685.

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

Pérez-Enciso, M., J. C. Rincón and A. Legarra, 2015 Sequence-vs. chip-assisted genomic selection: accurate biological information is advised. Genet. Sel. Evol. 47: 43.

Pryce, J., W. Wales, Y. De Haas, R. Veerkamp and B. Hayes, 2014 Genomic selection for feed efficiency in dairy cattle. Animal 8: 1-10.

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

Schaeffer, L. R., 1994 Multiple-country comparison of dairy sires. J. Dairy Sci. 77: 2671-2678.

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

Su, G., O. F. Christensen, T. Ostersen, M. Henryon and M. S. Lund, 2012 Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. PLoS ONE 7: e45293.

Van Binsbergen, R., M. Calus, M. Bink, F. Van Eeuwijk, C. Schrooten*, et al.*, 2015 Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. Genet. Sel. Evol. 47: 71.

Van den Berg, S., M. P. L. Calus, T. H. E. Meuwissen and Y. C. J. Wientjes, 2015 Across population genomic prediction scenarios in which Bayesian variable selection outperforms GBLUP. Submitted to BMC Genet.

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

VanRaden, P. M., K. Olson, D. Null, M. Sargolzaei, M. Winters*, et al.*, 2012 Reliability increases from combining 50,000-and 777,000-marker genotypes from four countries. Interbull Bull. 46: 75-79.

Wiggans, G. R., P. M. VanRaden and T. A. Cooper, 2011 The genomic evaluation system in the United States: Past, present, future. J. Dairy Sci. 94: 3202-3211.

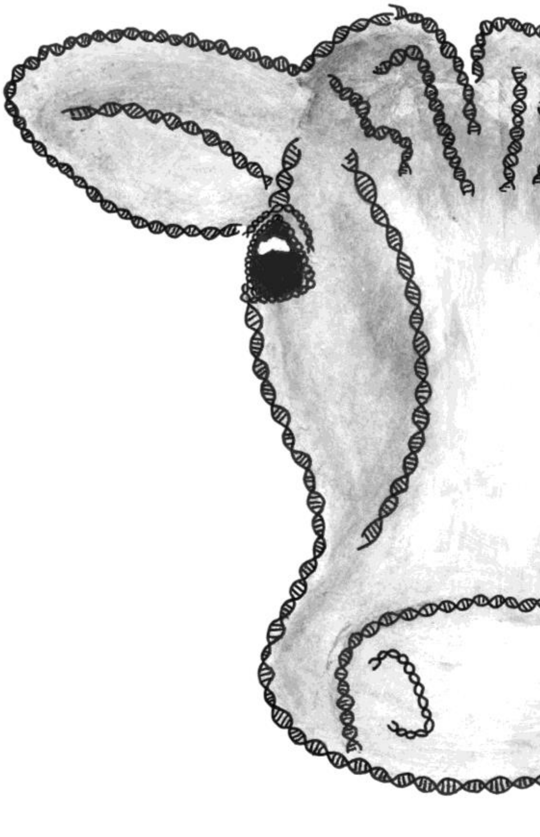
Wiggans, G. R., G. Su, T. A. Cooper, U. S. Nielsen, G. P. Aamand*, et al.*, 2015 Short communication: Improving accuracy of Jersey genomic evaluations in the United States and Denmark by sharing reference population bulls. J. Dairy Sci. 98: 3508-3513.

Zhou, L., X. Ding, Q. Zhang, Y. Wang, M. S. Lund*, et al.*, 2013 Consistency of linkage disequilibrium between Chinese and Nordic Holsteins and genomic prediction for Chinese Holsteins using a joint reference population. Genet. Sel. Evol. 45: 7.

Zhou, L., B. Heringstad, G. Su, B. Guldbrandtsen, T. Meuwissen*, et al.*, 2014 Genomic predictions based on a joint reference population for the Nordic Red cattle breeds. J. Dairy Sci. 97: 4485-4496.

Zumbach, B., H. Jorjani and J. Dürr, 2010 Brown Swiss genomic evaluation. Interbull Bull. 42: 44-51.

SUMMARY



In livestock breeding programs, genotype information is widely used to identify the genetically best animals to produce the next generation. For identifying those animals, genomic estimated breeding values are calculated for selection candidates using genotype information of many single-nucleotide polymorphism (SNP) markers spread across the genome. This information is combined with SNP effects, estimated in a reference population containing individuals with both phenotypes and SNP genotypes. For numerically small populations, the size of the reference population is often limited, which restricts the accuracy of genomic estimated breeding values for those populations as well as the response to selection. An attractive approach to increase the size of the reference population for numerically small populations is to add individuals from other populations, for example individuals from different countries, breeds, or lines. The differences between populations, such as differences in linkage disequilibrium (LD) between the SNPs and quantitative trait loci (QTL) underlying the trait, differences in allele frequencies of SNPs and QTL, differences in allele substitution effects of QTL, and the absence of close family relationships between populations, however, reduce the suitability of individuals from another population to increase the accuracy of genomic prediction.

Chapter 2 investigated the effect of absence of close family relationships between reference and selection individuals. The reference population for this study consisted of individuals with real genotype information. Five groups of selection candidates were simulated, using increasing amounts of information from the reference population: allele frequencies, LD pattern, haplotypes, haploid chromosomes, and family relationships. The results showed that the level of family relationships between reference and selection individuals has a higher effect on the accuracy of genomic prediction than LD *per se*. Moreover, the results showed that a deterministic equation using population parameters can accurately predict the accuracy for populations with complex family structures by estimating the effective number of chromosome segments (M_e) across reference and selection individuals, based on the genomic and pedigree based relationship matrix.

In **Chapter 3**, two different deterministic equations were derived to predict the accuracy of across-population genomic prediction. One equation was based on the genomic relationships within the reference population and between reference and selection individuals, the other equation was based on population parameters such as the M_e across populations. The equations were validated using real genotypes of three different cattle breeds and simulated phenotypes. It was shown that the equation based on genomic relationships was able to accurately estimate the accuracy. The equation based on population parameters overestimated the

accuracy by about 25 to 30%. Genetic correlations between populations lower than 1 reduced the accuracy of across-population genomic prediction, proportional to the genetic correlation. Therefore, the genetic correlation was an important input parameter for both equations. Moreover, it was shown that the number of QTL underlying the trait had no effect on the accuracy when a GBLUP type of model was used.

The same genotypes of the three different cattle breeds and simulated phenotypes were used in **Chapter 4** to investigate the consistency of multi-locus LD across populations, and its relationship with the accuracy of across-population genomic prediction. Since genomic prediction models are distributing the effect of a QTL among a number of SNPs, multi-locus LD was expected to be a better predictor for the potential of combining populations than consistency of LD between neighboring loci. The results showed that it was possible to estimate the consistency of multi-locus LD using a selection index approach, and that it could be seen as a characteristic of the properties of the QTL for the investigated populations. Consistency of multi-locus LD was highly related to the accuracy of across-population genomic prediction and can, therefore, be used to provide more insight in underlying reasons for a low empirical accuracy of across-population genomic prediction. By focusing only on SNPs closely located to a QTL, the consistency of multi-locus LD across populations increased. This indicates that the accuracy of across- and multi-population genomic prediction could be increased by focusing only on the neighboring SNPs of a QTL, for which the consistency of LD is higher across populations.

The effect of QTL properties, such as allele frequency pattern and distribution of allele substitution effects, on accuracy of multi-breed genomic prediction was investigated in **Chapter 5**. In this study, real genotype information of Holstein Friesian and Jersey cows was used. For all those individuals, three classes of variants obtained from whole-genome sequence data were imputed. Those classes of variants differed in their allele frequency pattern, ranging from moderately low to extremely low average minor allele frequencies (MAF), and amount of breed-specific variants. Phenotypes were simulated by sampling QTL from one of the classes of variants and by either randomly sampling an allele substitution effect for each QTL or by assigning larger effects to QTL with a low MAF. The accuracy of both single- and multi-population genomic prediction was shown to be lower when the average MAF of QTL underlying the trait was lower, especially when rare alleles were given a larger effect. It was demonstrated that QTL properties are key parameters determining the accuracy of genomic prediction. Those results show that the properties of QTL that underlie a trait can explain the limited benefit or the absence of benefit of combining information from multiple breeds that is described in empirical studies as opposed to the substantial benefit that is obtained in simulation studies.

In **Chapter 6**, a deterministic equation was developed to predict the accuracy of multi-population genomic prediction when populations from different breeds, lines or environments, or populations measured for different traits are combined in the reference population. The equation is using population parameters such as the M_e across populations and the genetic correlation between populations. Validation was performed using real genotypes and simulated phenotypes of Holstein Friesian cows, that were divided in three different populations by keeping half-sib families in the same population. Results showed that the derived equation can accurately predict the accuracy for different scenarios of multi-population genomic prediction, representing multi-environment and multi-trait genomic prediction. Therefore, the derived equation can be used to investigate the potential accuracy of different multi-population genomic prediction scenarios and to decide on the most optimal design of reference populations.

The general discussion of this thesis, presented in **Chapter 7**, discusses five different topics. As a first topic, the potential of multi-population genomic prediction is discussed by considering different scenarios, such as combining populations from the same breed from different countries, closely related breeds, or distantly related breeds. It is shown that combining populations in one reference population is likely to result in an increase in accuracy when; 1) the combined populations are closely related, 2) the population of the selection candidates in the reference population is small, and 3) the number of individuals added from the other population is very large. Therefore, the most optimal design to increase the accuracy of genomic prediction for numerically small populations would be to add a large number of individuals from the same breed from another country. Whenever that is not possible, it might help to add a large number of individuals from a closely related breed. Adding individuals from a distantly related breed is not expected to result in an increase in accuracy, due to the large differences between the populations.

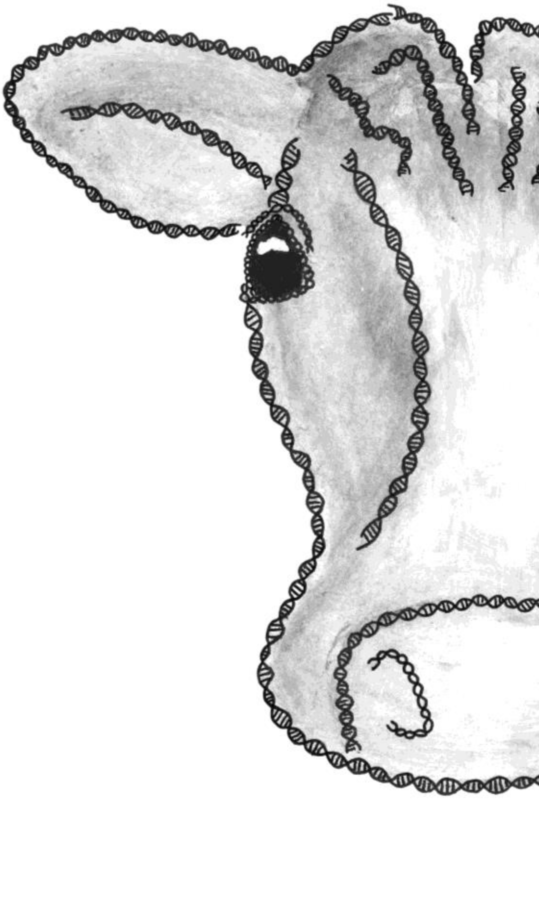
As a second topic, the impact of the model used to estimate genomic breeding values on the accuracy of multi-population genomic prediction is discussed. It is hypothesized that Bayesian variable selection models are better able to use information across closely related populations compared to GBLUP, especially for traits influenced by a low number of QTL or by a few QTL with large effect. This is a result of focusing more on the SNPs close to a QTL in a Bayesian variable selection model compared to GBLUP, which reduces the number of effects that have to be estimated and increases the proportion of the genetic variance captured by the SNPs in another population.

As a third topic, the estimation of the genetic correlation using SNP information is discussed. In this part, it is hypothesized that the genetic correlation can be accurately estimated using a multi-trait GBLUP model, where the same trait in the different populations is measured as a different trait. It is shown that the most accurate estimate of the genetic correlation can be obtained when the genomic relationship matrix is set-up using the population just before the populations split as base population.

As a fourth topic, the relation between the M_e across populations and the consistency of multi-locus LD across populations is discussed. Both measures reflect on the consistency of LD across populations, which is an important parameter influencing the accuracy of multi-population genomic prediction. It is shown that the M_e across reference and selection individuals is directly related to the consistency of multi-locus LD between the same individuals. So, when the M_e within the reference population is known and one of the measures for the consistency of LD across populations, the other measure can be calculated directly.

As a fifth topic, research directions for multi-population genomic prediction are discussed, focusing on the use of sequence data in genomic prediction, the identification and use of significant regions across populations, and the potential of including non-additive effects in genomic prediction models. The research direction which is suggested to have the highest potential to increase the accuracy of multi-population genomic prediction in the coming 10 years is the identification and use of significant regions across populations. In this research direction, it is assumed that even though the QTL or the effects of the QTL underlying the trait might differ across populations, the QTL are located in the same regions on the genome. Moreover, it is discussed that for optimizing the use of sequence data as well as for including non-additive effects for genomic prediction, more information about the genetic architecture of the trait should become available.

SAMENVATTING



In fokkerij-programma’s van landbouwhuisdieren is het steeds gebruikelijker om DNA-profielen, bestaande uit duizenden merkers op het DNA, te gebruiken om de genetisch beste dieren te selecteren voor het voortbrengen van de volgende generatie. Hiervoor wordt eerst in een referentie-populatie, oftewel een groep dieren met bekende DNA-profielen en productiegegevens (ook wel fenotypes genoemd), voor ieder van de merkers een effect op een bepaald kenmerk geschat. Deze geschatte merker-effecten worden gebruikt om genomische fokwaardes te berekenen voor jonge dieren waarvan de fenotypes nog niet bekend zijn, maar de DNA-profielen wel. Op basis van deze genomische fokwaardes worden de beste jonge dieren geselecteerd uit een groep van selectie-kandidaten. Aangezien een groot aantal merkers wordt gebruikt, heeft ieder gen wel een relatie met een paar merkers, waardoor de merkers het effect van de genen op een kenmerk kunnen verklaren.

Voor populaties met een klein aantal dieren is de referentie-populatie meestal te klein om de merker-effecten betrouwbaar te kunnen schatten, met een lage betrouwbaarheid van de genomische fokwaardes als gevolg. Dit maakt het lastiger om de beste dieren te selecteren en beperkt de genetische vooruitgang. Een manier om de referentie-populatie te vergroten is door dieren van een andere populatie toe te voegen, bijvoorbeeld dieren van een ander ras of uit een ander land. De verschillen tussen populaties maken het echter moeilijker om informatie van andere populaties te gebruiken. Zo zijn er tussen populaties geen sterke familierelaties, kunnen andere merkers een relatie hebben met een bepaald gen en kan het zijn dat de genen andere effecten hebben. In dit proefschrift is gekeken of en hoeveel de betrouwbaarheid van genomische fokwaardes verhoogd kan worden door informatie van verschillende populaties te combineren. De focus lag hierbij op het vergroten van de kennis over dit onderwerp, niet op het verhogen van de betrouwbaarheid van genomische fokwaardes voor een specifiek ras of kenmerk.

Hoofdstuk 2 beschrijft het effect van de afwezigheid van sterke familierelaties tussen de referentie-populatie en de selectie-kandidaten. De resultaten van deze studie laten zien dat het aantal familierelaties tussen de referentie-populatie en de selectie-kandidaten een grote invloed heeft op de betrouwbaarheid van het berekenen van genomische fokwaardes. Als familierelaties tussen de referentie-populatie en de selectie-kandidaten afwezig waren, was de betrouwbaarheid laag, zelfs als op DNA-niveau de relatie tussen de merkers en genen hetzelfde was. Daarnaast laat deze studie zien dat het mogelijk is om de behaalde betrouwbaarheid nauwkeurig te voorspellen door het ‘aantal effectieve chromosoom segmenten’ (M_e) tussen de referentie-populatie en de selectie-kandidaten te berekenen. De M_e geeft aan hoe verschillend de DNA-profielen van verschillende groepen dieren zijn.

In **Hoofdstuk 3** zijn twee formules afgeleid om de betrouwbaarheid te voorspellen van genomische fokwaardes als de referentie-populatie bestaat uit dieren van een andere populatie dan de selectie-kandidaten. Beide formules zijn gevalideerd door fenotypes te simuleren van drie verschillende melkveerassen op basis van echte DNA-profielen. De eerste formule gebruikt de genomische relaties tussen de referentie-populatie en selectie-kandidaten, waarbij de genomische relatie tussen twee dieren aangeeft hoeveel gelijkenis hun DNA-profielen vertonen. Deze eerste formule berekent de betrouwbaarheid van genomische fokwaardes erg nauwkeurig. De tweede formule gebruikt populatie-parameters, zoals de eerder genoemde M_e tussen referentie-populatie en selectie-kandidaten, en overschatte de betrouwbaarheid met 25 tot 30%. Voor beide formules was de genetische correlatie tussen de populaties, oftewel de correlatie tussen de effecten van de genen in de verschillende populaties, een belangrijke input-parameter.

Dezelfde DNA-profielen en gesimuleerde fenotypes van de drie melkveerassen zijn gebruikt in **Hoofdstuk 4**. Het doel van deze studie was om inzicht te krijgen in hoeverre de relatie tussen merkers en genen overeenkomt in verschillende populaties en wat de invloed hiervan is op de betrouwbaarheid van genomische fokwaardes als de referentie-populatie bestaat uit dieren van een andere populatie. De resultaten laten zien dat de relatie tussen merkers en genen gedeeltelijk anders is in verschillende populaties, waardoor de merkers die een bepaald kenmerk verklaren anders kunnen zijn in verschillende populaties. De mate waarin de relatie tussen merkers en genen verschilt, had een sterk verband met de betrouwbaarheid van de genomische fokwaardes van dieren uit een andere populatie dan de referentie-dieren. Daarnaast laat deze studie zien dat als alleen de merkers dichtbij een gen meegenomen worden, de relatie tussen merkers en genen meer overeenkomt tussen populaties, wat mogelijk de betrouwbaarheid van genomische fokwaardes voor dieren uit een andere populatie dan de referentie-dieren kan verhogen.

Het effect van de eigenschappen van de genen op de betrouwbaarheid van genomische fokwaardes is onderzocht in **Hoofdstuk 5**. Hier is gekeken naar enkele scenario’s waarbij twee verschillende rassen zijn samengevoegd in de referentie-populatie. In deze studie is ervan uitgegaan dat ieder gen en iedere merker voorkomt in twee varianten. Er is gekeken naar het effect van de frequentie van de minst voorkomende variant van een gen. De resultaten laten zien dat de betrouwbaarheid van genomische fokwaardes over het algemeen lager is als de genen een lagere frequentie van de minst voorkomende variant hebben, vooral als deze genen ook nog een groot effect op een kenmerk hebben. Dit geeft aan dat het lastig is om de effecten van genen met een lage frequentie van de minst voorkomende variant betrouwbaar te schatten en dat de eigenschappen van genen met een effect op een kenmerk een grote invloed hebben op de betrouwbaarheid van genomische fokwaardes.

In **Hoofdstuk 6** is een formule afgeleid om de betrouwbaarheid te voorspellen waarmee genomische fokwaardes berekend kunnen worden op basis van een referentie-populatie bestaande uit verschillende populaties. Deze formule is gebaseerd op populatie-parameters, zoals de eerder genoemde M_e tussen de referentie-populatie en selectie-kandidaten en de genetische correlatie. Deze formule is gevalideerd door Holstein Friesians in drie populaties in te delen. Van deze dieren waren DNA-profielen bekend en zijn de fenotypes gesimuleerd. De formule voorspelde de betrouwbaarheid erg nauwkeurig voor diverse scenario's. Dit geeft aan dat deze formule gebruikt kan worden om keuzes te maken over hoe de optimale referentie-populatie eruit moet zien, wat belangrijke informatie is voor het opstellen van fokprogramma's.

De algemene discussie van dit proefschrift, beschreven in **Hoofdstuk 7**, bediscussieert vijf verschillende onderwerpen. Als eerste wordt de potentie van het samenvoegen van populaties in de referentie-populatie besproken, door sterk verwante en ver verwante populaties te combineren. Op basis hiervan kan geconcludeerd worden dat het samenvoegen van populaties kan leiden tot een hogere betrouwbaarheid van genomische fokwaardes, wanneer: 1) de gecombineerde populaties nauw verwant zijn, 2) de populatie waartoe de selectie-kandidaten behoren klein is, en 3) het aantal toegevoegde dieren van een andere populatie groot is. Dit geeft aan dat het toevoegen van dieren van hetzelfde ras, maar uit een ander land, aan de referentie-populatie de beste manier is om de betrouwbaarheid van genomische fokwaardes te vergroten. Als dat niet mogelijk is, kan het helpen om dieren van een sterk verwant ras aan de referentie-populatie toe te voegen. Het toevoegen van dieren van een ver verwant ras heeft naar alle waarschijnlijkheid geen hogere betrouwbaarheid als gevolg, aangezien de rassen te verschillend zijn.

Als tweede wordt het effect van het gebruikte model voor het schatten van merker-effecten op de behaalde betrouwbaarheid van de fokwaardes bediscussieerd. Hier wordt de hypothese beschreven dat het beste model om DNA-informatie van verschillende populaties te gebruiken eerst een groep merkers selecteert met het meeste bewijs om effect te hebben op een kenmerk, en daarna alleen effecten schat voor deze groep merkers. Hierdoor hoeft dit model niet voor alle merkers een effect te schatten, waardoor ieder van de effecten nauwkeuriger kan worden geschat.

Als derde wordt het schatten van de genetische correlatie op basis van merker-informatie bediscussieerd. Met behulp van een zogenaamd multi-trait model, waarbij de fenotypes van de twee populaties als een verschillend kenmerk worden gemodelleerd, is het mogelijk om de genetische correlatie betrouwbaar te schatten. Hiervoor is het essentieel dat de genomische relaties tussen de rassen worden meegenomen.

Als vierde wordt bediscussieerd of de M_e tussen de referentie-populatie en selectie-kandidaten gerelateerd is aan de mate waarin de relatie tussen merkers en genen hetzelfde is in verschillende populaties. Beide parameters beschrijven namelijk de mate van verschil tussen DNA-profielen van verschillende populaties. In dit deel wordt aangetoond dat beide parameters sterk samenhangen en dat het mogelijk is om de waarde van de ene parameter uit te rekenen op basis van de andere parameter.

Als vijfde worden drie onderzoeksrichtingen bediscussieerd, welke de mogelijkheid hebben om beter gebruik te maken van informatie van andere populaties om de betrouwbaarheid van genomische fokwaardes te verhogen. De onderzoeksrichting met de hoogste potentie in de komende tien jaar is het identificeren van regio's op het DNA met een effect op een kenmerk en deze informatie in andere populaties te gebruiken. Hierbij wordt niet aangenomen dat de exacte locatie van een gen dat invloed heeft op een kenmerk hetzelfde is in verschillende populaties, maar wel dat ze in dezelfde gebieden voorkomen. Het gebruik van de hele DNA sequentie, oftewel alle merkers op het DNA, en het schatten van interactie-effecten tussen varianten van een gen of tussen verschillende genen kan in de toekomst ook voordelig zijn, maar dan is er eerst meer informatie nodig over de eigenschappen en locatie van de genen die effect hebben op een kenmerk.

Op basis van deze resultaten kan er geconcludeerd worden dat het combineren van informatie van verschillende populaties alleen in bepaalde gevallen leidt tot een hogere betrouwbaarheid van genomische fokwaardes. Met behulp van de afgeleide formules in dit proefschrift is het mogelijk geworden om te voorspellen voor welke scenario's de betrouwbaarheid zal stijgen en hoe groot de stijging zal zijn.

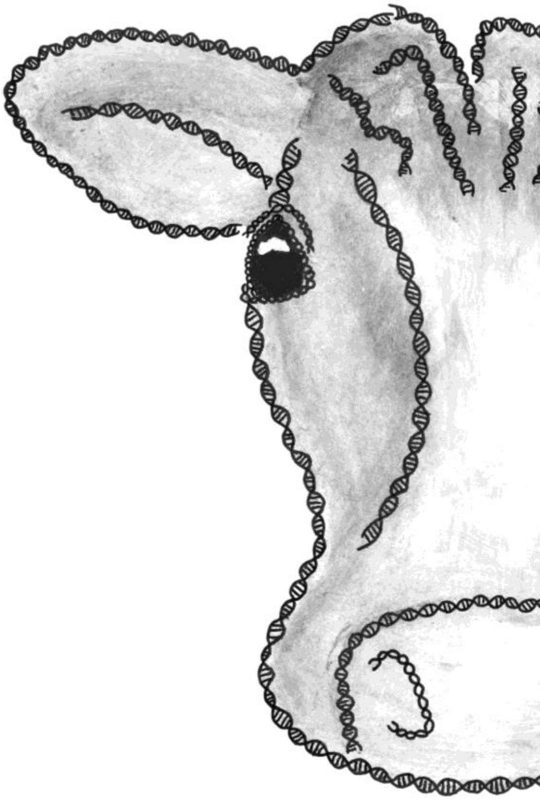
CURRICULUM VITAE

ABOUT THE AUTHOR

OVER DE AUTEUR

PUBLICATIONS

TRAINING AND SUPERVISION PLAN



About the author

Yvonne Cornelia Johanna Wientjes was born on the 19th of November 1987 in Boxmeer, and was raised on the dairy farm of her parents in Sint Anthonis, the Netherlands. In 2006, she graduated from high school Elzendaalcollege in Boxmeer. In the same year, she started her study Animal Sciences at Wageningen University. During her master, Yvonne specialized in Animal Nutrition and Animal Breeding and Genetics. For the specialization Animal Nutrition, she investigated the effect of health status on nitrogen retention and amino acid requirements in growing pigs. For the specialization Animal Breeding and Genetics, she studied the effect of mass selection for socially affected traits on the rate of inbreeding. The thesis about this topic was awarded the second prize for the NZV (Dutch Association of Animal Science Alumni) thesis award in 2012. Moreover, she spent four months at the TOPIGS Research Centre IPG for her internship, where she worked on diverse projects with a main focus on genomic analyses. After her cum laude graduation in 2011, Yvonne started her PhD research about the accuracy of multi-population genomic prediction at the Animal Breeding and Genomics Centre of Wageningen University and Research Centre. The results of her PhD research are described in this thesis. During her PhD, Yvonne spent three months at the Department of Environment and Primary Industries in Melbourne, Australia. Since November 2015, Yvonne is working as a postdoc on the 4-year project GenoMiX, focusing on accelerating genetic progress by utilizing crossbred information at the Animal Breeding and Genomics Centre of Wageningen University and Research Centre.

Over de auteur

Yvonne Cornelia Johanna Wientjes is geboren op 19 november 1987 te Boxmeer en opgegroeid op het melkveebedrijf van haar ouders in Sint Anthonis. In 2006 behaalde zij haar VWO-diploma aan het Elzendaalcollege in Boxmeer. In datzelfde jaar begon zij aan de studie Dierwetenschappen aan Wageningen Universiteit. Tijdens haar master heeft Yvonne zich gespecialiseerd in zowel Diervoeding als Fokkerij en Genetica. Voor de specialisatie Diervoeding heeft zij gekeken naar het effect van de gezondheidsstatus van vleesvarkens op de stikstok-retentie en aminozuurbehoefte. Voor de specialisatie Fokkerij en Genetica heeft zij gekeken naar het effect van het selecteren van de beste dieren op basis van eigen prestatie voor kenmerken die beïnvloed worden door sociale effecten op de inteelttoename. Hiermee behaalde Yvonne de tweede prijs bij de NZV (Nederlandse Zoötechnische Vereniging) thesis award in 2012. Daarnaast heeft zij voor deze specialisatie vier maanden stage gelopen bij TOPIGS Research Centre IPG, waar zij aan verschillende projecten heeft gewerkt met een focus op genomische analyses. Na haar cum laude afstuderen in 2011 is Yvonne gestart met haar promotieonderzoek naar de betrouwbaarheid van genomische fokwaardes geschat op basis van informatie van verschillende populaties bij het Animal Breeding and Genomics Centre aan Wageningen University and Research Centre. De resultaten van dit onderzoek zijn beschreven in dit proefschrift. Tijdens haar tijd als promovenda verbleef Yvonne drie maanden bij het Department of Environment and Primary Industries in Melbourne, Australië. Sinds november 2015 is Yvonne werkzaam als postdoc onderzoeker op het 4-jarige GenoMiX project bij het Animal Breeding and Genomics Centre van Wageningen University and Research Centre. Dit project heeft als doel om de genetische vooruitgang te versnellen door gebruik te maken van informatie van dieren welke een kruising zijn tussen twee rassen of lijnen.

Peer reviewed publications

Wientjes Y. C. J., P. Bijma, R. F. Veerkamp and M. P. L. Calus, An equation to predict the accuracy of genomic values by combining data from multiple traits, populations, or environments. Accepted in Genetics.

Van den Berg S., M. P. L. Calus, T. H. E. Meuwissen and Y. C. J. Wientjes, Across population genomic prediction scenarios in which Bayesian variable selection outperforms GBLUP. Submitted to BMC Genetics.

Wientjes, Y. C. J., R. F. Veerkamp and M. P. L. Calus, 2015 Using selection index theory to estimate consistency of multi-locus linkage disequilibrium across populations. BMC Genetics 16: 87.

Wientjes, Y. C. J., M. P. L. Calus, M. E. Goddard and B. J. Hayes, 2015 Impact of QTL properties on the accuracy of multi-breed genomic prediction. Genetics Selection Evolution 47: 42.

Wientjes, Y. C. J., R. F. Veerkamp, P. Bijma, H. Bovenhuis, C. Schrooten and M. P. L. Calus, 2015 Empirical and deterministic accuracies of across-population genomic prediction. Genetics Selection Evolution 47: 5.

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.

Conference proceedings, abstracts and presentations

Wientjes Y. C. J., P. Bijma, R. F. Veerkamp and M. P. L. Calus, 2015 Predicting the accuracy of multi-population genomic prediction. European Association of Animal Production, Warsaw, Poland.

Wientjes, Y. C. J., R. F. Veerkamp, P. Bijma, H. Bovenhuis, C. Schrooten and M. P. L. Calus, 2015 The potential of using genomic information across populations. Gordon Research Seminar and Conference, Lucca, Italy.

Calus M. P. L., H. Huang, J. W. M. Bastiaansen, J. ten Napel, M. D. Pryce, R. F. Veerkamp, A. Vereijken, Y. C. J. Wientjes and J. J. Winding, 2014 (A)cross-breed genomic prediction. World Congress on Genetics Applied to Livestock Production, Vancouver, Canada.

Wientjes, Y. C. J., M. P. L. Calus, M. E. Goddard and B. J. Hayes, 2015 Effect of genetic architecture on accuracy of multi breed genomic prediction. World Congress on Genetics Applied to Livestock Production, Vancouver, Canada.

Wientjes, Y. C. J., R. F. Veerkamp and M. P. L. Calus, 2014 Is it possible to use DNA information from Holstein Friesians to select the best animals in other breeds? WIAS Science Day, Wageningen, the Netherlands.

Wientjes, Y. C. J., R. F. Veerkamp, P. Bijma, H. Bovenhuis, C. Schrooten and M. P. L. Calus, 2014 Empirical and deterministic accuracies of across population genomic prediction. International Cattle Breeders Round Table, Odalgården, Sweden.

Calus M. P. L., Y. C. J. Wientjes and R. F. Veerkamp, 2013 Quantitative methods in genomics and animal breeding. Annual Meeting of the Brazilian Society of Animal Science, Campinas, Brazil.

Wientjes, Y. C. J., R. F. Veerkamp and M. P. L. Calus, 2013 Reliability of genomic prediction due to linkage disequilibrium and family relationships at different reference population sizes. Gordon Research Seminar and Conference, Galveston, United States of America.

Wientjes, Y. C. J., R. F. Veerkamp and M. P. L. Calus, 2012 Effect of linkage disequilibrium, haplotypes and family relations on accuracy of genomic prediction. European Association of Animal Production, Bratislava, Slovakia.

Wientjes, Y. C. J., R. F. Veerkamp and M. P. L. Calus, 2012 The effect of linkage disequilibrium, haplotypes and family relationships on the accuracy of direct genomic values. International Conference on Quantitative Genetics, Edinburgh, United Kingdom.

Training and supervision plan

The Graduate School WIAS WAGENINGEN INSTITUTE of ANIMAL SCIENCES

The Basic Package (3 ECTS)	**Year**
WIAS Introduction Course	2012
Ethics and Philosophy in Life Sciences	2013
Scientific Exposure (21 ECTS)	
International conferences (9 ECTS)	
4th International Conference on Quantitative Genetics (ICQG), Edinburgh, United Kingdom	2012
63rd Annual Meeting of the European Association of Animal Production (EAAP), Bratislava, Slovakia	2012
Gordon Research Seminar and Conference (GRS and GRC), Galveston, United States	2013
7th International Cattle Breeders Round Table (ICBRT), Odalgården, Sweden	2014
10th World Congress on Genetics Applied to Livestock Production (WCGALP), Vancouver, Canada	2014
Gordon Research Seminar and Conference (GRS and GRC), Lucca, Italy	2015
65th Annual Meeting of the European Association of Animal Production (EAAP), Warsaw, Poland	2015
Seminars and workshops (4 ECTS)	
WIAS Science Day, Wageningen, the Netherlands (4x)	2012-2015
WIAS Mini symposium on Advanced Genetics, Wageningen, the Netherlands	2012
Fokkerij en Genetica connectie dagen, Vught/Ellecom, the Netherlands (2x)	2012-2014
WIAS Seminar 'New opportunities for conservation genetics with genome wide information', Wageningen, the Netherlands	2012
Systems biology workshop, Melbourne, Australia	2013
Seminar 'How to Write a World-class Paper?', Wageningen, the Netherlands	2013
WIAS Seminar 'Genomic selection for novel traits', Wageningen, the Netherlands	2013
Wageningen PhD Symposium, Wageningen, the Netherlands	2013
Presentations (8 ECTS)	
Poster presentation at ICQG, Edinburgh, United Kingdom	2012
Oral presentation at EAAP, Bratislava, Slovakia	2012
Poster presentation at GRS and GRC, Galveston, United States	2013
Oral presentation at ICBRT, Odalgården, Sweden	2014

Oral presentation at WIAS Science Day, Wageningen, the Netherlands	2014
Oral presentation at WCGALP, Vancouver, Canada	2014
Poster presentation at GRS and GRC, Lucca, Italy	2015
Oral presentation at EAAP, Warsaw, Poland	2015

In-Depth Studies (8 ECTS)

Disciplinary and interdisciplinary courses (6 ECTS)

Genomic prediction, Edinburgh, United Kingdom	2012
Identity by descent approaches to genomic analyses of genetic traits, Wageningen, the Netherlands	2012
Advanced methods and algorithms in animal breeding with focus on genomic selection, Wageningen, the Netherlands	2012
Genetic analysis using ASReml4.0, Wageningen, the Netherlands	2014
Introduction to theory and implementation of Genomic Selection, Wageningen, the Netherlands	2014

PhD students' discussion groups (2 ECTS)

Quantitative Genetics Discussion Groups (QDG), Wageningen, the Netherlands	2011-2015

Professional Skills Support Courses (4 ECTS)

PhD Competence assessment, Wageningen, the Netherlands	2012
Techniques for Writing and Presenting a Scientific Paper, Wageningen, the Netherlands	2012
Project and Time management, Wageningen, the Netherlands	2012
Effective behaviour in your professional surroundings, Wageningen, the Netherlands	2014
WGS PhD Workshop Carousel, Wageningen, the Netherlands	2015

Research Skills Training (9 ECTS)

Preparing own PhD research proposal	2011-2012
External training period at Department of Environment and Primary Industries, Melbourne, Australia	2013
Reviewer Research Master Cluster	2014

Didactic Skills Training (3 ECTS)

One hour lecture in the course 'Introduction to theory and implementation of Genomic Selection', Wageningen, the Netherlands	2014
Supervising one MSc student	2014

Management Skills Training (8 ECTS)

WIAS Science Day Committee	2013
WIAS Associated PhD Students Council + Education Committee	2012-2014

Education and Training Total	**55 ECTS**

Dankwoord

Na vier jaar is het dan zo ver, mijn proefschrift is klaar! Wel is het met een dubbel gevoel dat ik dit schrijf. Van de ene kant ben ik blij dat het afgerond is, dat het doel van de afgelopen vier jaar om een proefschrift te schrijven is bereikt. Aan de andere kant vind ik het erg jammer dat het PhD kandidaat zijn nu ook is afgelopen, ik heb het namelijk een erg leuke en leerzame tijd gevonden. Met name van de sfeer binnen ABGC en van de samenwerking met verschillende mensen heb ik genoten. Iedereen bedankt die hieraan heeft bijgedragen! Een aantal mensen wil ik hiervoor in het bijzonder bedanken.

Als eerste wil ik mijn begeleiders Mario, Roel, Chris en Henk bedanken. Mario, ik ben er zeker van dat ik me geen betere dagelijkse begeleider had kunnen wensen. Altijd maakte je tijd voor mij en mijn proble... eh uitdagingen ;-), zelfs al zaten we aan verschillende kanten van de wereld. Tijdens mijn hele PhD tijd heb je mij het gevoel gegeven vertrouwen in mij te hebben, iets wat me zelf nog te vaak ontbreekt. Voor dit vertrouwen ben ik je erg dankbaar, want daardoor durfde ik het aan om mijn grenzen op te zoeken en zelfs te verleggen. Roel, ook jou wil ik bedanken voor je vertrouwen. Al was het af en toe een uitdaging om je opmerkingen te ontcijferen, ze waren altijd zinvol. Daarnaast was het zonder jouw advies om te focussen op de hoofdlijn nooit gelukt om de presentatie-prijs te winnen tijdens de WIAS Science Day. Chris, bedankt voor het delen van informatie over de praktijk van de rundveefokkerij en de daarbij behorende uitdagingen. En natuurlijk ook voor het kritisch doorlezen van mijn papers en het corrigeren van de taalfouten. Henk, bedankt voor jouw kritische blik op mijn werk. Dit heeft mij geholpen om anders naar dingen te kijken en hierdoor zijn veel papers beter geworden. Naast deze begeleiders, wil ik ook Piter bedanken voor het delen van zijn inzicht en enthousiasme over de theoretische kant van de fokkerij. Piter, je hebt mij geïnspireerd om meer deze theoretische kant op te gaan, iets waar ik zeker geen spijt van heb. Daarnaast ben ik zowel jou als Johan dankbaar om al tijdens het eindgesprek van mijn MSc-thesis te beginnen over de mogelijkheden voor het doen van een PhD. Tot slot wil ik zowel Mario als Piter bedanken voor het vertrouwen om onze samenwerking nog 4 jaar voort te zetten tijdens een postdoc-traject.

I also would like to thank the people at DEPI in Melbourne, Australia, for giving me the opportunity to work there for three months. Ben, thanks for your supervision. Even though your time was limited, you helped me a lot. Your enthusiasm and positivity works inspiring! Mike, thank you for the discussions, support and your very useful comments on the paper. I also would like to thank

Jennie for her support in organizing my stay in Melbourne. Moreover, I would like to thank the group of DEPI for their hospitality, help, nice conversations and sightseeing tips. Without the nice atmosphere at DEPI, I wouldn't have enjoyed my time in Melbourne and Australia as much as I did!!

Dit proefschrift is geschreven binnen het GenomXL project, een samenwerking tussen Wageningen UR Livestock Research en CRV. Binnen het GenomXL-team hebben we verschillende discussies en bijeenkomsten gehad. Naast de eerder genoemde personen, wil ik daar ook Sander en Ghyslaine, en later ook Marianne en Henk voor bedanken. Jullie feedback op mijn werk en jullie suggesties hebben zeker bijgedragen aan dit proefschrift en hebben mij geholpen om meer inzicht te verkrijgen in de rundveefokkerij. Bedankt daarvoor! Het GenomXL project was een project binnen het Breed4Food consortium. Daarom wil ik ook de andere partners van Breed4Food bedanken voor hun input en interesse voor mijn werk.

Many thanks to all the nice colleagues within ABGC! It was nice to get to know all of you, and I enjoyed the discussions during QDG and TLM, the conversations during coffee/lunch breaks, the birthday parties, the Sinterklaas celebrations and all other social activities. It was nice to be part of this group! Due to the international team around me, I learned a lot about other cultures and traditions. And even though some of you were making fun of my Dutch habits ("Oh, you're so Dutch!"), I hope that you learned that some of the Dutch habits are useful (like having an agenda ;-)). The number of colleagues that supported me is too high to mention all of them by name. However, a couple of colleagues deserve a special thank you, like my roommates, both in Triton and Radix. Besides our nice conversations, you were always there if I needed help or advice, or when I just wanted to express my frustrations when something didn't work out according to my plan. Ik wil ook graag de secretaresses van zowel ABG als WLR bedanken voor de administratieve ondersteuning. Lucia, jou wil ik bedanken voor je vertrouwen en je hulp bij het verleggen van mijn grenzen. Al was ik als PhD student een apart geval binnen de systemen van WLR, je hebt altijd je best gedaan om het voor mij zo makkelijk mogelijk te maken. Marcin, thanks for all your programming help, your guidance at the beginning of my PhD and all our discussions. It is nice to see that our friendship didn't end when you moved back to Poland. And it is getting time to plan my visit to Poznan ;-). Aniek, bedankt voor de gezelligheid tijdens de pauzes en je advies over allerlei zaken. Door onze discussies over inhoudelijke zaken is mijn inzicht in de genetica zeker gegroeid! Yvette, jou wil ik bedanken voor de gezelligheid, de wandelingen, de adviezen, de paardrijd-begeleiding, de leuke theaterbezoekjes, de final check van mijn thesis en ga zo maar door. Je had altijd snel in de gaten als ik ergens mee zat en stond dan meteen voor me klaar, zelfs al was je druk. Aniek en Yvette, het is fijn om zo'n collega's te hebben!! Al hadden jullie me wel een keer mogen laten winnen met spelletjes doen ;-).

Rianne, bijzonder om zowel onze studententijd als onze PhD-tijd samen door te maken. Erg gezellig natuurlijk, maar ook fijn om de ervaringen tijdens het doen van een PhD met iemand te kunnen delen. Alleen is het af en toe wel jammer dat je precies weet hoe je mij de slappe lach kunt bezorgen... Succes nog met de laatste loodjes van jouw PhD, maar ik heb er alle vertrouwen in dat dit goed gaat komen! Natasja, kei bedankt voor de gezellige tijd in Melbourne/Australië! Zonder jou had ik deze stap nooit durven zetten, en daar zal ik je altijd dankbaar voor blijven. Daarnaast natuurlijk ook bedankt voor alle Brabantse gezelligheid in Nederland, fijn om te weten dat ik altijd bij je terecht kan. Rianne en Natasja, ik ben blij dat jullie tijdens deze belangrijke dag naast me willen zitten op het podium (dit heeft natuurlijk ook nog als extra voordeel dat jullie me niet af kunnen leiden door het trekken van gekke bekken ;-)).

Meiden (en mannen) van de Maliboes, bedankt voor jullie vriendschap, jullie steun en de ontspannende dagen, avonden, weekenden en vakanties. Al zien we elkaar niet meer iedere dag, zoals tijdens onze studietijd, het is altijd gezellig om elkaar weer te zien, bij te kletsen en om (vaak om eigenlijk niets) te kunnen lachen. Bedankt voor jullie interesse in mijn onderzoek, al was het af en toe lastig uit te leggen waar ik nu precies mee bezig was. Hopelijk dat het na het lezen van mijn boekje iets duidelijker is geworden ;-). Daarnaast worden jullie lieve berichtjes op momenten dat ik het nodig heb zeker gewaardeerd! Laura, jou wil ik nog speciaal bedanken voor het ontwerpen (en tekenen) van de omslag van het boekje! Jouw creativiteit kwam daarvoor zeker van pas en ik ben echt blij met het resultaat!

Tot slot wil ik ook alle andere vrienden en familieleden bedanken voor alle steun, ontspanning, afleiding, en vriendschap. Pap en mam, bedankt voor de fijne thuisbasis op de Zandkant en de Brabantse no-nonsense opvoeding. Jullie opvoeding en het opgroeien tussen de koeien hebben me zeker geholpen tijdens mijn studietijd en mijn PhD. Bedankt dat jullie me altijd steunen en willen helpen. Samen met Anne & Rinus, Rob & Moniek en de kleine Cas, wil ik jullie bedanken voor jullie interesse, en de gezelligheid en ontspanning waar jullie voor gezorgd hebben. Alhoewel ik vaak gehoord heb "Ze doet iets met DNA en koeien, maar wat precies...?", hebben jullie wel altijd interesse getoond in mijn onderzoek. En het belangrijkste om te onthouden is in ieder geval dat ik het leuk vind! Hopelijk kan ik jullie in de toekomst nog eens overtuigen van de voordelen om de koeien te laten genotyperen ;-). Anne, jou wil ik nog speciaal bedanken voor je hulp en advies tijdens mijn PhD (je weet immers hoe het moet ;-)) en het doorlezen/doorkijken van de stukken.

Nogmaals bedankt allemaal!! Mede dankzij jullie hulp is mijn PhD-tijd een onvergetelijke en leuke tijd geworden. Hopelijk kom ik jullie in de toekomst nog vaak tegen en kan ik dan nog steeds op jullie hulp, steun, vertrouwen, vriendschap en humor rekenen!!

Colophon

The research described in this thesis was financially supported by CRV B.V. (Arnhem, The Netherlands), and was part of the Breed4Food consortium, a public-private partnership in the domain of animal breeding and genomics. The research visit of Yvonne Wientjes at the Department of Environment and Primary Industries in Victoria, Australia, was funded by EU FP7 IRSES SEQSEL (Grant no. 317697).

The data used in this study was provided by the 1000 bulls genomes consortium, the gDMI consortium, and by the projects RobustMilk, the Dutch Milk Genomics Initiative and 'Melk op Maat'. Those projects were funded by the National Institute of Food and Agriculture (NIFA, Washington DC, USA), Wageningen University (Wageningen, the Netherlands), the Dutch Dairy Association (NZO, Zoetermeer, the Netherlands), CRV B.V. (Arnhem, the Netherlands), the Dutch Technology Foundation (STW, Utrecht, the Netherlands), the Dutch Ministry of Economic Affairs (The Hague, the Netherlands) and the Provinces of Gelderland and Overijssel (Arnhem, the Netherlands).

The cover of this thesis was designed by Laura van der Woude.

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