Propositions

1. Modelling of biological information improves the accuracy of genomic prediction.
2. Utilization of Genome-wide association study is the next step for genomic prediction.
3. Discipline and character are more important to the process of science than intellect.
4. “Gut feelings” have a scientific value and should not be quickly dismissed.
5. More important than the ability to plan, is the ability to quickly adapt to deviations from a plan.
6. Moral and ethical values are conditional on the quality of life.

Propositions belonging to the thesis entitled:

Use of whole-genome sequence data for genomic prediction across populations and species

Biaty Raymond

Wageningen, 11 September 2020

Use of whole-genome sequence data for genomic prediction across populations and species

Biaty Raymond

Thesis committee

Promotor

Prof. Dr R.F. Veerkamp
Special professor Animal Breeding and Genomics
Wageningen University & Research

Co- promotor

Dr A.C. Bouwman
Researcher, Animal Breeding and Genomics Group
Wageningen University & Research

Other members

Prof. Dr L. Trindade, Wageningen University & Research
Prof. Dr D.J. de Koning, Swedish University of Agricultural Science, Uppsala, Sweden
Prof. Dr T. Meuwissen, Norwegian University of Life Sciences, Ås, Norway
Prof. Dr D. Berry, TEAGASC - The Agriculture and Food Development Authority,
Cork, Ireland

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

Use of whole-genome sequence data for genomic prediction across populations and species

Biaty Raymond

Thesis

submitted in fulfilment 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 September 11 2020
at 4 p.m. in the Aula

Raymond, B.

Use of whole-genome sequence data for genomic prediction across populations and species.

PhD thesis, Wageningen University, the Netherlands (2020)

With references, with summary in English

DOI: <https://doi.org/10.18174/526865>

ISBN: 978-94-6395-465-5

Abstract

Raymond, B. (2020). Use of whole-genome sequence data for genomic prediction across populations and species. PhD thesis, Wageningen University, the Netherlands

The availability of whole genome sequence data presents an opportunity to improve the accuracy of genomic prediction (GP), given the expectation that the data contains the causal variants that underlie any given complex trait. This thesis investigated the benefit of using whole genome sequence data for GP across populations and species. Results show that the accuracy of GP across populations does not improve by simply increasing marker density up to whole genome sequence level (**Chapter 2**). However, accuracy can be improved if GP across populations is based on a few variants that are pre-selected from whole genome sequence based on the significance level of their effect on the trait from a genome-wide association study (GWAS). The result highlights the relevance of GWAS for GP. In **Chapter 3**, it was demonstrated that the accuracy of prediction can further be improved using a multi-population GP model in which important pre-selected variants are used to create a genomic relationship matrix (GRM), other available and unselected variants are used to create a second GRM, and both GRMs are fitted simultaneously. While the pre-selected variants in the first GRM are isolated from the noise effect of neutral variants, and as such their effects are more accurately estimated, the variants in the second GRM captures genetic variance for the trait that cannot be captured by the pre-selected variants in the first GRM. In terms of accuracy, it was shown that the multi-population, multiple GRMs (MPMG) GP model outperforms within-population and multi-population GP models in which either the pre-selected or all available variants are equally weighted in a single GRM. In **Chapter 4**, the predictive performance of the MPMG model was theoretically underpinned by deriving and validating a deterministic prediction equation for its accuracy. Using the derived prediction equation, it was found that the predictive performance of the model is due to its ability to benefit from the low number of effective chromosomal segments (Me) represented by the few pre-selected variants in the first GRM. However, the low values for Me due to the pre-selected variants is an advantage for the MPMG model only if variant pre-selection is accurate, such that the pre-selected variants explain some genetic variance for the trait of interest. In addition to its use as a tool to gain theoretical insights into the performance of biology-informed GP models, the derived prediction equation can be used to, a priori, estimate expected accuracy if the MPMG model were to be implemented for GP. In **Chapter 5**, the usefulness of summary-level GWAS result

for human height as prior information for identifying genes and gene-associated variants that affect stature in cattle, was investigated. Results show that in some cases, for example in the absence of stature GWAS, human height GWAS results can be useful in identifying cattle stature genes and associated variants.

Contents

Chapter 1 – General introduction

Chapter 2 – Utility of whole-genome sequence data for across-breed genomic prediction

Chapter 3 – Genomic prediction for numerically small breeds, using models with pre-selected and differentially weighted markers

Chapter 4 – A deterministic equation to predict the accuracy of multi-population genomic prediction with multiple genomic relationship matrices

Chapter 5 – Using prior information from humans to prioritize genes and gene-associated variants for complex traits in livestock

Chapter 6 – General discussion

Summary

Curriculum vitae

Acknowledgement

Colophon

1

CHAPTER 1.

General introduction

1.1 Genetic improvement of livestock

The practice of selective breeding of livestock for economically important traits has traditionally been based on the phenotypic records of selection candidates and or those of their relatives. Phenotype and pedigree records are combined in the so called best linear unbiased prediction (BLUP) model (Henderson 1975; Henderson 1976; Henderson 1982), to estimate the breeding values of selection candidates. Estimated breeding values (EBVs) are then used as a criterium to rank the selection candidates, with the highest ranked individuals selected as parents for the next generation. Since its introduction, pedigree-based BLUP (A-BLUP) has been employed with great success in most genetic evaluation programs around the world (e.g. Parnell *et al.* 1997; Jones *et al.* 2001; Barwick and Henzell 2005; Hare *et al.* 2006; Van Pelt *et al.* 2016). A genetic selection program based on A-BLUP has a number of limitations. Firstly, the total breeding value of an individual at birth consists of two components, namely, the parent average and the mendelian sampling component (Woolliams *et al.* 1999). The A-BLUP model is able to accurately estimate the parent average component, given adequate phenotype and pedigree data. The A-BLUP model is also able to estimate the mendelian sampling component. To do so, however, the A-BLUP will require own performance or offspring records. Thus, the A-BLUP cannot estimate the mendelian sampling component for young individuals with no own performance or offspring records. Secondly, given its reliance on pedigree records, the A-BLUP model cannot utilize information on individuals that do not have a pedigree relationship with the selection candidates. Thirdly, the A-BLUP model requires phenotypic records on selection candidates and or their relatives, which for some traits are either difficult, takes very long time, or are costly to measure. The implication is a long generation interval, as selection decisions cannot be made at the time the selection candidates are born.

From around 1980 onwards, advances in molecular genetics research led to the discovery of large number of polymorphisms in the genome, and that it was possible to identify and utilize information on Quantitative trait loci (QTL), for the purpose of breeding values estimation (e.g. Soller and Beckmann 1982; Smith and Simpson 1986; Andersson 2001). This discovery ushered in a new area of research called marker-assisted selection (MAS), which held great promise to improve the accuracy of estimating EBVs and the rate of genetic gain, as selection could be performed directly on the QTL, and as such, some of the limitations of the A-BLUP model could be, partly, circumvented (Lande and Thompson 1990; Goddard 1996; Dekkers and Hospital 2002). With MAS, QTL information could be included as fixed or random

effects in A-BLUP. Another approach was to combine QTL-based EBVs with polygenic EBVs from the traditional A-BLUP, in a selection index (e.g. Lande and Thompson 1990). Despite the early optimism about MAS, it only resulted in either non or negligible improvement in genetic gain as compared with A-BLUP (e.g. Dekkers and Hospital 2002; Dekkers 2004). This was mainly because the few QTL found could only explain small proportions of the total genetic variance for complex traits, and thus, could not contribute significantly to improving the accuracy of prediction. Rather than focusing on the identification and utilization of a few QTL for breeding value estimation, Meuwissen *et al.* (2001) proposed a method to utilize information on all available loci in the genome, a method now known as genomic prediction (GP).

The basic principle behind GP is the estimation of the effects of DNA markers such as single nucleotide polymorphisms (SNPs) in a reference population, and use of the estimated marker effects to predict the genomically estimated breeding values (GEBVs) of selection candidates. Genomic prediction as originally proposed (Meuwissen *et al.* 2001; Vanraden 2008) does not depend on the identification of true causal variants for the traits, but its accuracy depends on the extent to which the effect of the causal variants are tagged by markers through linkage disequilibrium (LD). The application of GP for selection decisions has been very successful in both plant and animal breeding (e.g. Dalton 2009; Hayes *et al.* 2009a; Heffner *et al.* 2011). In dairy cattle for example, the adoption of GP as against pedigree-based methods has led to approximately double the rate of genetic gain (see Hayes *et al.* 2009b for a review). Therefore, the major development in animal breeding over the past decade has been the implementation of routine genomic prediction.

1.2 Statistical models for GP

There are several models that have been proposed for GP (e.g. Meuwissen *et al.* 2001; Vanraden 2008; Goddard 2009; Strandén and Garrick 2009; Habier *et al.* 2011; Gianola 2013). The models can be categorized into two, namely, linear and non-linear models. In linear models, phenotypes are modelled as a function of the linear combination of marker effects. An example of a linear model is a random marker regression model, also known as SNP-BLUP. The method is called SNP-BLUP because the SNP or marker effects are simultaneously estimated using BLUP. An alternative but mathematically equivalent model to SNP-BLUP is the G-BLUP model (Habier *et al.* 2007b; Goddard 2009; Strandén and Garrick 2009). The G-BLUP is the same as the traditional BLUP (Henderson 1975), except that pedigree relationship matrix is

replaced by genomic relationship matrix. An assumption of the classical linear GP models is that each marker has an infinitesimally small effect on the trait, with the variance of each marker effect equal to $\frac{\sigma_a^2}{n}$, where n is the total number of SNPs and σ_a^2 is the genetic variance for the trait. This assumption is also called the infinitesimal model assumption.

The application of linear model for GP has been very successful, especially in animal breeding (Vanraden *et al.* 2009). The success of the classical linear GP models in animal breeding can be attributed to a number of reasons. First, most traits that are bred for in livestock have a complex genetic architecture. For example, milk production in cattle is currently estimated to be affected by up to 10,000 genes (Meuwissen *et al.* 2016). This means that there are so many loci in the genome affecting complex traits such that the infinitesimal model assumption that each variant contributes to explaining the total genetic variance for the trait comes close to biological reality. Secondly, the infinitesimal model assumption is only an assumption and not a restraint on the output from the models. In practice, estimated marker effects may deviate from the infinitesimal model assumption especially with the availability of large sample sizes. With large sample sizes, linear models are able to accurately estimate SNP effects, and thus are able to disentangle important variants from neutral variants in the genome, even though the underlying assumption is that all variants explain equal variances. A third reason that may underlie the success of classical linear GP models in animal breeding is the fact that LD extends over long distances in livestock populations such that almost every causal variant is tagged by at least one SNP. The effect of a causal variant can then be captured by a linear combination of SNPs that are in LD with it (Hayes *et al.* 2009c). Fourthly, the linear GBLUP model is the same as the A-BLUP, with the only difference being the replacement of the pedigree relationship matrix with a genomic relationship matrix. As such linear GP model could easily be implemented with the existing software that were used for the A-BLUP.

The second category of models are the non-linear models, like Bayesian models, in which phenotypes are modelled as a function of the non-linear combination of SNP effects. Unlike in the classical linear models, the assumption in non-linear models is that a proportion (π) of the model parameters (SNPs) have a zero effect on the trait, while only a proportion ($1-\pi$) have a non-zero effect on the trait. Intuitively, this assumption matches more with biological reality, given that some variants in the genome are expected to be neutral in their effect. The first non-linear models applied

to GP: BayesA and BayesB, were proposed by Meuwissen *et al.* (2001). In their study, Meuwissen *et al.* (2001) showed that the non-linear models resulted in a higher accuracy of GP (0.85) as compared with the BLUP alternative (0.73), although they used chromosomal segments instead of SNPs as predictor variables. Since the landmark paper of Meuwissen *et al.* (2001), there has been a growing number of Bayesian alphabets that differ mainly in the specification of priors but not in their sampling approach (Gianola *et al.* 2009; De Los Campos *et al.* 2013; Gianola 2013). For example, BayesR extended the BayesB method of Meuwissen *et al.* (2001) by allowing SNP effects to be sampled from several normal distributions, starting with a distribution with a variance of zero to one with variance of up to 1% of the total genetic variance for the trait (Erbe *et al.* 2012). In general, flat or uninformative priors are used in the classical non-linear models. Thus, the model uses the data to find evidence as to whether a certain variant belongs to the proportion π or $1 - \pi$.

Results in literature suggest that non-linear models have somewhat a superior predictive ability than the linear models for traits that are controlled by a limited number of QTL with large effects (e.g. Habier *et al.* 2007b; Daetwyler *et al.* 2010; Meuwissen and Goddard 2010; Tiezzi and Maltecca 2015) and for GP with WGS data instead of lower density SNP panels (e.g. Meuwissen and Goddard 2010). For highly polygenic traits, and in animal breeding populations that are highly structured, the infinitesimal model assumption hold very well. A practical limitation of the non-linear models is their high computation time and resource requirement (Meuwissen *et al.* 2001), which hinder their adoption in routine genetic evaluation.

1.3 Genomic prediction across populations

For GP to be accurate, it is necessary to have a large reference population such that the effect of SNPs can be accurately estimated (e.g. Goddard 2009; Vanraden *et al.* 2009). For numerically small populations, it is difficult or impossible to assemble a large reference population. Thus, in such populations, the accuracy of GP is low, with consequent lower rate of genetic gain as compared to what is obtainable in numerically large populations. An option to increase the accuracy of GP in numerically small populations is to use a large reference population made up of individuals from a different population, for example, individuals from a different breed or line (e.g. Harris *et al.* 2008; Hayes *et al.* 2009c; Schrooten *et al.* 2013). This method is called across-population GP. Results in literature indicate that the accuracy of across-population GP is low (e.g. Harris *et al.* 2008; Erbe *et al.* 2012; Kachman *et al.* 2013; Calus *et al.* 2014). There are three possible reasons that

underlie the low accuracy of GP across populations. First, the pattern of LD between SNPs and causal variants may not be consistent across different populations (Gautier *et al.* 2007; De Roos *et al.* 2008; Thomasen *et al.* 2012). It is possible that a given SNP is in LD with a causal variant in population A, while the LD between the same SNP and the same causal variant is broken down due to recombination in population B. In the extreme case that LD is broken down to zero for example, estimating the effect of the SNP in population A and using the estimated SNP effect to predict EBV for individuals from population B will result in a prediction accuracy of zero. Furthermore, given a different pattern of LD between SNPs and causal variants across populations, the relationship between individuals across populations estimated using SNPs are less likely to represent the relationship at the causal variants, which is a requirement for accurate GP. Second, there may be differences of allele substitution effect of causal variants in different populations (Spelman *et al.* 2002). If the effects of the causal variants differ in different populations, meaning that the genetic correlation between the populations is less than one, it becomes difficult to translate the effect of a causal variant in one population to another. A third reason that may cause the low accuracy of GP across populations is possible differences in allele frequency of variants across populations. The proportion of genetic variance explained by a given variant is dependent on its allele substituting effect and its allele frequency. Say a given variant has the same allele substitution effect in two populations A and B, but the variants segregates at a high frequency in A, but at low frequency in B, the variant will explain more genetic variance for the trait in population A than in population B. This difference negatively impacts the accuracy of prediction across the populations A and B.

A suggestion to improve the accuracy of GP across populations is to first identify either the causal variant or SNPs that are physically close to and in very high LD with causal variants across-populations, and use only the identified variants for prediction (e.g. Zhong *et al.* 2009; Boichard *et al.* 2012; Hoze *et al.* 2014; Saatchi and Garrick 2014). Another suggestion is to increase the density of SNPs such that every causal variant is in LD with at least one SNPs across populations (De Roos *et al.* 2008). Following this reasoning, across population with whole genome sequence (WGS) data is expected to be more accurate than with low density SNP arrays. In this thesis, the impact of using all WGS variants on the accuracy of GP across populations was further investigated.

1.4 Multi-population GP

An alternative approach to across-population GP is to build a reference population with individuals from multiple populations, including individuals from the target population, a method called multi-population GP, see (Lund *et al.* 2014) for a review of the method. Just as in across-population GP, the success of multi-GP depends on the pattern of LD between SNPs and causal variants, differences in allele substitution effects and allele frequencies in the different populations (Wientjes *et al.* 2015; Wientjes *et al.* 2016). Several studies in literature reported only marginal or not benefit at all of implementing multi-population GP as against within-population GP (e.g. Hayes *et al.* 2009c; Harris and Johnson 2010; Kizilkaya *et al.* 2010; Erbe *et al.* 2012; Lund *et al.* 2014; Zhou *et al.* 2014b). However, there are a few studies that reported an increase in accuracy of prediction by implementing a multi-population GP instead of within-population GP (Lund *et al.* 2014). These studies, however, combined individuals from different sub-populations of the same breed (e.g. Schenkel *et al.* 2009; Vanraden *et al.* 2012; Zhou *et al.* 2013), or individuals from closely related breeds, that have had frequent exchanges of genetic material (e.g. Brøndum *et al.* 2011; Zhou *et al.* 2014a). For example, Zhou *et al.* (2013) compared the accuracy of prediction for Chinese Holstein population using either a Chinese Holstein reference population or a combined Chinese and Nordic Holstein reference population. By using the combined reference population instead of the reference population made of only Chinese Holsteins, they showed an increase in accuracy of prediction from 0.22, 0.15 and 0.11 to 0.51, 0.47 and 0.36 for milk yield, fat yield and protein yield, respectively. In general, the more close the populations are, the more benefit there is in combining them in a reference population for multi-population GP.

There are two approaches that are used for multi-population GP. The first and the most simple approach is to pool the data from different populations in a univariate analysis. This is the same as assuming a genetic correlation of 1 between all the populations. While this strategy may work when different sub-populations of the same breed are combined in the reference population, assuming a genetic correlation of 1 between distantly related populations in multi-population GP can result in even poorer prediction accuracies than a within-population GP (e.g. Olson *et al.* 2012). The second approach for multi-population GP is the multi-trait approach that takes into account the genetic correlation between the populations in the reference. In the multi-trait approach, phenotypes from the different populations are considered as those for different but correlated traits (e.g. Karoui *et al.* 2012;

Olson *et al.* 2012; Huang *et al.* 2014; Wientjes *et al.* 2016). The advantage of this approach is the information of individuals in the reference population are weighted by their genetic correlation with the validation population. Thus, individuals in the reference population can only contribute towards the prediction if and only if the genetic correlation with the validation population is different than zero. With the multi-trait approach, the worst case scenario is that the accuracy of prediction is the same as within-population GP accuracy, but not lower (Wientjes *et al.* 2016).

An implicit assumption made in the multi-trait, multi-population GP model is that the genetic correlation between populations at the causal variants is the same as the genetic correlation at the non-causal variants. In their study, Wientjes *et al.* (2018) showed that this assumption can be safe to make if the non-causal variants have similar properties, for example, the same patterns of allele frequencies as the causal variants. In situations where SNPs have different properties as the causal variants, assuming that there is only one genetic correlation between population across the whole genome is sub-optimal (Calus *et al.* 2018). As more biological information about complex traits becomes available, and as causal or predictive variants are identified, it will become important to have a multi-population GP model that takes into account the possibility that genetic correlation between populations at such variants maybe different from the genetic correlation at other variants in the genome.

1.5 The accuracy of GP

The accuracy of GP depends on four main factors. The first is the extent of linkage disequilibrium (LD) between SNPs and causal variants, as that determines the proportion of genetic variance for the trait that can be explained by SNPs (e.g. Calus and Veerkamp 2007; Calus *et al.* 2008; Solberg *et al.* 2008; Weir 2008; Hayes *et al.* 2009b). In dairy cattle for example, It has been estimated that SNP density must be high enough such that LD between adjacent SNPs as measured by r^2 (Hill 1981) must be equal to or greater than 0.2 to achieve a genomic prediction accuracy of 0.8 (Calus and Veerkamp 2007). The second factor affecting the accuracy of GP is the size of the reference population, as large sample sizes are required to accurately estimate the effects of variants and obtain accurate prediction of GEBVs (Goddard 2009; Vanraden *et al.* 2009). The third factor is the heritability of the trait. Given the same reference population size, GP prediction will be more accurate for a trait with a high heritability than for a trait will low heritability (e.g. Daetwyler *et al.* 2008; Goddard

2009). The fourth factor is the polygenicity of the trait. In general, larger reference population size will be required to obtain accurate GP of a highly polygenetic trait than a trait that is affected by only a few causal variants with large effects (Hayes *et al.* 2009b).

Two of the factors affecting the accuracy of GP, namely, the heritability and polygenicity of the trait cannot be influenced to increase the accuracy of GP. To increase the accuracy of GP, only the size of the reference population and the extent of LD between SNPs and causal variants can be influenced. Propositions to increase the size of the reference population include combining and sharing of data across countries as currently done through Interbull (Palucci *et al.* 2018), the genotyping and inclusion of cows in the reference population (e.g. Buch *et al.* 2012; Thomasen *et al.* 2014; Su *et al.* 2016; Uemoto *et al.* 2017), and the combination of individuals from different breeds in a reference population (e.g. Pryce *et al.* 2011; Wientjes *et al.* 2015; Wientjes *et al.* 2016). To increase the LD between SNPs and causal variants, high density SNP panels can be used such that every causal variant affecting the trait is tagged by at least one SNP. Given that whole genome sequence (WGS) data is the ultimate high-density panel, the use of WGS instead of low density SNP panels was expected to improve the accuracy of GP (e.g. Meuwissen and Goddard 2010). Furthermore, given the expectation that causal variants are included in WGS data, the accuracy of GP with WGS data no longer depend on LD between SNPs and causal variants, as the effect of the causal variants themselves can be directly estimated.

1.6 Use of WGS data in genomic prediction

The possibility to use WGS data instead of SNP arrays, and the potential impact on the accuracy of GP were first investigated by simulation studies. For example, the study by Meuwissen and Goddard (2010) reported an increase in the accuracy of GP by more than 40% when WGS data was used instead of a 30k SNP chip. Using WGS data, they also showed that accuracy of GP when both the reference and test population come from the same generation was the same with that in which the reference and test population were separated by 10 generations. Their results supports the theory that the accuracy of GP no longer depends on the existence of LD between SNPs and causal variants when WGS data is used in GP. The study by Macleod *et al.* (2014) also reported an increase in the accuracy of GP by 22% when WGS was used compared to a 600k SNP chip. This increase in accuracy, however,

was achieved under the condition that the effective population size was large, and therefore average LD was low in the population. Other studies (e.g. Druet et al. 2014; Clark et al. 2011) obtained similar results.

Given the optimistic results regarding the use of WGS data in GP from simulation studies, a number of studies using real data followed, which mostly did not support the results from simulation studies (e.g. Ober et al. 2012; Iheshiulor et al. 2014; Brøndum et al. 2015; Van Binsbergen et al. 2015). For example, Ober et al. (2012) used WGS data to predict quantitative traits in *Drosophila melanogaster*. They found that WGS added little or no benefit to the accuracy of GP when compared with HD SNP chip, although they had a very small sample size of 157. In dairy cattle, Van Binsbergen et al. (2015) used imputed WGS data for GP with a relatively bigger sample size (3,416 bulls in the reference population, and 2,087 bulls in the validation set). They also did not find any advantage for using WGS over HD SNP chip. While theoretically, the use of WGS held the promise to greatly improve the accuracy of GP, results from studies using real data were suggesting otherwise.

One of the hypotheses put forward with respect to the results from studies that used real data was that the lack of increase in accuracy is because the studies were carried out within-breed where LD extends over long distances. Given that GP is dependent on the existence of LD (Habier et al. 2007a; Gianola et al. 2009; Habier et al. 2010), HD SNP chips are already sufficient to tag the effect of causal variants within-breed, especially in breeds with very small effective population sizes. In that sense, simply going from HD SNP chips to WGS level is unlikely to improve GP accuracy. The benefit of using WGS for GP is therefore expected to be highest when applied in situations where LD is conserved only across short distances, for example in across or multi-breed GP scenarios (Van Den Berg et al. 2016).

Another hypothesis is that the sample sizes used in the studies were not large enough to be able to simultaneously and accurately estimate the effect of millions of SNPs in WGS data (Gianola 2017). This is referred to as the $N \ll p$ problem, where N is the size of the sample size and p is the number of variants whose effects are estimated. Thus, it is expected that as sample sizes increases, evidence for the benefit of WGS data for GP will become stronger. It must be noted, however, that animal breeding populations are highly structured and have small effective population sizes. Thus, simply increasing sample size is not the same as increasing the effective sample size, especially if the added samples are from the same population as the original samples. The pre-selection of variants in WGS data based

on biological information may help to alleviate the $N \ll p$ problem and improve the accuracy of GP. The pre-selection of markers from WGS based on biological information and its impact on the accuracy of GP was investigated in this thesis.

1.7 Objective and outline of this thesis

The overall objective of this thesis was to investigate the benefit of WGS data for GP across populations and species. The overall objective was divided into three sub-objectives. The first sub-objective was to investigate strategies for the efficient use of WGS data in across-population GP. Mainly, I investigated the benefit pre-selection of WGS variants based on GWAS results and the subsequent use of the pre-selected variants in across- and multi-population GP. The second sub-objective of the thesis was to theoretically underpin the advantages and limitations of multi-population genomic prediction model that weights pre-selected variants from GWAS differently. The third sub-objective of the thesis was to investigate the usefulness of results from a WGS-based genome-wide association study (GWAS) for complex traits in humans, in the identification of genes and variants in or close to the genes that affect equivalent traits in cattle.

In **Chapter 2**, I investigated whether using WGS data instead of low-density SNP arrays can improve GP across breeds, especially when variants in WGS data are pre-selected based on their estimated effect or conditional and joint (COJO) effect on the trait from a genome-wide association study (GWAS). In **Chapter 3**, strategies for the incorporation of prior biological information from WGS data in multi-population GP were investigated. For this purpose, I tested the performance of a multi-breed, multiple genomic relationship matrices GP (MBMG) model that weight pre-selected variants separately, uses the remaining unselected variants to explain the residual genetic variance that is not explained by pre-selected variants, and weight information of breeds in the reference population by their genetic correlation with the validation breed. In **Chapter 4**, selection index theory was used to derived a deterministic prediction equation for the accuracy of the MPMG model. The purpose of the equation was to theoretically underpin the advantages and limitations of the MBMG model over multi-population GP model that equally weight all variants. In **Chapter 5**, I investigated, if instead of using a within-species GWAS to identify genes and gene-associated variants for complex traits in livestock, a GWAS results for the same trait but from a different species could be utilized. Thus, I investigate the

usefulness of GWAS results on human height as a prior information for identifying genes and gene-associated variants that affect stature in cattle.

In **Chapter 6** (General discussion), I discussed two main topics related to the results of the previous chapters. The first topic discussed was the methods for variants pre-selection based on biological information, and the impact of pre-selection of variants on the accuracy of GP. This included discussion on the use of GWAS to pre-select relevant variants for GP, the combined use of GWAS and gene-expression analysis and other methods such as the combined annotation dependent depletion (CADD) scores and the functional and evolutionary trait heritability (FAETH) scores. The second topic that was discussed was the statistical and modelling methods for the implementation of biology-driven GP.

1.8 References

- Andersson, L., 2001 Genetic dissection of phenotypic diversity in farm animals. *Nature Reviews Genetics* 2: 130.
- Barwick, S., and A. L. Henzell, 2005 Development successes and issues for the future in deriving and applying selection indexes for beef breeding. *Australian Journal of Experimental Agriculture* 45: 923-933.
- Boichard, D., F. Guillaume, A. Baur, P. Croiseau, M.-N. Rossignol *et al.*, 2012 Genomic selection in French dairy cattle. *Animal Production Science* 52: 115-120.
- Brøndum, R. F., E. Rius-Vilarrasa, I. Strandén, G. Su, B. Guldbrandtsen *et al.*, 2011 Reliabilities of genomic prediction using combined reference data of the Nordic Red dairy cattle populations. *Journal of Dairy Science* 94: 4700-4707.
- Brøndum, R. F., G. Su, L. Janss, G. Sahana, B. Guldbrandtsen *et al.*, 2015 Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. *Journal of dairy science* 98: 4107-4116.
- Buch, L. H., M. Kargo, P. Berg, J. Lassen and A. C. Sørensen, 2012 The value of cows in reference populations for genomic selection of new functional traits. *Animal* 6: 880-886.
- Calus, M., A. De Roos and R. Veerkamp, 2008 Accuracy of genomic selection using different methods to define haplotypes. *Genetics* 178: 553-561.
- Calus, M., M. Goddard, Y. Wientjes, P. Bowman and B. Hayes, 2018 Multibreed genomic prediction using multitrait genomic residual maximum likelihood and multitask Bayesian variable selection. *Journal of dairy science* 101: 4279-4294.
- Calus, M., and R. Veerkamp, 2007 Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. *Journal of Animal Breeding and Genetics* 124: 362-368.
- Calus, M. P., 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. *Genetics Selection Evolution* 46: 57.
- Clark, S. A., J. M. Hickey and J. H. Van der Werf, 2011 Different models of genetic variation and their effect on genomic evaluation. *Genet Sel Evol* 43: 10.1186.
- 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., B. Villanueva and J. A. Woolliams, 2008 Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PloS one* 3.
- Dalton, R., 2009 No bull: genes for better milk. *Nature News* 457: 369-369.
- de los Campos, G., J. M. Hickey, R. Pong-Wong, H. D. Daetwyler and M. P. Calus, 2013 Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* 193: 327-345.

- De Roos, A., B. J. Hayes, R. Spelman and M. E. Goddard, 2008 Linkage disequilibrium and persistence of phase in Holstein–Friesian, Jersey and Angus cattle. *Genetics* 179: 1503-1512.
- Dekkers, J. C., 2004 Commercial application of marker-and gene-assisted selection in livestock: strategies and lessons. *Journal of animal science* 82: E313-E328.
- Dekkers, J. C., and F. Hospital, 2002 Multifactorial genetics: The use of molecular genetics in the improvement of agricultural populations. *Nature Reviews Genetics* 3: 22.
- Druet, T., I. Macleod and B. Hayes, 2014 Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. *Heredity* 112: 39-47.
- e.g. Parnell, P., P. Arthur and R. Barlow, 1997 Direct response to divergent selection for yearling growth rate in Angus cattle. *Livestock Production Science* 49: 297-304.
- Erbe, M., B. Hayes, L. Matukumalli, S. Goswami, P. Bowman *et al.*, 2012 Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *Journal of dairy science* 95: 4114-4129.
- 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.
- Gianola, D., 2013 Priors in whole-genome regression: the Bayesian alphabet returns. *Genetics* 194: 573-596.
- Gianola, D., 2017 It is unlikely that genomic selection will ever be 100% accurate. *Journal of animal breeding and genetics= Zeitschrift fur Tierzuchtung und Zuchtungsbiologie* 134: 438.
- Gianola, D., G. de los Campos, W. G. Hill, E. Manfredi and R. Fernando, 2009 Additive genetic variability and the Bayesian alphabet. *Genetics* 183: 347-363.
- Goddard, M., 1996 The use of marker haplotypes in animal breeding schemes. *Genetics Selection Evolution* 28: 161.
- Goddard, M., 2009 Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* 136: 245-257.
- Habier, D., R. Fernando and J. Dekkers, 2007a The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177: 2389-2397.
- Habier, D., R. L. Fernando and J. C. M. Dekkers, 2007b The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. *Genetics* 177: 2389-2397.
- Habier, D., R. L. Fernando, K. Kizilkaya and D. J. Garrick, 2011 Extension of the Bayesian alphabet for genomic selection. *BMC bioinformatics* 12: 186.
- Habier, D., J. Tetens, F.-R. Seefried, P. Lichtner and G. Thaller, 2010 The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genetics Selection Evolution* 42: 5.

- Hare, E., H. Norman and J. Wright, 2006 Trends in calving ages and calving intervals for dairy cattle breeds in the United States. *Journal of dairy science* 89: 365-370.
- Harris, B., and D. Johnson, 2010 Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. *Journal of dairy science* 93: 1243-1252.
- Harris, B., D. Johnson, R. Spelman and J. Sattler, 2008 Genomic selection in New Zealand and the implications for national genetic evaluation, pp. in *Proc. Interbull Meeting, Niagara Falls, Canada*.
- Hayes, B., P. Bowman, A. Chamberlain and M. Goddard, 2009a Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of dairy science* 92: 433-443.
- Hayes, B. J., P. J. Bowman, A. Chamberlain and M. Goddard, 2009b Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of dairy science* 92: 433-443.
- Hayes, B. J., P. J. Bowman, A. C. Chamberlain, K. Verbyla and M. E. Goddard, 2009c Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genetics Selection Evolution* 41: 51.
- Heffner, E. L., J.-L. Jannink and M. E. Sorrells, 2011 Genomic selection accuracy using multifamily prediction models in a wheat breeding program. *The Plant Genome* 4: 65-75.
- Henderson, C., 1982 Best linear unbiased prediction in populations that have undergone selection, pp. in *Proceedings of the world congress on sheep and beef cattle breeding*.
- Henderson, C. R., 1975 Best linear unbiased estimation and prediction under a selection model. *Biometrics*: 423-447.
- Henderson, C. R., 1976 A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics*: 69-83.
- Hill, W. G., 1981 Estimation of effective population size from data on linkage disequilibrium. *Genetics Research* 38: 209-216.
- Hoze, C., S. Fritz, F. Phocas, D. Boichard, V. Ducrocq *et al.*, 2014 Genomic evaluation using combined reference populations from Montbéliarde and French Simmental breeds, pp. 17-22 in *Proceedings of the 10th world congress of genetics applied to livestock production*.
- 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. *Genetics Selection Evolution* 46: 75.
- Iheshiulor, O., J. Woolliams, X. Yu and R. Wellmann, 2014 Genomic Predictions Using Whole Genome Sequence Data and Multi-breed Reference Populations, pp. in *10th World Congress on Genetics Applied to Livestock Production*. Asas.
- Jones, D., K. Anderson and G. Davis, 2001 The effects of genetic selection on production parameters of single comb White Leghorn hens. *Poultry science* 80: 1139-1143.

- 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. *Genetics Selection Evolution* 45: 30.
- Karoui, S., M. J. Carabaño, C. Díaz and A. Legarra, 2012 Joint genomic evaluation of French dairy cattle breeds using multiple-trait models. *Genetics Selection Evolution* 44: 39.
- Kizilkaya, K., R. Fernando and D. Garrick, 2010 Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. *Journal of animal science* 88: 544-551.
- Lande, R., and R. Thompson, 1990 Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* 124: 743-756.
- Lund, M. S., G. Su, L. Janss, B. Guldbrandtsen and R. F. Brøndum, 2014 Genomic evaluation of cattle in a multi-breed context. *Livestock Science* 166: 101-110.
- MacLeod, I. M., B. J. Hayes and M. E. Goddard, 2014 The effects of demography and long-term selection on the accuracy of genomic prediction with sequence data. *Genetics* 198: 1671-1684.
- Meuwissen, T., and M. Goddard, 2010 Accurate prediction of genetic values for complex traits by whole-genome resequencing. *Genetics* 185: 623-631.
- Meuwissen, T., B. Hayes and M. Goddard, 2016 Genomic selection: A paradigm shift in animal breeding. *Animal frontiers* 6: 6-14.
- 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.
- 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., P. VanRaden and M. Tooker, 2012 Multibreed genomic evaluations using purebred Holsteins, Jerseys, and Brown Swiss. *Journal of Dairy Science* 95: 5378-5383.
- Palucci, V., H. Jorjani, H. Benhajali, E. Hjerpe, J. Sendecka *et al.*, 2018 Interbull Centre-Special Issue. *Interbull Bulletin*.
- Pryce, J., B. Gredler, S. Bolormaa, P. Bowman, C. Egger-Danner *et al.*, 2011 Genomic selection using a multi-breed, across-country reference population. *Journal of Dairy Science* 94: 2625-2630.
- Saatchi, M., and D. Garrick, 2014 Across breed genomic predictions in beef cattle, pp. 17-22 in *Proceedings of the 10th world congress on genetics applied to livestock production*.
- Schenkel, F., M. Sargolzaei, G. Kistemaker, G. Jansen, P. Sullivan *et al.*, 2009 Reliability of genomic evaluation of Holstein cattle in Canada. *Interbull Bulletin*: 51-51.
- Schrooten, C., G. Schopen and P. Beatson, 2013 Across-breed genomic evaluation based on BovineHD genotypes, and phenotypes of bulls and cows, pp. 30 in *Proceedings of the 64th EAAP annual meeting*.

- Smith, C., and S. Simpson, 1986 The use of genetic polymorphisms in livestock improvement. *Journal of Animal Breeding and Genetics* 103: 205-217.
- Solberg, T., A. Sonesson, J. Woolliams and T. Meuwissen, 2008 Genomic selection using different marker types and densities. *Journal of animal science* 86: 2447-2454.
- Soller, M., and J. Beckmann, 1982 Restriction fragment length polymorphisms and genetic improvement, pp. 396-404 in *Proceedings of the 2nd World Congress on Genetics Applied to Livestock production*.
- Spelman, R., C. Ford, P. McElhinney, G. Gregory and R. Snell, 2002 Characterization of the DGAT1 gene in the New Zealand dairy population. *Journal of Dairy Science* 85: 3514-3517.
- Strandén, I., and D. Garrick, 2009 Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. *Journal of dairy science* 92: 2971-2975.
- Su, G., P. Ma, U. Nielsen, G. Aamand, G. Wiggans *et al.*, 2016 Sharing reference data and including cows in the reference population improve genomic predictions in Danish Jersey. *Animal* 10: 1067-1075.
- Thomassen, 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.
- Thomassen, J. R., A. C. Sørensen, M. S. Lund and B. Guldbrandtsen, 2014 Adding cows to the reference population makes a small dairy population competitive. *Journal of Dairy Science* 97: 5822-5832.
- Tiezzi, F., and C. Maltecca, 2015 Accounting for trait architecture in genomic predictions of US Holstein cattle using a weighted realized relationship matrix. *Genetics Selection Evolution* 47: 24.
- Uemoto, Y., T. Osawa and J. Saburi, 2017 Effect of genotyped cows in the reference population on the genomic evaluation of Holstein cattle. *animal* 11: 382-393.
- van Binsbergen, R., M. P. Calus, M. C. Bink, F. A. van Eeuwijk, C. Schrooten *et al.*, 2015 Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. *Genetics Selection Evolution* 47: 1-13.
- van den Berg, I., D. Boichard, B. Guldbrandtsen and M. S. Lund, 2016 Using sequence variants in linkage disequilibrium with causative mutations to improve across-breed prediction in dairy cattle: a simulation study. *G3: Genes, Genomes, Genetics* 6: 2553-2561.
- Van Pelt, M., G. De Jong and R. Veerkamp, 2016 Changes in the genetic level and the effects of age at first calving and milk production on survival during the first lactation over the last 25 years. *animal* 10: 2043-2050.
- VanRaden, P., C. Van Tassell, G. Wiggans, T. Sonstegard, R. Schnabel *et al.*, 2009 Invited review: Reliability of genomic predictions for North American Holstein bulls. *Journal of dairy science* 92: 16-24.
- VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. *Journal of dairy science* 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 Bulletin*.
- Weir, B., 2008 Linkage disequilibrium and association mapping. *Annu. Rev. Genomics Hum. Genet.* 9: 129-142.
- Wientjes, Y. C., M. P. Calus, P. Duenk and P. Bijma, 2018 Required properties for markers used to calculate unbiased estimates of the genetic correlation between populations. *Genetics Selection Evolution* 50: 1-16.
- Wientjes, Y. C., M. P. 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., P. Bijma, R. F. Veerkamp and M. P. L. Calus, 2016 An Equation to Predict the Accuracy of Genomic Values by Combining Data from Multiple Traits, Populations, or Environments. *Genetics* 202: 799-823.
- Woolliams, J. A., P. Bijma and B. Villanueva, 1999 Expected Genetic Contributions and Their Impact on Gene Flow and Genetic Gain. *Genetics* 153: 1009-1020.
- Zhong, S., J. C. 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.
- 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. *Genetics Selection Evolution* 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. *Journal of Dairy Science* 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. *Journal of animal breeding and genetics* 131: 249-257.

2

CHAPTER 2.

Utility of whole-genome sequence data for across-breed genomic prediction

Biaty Raymond^{1,2}, Aniek C. Bouwman¹, Chris Schrooten³, Jeanine Houwing-Duistermaat^{4,5} and Roel F. Veerkamp¹

¹Animal Breeding and Genomics, Wageningen University and Research, P.O. Box 338, 6700 AH Wageningen, The Netherlands; ²Biometris, Wageningen University and Research, 6700AA Wageningen, The Netherlands; ³CRV BV, P.O. Box 454, 6800 AL Arnhem, The Netherlands; ⁴Department of Medical Statistics and Bioinformatics, Leiden University Medical Centre, Leiden, The Netherlands; ⁵School of Mathematics, Faculty of Mathematics and Physical Sciences, University of Leeds, United Kingdom

Genetics Selection Evolution (2018) 50:2

Abstract

Background: Genomic prediction (GP) across breeds has so far resulted in low accuracies of the predicted genomic breeding values. Our objective was to evaluate whether using whole-genome sequence (WGS) instead of low-density markers can improve GP across breeds, especially when markers are pre-selected from a genome-wide association study (GWAS), and to test our hypothesis that many non-causal markers in WGS data have a diluting effect on accuracy of across-breed prediction.

Methods: Estimated breeding values (EBV) for stature and bovine high-density (HD) genotypes were available for 595 Jersey bulls from New Zealand, 957 Holstein bulls from New Zealand and 5553 Holstein bulls from the Netherlands. BovineHD genotypes for all bulls were imputed to WGS using Beagle4 and Minimac2. Genomic prediction across the three populations was performed with ASReml4, with each population used as single reference and as single validation sets. In addition, the 50k, HD and WGS markers that were significantly associated with stature in a large meta-GWAS analysis were selected and used for prediction, resulting in 10 prediction scenarios. Furthermore, we estimated the proportion of genetic variance captured by markers in each scenario.

Results: Across breeds, 50k, HD and WGS markers resulted in very low accuracies of prediction ranging from -0.04 to 0.13. Accuracies were higher in scenarios with pre-selected markers from a meta-GWAS. For example, using only the 133 most significant markers in 133 QTL regions from the meta-GWAS yielded accuracies ranging from 0.08 to 0.23, while 23,125 markers with a $-\log_{10}(p)$ higher than 7 resulted in accuracies of up to 0.35. Using WGS data did not significantly improve the proportion of genetic variance captured across breeds compared to scenarios with few but pre-selected markers.

Conclusions: Our results demonstrated that the accuracy of across-breed GP can be improved by using markers that are pre-selected from WGS based on their potential causal effect. We also showed that simply increasing the number of markers up to the WGS level does not increase the accuracy of across-breed prediction, even when markers that are expected to have a causal effect are included.

Key words: Across-breed genomic prediction, whole genome sequence, marker pre-selection, cattle, stature

2.1 Background

Today genomic prediction (GP) is a routine procedure in livestock breeding schemes (Dalton 2009; Hayes *et al.* 2009a), and its adoption has resulted in a significant increase in the rate of genetic gain. This is especially evident in dairy cattle, where its application has led to a doubling of the rate of genetic gain (Schaeffer 2006), as compared to the traditional progeny-testing scheme. Genomic prediction has been very successful mainly because it enables the estimation of genomic estimated breeding values (GEBV) for young bulls, thereby drastically reducing the generation interval. This method has been implemented mainly for within-breed genetic evaluation for which both the reference and the validation sets come from the same breed.

The use of the effect of markers that are trained in one breed to obtain GEBV for animals from a different breed, a method called across-breed GP has been rather unsuccessful. Reported accuracies for across-breed GP using the genomic best linear unbiased prediction (GBLUP) model are clustered around 0 (e.g. Hayes *et al.* 2009b; Erbe *et al.* 2012; Kachman *et al.* 2013; Calus *et al.* 2014), except in rare cases in which the quantitative trait loci (QTL) that underlie the trait are known and information about these QTL is used in the prediction model (e.g. Boichard *et al.* 2012; Hoze *et al.* 2014; Saatchi and Garrick 2014). The poor performance of across-breed GP could be due to a number of reasons, some of which include differences in linkage disequilibrium (LD) phase between markers and QTL across breeds, QTL segregating in only one breed and not the other, differences in minor allele frequencies (MAF) and allele substitution effects due to the different genetic background of the breeds (Spelman *et al.* 2002; Thaller *et al.* 2003; Zhong *et al.* 2009). An earlier hypothesis was that accuracies of across-breed prediction increase when higher density SNP panels are used (e.g. De Roos *et al.* 2008). An important question in this regard is how much gain in accuracy of across-breed prediction can be achieved if whole-genome sequence (WGS) markers are used instead of the lower density single nucleotide polymorphisms (SNPs) panels.

Given that multi-locus LD across breeds is conserved only over short distances (De Roos *et al.* 2008), the expectation is that for a large proportion of the markers, LD with causal loci changes across breeds, thus they become independent of the causal loci. Our hypothesis is that independent, non-causal markers from WGS data exhibit a dilution effect on accuracy of across-breed prediction. In other words, these markers have no explanatory value and only serve as additional sources of noise in prediction models, which in turn negatively impact the accuracy of predicting GEBV

across breeds. For example, in their simulation study, Wientjes *et al.* (2015) obtained an increase in the accuracy of across-breed GP of approximately 53% by selecting only the markers that surround the causal mutations rather than using all those available, effectively eliminating the non-causal markers that are not in LD with the causal loci. Similar results were obtained by Van Den Berg *et al.* (2016). To fully use WGS data for across-breed GP, it may become necessary to first pre-select markers with a potential causal effect on the trait of interest and use only these for prediction.

Thus, our objective was to investigate the effect of using WGS data instead of the lower density SNP chips on the accuracy of across-breed prediction. In addition, we assessed the importance of pre-selecting SNPs, using the significance and the conditional and joint (COJO) effects of SNPs in a meta-GWAS as selection criteria. We investigated this objective using 595 New Zealand Jersey and 957 New Zealand Holstein bulls, as well as 5553 Dutch Holstein bulls, all of which had deregressed proofs (DRP) for stature and bovine HD genotypes that were imputed to WGS.

2.2 Methods

2.2.1 Phenotype

Stature was used as a model trait in all analyses. Estimated breeding values (EBV) and number of daughters were available for 735 Jersey (NZJ) and 1239 Holstein (NZH) bulls from New Zealand and 5553 Dutch Holstein (DH) bulls. Data were provided by CRV BV (Cooperative Cattle Improvement Organization, Arnhem, the Netherlands). Estimated breeding values for stature for all bulls were deregressed to obtain DRP using the iterative procedure described in Calus *et al.* (2016a). This approach, called matrix deregression corrects for information of the parents of an individual on its EBV and shrinkage to the mean. To avoid double-counting of information, matrix deregression also corrects for information of other relatives that are included in the same reference population as the individual whose EBV is being deregressed. The output from this procedure were the DRP and deregressed effective daughter contributions (EDC). Animals with deregressed EDC of 0 were removed, resulting in 595 NZJ, 957 NZH and 5553 DH bulls in the final dataset. Deregressed EDC were used as weights for the DRP in subsequent analyses. For the DH set, the mean EDC was equal to 52, while for the NZH and NZJ sets, it was equal to 16 and 17, respectively. The set of DH individuals included in the meta-GWAS of Bouwman *et al.* (2018) was the exact same set of DH individuals used in our study.

2.2.2 Genotypes and imputation to WGS

The bulls were either genotyped with the Illumina BovineHD Bead chip (HD; 734,403 SNPs; Illumina Inc., San Diego, USA), or with a 50K SNP chip and imputed to HD. The HD genotypes were subsequently imputed up to WGS using the sequenced population from the 1000 Bull Genomes Project Run 4 as reference population. This multi-breed reference population contained 1147 sequenced animals with on average 11-fold coverage. The reference population contained 311 Holstein and 61 Jersey animals, but all individuals were used as reference because earlier studies showed that a multi-breed sequenced reference population can be beneficial for imputation accuracy, especially for SNPs with a low minor allele frequency (MAF) (e.g. Bouwman and Veerkamp 2014; Brøndum *et al.* 2014; Daetwyler *et al.* 2014). Polymorphic sites, including SNPs and short insertions and deletions (indels) were identified in the 1147 individuals simultaneously as described in Daetwyler *et al.* (2014). The genotype calls of the 1000 bull genomes reference population were improved with BEAGLE 4.0 (Browning and Browning 2009). The sequence data contained 30,339,468 bi-allelic markers (SNPs and indels) with four or more copies of the minor allele in the reference population, which were used for imputation. Further details about the 1000 bull genomes reference population, variant calling and filtering of variants are in Daetwyler *et al.* (2014).

Imputation of HD genotypes to WGS was done using standard settings in MINIMAC2 (Fuchsberger *et al.* 2015) and the pre-phased reference genotypes (SNPs and indels) that resulted from BEAGLE 4.0 (default settings) (Browning and Browning 2009). The imputed data were allele dosages that were subsequently converted to most probable genotypes. The output from MINIMAC2 included an r^2 value, which is the estimated correlation between imputed and true genotypes. Markers that had an r^2 lower than 0.5 and those that were either monomorphic or had only opposing homozygous genotypes were filtered out. We also filtered out markers that had less than 10 copies of the minor allele in the populations. This criterion was the same as setting a MAF threshold of $\frac{10}{N*2}$, where N is the number of genotyped individuals. Filtering based on MAF was done separately for the NZH, NZJ and DH sets. Furthermore, in the final dataset we retained only markers that segregated in all three populations.

The results of a meta-GWAS analysis conducted on 58,265 individuals with imputed WGS were used to select variants associated with stature (Bouwman *et al.* 2018). This meta-analysis was performed on 17 populations from eight breeds, including Holstein and Jersey, from different countries. The DH population was included in the

meta-GWAS, whereas both the NZH and NZJ populations were not included. The meta-analysis across the populations found 24,230 genome-wide significant ($p < 5 \times 10^{-8}$) WGS markers in 163 distinct QTL regions that were spread across 27 autosomes (excluding chromosomes 24 and 27). For each of the 163 regions, the lead marker (160 SNPs and three indels) was determined based on its highest significance.

2.2.3 Scenarios and selection of markers

In our first scenario, we used all available markers in the WGS data for GP to test the hypothesis that the number of markers used determines the accuracy of across-breed predictions. We also created scenarios that contained only markers that showed statistical significance in the previously conducted meta-GWAS analysis to evaluate whether the non-significant markers from the meta-GWAS only add noise in a GP model. Since markers can show spurious associations based on the effect of LD and family structure, simply selecting markers based on statistical significance is not sufficient to filter out all the potentially non-causal markers. Hence, we performed a forward selection of markers to create other scenarios using the conditional and joint (COJO) effect method described in Yang *et al.* (2012) and implemented in the GCTA software Yang *et al.* (2011). The COJO option in GCTA is a stepwise marker selection procedure that selects independently associated markers, given some parameter thresholds and the summary level result from a GWAS. As outlined in the main paper of Yang *et al.* (2011), the marker selection model is initiated with the most significant marker that is below the threshold, followed by a calculation of the conditional p values for the remaining markers i.e. conditional on the marker that is already in the model. The marker with the lowest conditional p value is then selected and added to the model, and the process is repeated. It is possible that the marker being tested is in high LD with other markers that are already in the model. To avoid having such markers in the model, a collinearity threshold is set. If the threshold is exceeded, the marker being tested is dropped and the next one is considered instead. A window size in mega base pairs (Mb) is defined, outside of which markers are assumed to be in linkage equilibrium. Once all the markers have been considered and those that are independently significant have been identified by the model and selected, the selected markers are then jointly fitted in a model and those with a p value higher than the threshold are dropped. These steps are repeated iteratively until no more markers can be added or removed from the model. In total, 10 scenarios were created depending on whether or not variants were pre-selected from WGS data and the criteria of selection that were

used. All the scenarios are described in Table 2.1. The total number of markers in each scenario and their mean MAF are in Table 2.2.

Table 2.1 Description of scenarios used in the study

Scenario	Description
Scenarios with unselected markers	
Full_seq	All variants in WGS data that met the quality check criteria were included, following the assumptions of infinitesimal model; this was the default scenario
HD	SNPs on the commercial BovineHD SNP chip
50k	SNPs on the commercial 50k chip
Scenarios with selected marker + unselected markers	
HD_Top	Top_markers added to HD
50k_Top	Top_markers added to 50k
Scenarios with only selected markers	
Pval5	All variants that had a $-\log_{10}(p)$ value higher than 5 from the meta-GWAS analysis
Pval7	All variants that had a $-\log_{10}(p)$ value higher than 7 from the meta-GWAS analysis
COJO3	Variants selected using the following COJO model parameters thresholds: conditional and joint p-value threshold (p) = $5e-3$, collinearity between selected SNPs = 0.9, window size = 10 Mb
COJO8	Variants selected using the following COJO model parameters thresholds: conditional and joint p-value threshold (p) = $5e-8$, collinearity between selected SNPs = 0.9, window size = 10 Mb
Top_markers	The most significant variants identified in each defined QTL region from the meta-GWAS analysis; only those that passed QC are included here

Table 2.2 Number of markers in each scenario and their mean minor allele frequency in Dutch Holsteins (DH), New Zealand Holsteins (NZH) and New Zealand Jerseys (NZJ)

Scenario	Number of SNPs	Mean MAF		
		DH	NZH	NZJ
Scenarios with unselected markers				
Full_seq	14,341,737	0.18	0.17	0.15
HD	583,078	0.27	0.28	0.24
50k	48,912	0.25	0.26	0.21
Scenarios with unselected + selected markers				
HD_TOP	583,194	0.27	0.28	0.24
50k_TOP	49,045	0.25	0.25	0.21
Scenarios with only selected markers				
Pval5	59,828	0.25	0.23	0.17
Pval7	23,125	0.24	0.23	0.17
COJO3	1570	0.18	0.17	0.13
COJO8	360	0.20	0.21	0.13
Top_markers	133	0.23	0.24	0.16

2.2.4 Proportion of genetic variance explained

In each scenario, a genomic relationship matrix (GRM) was calculated for all the animals according to the first method of Vanraden (2008). This was done using `calc_grm` (Calus 2013) as:

$$\text{GRM} = \frac{ZZ'}{2 \sum p_j(1-p_j)},$$

where p_j is the frequency of the allele of marker j with the homozygous genotype being coded as 2, and Z is the centred ($Z_{ij} = X_{ij} - 2p_j$) genotype matrix. X_{ij} is the genotype of marker j for animal i , with the elements of X_{ij} coded as 0, 1, 2. Using the estimated GRM and the DRP of the reference animals as phenotypes, a genomic restricted maximum likelihood (GREML) analysis was run in ASReml (Gilmour et al. 2009) to estimate the proportion of genetic variance captured by the markers in each scenario. In the GREML model, deregressed EDC were used to scale the residual of each animal in the model such that animals with a high deregressed EDC has effectively a small residual. The weighting procedure is described in more detail below in the genomic prediction sub-section. Normally, the residual variance obtained from such analyses is defined for a weight of 1, but because the deregressed EDC that were used as weights were considerably higher than 1, the

resulting residual variances were inflated. To correct for this, we scaled the residual variances by the mean of the deregressed EDC in the data (Garrick *et al.* 2009). For example, if the mean of deregressed EDC in the data is k , we computed the phenotypic variance as: $var(g) + \frac{var(e)}{k}$, where $var(g)$ is the genetic variance and $var(e)$ is the residual variance.

2.2.5 Genomic prediction with variants in the different scenarios

For GP, the composition of reference and validation sets are in Table 2.3. Estimation of GEBV and the proportion of genetic variance captured by markers as explained in the previous section was carried out simultaneously using the same GRM, phenotype data and model. A single-trait animal model was fitted in ASReml (Gilmour *et al.* 2009) as follows:

$$y = 1\mu + Wg + e,$$

where y is a vector containing stature DRP for the reference set (missing for the validation set), 1μ is the mean of the DRP, g is a vector of additive genetic effects for all animals, W is the design matrix that links g to the DRP in y and e is a vector containing the residuals. Both g and e are assumed to be normally distributed as $g = N(0, GRM\sigma_g^2)$ and $e = N(0, D\sigma_e^2)$, where σ_g^2 the genetic variance, σ_e^2 the residual variance and D is a diagonal matrix that contains the inverse of deregressed EDC, which in this case were used as weights for the DRP in the model. We measured the accuracy of GP as the correlation between the GEBV and the DRP of the validation animals. The standard error (SE) of accuracy was calculated based on the sampling variance of a correlation as $SE = \frac{1-r^2}{\sqrt{N-2}}$, where r is the accuracy and N is the sample size (number of validation candidates).

Table 2.3 Composition of reference and validation sets used in the study

Reference set (size)	Validation set (size)
New Zealand Holsteins (957)	New Zealand Jerseys (595) Dutch Holsteins (5,553)
New Zealand Jerseys (595)	New Zealand Holsteins (957) Dutch Holsteins (5553)
Dutch Holsteins (5553)	New Zealand Holsteins (957) New Zealand Jerseys (595)

Furthermore, prediction bias was assessed by regressing DRP of validation animals on their GEBV. Stature is measured on a different scale in New Zealand and in the Netherlands. As a result, the GEBV of the NZH and NZJ bulls that were obtained by using a DH reference population will be on the DH scale, while their DRP are on a different scale. To estimate the slope and intercept of regression, we rescaled the DRP and GEBV so that they are on the same scale and expressed in their standard deviation units as:

$$\text{Scaled_DRP} = \frac{1}{(SD_{\text{DRP}})} * (\text{DRP} - \overline{\text{DRP}}),$$

$$\text{Scaled_GEBV} = \frac{1}{(SD_{\text{DRP_Ref}})} * (\text{GEBV}),$$

where Scaled_DRP and Scaled_GEBV are the rescaled DRP and GEBV of the validation animals, DRP is the DRP of the validation animals on their original scale, SD_{DRP} is the standard deviation of DRP and $SD_{\text{DRP_Ref}}$ is the standard deviation of the DRP of reference population.

2.3 Results

2.3.1 Proportion of genetic variance explained

We estimated the proportion of genetic variance explained by markers in the different scenarios. This was done separately for the DH, NZH and NZJ datasets and the results are presented in Fig. 2.1. Across the three populations, the proportion of genetic variance captured increases almost linearly with the number of markers up to 50k. Above 50k, the proportion of genetic variance captured plateaus. Consequently, using WGS markers did not lead to any significant gain in the proportion of genetic variance explained compared to the lower density panels.

Fig. 2.1 Proportion of genetic variance for stature explained by the different marker sets in Dutch Holstein, New Holstein and New Zealand Jersey populations.

Except for the extreme cases of the Top-markers, COJO8 and COJO3 scenarios that contained only 133, 360 and 1570 markers, respectively, the proportion of captured genetic variance did not differ markedly between scenarios. With only 133 markers that were identified as lead markers in a meta-GWAS analysis [18] and across the three populations, 41 to 59% of the genetic variance was captured, which is more than half of the genetic variance captured by the WGS data, which include approximately 14 million markers (76 to 97%). Across all the scenarios, more genetic variance was explained in the DH population followed by the NZH population. This was expected, given that the DH bulls were included in the meta-GWAS that was used to select the markers, whereas the New Zealand populations were not included. Moreover, about 40% of the animals used for the meta-GWAS were Holsteins whereas only about 5% of the animals were Jerseys from Australia. Another important reason in this regard is the large size of the DH population compared to those of NZH and NZJ.

2.3.2 Accuracy of across-breed genomic prediction

Our main objective was to investigate how much gain in accuracy of genomic prediction can be achieved across breed and across country by using WGS data instead of the lower density marker panels. As presented in Table 2.4, using WGS data did not lead to any gain in the accuracy of GP compared to using HD or 50k markers. Accuracy of prediction was close to 0 (0.02 to 0.08) in the across-breed, across-country (DH and NZJ) GP scenario. Similarly, accuracy of prediction obtained using WGS data ranged from 0.06 to 0.09 in the across-breed but within-country (NZH and NZJ) GP scenario. In the within-breed, across-country (NZH and DH) GP scenario, accuracies of prediction were considerably higher, ranging from 0.41 to 0.51. However, even in this case the accuracies obtained using WGS data were lower than those obtained using the 50k and HD marker panels.

The remaining scenarios, apart from those with WGS data which contained unselected markers (50k and HD), also performed poorly in across-breed prediction with accuracies ranging from -0.04 to 0.13. The top markers were subsequently added to the 50k and HD markers to form new scenarios (50k_Top and HD_Top, respectively). However, using the combined markers (50k_Top and HD_Top) in across-breed GP resulted in low accuracies of prediction (-0.04 to 0.13), similar to those obtained by using only the 50k and HD markers. Using the 50k_top and HD_Top markers in within-breed GP, but across countries, also resulted in similar accuracies as those obtained using only 50k and HD. Overall, adding the potentially causal markers to the 50k and HD markers did not result in any gain in the accuracy of GP.

The prior-selection of markers from WGS data, which was based on their potential causal effect, and the benefit of using such markers for across-breed GP were also investigated in this study. Accuracies of across-breed prediction obtained by using the pval5 markers ranged from 0.14 to 0.32, whereas accuracies obtained with the pval7 markers ranged from 0.07 to 0.35. Accuracies of prediction using the Pval5 and Pval7 markers are already 3 to 4 times higher than those obtained using WGS data although they contained only very small fractions of the number of markers that are in WGS data. In the within-breed, across-country GP scenario, we also observed a slight increase in the accuracy of GP using pval5 and pval7 markers compared to all the scenarios with unselected markers.

Table 2.4 Accuracies of prediction for all prediction scenarios in the study

Reference Validation	Across-breed, across-country				Across-breed, within-country				Within-breed, across-country			
	DH	NZJ	DH	NZJ	NZH	NZJ	NZH	DH	NZH	DH	NZH	DH
Scenarios with unselected markers (number of markers)												
Full_seq (14,341,737)	0.08	0.02			0.09	0.06			0.41			0.51
HD (583,078)	0.10	0.05			0.07	-0.04			0.43			0.55
50k (48,912)	0.06	0.05			0.13	-0.04			0.42			0.52
Scenarios with unselected markers + selected markers (number of markers)												
HD_Top (583,194)	0.10	0.05			0.07	-0.04			0.43			0.55
50k_Top (49,045)	0.07	0.06			0.13	-0.03			0.43			0.53
Scenarios with only selected markers (number of markers)												
Pval5 (59,828)	0.19	0.19			0.14	0.32			0.52			0.58
Pval7 (23,125)	0.01	0.15			0.07	0.35			0.46			0.47
COJO3 (1570)	0.20	0.22			0.14	0.31			0.45			0.52
COJO8 (360)	0.21	0.25			0.16	0.18			0.47			0.44
Top_markers (133)	0.23	0.18			0.08	0.27			0.41			0.47
SE*	0.04	0.01			0.04	0.03			0.01			0.01

The populations are Dutch Holsteins (DH), New Zealand Holsteins (NZH) and New Zealand Jerseys (NZJ). *SE: Standard error of estimates.

One can argue that selecting markers purely on significance in a GWAS is naive, in the sense that some of these markers are selected simply because they are in LD with causal mutations and that they provide redundant information in GP. It was for this reason that we performed forward marker selection to limit the effect of LD, thus resulting in the COJO3 and COJO8 scenarios. The COJO scenario contained independently significant markers from the meta-GWAS at a significance level of $5e-3$ for COJO3 and $5e-8$ for COJO8. Accuracy of across-breed prediction using the markers in the COJO3 scenario ranged from 0.14 to 0.31, whereas that in the COJO8 scenario ranged from 0.16 to 0.25. This indicates that the COJO scenarios have about the same level of accuracy as the Pval scenarios, in spite of having fewer markers. The accuracy of across-breed prediction using only the 133 top markers from the meta-GWAS ranged from 0.08 to 0.27, which for some across-breed prediction scenarios (NZJ as reference population or DH as reference population to predict GEBV for NZJ) is approximately 3 times higher than the accuracy obtained using WGS data.

2.3.3 Prediction bias

In this study, we assessed prediction bias based on the slope of regression when the DRP of the validation animals were regressed onto their GEBV; the results are in Table 2.5. For across-breed, across-country GP, the slope ranged from 0.01 to 0.81, and there was no apparent difference between the scenarios with selected and unselected markers sets. For across-breed, within-country GP, predictions were also biased. GEBV for the NZJ predicted from the NZH were underestimated, whereas GEBV for the NZH predicted from the NZJ were overestimated. Exceptions were the scenarios HD, HD_Top, 50k and 50k_Top, which all had negative slopes, although these estimates had high standard errors and were not significantly different from zero. These scenarios using unselected marker sets, also had close to zero (and slightly negative) prediction accuracies (Table 4). For within-breed, across-country GP, when DH was used as reference population to predict GEBV for NZH, slopes deviated more from 1 in the scenarios with pre-selected markers (0.55 to 0.72) than those with unselected markers (0.77 to 0.85). When GEBV for DH were predicted from NZH, the use of pre-selected markers or unselected markers made a difference in either over- or under-predicting the breeding values.

Table 2.5 Intercept and slope of regression when deregressed breeding values of validation animals were regressed onto their predicted genomic breeding values. The populations are Dutch Holsteins (DH), New Zealand Holsteins (NZH) and New Zealand Jerseys (NZJ). Standard errors for the slopes are given in parentheses.

Across-breed, across-country				
Scenarios	DH reference, NZJ validation		NZJ reference, DH validation	
	Intercept	Slope	Intercept	Slope
Full_seq	0.28	0.38 (0.20)	-0.01	0.36 (0.22)
HD	0.39	0.41 (0.16)	-0.02	0.73 (0.20)
50k	0.24	0.20 (0.14)	-0.01	0.63 (0.16)
HD_Top	0.39	0.41 (0.16)	-0.02	0.74 (0.20)
50k_Top	0.28	0.23 (0.13)	-0.02	0.73 (0.16)
Pval5	0.30	0.36 (0.08)	-0.03	0.59 (0.04)
Pval7	0.01	0.01 (0.06)	-0.04	0.42 (0.03)
COJO3	0.40	0.35 (0.07)	-0.05	0.71 (0.04)
COJO8	0.13	0.40 (0.07)	-0.03	0.81 (0.04)
Top_markers	0.46	0.43 (0.08)	-0.06	0.48 (0.03)
Across-breed, within-country				
Scenarios	NZH reference, NZJ validation		NZJ reference, NZH validation	
	Intercept	Slope	Intercept	Slope
Full_seq	0.37	1.17 (0.01)	-0.04	0.87 (0.47)
HD	0.47	1.17 (0.01)	0.02	-0.58 (0.43)
50k	0.45	1.17 (0.01)	0.02	-0.43 (0.36)
HD_Top	0.47	1.17 (0.01)	0.02	-0.57 (0.43)
50k_Top	0.46	1.19 (0.01)	0.02	-0.33 (0.36)
Pval5	0.32	1.19 (0.02)	0.01	0.90 (0.08)
Pval7	0.40	1.16 (0.03)	0.05	0.87 (0.07)
COJO3	0.52	1.20 (0.02)	0.05	0.89 (0.09)
COJO8	0.54	1.13 (0.03)	0.00	0.56 (0.10)
Top_markers	0.51	1.02 (0.04)	0.03	0.70 (0.08)
Within-breed, across-country				
Scenarios	DH reference, NZH validation		NZH reference, DH validation	
	Intercept	Slope	Intercept	Slope
Full_seq	0.45	0.85 (0.04)	-0.24	1.16 (0.03)
HD	0.40	0.81 (0.04)	-0.27	1.17 (0.03)
50k	0.36	0.77 (0.04)	-0.27	1.13 (0.03)
HD_Top	0.40	0.81 (0.04)	-0.27	1.17 (0.03)

50k_Top	0.38	0.78 (0.04)	-0.27	1.14 (0.03)
Pval5	0.47	0.72 (0.03)	-0.27	1.09 (0.02)
Pval7	0.24	0.55 (0.03)	-0.27	0.98 (0.03)
COJO3	0.44	0.64 (0.03)	-0.22	0.91 (0.02)
COJO8	0.33	0.55 (0.04)	-0.22	0.91 (0.02)
Top_markers	0.47	0.69 (0.04)	-0.16	0.74 (0.02)

2.4 Discussion

The objective of this study was to investigate the potential gain in accuracy of across-breed prediction when WGS data are used instead of lower density marker panels. Our results showed that simply increasing the number of markers up to WGS does not result in an increased accuracy of across-breed prediction and that, in some cases, it can even result in lower accuracies compared to when 50k or HD marker panels are used. These results are in agreement with previous studies that showed no added benefit of using WGS data for GP (e.g. Ober *et al.* 2012; Hayes *et al.* 2014; Van Binsbergen *et al.* 2015; Calus *et al.* 2016b), although in these studies, GP was carried out within breed.

One of the explanations for the low accuracy of across-breed GP is the inconsistency of LD between markers and causal mutations across breeds (Meuwissen *et al.* 2001). On the one hand, this challenge can be circumvented by simply increasing the number of markers up to the point that each causal locus is in LD with at least one marker across breeds. For example, in cattle it was shown that about 300,000 SNPs are sufficient to ensure the consistency of LD phase across breeds (De Roos *et al.* 2008), which is far less than the 14 million markers that we used in this study. On the other hand, increasing the number of markers up to WGS increases the chance that causal mutations are in LD with a large number of non-causal markers, even across breeds. Thus, the effect of the causal markers are picked up by a linear combination of many other non-causal markers, with which the causal markers are in LD. In other words, with a GREML model, the effect of the causal markers is shrunk towards the mean to obtain a uniform distribution of marker effects, as is the assumption in the infinitesimal model. As a result, the model is not able to pinpoint appropriately the true effect of the causal markers across breeds, thereby resulting in low accuracies of prediction.

Another possible explanation for our result, which was also put forward by Wientjes *et al.* (2015), is that the accuracy of estimating marker effects is low. The low accuracy of estimating marker effects can be due to differences in allele substitution effects at the causal loci between the breeds (Spelman *et al.* 2002), resulting from differences in genetic background, differences in MAF between breeds, possible gene-by-environment interactions, different contributions of non-additive gene effects and the limited size of available data. The differences in allele substitution effects at the causal loci reflect the genetic correlation between breeds (Bohren *et al.* 1966; Falconer and Mackay 1996), which in theory sets an upper limit for the accuracy of across-breed GP. Although the low accuracies that we observed for across-breed prediction might have been exacerbated by the low accuracies of estimating marker effects, we believe that another major underlying reason is the increased model complexity and over-parameterization that comes along with the use of WGS data in GP.

In this study, we investigated the effect of prior selection of markers from a meta-GWAS study on the accuracy of across-breed GP. Our results clearly demonstrated that when markers in the WGS data are selected a priori based on their potential causal effect and that the potentially non-causal marker are excluded, the accuracy of across-breed prediction can be increased significantly. We believe that the non-causal markers in WGS data exhibit some kind of dilution effect on the accuracy of genomic prediction. Although a slight dilution effect of non-causal markers was noted for within-breed, across-country GP i.e., when we used DH to predict NZH and the other way round, this effect was more pronounced in an across-breed GP scenario. A possible explanation is that within breeds, LD extends over long distances, which increases the chance that every marker in the dataset is in LD with a causal locus and with other markers in LD with the same causal locus. Thus, many of the markers, although having no causal effect on their own, still have an explanatory value due to LD. However, LD between markers and causal loci across breeds is conserved only across short distances (De Roos *et al.* 2008). This means that across breeds, there are probably many markers that are not linked to any causal locus, and have different LD patterns in different breeds, which in a way make them serve as sources of noise and dilution factors in across-breed GP. The proportion of these 'lone standing' markers and the corresponding noise effect increase when for example instead of 50k or HD marker panels, WGS data are used for GP across distantly related breeds. Our results agree completely with those of Van Den Berg *et al.* (2016) who found in their simulation studies that GP using markers that are in LD with causal loci was more accurate than when all markers

(both those in LD with causal loci and those that are not) were used together. Similar results were obtained Brøndum *et al.* (2014). We have empirically shown that there is a good prospect for the use of WGS data in across-breed GP, provided the potentially causal markers in such data are accurately identified and used for prediction.

In this study, we benefitted from the results of a very powerful meta-GWAS (Bouwman *et al.* 2018). Thus, there is a high level of confidence that the pre-selected markers from the meta-GWAS either are the causal variants themselves or are in high LD with the causal variants. This explains the significant improvement in accuracies of across-breed GP resulting from the use of pre-selected markers compared to the use of unselected marker sets (Table 4). Such a powerful meta-GWAS for pre-selection of markers is currently rather unique, hence pre-selection of markers for other traits is not a trivial task. Other ways to pre-select markers from whole-genome sequence based on their significance for the trait of interest include GWAS using (a subset of) the reference population or variable selection models. Performing a GWAS using data on the reference population and using preselected markers based on their significance from the GWAS for GP in the same population was shown to be less optimal than using unselected markers sets for GP (Veerkamp *et al.* 2016). In their study, Veerkamp *et al.* (2016) found that the use of pre-selected markers from a GWAS for GP in the same population than that used for the GWAS resulted in lower accuracy of prediction and increased prediction bias compared to the use of unselected marker sets. If the use of such pre-selected markers results in decreased accuracy of within-breed prediction, the expectation is that using such markers for across-breed GP will result in little to no benefit, given that there may be differences in the properties of the markers across breeds.

In theory, similar or better prediction accuracies than those reported in this study can be obtained if data on all the individuals in the meta-GWAS are available, and if they are combined in a single multi-breed reference population to predict the GEBV of validation animals, using a Bayesian variable selection model. There are, however, two major limitations to this approach, the first being the willingness of all parties involved to share both phenotype and genotype data. In the meta-GWAS, this limitation was non-existent because the partners in the consortium shared only the summary level results of their 'in-house' GWAS (Bouwman *et al.* 2018) and did not have to share their phenotype or genotype data. Second, implementing a Bayesian variable selection model with 58,265 individuals all of which have WGS data is computationally not practical (e.g. Van Binsbergen *et al.* 2015; Calus *et al.* 2016b).

Alternatively, a Bayesian variable selection model can be used to analyse the WGS data of the reference population in this study and the results could be applied for across-breed GP. However, with Bayesian variable selection models, like most other models, inference about marker effects or variances depends highly on sample size (Gianola *et al.* 2009; Habier *et al.* 2011) i.e., the size of the reference population. In that sense, the inferential power or accuracy of a Bayesian variable selection model using a training population of at most 5,553 individuals is expected to be much lower compared than that of a meta-GWAS carried out using 58,265 individuals with WGS. Moreover, with a single breed reference population, strong family relationships and long stretches of LD can hinder the accurate pinpointing of the causal regions underlying complex traits, irrespective of the type of model applied (e.g. Habier *et al.* 2007; Habier *et al.* 2010; Calus *et al.* 2016b). For example, Habier *et al.* (2007) showed that, in the presence of LD, prediction accuracies obtained using different models including BayesB decayed rapidly across generations. If the argument holds that such a decay in accuracy is a result of differences in LD patterns across generations, then difference in LD patterns across breeds will hinder the accurate prediction of GEBV across breeds using a single breed reference population, even with a Bayesian variable selection model.

In GP, prediction bias is commonly estimated as the regression coefficient (β) when DRP of validation animals are regressed onto their GEBV. The expectation of 1 for β does not hold in an across-breed GP scenario, mainly because the prediction model does not take all the uncertainty into account i.e. differences in LD phase between breeds and differences in allele substitution effect at the causal locus (genetic correlation less than 1) between breeds (Van Den Berg *et al.* 2016). When markers are in high LD with causal mutations within the reference population, their estimated effects are likely to be close to the true effect of the causal loci, at least proportional to the extent of LD with the causal loci (Zhu *et al.* 2015). However, if the LD between these markers and the causal mutation does not hold across breeds, then these markers lose their effect across breeds. This loss in the effect of markers across breeds is not accounted for in the prediction model and therefore can result in the inflation of the variance of GEBV of the validation animals (Wientjes *et al.* 2016) and thus in lower slope values, as observed in most of the scenarios in our study (Table 2.5).

The DH population used in our study was included in the meta-GWAS that was used to pre-select the markers. When the DH population was used as a reference population to predict the GEBV of another Holstein population (NZH), predictions

were more biased downwards in the scenarios with pre-selected markers than in the scenarios with unselected markers. One of the reasons for this might be that the markers are discovered in (partly) the same population as the one used to train the effects of the markers for GP, resulting in an overestimation of the effects of the markers and consequently of the variance of GEBV, a phenomenon commonly referred to as the “Beavis effect” (Xu 2003). For example, the variance of the GEBV estimated by using WGS data was equal to 0.36, while it was 0.46 when the pre-selected top-markers were used instead. Moreover, the DH and NZH populations are related, since some of the DH individuals were used as sires for the NZH population. Therefore, using DH as a reference population to predict the GEBV of NZH could have resulted in prediction error covariance between the DRP and GEBV of the validation animals (Wray *et al.* 2013), hence the biased predictions. Veerkamp *et al.* (2016) also showed that predictions were more biased when they used pre-selected markers from WGS data based on the result of a GWAS, than when unselected marker sets were used. Moreover, in their study, the discovery (GWAS) population was the same as the reference population used for GP.

When GEBV are estimated with low accuracy, they are strongly regressed towards the mean, resulting in a narrow range of values for GEBV, which means that the slope of regression (bias) becomes very sensitive to even slight changes in GEBV between scenarios. For example, when the numerically small NZJ population was used as a reference population to predict GEBV for NZH, the slope of regression drastically changed from 0.87 in Full_seq to -0.58 in HD. These scenarios have very low prediction accuracies (Table 4) and very narrow ranges of values for GEBV (-0.22 to 0.26 for Full_seq and -0.26 to 0.29 for HD). Because the DRP remained the same between the scenarios (-3.37 to 3.2), only a slight change in GEBV between these scenarios could have resulted in such a huge difference in the estimated slope, accompanied by high standard errors (Table 2.5).

With higher accuracies, for example in the scenarios with pre-selected markers, the GEBV become less regressed to the mean compared to the scenarios with very low prediction accuracies. For example, GEBV ranged from -1.26 to 1.33 when using the Top-markers and from -1.1 to 1.1 when using the COJO8 markers. As a result, the slopes are estimated more precisely in these scenarios, i.e., with low standard errors and the slopes do not change drastically between the scenarios (Table 5). An important message from our results is that expectation for the bias is not 1, when across-breed differences are not properly accounted for, especially when a numerically small reference population is used. Furthermore, with low accuracies of

across-breed prediction, the range of GEBV for the validation animals becomes very small due to shrinkage to the mean, and estimated bias in such cases can be very unreliable, since they are estimated with high standard errors.

2.5 Conclusions

The use of whole-genome sequence data holds the potential to increase the accuracy of across-breed GP because it is expected to contain all the causal mutations that underlie the traits of interest. Contrary to common belief (e.g. Meuwissen and Goddard 2010; Macleod *et al.* 2014), our results showed that simply using all markers in the WGS data in a linear GREML model results in a very low accuracy of across-breed GP. In addition to possible differences in allele substitution effects at the causal loci between breeds, this could be due to the increased model complexity and over-parameterization that accompany the use of WGS data. Furthermore, the results of this study showed that the problem of over-parametrization can be avoided partly by prior selection of markers in WGS data to contain only those with a potential causal effect. This approach enables the model to focus directly on the important regions and can result in an increase in accuracy of across-breed prediction that is more than twice that obtained by using 50k and HD marker panels or unselected WGS data. However, for some scenarios in our study, accuracies of prediction were still low even when doubled by the use of pre-selected markers instead of WGS.

2.6 Acknowledgements

This study was financially supported by NWO-TTW, the Dutch Ministry of Economic Affairs (TKI Agri & Food project 16022, code BO-22.04-011-001-ASG-LR), and the Breed4Food partners Cobb Europe, CRV, Hendrix Genetics and Topigs Norsvin. The use of the HPC cluster was made possible by CAT-AgroFood (Shared Research Facilities, Wageningen University and Research).

2.7 References

- Bohren B, Hill W, Robertson A. Some observations on asymmetrical correlated responses to selection. *Genet Res.* 1966;7:44–57.
- Boichard D, Guillaume F, Baur A, Croiseau P, Rossignol MN, Boscher MY, et al. Genomic selection in French dairy cattle. *Anim Prod Sci.* 2012;52:115–20.
- Bouwman AC, Daetwyler HD, Chamberlain AJ, Ponce CH, Sargolzaei M, Schenkel FS, et al. Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. *Nat Genet.* 2018;50:362–7.
- Bouwman AC, Veerkamp RF. Consequences of splitting whole-genome sequencing effort over multiple breeds on imputation accuracy. *BMC Genet.* 2014;15:105.
- Brondum RF, Guldbrandsen B, Sahana G, Lund MS, Su GS. Strategies for imputation to whole genome sequence using a single or multi-breed reference population in cattle. *BMC Genomics.* 2014;15:728.
- 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. *The American Journal of Human Genetics* 84: 210-223.
- Calus M, Vandenplas J, Ten Napel J, Veerkamp RF. Validation of simultaneous deregression of cow and bull breeding values and derivation of appropriate weights. *J Dairy Sci.* 2016;99:6403–19.
- Calus, M. *calc_grm*. A programme to compute pedigree, genomic, and combined relationship matrices. Animal Breeding and Genomics Centre, Wageningen UR Livestock Research; 2013.
- Calus MPL, Bouwman AC, Schrooten C, Veerkamp RF. Efficient genomic prediction based on whole-genome sequence data using split-and-merge Bayesian variable selection. *Genet Sel Evol.* 2016;48:49.
- Calus MP, Huang H, Vereijken A, Visscher J, Ten Napel J, Windig JJ. Genomic prediction based on data from three layer lines: a comparison between linear methods. *Genet Sel Evol.* 2014;46:57.
- Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brøndum RF, et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nat Genet.* 2014;46:858–65.
- De Roos APW, Hayes BJ, Spelman R, Goddard ME. Linkage disequilibrium and persistence of phase in Holstein–Friesian Jersey and Angus cattle. *Genetics.* 2008;179:1503–12.
- Dalton R. No bull: genes for better milk. *Nat News.* 2009;457:369.
- Erbe M, Hayes BJ, Matukumalli LK, Goswami S, Bowman PJ, Reich CM, et al. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *J Dairy Sci.* 2012;95:4114–29.
- Falconer D, Mackay T. *Introduction to quantitative genetics.* Harlow: Pearson Education Ltd; 1996.

- Fuchsberger C, Abecasis GR, Hinds DA. minimac2: faster genotype imputation. *Bioinformatics*. 2015;31:782–4.
- Garrick DJ, Taylor JF, Fernando RL. Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genet Sel Evol*. 2009;41:55.
- Gianola D, de los Campos G, Hill WG, Manfredi E, Fernando R. Additive genetic variability and the Bayesian alphabet. *Genetics*. 2009;183:347–63.
- Gilmour AR, Gogel RBJ, Cullis BR, Thompson R. ASReml user guide release 3.0. Hemel Hempstead: VSN International Ltd; 2009.
- Habier D, Fernando RL, Dekkers JCM. The impact of genetic relationship information on genome-assisted breeding values. *Genetics*. 2007;177:2389–97.
- Habier D, Fernando RL, Kizilkaya K, Garrick DJ. Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics*. 2011;12:186.
- Habier D, Tetens J, Seefried FR, Lichtner P, Thaller G. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genet Sel Evol*. 2010;42:5.
- Hayes B, Bowman PJ, Chamberlain AJ, Goddard ME. Invited review. Genomic selection in dairy cattle: progress and challenges. *J Dairy Sci*. 2009;92:433–43.
- Hayes BJ, Bowman PJ, Chamberlain AC, Verbyla K, Goddard ME. Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genet Sel Evol*. 2009;41:1.
- Hayes BJ, MacLeod IM, Daetwyler HD, Bowman PJ, Chamberlain AJ, Vander Jagt CJ, et al. Genomic prediction from whole genome sequence in livestock: the 1000 bull genomes project. In: *Proceedings of the 10th world congress on genetics applied to livestock production: 17–22 August 2014; Vancouver; 2014*.
- Hoze C, Fritz S, Phocas F, Boichard D, Ducrocq V, Croiseau P. Genomic evaluation using combined reference populations from Montbéliarde and French Simmental breeds. In: *Proceedings of the 10th world congress of genetics applied to livestock production: 17–22 August 2014; Vancouver; 2014*.
- Kachman SD, Spangler ML, Bennett GL, Hanford KJ, Kuehn LA, Snelling WM, et al. Comparison of molecular breeding values based on within- and across-breed training in beef cattle. *Genet Sel Evol*. 2013;45:30.
- MacLeod IM, Hayes BJ, Goddard ME. The effects of demography and long-term selection on the accuracy of genomic prediction with sequence data. *Genetics*. 2014;198:1671–84.
- Meuwissen T, Goddard M. Accurate prediction of genetic values for complex traits by whole-genome resequencing. *Genetics*. 2010;185:623–31.
- Meuwissen TH, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*. 2001;157:1819–29.
- Ober U, Ayroles JF, Stone EA, Richards S, Zhu D, Gibbs RA, et al. Using whole-genome sequence data to predict quantitative trait phenotypes in *Drosophila melanogaster*. *PLoS Genet*. 2012;8:e1002685.

- Saatchi M, Garrick D. Across breed genomic predictions in beef cattle. In: Proceedings of the 10th world congress on genetics applied to livestock production: 17–22 August 2014, Vancouver; 2014.
- Schaeffer LR. Strategy for applying genome-wide selection in dairy cattle. *J Anim Breed Genet.* 2006;123:218–23.
- Spelman R, Ford C, McElhinney P, Gregory GC, Snell RG. Characterization of the *DGAT1* gene in the New Zealand dairy population. *J Dairy Sci.* 2002;85:3514–7.
- Thaller G, Krämer W, Winter A, Kaupe B, Erhardt G, Fries R. Effects of variants on milk production traits in German cattle breeds. *J Anim Sci.* 2003;81:1911–8.
- van Binsbergen R, Calus MP, Bink MC, van Eeuwijk FA, Schrooten C, Veerkamp RF. Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. *Genet Sel Evol.* 2015;47:71.
- van den Berg I, Boichard D, Guldbrandtsen B, Lund MS. Using sequence variants in linkage disequilibrium with causative mutations to improve across-breed prediction in Dairy cattle: a simulation study. *G3 (Besthesda).* 2016;6:2553–61.
- VanRaden PM. Efficient methods to compute genomic predictions. *J Dairy Sci.* 2008;91:4414–23.
- Veerkamp RF, Bouwman AC, Schrooten C, Calus MP. Genomic prediction using preselected DNA variants from a GWAS with whole-genome sequence data in Holstein-Friesian cattle. *Genet Sel Evol.* 2016;48:95.
- Wientjes YC, Bijma P, Veerkamp RF, Calus MP. An equation to predict the accuracy of genomic values by combining data from multiple traits, populations, or environments. *Genetics.* 2016;202:799–823.
- Wientjes YC, Veerkamp RF, Calus MP. Using selection index theory to estimate consistency of multi-locus linkage disequilibrium across populations. *BMC Genet.* 2015;16:87.
- Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM. Pitfalls of predicting complex traits from SNPs. *Nat Rev Genet.* 2013;14:507–15.
- Xu S. Theoretical basis of the Beavis effect. *Genetics.* 2003;165:2259–68.
- Yang J, Ferreira T, Morris AP, Medland SE, Madden PA, Heath AC, Martin NG, et al. Conditional and Joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet.* 2012;44:369–75.
- Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet.* 2011;88:76–82.
- Zhong S, Dekkers JC, Fernando RL, Jannink JL. Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. *Genetics.* 2009;182:355–64.
- Zhu Z, Bakshi A, Vinkhuyzen AA, Hemani G, Lee SH, Nolte IM, et al. Dominance genetic variation contributes little to the missing heritability for human complex traits. *Am J Hum Genet.* 2015;96:377–85.

3

CHAPTER 3.

Genomic prediction for numerically small breeds,
using models with pre-selected and differentially
weighted markers

Biaty Raymond^{1,2}, Aniek C. Bouwman¹, Yvonne C.J. Wientjes¹, Chris Schrooten³,
Jeanine Houwing-Duistermaat^{4,5} and Roel F. Veerkamp¹

¹Animal Breeding and Genomics, Wageningen University and Research, P.O. Box 338, 6700 AH Wageningen, The Netherlands; ²Biometris, Wageningen University and Research, 6700AA Wageningen, The Netherlands; ³CRV BV, P.O. Box 454, 6800 AL Arnhem, The Netherlands; ⁴Department of Medical Statistics and Bioinformatics, Leiden University Medical Centre, Leiden, The Netherlands; ⁵School of Mathematics, Faculty of Mathematics and Physical Sciences, University of Leeds, United Kingdom

Genetics Selection Evolution (2018) 50:49

Abstract

Background: Genomic prediction (GP) accuracy in numerically small breeds is limited by the small size of the reference population. Our objective was to test a multi-breed multiple genomic relationship matrices (GRM) GP model (MBMG) that weighs pre-selected markers separately, uses the remaining markers to explain the remaining genetic variance that can be explained by markers, and weighs information of breeds in the reference population by their genetic correlation with the validation breed.

Methods: Genotype and phenotype data were used on 595 Jersey bulls from New Zealand and 5,503 Holstein bulls from the Netherlands, all with deregressed proofs for stature. Different sets of markers were used, containing either pre-selected markers from a meta-genome-wide association analysis (GWAS) on stature, remaining markers or both. We implemented a multi-breed bivariate GREML model in which we fitted either a single multi-breed GRM (MBSG), or two distinct multi-breed GRM (MBMG), one made with pre-selected markers and the other with remaining markers. Accuracies of predicting stature for Jersey individuals using the multi-breed models (Holstein and Jersey combined reference population) was compared to those obtained using either the Jersey (within-breed) or Holstein (across-breed) reference population. All the models were subsequently fitted in the analysis of simulated phenotypes, with a simulated genetic correlation between breeds of 1, 0.5, and 0.25.

Results: The MBMG model always gave better prediction accuracies for stature compared to MBSG, within-, and across-breed GP models. For example, with MBSG, accuracies obtained by fitting 48,912 unselected markers (0.43), 357 pre-selected markers (0.38) or a combination of both (0.43), were lower than accuracies obtained by fitting pre-selected and unselected markers in separate GRM in MBMG (0.49). This improvement was further confirmed by results from a simulation study, with MBMG performing on average 23% better than MBSG with all markers fitted.

Conclusions: With the MBMG model, it is possible to use information from numerically large breeds to improve prediction accuracy of numerically small breeds. The superiority of MBMG is mainly due to its ability to use information on pre-selected markers, explain the remaining genetic variance and weigh information from a different breed by the genetic correlation between breeds.

Key words: multi-breed genomic prediction, marker pre-selection, cattle, stature, multiple genomic relationship matrices.

3.1 Background

The accuracy of genomic prediction (GP) depends on the size of the reference population. Therefore, accuracy of GP is limited in numerically small populations (e.g. Vanraden 2008; Goddard 2009). Potentially, information from numerically larger breeds can be used to predict genomic breeding values (GEBV) of animals in numerically small breeds, through a method called across-breed GP. However, in practice, it has been shown that across-breed GP does not result in significant improvement in prediction accuracy, as compared with within-breed GP (e.g. Hayes *et al.* 2009; Erbe *et al.* 2012; Kachman *et al.* 2013; Calus *et al.* 2014). In some cases, across-breed GP can result even in negative prediction accuracies (e.g. Hayes *et al.* 2009; Calus *et al.* 2014). One of the suggested reasons for poor prediction accuracies across breeds is that breeds differ in patterns of linkage disequilibrium (LD) between quantitative trait loci (QTL) and markers. Some studies suggested that increasing marker density, increases the probability that some of the markers are close to the QTL and have a consistent LD across breeds. Therefore, GP across breeds is expected to improve with increasing marker density (De Roos *et al.* 2008; Kizilkaya *et al.* 2010). In contrast to these expectations, Van Den Berg *et al.* (2016a) and Raymond *et al.* (2018) showed that simply increasing the density of markers up to whole-genome sequence does not improve accuracy of GP across breeds.

As an alternative to a simple increase in marker density, some studies suggested that across-breed prediction should be based on the QTL themselves, excluding potential non-causal markers (e.g. Boichard *et al.* 2012; Hoze *et al.* 2014). While the use of pre-selected markers or known QTL has been shown to significantly increase the accuracy of across-breed GP (Van Den Berg *et al.* 2016a), the magnitude of the increase in accuracy depends on the similarity of the effects of true QTL between the breeds (Wientjes *et al.* 2015a; Brown 2016). This similarity can be measured by the genetic correlation (r_g) between breeds (Fisher 1919; Falconer 1952), that can be estimated using a multi-breed genomic relationship matrix (GRM) (Wientjes *et al.* 2017; Wientjes *et al.* 2018). Accurately pinpointing QTL in the genome is often not trivial, and in most cases, the identified QTL explain only a fraction of the total genetic variance for the traits of interest.

Our hypothesis is that the accuracy of GP in numerically small breeds can be improved using a model that uses all available markers, but weighs markers that have a significant effect on the trait differently from other markers. The model must also account for the r_g between breeds in the reference and validation breeds, and

use it to weigh the contributions of the breeds in the reference population to predict the breeding values for individuals from the breed of interest (Wientjes *et al.* 2017). The objective of this study was to test a multi-breed multi-GRM model (MBMG) that weighs pre-selected significant markers separately from unselected markers, uses unselected markers to explain the remaining genetic variance that can be explained by markers (RGV_m), and weighs the information of breeds in the reference population by the genetic correlation between the reference and target breeds. To validate the performance of MBMG model in the analysis of real phenotypes, we also implemented the model for the analysis of simulated phenotypes.

3.2 Methods

3.2.1 Genotype data

For this study, we used data on 595 New Zealand Jersey bulls (NZJ) and 5,503 Dutch Holstein bulls (DH), all of which had genotypes for single nucleotide polymorphisms (SNPs) on the Illumina Bovine snp50 beadchip (Illumina Inc., San Diego, CA, USA) with 48,912 SNPs remaining after quality control. These SNPs had at least ten copies of the minor allele in each of the considered populations. In a combined dataset of the DH and NZJ bulls, minor allelic frequencies (MAF) ranged from 0.009 to 0.5. This SNP set will be referred to as 50k. We used two additional sets of SNPs that were pre-selected based on their significance for stature in a meta-genome-wide association analysis (GWAS) analysis using imputed whole-genome sequence (Bouwman *et al.* 2018). This meta-GWAS was performed on 17 populations from different countries, comprising eight breeds, including Holstein and Jersey. The DH population was included in the meta-GWAS, whereas the NZJ population was not included. The meta-GWAS across the populations identified 24,230 genome-wide significant ($p < 5 \times 10^{-8}$) SNPs in 163 distinct QTL regions spread across 27 autosomes. The first set of pre-selected SNPs were the so-called TOP SNPs, which showed the highest level of significance in each of the 163 QTL regions identified in the meta-GWAS analysis. These TOP SNPs accounted for 13.8% of the phenotypic variance for stature in the meta-GWAS. In our study, some of the TOP SNPs had either a very low MAF or segregated only in one of the two breeds and thus, after quality control, only 133 of these TOP SNPs remained. The second set of pre-selected SNPs were the COJO8 SNPs, which were those that showed independent significant effects ($p < 10^{-8}$) in the meta-GWAS study. We selected these SNPs using the conditional and joint effect analysis as described in Yang *et al.* (2012) and implemented in the GCTA software (Yang *et al.* 2011). From the 24,230 significant SNPs in the meta-GWAS, 357 COJO8 SNPs were selected. TOP and COJO8 SNPs were selected from imputed whole-

genome sequence data, and the 50k SNP set used for our analyses did not contain any TOP or COJO8 SNPs. Sixty-five SNPs overlapped between the COJO8 and TOP SNPs.

For the analysis using real phenotypes, 50k, TOP and COJO8 SNP sets were each fitted, one at a time, with a separate single GRM fitted in the prediction model. SNP sets were evaluated in terms of the estimated genetic parameters and prediction accuracy for stature. Furthermore, the TOP SNPs were combined with the 50k SNPs into a single set (50k+TOP), and the 50k+COJO8 SNP set was a combination of COJO8 and 50k SNPs. These combined sets; 50k+TOP and 50k+COJO8, were fitted via one GRM matrix and evaluated for the estimated genetic parameters and prediction accuracy for stature.

3.2.2 Empirical phenotypes

For the analysis of real phenotypes, estimated breeding values (EBV) for stature and the effective daughter contributions (EDC), which were the number of daughters on which the EBV were based, were available for all 595 NZJ and 5503 DH bulls. Data were provided by CRV BV (Cooperative Cattle Improvement Organization, Arnhem, the Netherlands). EBV and EDC for stature for all bulls were deregressed to obtain deregressed proofs (DRP) and deregressed effective daughter contributions (dEDC), respectively, according to Calus *et al.* (2016). To obtain DRP, the deregression procedure, called matrix deregression, corrects for the contribution of information on parents, which also includes information on sibs, to an individual's EBV. Similarly, dEDC were obtained by removing any EDC that are contributed by relatives in the dataset. The dEDC were used as weights for the DRP in subsequent analyses. Mean dEDC were equal to 52 for DH and 17 for NZJ. In the meta-GWAS (Bouwman *et al.* 2018), 3047 Jerseys from Australia were included. However, Australia and New Zealand do not have a joint genetic evaluation procedure for Jerseys, thus, the DRP of the NZJ bulls do not depend on information that went into the meta-GWAS.

3.2.3 Simulated phenotypes

To validate the performance of the MBMG model in terms of prediction accuracy and genetic parameter estimation, we also performed a simulation study. Phenotypes for all 595 NZJ and 5,503 DH were simulated, using the real genotypes for 49,045 SNPs (50k and TOP SNPs combined) in both breeds. A quantitative trait with a heritability (h^2) of 0.8 in both breeds was simulated. The TOP SNPs ($n = 133$) and 150 randomly selected SNPs were assumed to be causal in both breeds. Allele substitution effects of the causal SNPs were sampled from a bi-variate normal

distribution with a mean of 0, a variance of 1, and a correlation of 1, 0.5 and 0.25 between the breeds. Allele substitution effects were sampled independently from the allele frequency of causal SNPs. Therefore, the correlation between allele substitution effects was similar to the correlation between breeding values for performance in DH and NZJ, which is defined as the genetic correlation (r_g) between DH and NZJ. In both breeds, a true breeding value (TBV) for individual i was calculated as $\sum(x_{i,j} * a_j)$, where $x_{i,j}$ is the genotype of individual i at causal locus j (coded as 0, 1, 2), and a_j is the allele substitution effect of causal variant j . The corresponding phenotype was computed as $TBV_i + e_i$, where e_i is the residual effect of individual i , sampled from a standard normal distribution with a mean of 0 and a variance equal to $\sigma_{a_k}^2 * \left(\frac{1}{h^2} - 1\right)$, where $\sigma_{a_k}^2$ is the genetic variance of TBV for breed k . Simulation of phenotypes was carried out in R (R Core Team 2013) and was replicated 100 times.

For the analysis of simulated phenotypes, the following SNP sets were selected:

ALL (49,045 SNPs): we considered this set of SNPs to be the default set, which included 283 causal and 48,762 non-causal SNPs.

50k (48,912 SNPs): this set included the 150 randomly selected causal and 48,762 non-causal SNPs.

TOP (133 SNPs): this set consisted of the 133 causal SNPs that were the TOP SNPs in the empirical analysis, which were identified in the meta-GWAS analysis.

TOP+RN (283 SNPs): this set included 133 TOP SNPs and 150 additional randomly selected non-causal SNPs that represented random noise (RN).

CAUSAL (283 SNPs): this set consisted of all the 283 causal SNPs used in this study.

NON CAUSAL (48,762 SNPs): this set consisted of all non-causal SNPs.

3.2.4 Statistical models

For both the empirical analysis and simulation, multi-breed GRM were calculated for all SNP sets according to the method proposed by Wientjes et al. (Wientjes *et al.* 2017) as:

$$GRM = \begin{bmatrix} \frac{Z_{DH}Z'_{DH}}{\sum 2p_{DHj}(1-p_{DHj})} & \frac{Z_{DH}Z'_{NZJ}}{\sqrt{\sum 2p_{DHj}(1-p_{DHj})}\sqrt{\sum 2p_{NZJj}(1-p_{NZJj})}} \\ \frac{Z_{NZJ}Z'_{DH}}{\sqrt{\sum 2p_{DHj}(1-p_{DHj})}\sqrt{\sum 2p_{NZJj}(1-p_{NZJj})}} & \frac{Z_{NZJ}Z'_{NZJ}}{\sum 2p_{NZJj}(1-p_{NZJj})} \end{bmatrix},$$

where Z_{DH} and Z_{NZJ} are matrices containing centred genotype codes for all individuals from DH and NZJ populations, respectively, which were centred with breed-specific allele frequencies for all loci, p_{DHj} is the allele frequency for locus j in the DH population and p_{NZJj} is the allele frequency for locus j in the NZJ population. Two types of bivariate GREML models were implemented in the MTG2 (Lee and Van Der Werf 2016) software using the calculated GRM and phenotypes.

3.2.4.1 MBSG model

This model is a bivariate, multi-breed, single GRM (MBSG) model. The bivariate model considers the phenotypes of DH and NZJ for the same trait as those from two different correlated traits and uses one GRM combining all markers. The model was as follows:

$$\begin{bmatrix} y_{DH} \\ y_{NZJ} \end{bmatrix} = \begin{bmatrix} 1_{DH} & 0 \\ 0 & 1_{NZJ} \end{bmatrix} \begin{bmatrix} \mu_{DH} \\ \mu_{NZJ} \end{bmatrix} + \begin{bmatrix} W_{DH} & 0 \\ 0 & W_{NZJ} \end{bmatrix} \begin{bmatrix} g_{DH} \\ g_{NZJ} \end{bmatrix} + \begin{bmatrix} e_{DH} \\ e_{NZJ} \end{bmatrix},$$

where y is a vector of phenotypes, μ is the trait mean, W is an incidence matrix linking observations in y to genetic effects in g , and e is the residual. Genetic effects were assumed to be normally distributed as $\begin{bmatrix} g_{DH} \\ g_{NZJ} \end{bmatrix} \sim N(0, K \otimes \text{GRM})$ and $K = \begin{bmatrix} \sigma_{g_{DH}}^2 & \sigma_{g_{DH,NZJ}} \\ \sigma_{g_{DH,NZJ}} & \sigma_{g_{NZJ}}^2 \end{bmatrix}$, where $\sigma_{g_{DH}}^2$ and $\sigma_{g_{NZJ}}^2$ are genetic variances in DH and NZJ populations, respectively, and $\sigma_{g_{DH,NZJ}}$ is the genetic covariance between the breeds. The GRM fitted in this model were those computed based on SNPs in the 50k, TOP, COJO8, 50k+TOP and 50k+COJO8 sets for the empirical analysis. For the simulated scenarios, GRM based on 50k, TOP, TOP+RN, CAUSAL, NON CAUSAL and ALL sets were fitted.

3.2.4.2 MBMG model

This model is a bivariate, multi-breed, multi-GRM (MBMG) model and is an extension of MBSG, in which instead of a single multi-breed GRM, two separate multi-breed GRM (formed from two different marker sets) were fitted simultaneously. The model was as follows:

$$\begin{bmatrix} y_{DH} \\ y_{NZJ} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_{DH} \\ \mu_{NZJ} \end{bmatrix} + \begin{bmatrix} W_{1DH} & 0 \\ 0 & W_{1NZJ} \end{bmatrix} \begin{bmatrix} g_{1DH} \\ g_{1NZJ} \end{bmatrix} + \begin{bmatrix} W_{2DH} & 0 \\ 0 & W_{2NZJ} \end{bmatrix} \begin{bmatrix} g_{2DH} \\ g_{2NZJ} \end{bmatrix} + \begin{bmatrix} e_{DH} \\ e_{NZJ} \end{bmatrix},$$

where subscripts 1 and 2 represents the first and second GRM fitted in the model, respectively, W_1 and W_2 are identical incidence matrices linking observations in y to the two genetic effects, g_1 and g_2 . Here, genetic effects were also assumed to be normally distributed as:

$$\begin{bmatrix} g_{1DH} \\ g_{1NZJ} \end{bmatrix} \sim N(0, K_1 \otimes \text{GRM1}), \begin{bmatrix} g_{2DH} \\ g_{2NZJ} \end{bmatrix} \sim N(0, K_2 \otimes \text{GRM2}),$$

with $K_1 = \begin{bmatrix} \sigma_{g_{1DH}}^2 & \sigma_{g_{1DH},NZJ} \\ \sigma_{g_{1DH},NZJ} & \sigma_{g_{1NZJ}}^2 \end{bmatrix}$ and $K_2 = \begin{bmatrix} \sigma_{g_{2DH}}^2 & \sigma_{g_{2DH},NZJ} \\ \sigma_{g_{2DH},NZJ} & \sigma_{g_{2NZJ}}^2 \end{bmatrix}$.

For the empirical analysis, the sets of GRM fitted together in the MBMG model were 50k and TOP, 50k and COJO8. For the simulation analysis, the sets of GRM fitted together were TOP and 50k, TOP and NON CAUSAL, CAUSAL and NON CAUSAL.

For both the MBSG and MBMG models, residual terms were assumed to be normally distributed as $e \sim N(0, D\sigma_e^2)$ in the empirical analysis, where σ_e^2 is the residual variance and D is a diagonal matrix that contains the inverse of dEDC, which were used as weights for the DRP. The dEDC in D were used to scale the residual of the animals in the model, such that animals with a high dEDC have effectively a small residual. For the analysis using simulated phenotypes, we assumed $e \sim N(0, I\sigma_e^2)$ where I is an identity matrix.

For the multi-breed GP models described above, the reference population contained both DH and a subset ($n = 476$) of NZJ bulls. NZJ bulls were always used as validation individuals using 5-fold cross-validation, where the NZJ bulls were randomly split into five sets of 119 individuals. The same validation sets were used for both the analyses using real and simulated data. The GEBV of bulls in a particular validation set were estimated based on the model obtained from the rest of the data, i.e. phenotypes of all DH bulls and the remaining 476 NZJ bulls. This was repeated five times (once for each validation set). In all models, accuracy of prediction was computed as the correlation between the GEBV of NZJ bulls in each validation set and their TBV, which in the analysis using real data were approximated by the DRP. The mean correlation across the five validation sets was used as the prediction accuracy.

To assess the benefit of combining both breeds in the reference population compared to only a single breed in the reference population, we implemented alternative models to MBSG and MBMG in which only DH (across-breed) or only NZJ

(within-breed) bulls were included in the reference population. The single-breed equivalent to MBSG was:

$$y = 1\mu + Wg + e,$$

where y is a vector of phenotypes for either DH (across-breed), in which case the model is referred to as across-breed, single-GRM (ABSG), or NZJ (within-breed), in which case the model is referred to as within-breed, single-GRM (WBSG). All other model parameters are as described under MBSG. The single-breed equivalent to MBMG was:

$$y = 1\mu + W_1g_1 + W_2g_2 + e,$$

where y is a vector of phenotypes for either DH (across-breed), in which case the model is referred to as across-breed, multiple-GRM (ABMG), or NZJ (within-breed), in which case the model is referred to as within-breed, multiple-GRM (WBMG). All other model parameters are as described under MBMG.

The models ABSG, ABMG, WBSG, and WBMG were fitted using both real and simulated phenotypes and NZJ bulls were the validation candidates. In the analyses of real data, the calculated accuracies of prediction, which were the correlations between DRP and GEBV, were scaled by the mean accuracy of the DRP. The dEDC for all NZJ were converted into reliabilities (r^2) of DRP as $r^2 = \frac{dEDC}{dEDC + \frac{1}{h^2}}$, where h^2 , in this case, is the heritability of stature in NZJ. The mean r^2 for the validation candidates (NZJ) was equal to 0.76.

3.3 Results

3.3.1 Empirical phenotypes

3.3.1.1 Estimated genetic parameters

In the analysis using DRP for stature, we estimated h_{DRP}^2 for both DH and NZJ using the MBSG and MBMG models, which in this case reflect the proportion of the explained variance of DRP. These estimates are in Table 1. The estimated h_{DRP}^2 for DH and NZJ were similar between the bivariate multi-breed models and their equivalent univariate models (results not shown). Thus, estimates of h_{DRP}^2 obtained with the MBSG model were similar to those obtained with the ABSG model for DH, and WBSG for NZJ. Likewise, estimates of h_{DRP}^2 obtained with the MBMG model were similar to those obtained with ABMG for DH and WBSG for NZJ. In general, estimated h_{DRP}^2 were lower for NZJ than DH, which reflected the smaller number of NZJ individuals in the reference population and their lower mean dEDC (17). The mean of dEDC for the DH bulls was equal to 52.

Table 3.1 Estimated heritability (h^2_{DRP}) obtained by fitting different genomic relationship matrices (GRM) formed from different sets of markers and the corresponding genetic correlation (r_g) between Dutch Holstein (DH) and New Zealand Jersey (NZJ) breeds

GRM fitted	h^2_{DRP} DH	h^2_{DRP} NZJ	Estimated r_g (DH vs. NZJ)
MBSG			
50k	0.97 (0.00)	0.71 (0.02)	0.22 (0.17)
TOP	0.26 (0.01)	0.27 (0.01)	0.41 (0.15)
COJO8	0.75 (0.01)	0.38 (0.02)	0.44 (0.11)
50k+TOP	0.97 (0.00)	0.71 (0.02)	0.25 (0.17)
50k+COJO8	0.97 (0.00)	0.71 (0.03)	0.29 (0.17)
MBMG			
TOP and 50k	0.25 (0.01) and 0.71 (0.01)	0.11 (0.01) and 0.61 (0.02)	0.64 (0.17) and 0.16 (0.19)
COJO8 and 50k	0.26 (0.01) and 0.70 (0.01)	0.17 (0.02) and 0.57 (0.02)	0.88 (0.14) and 0.21 (0.19)

Standard error of estimates are given in parentheses. In the multi-breed, single-GRM model (MBSG), a single multi-breed GRM was fitted in a prediction model, while in the multi-breed, multiple-GRM model (MBMG), two separate multi-breed GRM, formed from different SNP sets, were fitted simultaneously in a prediction model.

When the GRM based on the TOP 133 SNPs was fitted in MBSG, estimated h_{DRP}^2 were low, i.e. 0.26 and 0.27 for DH and NZJ, respectively. However, when using the GRM based on the 357 COJO8 SNPs, estimated h_{DRP}^2 were significantly higher, i.e. 0.75 and 0.38 for DH and NZJ, respectively. Estimated h_{DRP}^2 were highest with the 50k GRM, i.e. 0.97 and 0.71 for DH and NZJ, respectively. h_{DRP}^2 did not increase further by adding either the TOP or COJO8 SNPs to the 50k set. Total h_{DRP}^2 estimated when fitting two separate GRM simultaneously in the MBMG model did not differ significantly from those obtained by fitting only the 50k GRM. For example, for DH, the TOP and COJO8 SNPs fitted as the first GRM resulted in estimated h_{DRP}^2 of 0.25 and 0.26, respectively, while the 50k SNPs fitted as the second GRM resulted in an h_{DRP}^2 of about 0.70.

Estimated r_g for stature between the DH and NZJ breeds, which can be interpreted as the correlation between the breeding values of individuals from the two breeds, are also in Table 1. With the MBSG model, in which a single GRM was fitted in a bivariate model, r_g were higher when only the pre-selected SNPs (TOP or COJO8) were fitted in the model compared to when the 50k or 50k plus pre-selected SNPs were fitted. However, given the large standard errors of the r_g , the differences in estimated r_g between the different GRM were not significant. With the MBMG model, estimated r_g resulting from the pre-selected SNP sets were significantly higher than those obtained with MBSG. For the TOP SNPs, r_g increased from 0.41 when fitted alone in MBSG to 0.64 when fitted together with the 50k SNPs in MBMG. For the COJO8 SNPs, r_g increased from 0.44 when fitted alone in the model, to 0.88 when fitted together with the 50k SNPs in MBMG. While the estimated r_g due to pre-selected SNPs increased when fitted simultaneously with the 50k SNPs in MBMG, the estimated r_g obtained for the 50k set did not change significantly, but tended to decrease compared to that obtained when the 50k SNPs were fitted alone in MBSG. In general, estimated r_g between breeds were much higher for the pre-selected SNPs than for the 50k or 50k plus pre-selected SNPs.

3.3.1.2 Accuracy of genomic breeding values

The accuracy of GEBV for the NZJ bulls using single or multi-breed reference populations is shown in Fig. 3.1. When using a single GRM and a NZJ reference population (WBSG model), accuracies of prediction were lower using the pre-selected marker sets (0.31 for the TOP and 0.34 for the COJO8 sets) than using the unselected 50k SNPs (0.43). Accuracy did not change by adding either the TOP or COJO8 SNPs to the 50k set. When using two separate GRM (WBMG model),

accuracies increased slightly compared to a single GRM, i.e. 0.46 when fitting the TOP and 50k SNPs in separate GRM, and 0.47 when fitting the COJO8 and 50k SNPs in separate GRM.

When using a single GRM and a DH reference population (ABSG model), the highest prediction accuracies were obtained using the pre-selected marker sets, i.e. 0.27 when the TOP SNPs were fitted and 0.25 when the COJO8 SNPs were fitted. With the unselected marker sets (50k, 50k+TOP, 50k+COJO8), accuracies were low (~0.14). When using two separate GRM (ABMG model), accuracies were lower when fitting TOP and 50k SNPs in separate GRM (0.21) compared to the model with only the TOP SNPs fitted (0.27). Fitting COJO8 and 50k SNPs in separate GRM improved prediction accuracy slightly (0.27) compared to fitting only the COJO8 SNPs (0.25).

Figure 3.1 Accuracy of predicting genomic breeding values (GEBV) for stature of New Zealand Jersey (NZJ) bulls from a reference population consisting of only NZJ or only Dutch Holstein (DH) or a combination of NZJ and DH bulls.

When using a single GRM and combined reference population of DH and NZJ bulls (MBSG model), prediction accuracies did not differ much from those obtained by using only NZJ in the reference population for the 50k, 50k+TOP and 50k+COJO8 sets. For the scenarios with pre-selected markers, accuracy of prediction increased

compared to when only the Jersey bulls were included in the reference population. For the TOP and COJO8 sets, prediction accuracies increased from 0.31 to 0.32 and from 0.34 to 0.38, respectively, with a multi-breed reference population (MBSG) compared to only NZJ bulls (WBSG). When using two separate GRM in the MBMG model, prediction accuracies were always higher compared to the model with only one GRM or to the model with only NZJ in the reference population. For example, TOP and 50k GRM fitted in the MBMG model resulted in a prediction accuracy of 0.47, and COJO8 and 50k GRM fitted in the MBMG model resulted in an accuracy of 0.49.

3.3.2 Simulated phenotypes

3.3.2.1 Estimated genetic parameters

In the simulation study, a quantitative trait with an h^2 of 0.8 was simulated in the DH and NZJ populations. The estimates of h^2 are in Table 3.2. In general, estimated h^2 were lower in NZJ than in DH, as was also observed for a real phenotype. With a simulated r_g of 1 for the trait in DH and NZJ, the unselected marker sets (50k, NON CAUSAL, ALL) captured almost the entire h^2 in DH, ranging from 0.77 to 0.8, while these values were slightly lower in NZJ ranging from 0.71 to 0.76.

Table 3.2 Estimated heritability (h^2) in Dutch Holstein (DH) and New Zealand Jersey (NZJ), and the estimated genetic correlation between breeds (r_g) estimated using the different genomic relationship matrices (GRM) and simulated phenotypes (100 replicates). Simulated heritability for the trait was 0.8 and simulated r_g between breed were 1, 0.5 and 0.25.

GRM fitted	h^2 DH	h^2 NZJ	r_g
Simulated $r_g = 1$			
MBSG (1 GRM fitted in a bivariate model)			
ALL	0.80 (0.01)	0.76 (0.06)	0.98 (0.19)
50k	0.79 (0.01)	0.76 (0.07)	0.67 (0.18)
TOP	0.44 (0.03)	0.38 (0.06)	0.85 (0.05)
TOP+RN	0.48 (0.03)	0.44 (0.05)	0.77 (0.07)
CAUSAL	0.80 (0.02)	0.76 (0.02)	1.00 (0.00)
NON CAUSAL	0.77 (0.01)	0.71 (0.20)	0.20 (0.20)
MBMG (2 separate GRMs fitted simultaneously in a bivariate model)			
TOP & 50k	0.37 (0.04) & 0.41 (0.04)	0.29 (0.05) & 0.48 (0.07)	1.01 (0.03) & 0.98 (0.21)
TOP & NON CAUSAL	0.37 (0.04) & 0.41 (0.03)	0.31 (0.05) & 0.47 (0.07)	1.02 (0.03) & 0.19 (0.22)
CAUSAL & NON CAUSAL	0.80 (0.02) & 0.00 (0.00)	0.76 (0.04) & 0.00 (0.02)	1.01 (0.01) & NA
Simulated $r_g = 0.5$			
MBSG (1 GRM fitted in a bivariate model)			
ALL	0.80 (0.01)	0.80 (0.06)	0.50 (0.18)
50k	0.79 (0.01)	0.79 (0.06)	0.34 (0.18)
TOP	0.44 (0.03)	0.39 (0.05)	0.43 (0.12)
TOP+RN	0.48 (0.03)	0.44 (0.05)	0.41 (0.11)

CAUSAL	0.80 (0.01)	0.80 (0.02)	0.51 (0.05)
NON CAUSAL	0.78 (0.01)	0.79 (0.06)	0.11 (0.18)
MBMG (2 separate GRMs fitted simultaneously in a bivariate model)			
TOP & 50k	0.37 (0.03) & 0.43 (0.04)	0.31 (0.05) & 0.49 (0.07)	0.52 (0.11) & 0.53 (0.21)
TOP & NON CAUSAL	0.38 (0.03) & 0.41 (0.03)	0.31 (0.05) & 0.48 (0.07)	0.52 (0.12) & 0.12 (0.21)
CAUSAL & NON CAUSAL	0.80 (0.01) & 0.00 (0.00)	0.80 (0.02) & 0.00 (0.02)	0.51 (0.05) & NA
Simulated $r_g' = 0.25$			
MBSG (1 GRM fitted in a bivariate model)			
ALL	0.80 (0.01)	0.80 (0.06)	0.27 (0.19)
50k	0.79 (0.01)	0.79 (0.06)	0.18 (0.18)
TOP	0.44 (0.03)	0.39 (0.05)	0.22 (0.13)
TOP+RN	0.48 (0.03)	0.44 (0.05)	0.22 (0.12)
CAUSAL	0.80 (0.01)	0.80 (0.02)	0.26 (0.06)
NON CAUSAL	0.78 (0.01)	0.76 (0.06)	0.06 (0.18)
MBMG (2 separate GRMs fitted simultaneously in a bivariate model)			
TOP & 50k	0.37 (0.03) & 0.43 (0.04)	0.31 (0.05) & 0.49 (0.07)	0.27 (0.14) & 0.29 (0.21)
TOP & NON CAUSAL	0.38 (0.03) & 0.41 (0.03)	0.31 (0.05) & 0.48 (0.07)	0.27 (0.13) & 0.08 (0.21)
CAUSAL & NON CAUSAL	0.80 (0.01) & 0.00 (0.00)	0.80 (0.02) & 0.00 (0.03)	0.26 (0.06) & NA

The TOP SNPs resulted in an estimated h^2 of 0.44 in DH and 0.38 in NZJ. Adding 150 randomly chosen non-causal SNPs, representing random noise, to the TOP (TOP+RN) set resulted in increased h^2 , i.e. 0.48 in DH and 0.44 in NZJ. When the TOP SNPs were fitted together with either the 50k or NON CAUSAL sets in MBMG, estimated h^2 using the TOP SNPs decreased to 0.37 in DH and ~ 0.30 in NZJ. The CAUSAL SNPs resulted in an estimated h^2 of 0.8 in DH, both when fitted alone or together with NON CAUSAL SNPs. The NON CAUSAL SNPs resulted in an h^2 of 0.76, when fitted with the CAUSAL SNPs. In general, there was no significant difference in the percentage of h^2 captured between the scenarios with a simulated r_g of 1 and those with a simulated r_g of 0.5 or 0.25 (Table 3.2).

Estimated r_g using simulated phenotypes with a simulated r_g of 1 are also in Table 3.2. The estimated r_g obtained by fitting the CAUSAL GRM in MBSG was equal to 1. When all the SNPs (CAUSAL + NON CAUSAL) were fitted in MBSG, the estimated r_g was equal to 0.98 but decreased to 0.67 when only the 50k SNPs was fitted. When the TOP SNPs were fitted, the estimated r_g was equal to 0.85 and decreased to 0.77 when some 150 random noise were added to the TOP SNPs (TOP+RN). In general, estimates of r_g were more precise when the causal markers were isolated from the non-causal markers in a separate GRM (MBMG model), but also when all the causal markers were present in the GRM. In general, estimated r_g were proportional to the simulated r_g (Table 3.2).

3.3.2.2 Accuracy of genomic breeding values

3.3.2.2.1 Simulated genetic correlation of 1 between breeds

In the WBSG model, the unselected marker sets (50k, NON CAUSAL, and ALL) resulted in accuracies of ~ 0.60 (Fig. 3.2). The TOP and TOP+RN sets also gave accuracies of ~ 0.60 , although they consisted only of a fraction of the SNPs in the unselected marker sets. Except when fitting the CAUSAL SNPs in within-breed, single-GRM model, accuracies were significantly higher when two separate GRM were fitted simultaneously rather than those in which only a single GRM was fitted.

In the ABSG model, in which a DH reference population was used to predict NZJ GEBV, accuracies were much lower than those obtained in the within breed single-GRM model for the unselected marker sets. However, accuracies were only slightly lower in ABSG for the TOP and TOP+RN SNP sets than those obtained in the within breed single-GRM model. The only scenarios in which across-breed prediction outperformed within-breed prediction were the scenarios in which the CAUSAL SNPs

were fitted in ABSG and the CAUSAL and NON CAUSAL SNPs were fitted simultaneously in ABMG, which reflects the larger reference population of DH than of NZJ and an r_g of 1. In general, across-breed prediction accuracies were by far higher when the pre-selected marker sets were used for prediction compared to when the unselected marker sets (50k, NON CAUSAL, and ALL) were used. Except when fitting the CAUSAL SNPs in ABSG, across-breed prediction accuracies were significantly higher with the ABMG model than with the ABSG model.

Figure 3.2 Accuracy with standard errors of predicting the genomic breeding values (GEBV) of Jersey bulls from a reference population made of only Jersey or only Holstein or a combination of Holstein and Jersey bulls. Simulated genetic correlation between breed was 1.

Combining DH and NZJ bulls in the reference population to obtain NZJ GEBV (multi-breed prediction) resulted in slightly better prediction accuracies than when only NZJ

were used in the reference population. As also observed with empirical phenotypes, the highest accuracies were obtained in MBMG, except in the unrealistic case that all causal SNPs could be identified with 100% accuracy and were all fitted in the MBSG model.

3.3.2.2.2 Simulated genetic correlation of 0.5 and 0.25 between breeds

Figures 3.3 and 3.4 show the prediction accuracies of the simulation with a r_g of 0.5 and 0.25 respectively, between DH and NZJ. In general, prediction accuracies of GEBV for the NZJ bulls were lower with a simulated r_g between DH and NZJ of 0.5 and 0.25 than with an r_g of 1, especially when only DH were included in the reference population. In almost all cases, accuracies were higher in scenarios in which two separate GRM were fitted simultaneously than those in which only a single GRM was fitted. One exception was when the CAUSAL GRM was fitted alone. In across-breed prediction, when only the DH reference population was used to obtain GEBV for NZJ, the accuracy of prediction was about 50% lower when the simulated r_g was 0.5 instead of 1 and 75% lower when the simulated r_g was 0.25 instead of 1. With a simulated r_g of 0.5 and 0.25, combining DH and NZJ in the reference population did not improve prediction accuracy compared to a reference population including only NZJ individuals, although the combined reference population was more than 10 times larger than the NZJ reference population.

Figure 3.3 Accuracy with standard error of predicting the genomic breeding values (GEBV) of Jersey bulls from a reference population made of only Jersey or only Holstein or a combination of Holstein and Jersey bulls. Simulated genetic correlation between breed was 0.5.

Figure 3.4 Accuracy with standard error of predicting the genomic breeding values (GEBV) of Jersey bulls from a reference population made of only Jersey or only Holstein or a combination of Holstein and Jersey bulls. Simulated genetic correlation between breed was 0.25.

3.4 Discussion

The objective of this study was to test the performance of a multi-breed, multi-GRM (MBMG) model in terms of estimated genetic parameters and accuracy of predicting GEBV of individuals from a numerically small population. The results of our analyses using both real and simulated phenotypes demonstrated the superiority of the MBMG model in terms of prediction accuracy in comparison with a single GRM fitted in within-, across-, and multi-breed prediction models. The MBMG model also outperformed a within- (WBMG) and across-breed (ABMG) GP model with two

separate GRM fitted simultaneously. Three main distinguishing properties of MBMG are: (1) prioritisation of pre-selected functional markers over other types of markers, (2) use of the unselected markers to account for RGV_m in such a way that they do not dilute the effect of the pre-selected functional markers, and (3) appropriate weighting of information provided by each breed in the reference population by their r_g with the validation breed.

Prioritisation of known QTL or pre-selected functional markers over other types of markers for the estimation of breeding value not an entirely new concept, and was already the main idea behind marker-assisted selection (MAS) (Dekkers and Hospital 2002; Meuwissen *et al.* 2016). With MAS, pre-selected significant markers from a genome-wide association study (GWAS) are used to predict breeding values for a certain trait. One major pitfall of MAS for complex traits was that the pre-selected significant markers accounted for only a small fraction of the total genetic variance for the traits of interest (Meuwissen *et al.* 2016). As a result, there was little to no gain in using (only) MAS over phenotypic selection (Dekkers and Hospital 2002), and therefore most applications have focussed on GP using many markers across the genome (Meuwissen *et al.* 2001). Even in this genomic selection era, a number of studies have advocated the prioritization of causal regions in GP models, especially as a result of the increasing availability and interest in the use of whole-genome sequence (WGS) data in breeding (e.g. Brøndum *et al.* 2015; Van Den Berg *et al.* 2016b). For example, Brøndum *et al.* (2015) showed that when significant SNPs from a GWAS using WGS data are added to a low-density SNP panel (50k), and used for GP, accuracy of within-breed prediction can be increased. Also, Van Den Berg *et al.* (2016b) showed that adding SNPs that are near QTL peaks in a GWAS to a 50k panel improved the accuracy of within- and multi-breed GP as compared to the use of a 50k panel only. In the absence of additional markers, using only pre-selected markers has been shown to result in little or no advantage for within-breed GP (Veerkamp *et al.* 2016), which was confirmed in this study. In general, when we fitted only the pre-selected (TOP and COJO8) SNPs in a GP model and using real phenotypes, accuracies of within-breed GP were lower compared to when we fitted the unselected SNP sets (50k, 50k+TOP, 50k+COJO8) (Fig. 1). One reason could be that, for within-breed GP in which the training set is the same breed as the validation set, the unselected markers can still explain part of the genetic variance due to LD between markers and QTL that are not yet identified (Table 3.1). Since long-range LD is present between markers and QTL within breed, it is less important to identify the true causal markers for within-breed GP.

Across breeds, the LD phase between markers and QTL is not necessarily consistent (e.g. De Roos *et al.* 2008; Wientjes *et al.* 2015b). Consequently, using non-causal markers that do not tag the QTL through LD, may result in poor across-breed GP accuracies, as was observed in this study for the unselected SNP sets and by others (e.g. Hayes *et al.* 2009; Erbe *et al.* 2012; Calus *et al.* 2014; Saatchi and Garrick 2014). This is mainly because the estimated effect of a non-causal marker is proportional to the extent of its LD with a QTL (Zhu *et al.* 2015). Thus, if LD phase between the non-causal marker and QTL is different between breeds, then the non-causal marker would have different estimated effects in different breeds, which is not accounted for in the prediction model. For across-breed GP, it is important to identify the QTL or markers close to QTL and use them for prediction (Wientjes *et al.* 2015b). Our results from the analyses of real and simulated phenotypes showed that when markers are pre-selected based on their causal effects and used for across-breed GP, the accuracy of prediction is significantly higher than when all markers, including non-causal or unselected markers are used. The obtained accuracies of across-breed GP agree with those of Saatchi and Garrick (2014) who found that across-breed genomic prediction resulted in accuracies close to 0, except for some traits due to the segregation of common large-effect QTL with conserved linkage phase among the different breeds.

Pre-selection of markers based on their potential causal effect is not trivial (Andersson 2001). This is especially the case when pre-selection of potential causal markers is carried out for traits with a complex genetic architecture within a single population using GWAS. In those situations, the power of GWAS methods to precisely identify the causative mutations for the traits of interest is hindered by the high relatedness between individuals, strong and long range LD between QTL and many non-causal markers within the genome, many QTL with very small effect sizes, low allele frequency of the QTL, and small sample size (Korte and Farlow 2013). In the current study, we benefitted from the results of a large international meta-GWAS (Bouwman *et al.* 2018), in which some of these limiting factors, such as small sample size, use of a single population, confounding effect of LD and family relationships were overcome by the use of multiple populations consisting of eight cattle breeds from different countries. Nevertheless, it is not guaranteed that all pre-selected SNPs are truly causal, but the ones that are not false positives are at least in high LD with and located close to the true causal loci. The large power in the meta-GWAS to identify the potential causal markers that we used in our study can probably explain the improved prediction accuracies in the across- and multi-breed GP scenarios using pre-selected markers.

Another possible pitfall of pre-selection of markers based on the results of GWAS is the likely event of false positive findings. To investigate the effect of including random noise on prediction accuracy, our simulation study included a scenario (TOP+RN) in which 150 false positive SNPs were added to 133 causal SNPs. For across-, and multi-breed GP, our results showed very little to no effect of adding the random noise (53% of the SNPs) to the causal SNPs on prediction accuracy. When the non-causal markers outweigh very much the number of true causal markers in a set of markers, as was the case in the 50k scenario, accuracy of prediction becomes considerably lower than when using only the true causal markers. Wang *et al.* (2018) showed that reducing the number of non-causal markers from 12,000 to 1000 that were added to a set of 10 true causal markers resulted in higher prediction accuracy. In practise, a much lower proportion of false positive markers can be assumed, e.g. a 5% false discovery rate is commonly used. Our results showed that it is unlikely that pre-selected false positive markers (less than 53% of all pre-selected markers) will have a significant negative effect on prediction accuracy.

In practice, pre-selected markers from a GWAS using whole-genome sequence data may not be found on the traditional 50k SNP chip. In order to use the pre-selected SNPs in routine genomic evaluations with the MBMG model, the pre-selected SNPs may have to be imputed in the populations of interest. However, some pre-selected SNPs, especially those with a low MAF, could be imputed with relatively low accuracy, which could, in turn, negatively impact the performance of the MBMG model. Still, the impact will most likely be small, given that, in general, SNPs with a low MAF explain only a small proportion of the genetic variance. One way that the imputation error of markers could be accounted for in GP is to use dosage scores rather than genotype calls, to calculate the GRM. A more practical solution for the breeding industry is to add known QTL or pre-selected SNPs to custom chips, as this will result in more accurate genotypes.

The MBMG model uses the unselected markers to explain RGV_m that is not explained by the pre-selected markers. Thus, instead of fitting only the pre-selected markers, we fitted both pre-selected and unselected markers in two separate GRM, by using the multi-breed reference population. In this way, the model benefits from improved tagging of QTL by the pre-selected markers across breeds and, at the same time, benefits from additional genetic variance explained by the unselected markers within each breed. If all the QTL that underlie a trait are identified with 100% accuracy, there would be no need to include other markers in the prediction model,

since the QTL could explain 100% of the genetic variance (the CAUSAL scenarios in our simulations). However, in practice it is unlikely to identify all the QTL for complex traits. For example, the top SNPs in the 163 QTL identified in the large meta-GWAS for stature explained only 13.8% of phenotypic variance for the trait (Bouwman *et al.* 2018). This shows the importance of including multiple GRM in the MBMG model to maximise the proportion of total genetic variance explained within breed.

Appropriate weighting of the information provided by each breed in the reference population by their r_g with the validation breed is an important feature of the MBMG model. Multi-breed GP can be implemented by simply pooling the information of different breeds in a univariate GP model, effectively assuming a r_g of 1 between the breeds in the reference population and between the reference and validation breeds. Assuming a r_g of 1 between breeds, especially between distantly related breeds such as Holstein and Jersey, can result in little to no advantage or even poorer prediction accuracies, as compared to a within-breed prediction model (Olson *et al.* 2012). The MBMG model is a multi-trait model in which the phenotypes from different breeds are treated as those from different but correlated traits (e.g. Karoui *et al.* 2012; Olson *et al.* 2012; Wientjes *et al.* 2016). The expectation is that multi-trait modelling only uses information from another breed that will improve accuracy or, in the worst case, will not affect the accuracy, but it will not decrease the accuracy either (Wientjes *et al.* 2016).

In the traditional multi-trait models used for multi-breed GP (MBSG), an implicit assumption that is made is that r_g is the same across the entire genome. Due to differences in genetic background and breeding practices between breeds, traits might have evolved differently (Thaller *et al.* 2003). In such cases, it is expected that marker effects, including the effect of causal markers, have different covariance structures in the different breeds. In MBMG, we assumed that the r_g between breeds at the causal markers can be different from the r_g between breeds at the unselected markers. We believe that the assumption of different r_g for different regions in the genome is one of the factors that make MBMG superior to MBSG in both our empirical and simulation analyses (Figs. 3.1, 3.2, and 3.3).

In addition to accounting for the RGV_m , fitting a second GRM made from the unselected markers in MBMG allowed for a more precise estimation of the r_g between breeds at the pre-selected markers compared to fitting only the pre-selected markers in MBSG (Tables 1 and 2). For example, in the scenario with a

simulated r_g of 1 between DH and NZJ, the estimated r_g at the TOP SNPs was 0.88 when only the TOP SNPs were fitted in the model (MBSG). However, by fitting the 50k GRM as a separate component in MBMG, the estimated r_g at the TOP SNPs was exactly 1, as simulated. The difference in the estimate of r_g at the pre-selected markers between MBSG and MBMG can probably be explained as follows: when the pre-selected markers are fitted alone in MBSG, they also explain RGV_m due to long-range LD within the breeds. However, when the pre-selected markers are fitted together with other markers in separate GRM (MBMG), the pre-selected markers account for the genetic variance that is mainly due to the pre-selected markers, and the RGV_m is explained by other markers in the second GRM. The covariance between breeds (the numerator) at the pre-selected markers remains similar between both models (MBSG and MBMG), since genetic covariance depends only on the conserved LD phase between breeds, which does not change between models (results not shown). Thus, the inflated genetic variance within breeds (the denominators) explained by the pre-selected markers in MBSG results in a lower estimated r_g at the pre-selected markers using MBSG than using MBMG. However, even in MBSG, we observed higher estimates of r_g at the pre-selected markers than at the unselected markers. This is because the precision of estimating r_g between breeds using markers depends on how accurately the marker-based relationships describes the relationships at the causal markers (Wientjes *et al.* 2018). Hence, the estimated r_g is more precise for causal markers than for non-causal markers.

The MBMG model follows a similar principle since it can be implemented in Bayesian models, for example, in the BayesRC method (Macleod *et al.* 2016) or the so-called multitask Bayesian learning model (Chen *et al.* 2014; Calus *et al.* 2018). Similar to MBMG, BayesRC incorporates prior biological knowledge in the prediction model by assigning different markers into different classes, for instance, a pre-selected marker class and unselected marker class, or even more classes. One of the differences between MBMG and BayesRC is that in MBMG, we assumed a priori that the effects of all markers in a certain class follow a normal distribution whereas in BayesRC, for each class of markers, there is a mixture of four normal distributions, which allows for pre-selected markers to have a large effect, an average effect, a small effect, or no effect (false positive). In addition to allowing for different effect sizes of markers and QTL as in BayesRC, the multitask Bayesian variable selection model for multi-breed GP allows for markers associated with a breed specific QTL to have a large effect in one breed and a small effect in another breed. Calus *et al.* (2018) showed that, compared to a multi-trait GREML model (MBSG) that assumes a single r_g across

the genome, the multitask Bayesian variable selection model had similar accuracies of prediction for traits with a relatively high estimated r_g between Holstein and Jersey, but outperformed the MBSG model for traits with a very low estimated r_g between Holstein and Jersey. Our expectation is that Bayesian models will be more suited than MBMG for prediction of traits that are influenced by a few QTL with large effects and many additional smaller QTL, such as milk fat percentage with the *DGAT1* QTL. Otherwise, when traits are influenced by many QTL with small effects, we expect the MBMG and the Bayesian models to perform similarly. This is mainly because, in such case, the assumption of the GREML model that the effect sizes of QTL or markers linked to QTL follow a normal distribution holds.

3.5 Conclusions

We presented a multi-breed multi-GRM model, MBMG, for genomic selection in numerically small breeds. The key features of MBMG are: (1) appropriate use of information on known QTL or pre-selected markers; (2) explanation of RGV_m by the unselected markers; and (3) use of the genetic correlation between reference and validation breeds to weigh the contribution of each breed in the reference population to the accuracy of prediction. Our proposed MBMG model can be applied to any case in which some form of prior information on markers exist, e.g., genome annotation information, experimental evidence, etc. By using this model, we are able to better combine information from numerically small and large breeds to improve prediction accuracy for the numerically small breed using both pre-selected and unselected markers. The MBMG model, together with the availability of whole-genome sequence data and accurate marker pre-selection strategies, can result in more accurate genomic prediction in numerically small breeds and thus improve their rate of genetic gain.

3.6 Acknowledgements

This study was financially supported by NWO-TTW, the Dutch Ministry of Economic Affairs (TKI Agri & Food project 16022, code BO-22.04-011-001-ASG-LR), and the Breed4Food partners Cobb Europe, CRV, Hendrix Genetics and Topigs Norsvin. The use of the HPC cluster was made possible by CAT-AgroFood (Shared Research Facilities, Wageningen University and Research).

3.7 References

- Andersson L. Genetic dissection of phenotypic diversity in farm animals. *Nat Rev Genet.* 2001;2:130–8.
- Boichard D, Guillaume F, Baur A, Croiseau P, Rossignol MN, et al. Genomic selection in French dairy cattle. *Anim Prod Sci.* 2012;52:115–20.
- Bouwman AC, Daetwyler HD, Chamberlain AJ, Ponce CH, Sargolzaei M, Schenkel FS, et al. Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. *Nat Genet.* 2018;50:362–7.
- Brøndum RF, Su G, Janss L, Sahana G, Guldbrandtsen B, Boichard D, et al. Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. *J Dairy Sci.* 2015;98:4107–16.
- Brown BC, Asian Genetic Epidemiology Network Type 2 Diabetes Consortium, Ye CJ, Price AL, Zaitlen N. Transethnic genetic-correlation estimates from summary statistics. *Am J Hum Genet.* 2016;99:76–88.
- Calus MPL, Goddard ME, Wientjes YCJ, Bowman PJ, Hayes BJ. Multibreed genomic prediction using multitrait genomic residual maximum likelihood and multitask Bayesian variable selection. *J Dairy Sci.* 2018;101:4279–94.
- Calus MPL, Vandenplas J, Ten Napel J, Veerkamp RF. Validation of simultaneous deregression of cow and bull breeding values and derivation of appropriate weights. *J Dairy Sci.* 2016;99:6403–19.
- Calus MP, Huang H, Vereijken A, Visscher J, ten Napel J, Windig JJ. Genomic prediction based on data from three layer lines: a comparison between linear methods. *Genet Sel Evol.* 2014;46:57.
- Chen L, Li C, Miller S, Schenkel F. Multi-population genomic prediction using a multi-task Bayesian learning model. *BMC Genet.* 2014;15:53.
- de Roos AP, Hayes BJ, Spelman RJ, Goddard ME. Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. *Genetics.* 2008;179:1503–1.
- Dekkers JCM, Hospital F. The use of molecular genetics in the improvement of agricultural populations. *Nat Rev Genet.* 2002;3:22–32.
- Erbe M, Hayes BJ, Matukumalli LK, Goswami S, Bowman PJ, Reich CM, et al. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *J Dairy Sci.* 2012;95:4114–29.
- Falconer DS. The problem of environment and selection. *Am Nat.* 1952;86:293–8.
- Fisher RAXV. The correlation between relatives on the supposition of Mendelian inheritance. *Earth Environ Sci Trans R Soc Edinb.* 1919;52:399–433.
- Goddard M. Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica.* 2009;136:245–57.

- Hayes BJ, Bowman PJ, Chamberlain AC, Verbyla K, Goddard ME. Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genet Sel Evol.* 2009;41:51.
- Hoze C, Fritz S, Phocas F, Boichard D, Ducrocq V, Croiseau P. Genomic evaluation using combined reference populations from Montbéliarde and French Simmental breeds. In: *Proceedings of the 10th world congress on genetics applied to livestock production: 17–22 August 2014, Vancouver; 2014 pp. 17–22.*
- Kachman SD, Spangler ML, Bennett GL, Hanford KJ, Kuehn LA, Snelling WM, et al. Comparison of molecular breeding values based on within- and across-breed training in beef cattle. *Genet Sel Evol.* 2013;45:30.
- Karoui S, Carabano MJ, Diaz C, Legarra A. Joint genomic evaluation of French dairy cattle breeds using multiple-trait models. *Genet Sel Evol.* 2012;44:39.
- Kizilkaya K, Fernando RL, Garrick DJ. Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotype. *J Anim Sci.* 2010;88:544–51.
- Korte A, Farlow A. The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods.* 2013;9:29.
- Lee SH, van der Werf JH. MTG2: An efficient algorithm for multivariate linear mixed model analysis based on genomic information. *Bioinformatics.* 2016;32:1420–2.
- MacLeod I, Bowman P, Vander Jagt C, Haile-Mariam M, Kemper K, Chamberlain AJ, et al. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. *BMC Genomics.* 2016;17:144.
- Meuwissen T, Hayes B, Goddard M. Genomic selection: A paradigm shift in animal breeding. *Anim Front.* 2016;6:6–14.
- Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. *Genetics.* 2001;157:1819–29.
- Olson KM, VanRaden PM, Tooker ME. Multibreed genomic evaluations using purebred Holsteins, Jerseys, and Brown Swiss. *J Dairy Sci.* 2012;95:5378–83.
- Raymond B, Bouwman AC, Schrooten C, Houwing-Duistermaat J, Veerkamp RF. Utility of whole-genome sequence data for across-breed genomic prediction. *Genet Sel Evol.* 2018;50:27.
- Saatchi M, Garrick D. Across breed genomic predictions in beef cattle. In: *Proceedings of the 10th World Congress on Genetics Applied to Livestock Production: 17–22 August 2014, Vancouver; 2014.*
- R Development Core Team. R: A language and environment for statistical computing. Vienna: the R Foundation for Statistical Computing; 2013.
- Thaller G, Krämer W, Winter A, Kaup B, Erhardt G, Fries R. Effects of *DGAT1* variants on milk production traits in German cattle breeds. *J Anim Sci.* 2003;81:1911–8.
- van den Berg I, Boichard D, Guldbrandtsen B, Lund MS. Using sequence variants in linkage disequilibrium with causative mutations to improve across-breed

- prediction in dairy cattle: A simulation study. *G3* (Bethesda). 2016a;6:2553-61.
- van den Berg I, Boichard D, Lund MS. Sequence variants selected from a multi-breed GWAS can improve the reliability of genomic predictions in dairy cattle. *Genet Sel Evol*. 2016b;48:83.
- VanRaden PM. Efficient methods to compute genomic predictions. *J Dairy Sci*. 2008;91:4414–23.
- Veerkamp RF, Bouwman AC, Schrooten C, Calus MP. Genomic prediction using preselected DNA variants from a GWAS with whole-genome sequence data in Holstein-Friesian cattle. *Genet Sel Evol*. 2016;48:95.
- Wang J, Zhou Z, Zhang Z, Li H, Liu D, et al. Expanding the BLUP alphabet for genomic prediction adaptable to the genetic architectures of complex traits. *Heredity* (Edinb). 2018; in press.
- Wientjes YC, Bijma P, Vandenplas J, Calus MP. Multi-population genomic relationships for estimating current genetic variances within and genetic correlations between populations. *Genetics*. 2017;207:503–15.
- Wientjes, Y.C.J., Calus, M.P.L., Duenk, P. *et al*. Required properties for markers used to calculate unbiased estimates of the genetic correlation between populations. *Genet Sel Evol*. 2018; 50:65.
- Wientjes YC, Veerkamp RF, Bijma P, Bovenhuis H, Schrooten C, Calus MP. Empirical and deterministic accuracies of across-population genomic prediction. *Genet Sel Evol*. 2015;47:5.
- Wientjes YC, Veerkamp RF, Calus MP. Using selection index theory to estimate consistency of multi-locus linkage disequilibrium across populations. *BMC Genet*. 2015;16:87.
- Wientjes YC, Bijma P, Veerkamp RF, Calus MP. An equation to predict the accuracy of genomic values by combining data from multiple traits, populations, or environments. *Genetics*. 2016;202:799–823.
- Yang J, Ferreira T, Morris AP, Medland SE, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet*. 2012;44:369-75.
- Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet*. 2011;88:76–82.
- Zhu Z, Bakshi A, Vinkhuyzen AA, Hemani G, Lee SH, Nolte IM, et al. Dominance genetic variation contributes little to the missing heritability for human complex traits. *Am J Hum Genet*. 2015;96:377–85.

4

CHAPTER 4.

A deterministic equation to predict the accuracy
of multi-population genomic prediction with multiple
genomic relationship matrices

Biaty Raymond^{1,2}, Aniek C. Bouwman¹, Yvonne C.J. Wientjes¹, Chris Schrooten³ and
Roel F. Veerkamp¹

¹Animal Breeding and Genomics, Wageningen University and Research, P.O. Box 338, 6700 AH Wageningen, The Netherlands; ²Biometris, Wageningen University and Research, 6700AA Wageningen, The Netherlands; ³CRV BV, P.O. Box 454, 6800 AL Arnhem, The Netherlands.

Genetics Selection Evolution (2020) 52:21

Abstract

Background: A multi-population genomic prediction (GP) model in which important pre-selected single nucleotide polymorphisms (SNPs) are differentially weighted (MPMG) has been shown to result in better prediction accuracy than a multi-population, single genomic relationship matrix (GRM) GP model (MPSG) in which all SNPs are weighted equally. Our objective was to underpin theoretically the advantages and limits of the MPMG model over the MPSG model, by deriving and validating a deterministic prediction equation for its accuracy.

Methods: Using selection index theory, we derived an equation to predict the accuracy of estimated total genomic values of selection candidates from population A ($r_{EGV_{A_T}}$), when individuals from two populations, A and B , are combined in the training population and two GRM, made respectively from pre-selected and remaining SNPs, are fitted simultaneously in MPMG. We used simulations to validate the prediction equation in scenarios that differed in the level of genetic correlation between populations, heritability, and proportion of genetic variance explained by the pre-selected SNPs. Empirical accuracy of the MPMG model in each scenario was calculated and compared to the predicted accuracy from the equation.

Results: In general, the derived prediction equation resulted in accurate predictions of $r_{EGV_{A_T}}$ for the scenarios evaluated. Using the prediction equation, we showed that an important advantage of the MPMG model over the MPSG model is its ability to benefit from the small number of independent chromosome segments (M_e) due to the pre-selected SNPs, both within and across populations, whereas for the MPSG model, there is only a single value for M_e , calculated based on all SNPs, which is very large. However, this advantage is dependent on the pre-selected SNPs that explain some proportion of the total genetic variance for the trait.

Conclusions: We developed an equation that gives insight into why, and under which conditions the MPMG outperforms the MPSG model for GP. The equation can be used as a deterministic tool to assess the potential benefit of combining information from different populations, e.g., different breeds or lines for GP in livestock or plants, or different groups of people based on their ethnic background for prediction of disease risk scores.

Key words: multi-population genomic prediction, deterministic prediction equation, multiple genomic relationship matrices, accuracy, selection index.

4.1 Background

Genomic prediction (GP) (Meuwissen *et al.* 2001) has become widely accepted and has been successfully implemented in both animal and plant breeding schemes (e.g. Dalton 2009; Hayes *et al.* 2009a; De Roos *et al.* 2011; Heffner *et al.* 2011). However, for accurate GP it is essential that the training population is large (e.g. Daetwyler *et al.* 2008; Vanraden *et al.* 2009; Zhong *et al.* 2009). For numerically small populations, e.g., numerically small breeds or lines in livestock or numerically small human ethnic groups, it is difficult or impossible to assemble a large enough training population that can accurately predict the genomic values. Therefore, the accuracy of GP in numerically small populations is limited (Thomasen *et al.* 2012).

A potential option to increase the accuracy of GP in numerically small populations is to use a large training population made up of individuals from multiple populations, including the target population, a method known as multi-population GP. Results from dairy cattle indicate that this approach can lead to substantial increases in the accuracy of GP for numerically small breeds, if the training population is made up of individuals from different but closely-related breeds that have recently had substantial exchanges of genetic material, and that a large number of individuals from the additional breed is included (Zhou *et al.* 2014). However, in cases in which distantly related breeds were combined in a single training population, increases in the accuracy of multi-population GP were limited compared to that of within-population GP (e.g. Hayes *et al.* 2009b; Pryce *et al.* 2011; Erbe *et al.* 2012).

Different statistical approaches have been adopted for multi-population GP. One approach, and the most straightforward, is the univariate single-trait approach in which individuals from multiple populations are pooled and treated as individuals from the same population in a training population. The underlying assumption of this approach is that the genetic correlation between the populations is equal to 1. Deviations from this assumption, for example, when distantly related populations are combined, can result in prediction accuracies that are even lower than the accuracy of within-population GP (Hayes *et al.* 2009b). Another approach used for multi-population GP is to consider phenotypes of individuals from different populations, e.g. phenotypes from different, but correlated traits (e.g. Karoui *et al.* 2012; Olson *et al.* 2012). The advantage of this multi-trait approach is that it can consider that the genetic correlation between populations can be less than 1. In the worst-case scenario, the accuracy of multi-population GP using a multi-trait

approach is expected to be the same as the accuracy of within-population GP, but not lower (Wientjes *et al.* 2016).

In both the single-trait and the multi-trait approaches for multi-population GP, DNA markers such as single nucleotide polymorphisms (SNPs) are equally weighted in the model. However, some studies have shown that accuracy of GP can be improved by prioritising certain SNPs that have a significant effect on the trait or by incorporating prior biological knowledge on SNPs in the prediction model (e.g. Brøndum *et al.* 2015; Macleod *et al.* 2016; Van Den Berg *et al.* 2016). Based on that rationale, and to improve the potential to use information from other populations, Raymond *et al.* (2018) proposed the so-called multi-breed, multiple genomic relationship matrices (GRM) GP model (MBMG), which in this study is generalised as MPMG, given that the model can be applied in other species, e.g., plant or humans. The three key features of this model are: (1) SNPs are pre-selected based on prior knowledge of potential causal effects and are used to build a GRM; (2) the remaining unselected SNPs are used to make a separate GRM, in order to explain the residual genetic variance for the trait; and (3) information of each population in the training population is weighed by their genetic correlation with the other populations and with the selection candidates. The MPMG model is equivalent to a model with a single GRM, in which different weights are applied to the two classes of SNPs. Using both real and simulated data, Raymond *et al.* (2018) showed that the MPMG model can result in significant increases in the accuracy of GP, as compared with a multi-trait approach in which all SNPs are pooled together in a single GRM (MPSG). Given the superior performance of the MPMG model over MPSG, the objective of this study was to underpin theoretically the advantages and limits of the MPMG model as compared to the MPSG model, by deriving and validating a deterministic prediction equation for the accuracy.

4.2 Methods

4.2.1 Multi-population, multiple genomic relationship matrices (MPMG) model

We assume that individuals from populations *A* and *B* are combined in the training population to predict the genomic value of selection candidates from population *A* using the MPMG model following Raymond *et al.* (2018). The MPMG model is a bivariate model that considers the phenotypes of individuals from populations *A* and *B* for the same trait as those from two different, but correlated traits. The prior biological knowledge that exists about the effect sizes of the SNPs is used to pre-select important SNPs that are used to build one GRM. The remaining SNPs are used

to build a second GRM to explain the residual genetic variance not explained by the pre-selected SNPs. Both GRM are fitted simultaneously in the bivariate model (Raymond *et al.* 2018). The model can be specified as:

$$\begin{bmatrix} y_A \\ y_B \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix} + \begin{bmatrix} W_{1A} & 0 \\ 0 & W_{1B} \end{bmatrix} \begin{bmatrix} EGV_{1A} \\ EGV_{1B} \end{bmatrix} + \begin{bmatrix} W_{2A} & 0 \\ 0 & W_{2B} \end{bmatrix} \begin{bmatrix} EGV_{2A} \\ EGV_{2B} \end{bmatrix} + \begin{bmatrix} e_A \\ e_B \end{bmatrix}, \quad (4.1)$$

where subscripts 1 and 2 represent the first and second GRM fitted in the model and subscripts A and B represent the populations A and B . y_A is a vector of phenotypes for individuals from population A and y_B is a vector of phenotypes for individuals from population B , μ is the trait mean, W_1 and W_2 are incidence matrices linking phenotypes to the two estimated genomic values, EGV_1 and EGV_2 , and e is the residual. Estimated genomic values are assumed to be normally distributed as:

$$\begin{bmatrix} EGV_{1A} \\ EGV_{1B} \end{bmatrix} \sim N(0, K_1 \otimes GRM_1), \text{ and } \begin{bmatrix} EGV_{2A} \\ EGV_{2B} \end{bmatrix} \sim N(0, K_2 \otimes GRM_2),$$

$$\text{with } K_1 = \begin{bmatrix} \sigma_{gA_1}^2 & \sigma_{gA,B_1} \\ \sigma_{gA,B_1} & \sigma_{gB_1}^2 \end{bmatrix} \text{ and } K_2 = \begin{bmatrix} \sigma_{gA_2}^2 & \sigma_{gA,B_2} \\ \sigma_{gA,B_2} & \sigma_{gB_2}^2 \end{bmatrix},$$

where σ_{gA}^2 and σ_{gB}^2 are genetic variances in populations A and B , respectively, and $\sigma_{gA,B}$ is the genetic covariance between the populations. The multi-population GRM fitted in the MPMG model are calculated according to Wientjes *et al.* (2017).

4.2.1 Theory

In the following derivation, the main interest is to predict the accuracy of the estimated total genomic value of selection candidates from population A ($r_{EGV_{AT}}$). $r_{EGV_{AT}}$ is a product of the accuracy of estimating SNP effects (r_{SNP}) and the square root of the proportion of total genetic variance explained by SNPs (ρ) (Dekkers 2007; Daetwyler *et al.* 2008; Goddard 2009). As a foundation, we will first derive an equation to predict $r_{EGV_{AT}}$ for within-population GP using two GRM, made from two separate sets of SNPs, that are fitted simultaneously in a GREML model. Subsequently, we will derive the full equation to predict $r_{EGV_{AT}}$ when individuals from two populations, A and B , are combined in the training population and two separate GRM are fitted simultaneously in a GREML model (MPMG). For the derivations, we will use selection index theory (Hazel 1943), and build upon works from Daetwyler *et al.* (2008) and Wientjes *et al.* (2016), who have done similar

derivations for within- and multi-population GP models in which all SNPs are equally weighted in a single GRM.

4.2.1.1 Accuracy of within-population genomic prediction with two separate GRM (WPMG)

The within-population, multiple GRM model (WPMG) can be represented as:

$$y_A = 1\mu_A + W_{1A}EGV_{1A} + W_{2A}EGV_{2A} + e_A. \quad (4.2)$$

With the WPMG model, two different sets of estimated genomic values are obtained for the validation candidates. These are EGV_{A_1} , due to GRM_1 , and EGV_{A_2} , due to GRM_2 . Both estimates of genomic values can be combined as sources of information in a selection index approach to obtain EGV_{A_T} as follows:

$$EGV_{A_T} = b_{A_1}EGV_{A_1} + b_{A_2}EGV_{A_2}, \quad (4.3)$$

where b_{A_1} and b_{A_2} are weighting factors for the two sources of information. In the context of selection index theory, the breeding goal (H) is TGV_A and the index trait (I) is EGV_{A_T} . The optimum values of b_{A_1} and b_{A_2} can be obtained as:

$$\begin{bmatrix} b_{A_1} \\ b_{A_2} \end{bmatrix} = P^{-1}g \quad (4.4)$$

where P is the (co)variance matrix of information sources EGV_{A_1} and EGV_{A_2} , and g is a vector containing the covariances between information sources EGV_{A_1} and EGV_{A_2} and the true genomic value (TGV_A). Thus:

$$P = \begin{bmatrix} var(EGV_{A_1}) & cov(EGV_{A_1}, EGV_{A_2}) \\ cov(EGV_{A_1}, EGV_{A_2}) & var(EGV_{A_2}) \end{bmatrix}, \quad (4.5)$$

and

$$g = \begin{bmatrix} cov(EGV_{A_1}, TGV_A) \\ cov(EGV_{A_2}, TGV_A) \end{bmatrix}. \quad (4.6)$$

For simplicity of the derivation, we will assume that TGV_A are scaled such that they have a variance of 1. Therefore, the variance of the estimated genomic values can be calculated as the reliability (r^2) of the estimated genomic values. Thus,

$$var(EGV_{A_1}) = r_{EGV_{A_1}}^2 \text{ and } var(EGV_{A_2}) = r_{EGV_{A_2}}^2. \quad (4.7)$$

We assume that there is no covariance between EGV_{A_1} and EGV_{A_2} , since the expectation is zero when both GRM are jointly fitted (Yang *et al.* 2011). Thus, when GRM are fitted simultaneously in a GREML model, only partial variances are explained by each of the GRM such that the sum of the variances explained by the two GRM equals the total genetic variance for the trait that can be explained by all SNPs. The partial variances explained by SNPs in each GRM can be viewed as partial regression coefficients in a multiple regression scenario.

$$\text{Hence, } P = \begin{bmatrix} r_{EGV_{A_1}}^2 & 0 \\ 0 & r_{EGV_{A_2}}^2 \end{bmatrix}.$$

The first element of the vector g is:

$$\begin{aligned} & cov(EGV_{A_1}, TGV_A) \\ &= cor(EGV_{A_1}, TGV_A) \sqrt{var(EGV_{A_1}) * var(TGV_A)} \\ &= (r_{EGV_{A_1}}) \sqrt{r_{EGV_{A_1}}^2 * 1} = r_{EGV_{A_1}}^2 \end{aligned} \quad (4.8)$$

Similarly, $cov(EGV_{A_2}, TGV_A) = r_{EGV_{A_2}}^2$.

$$\text{Therefore, } g = \begin{bmatrix} r_{EGV_{A_1}}^2 \\ r_{EGV_{A_2}}^2 \end{bmatrix}. \quad (4.9)$$

The accuracy of selection index, representing the accuracy of EGV_{A_T} is the correlation between the index values and the breeding goal (r_{IH}). Thus, $r_{IH} = \frac{cov_{IH}}{\sigma_I \sigma_H}$. As explained in Falconer and MacKay (Falconer and Mackay 1996) the selection index is constructed such that one unit of the index is equivalent to one unit of the

breeding goal. In other words, the selection index is constructed such that the regression of the breeding goal on the index (b_{HI}) is 1, resulting in the expression $cov_{IH} = \sigma_I^2$. Thus,

$$\begin{aligned}
 r_{IH} &= r_{EGV_{AT}} = \frac{\sigma_I}{\sigma_H} \\
 &= \sqrt{\frac{b'g}{var(H)}} = \sqrt{\frac{g'P^{-1}g}{var(TGV_A)}} = \sqrt{g'P^{-1}g} \\
 &= \sqrt{\begin{bmatrix} r_{EGV_{A_1}}^2 & r_{EGV_{A_2}}^2 \end{bmatrix} \begin{bmatrix} r_{EGV_{A_1}}^2 & 0 \\ 0 & r_{EGV_{A_2}}^2 \end{bmatrix}^{-1} \begin{bmatrix} r_{EGV_{A_1}}^2 \\ r_{EGV_{A_2}}^2 \end{bmatrix}} \\
 &= \sqrt{\begin{bmatrix} r_{EGV_{A_1}}^2 & r_{EGV_{A_2}}^2 \end{bmatrix} \begin{bmatrix} \frac{r_{EGV_{A_2}}^2}{r_{EGV_{A_1}}^2 r_{EGV_{A_2}}^2} & 0 \\ 0 & \frac{r_{EGV_{A_1}}^2}{r_{EGV_{A_1}}^2 r_{EGV_{A_2}}^2} \end{bmatrix} \begin{bmatrix} r_{EGV_{A_1}}^2 \\ r_{EGV_{A_2}}^2 \end{bmatrix}} \\
 &= \sqrt{r_{EGV_{A_1}}^2 + r_{EGV_{A_2}}^2}.
 \end{aligned}
 \tag{4.10}$$

The accuracies of estimated genomic values $r_{EGV_{A_1}}$ and $r_{EGV_{A_2}}$ can be calculated as $\rho_{A_1} r_{SNP_{A_1}}$ and $\rho_{A_2} r_{SNP_{A_2}}$ respectively, where $r_{SNP_{A_1}}$ and $r_{SNP_{A_2}}$ are the accuracies of estimated SNP effects in population A for SNPs in GRM₁ and GRM₂ respectively, ρ_{A_1} and ρ_{A_2} are the square root of the proportions of genetic variance explained in the validation population A by SNPs in GRM₁ and GRM₂, respectively. The accuracies of estimated SNP effects $r_{SNP_{A_1}}$ and $r_{SNP_{A_2}}$ can be deterministically predicted by Daetwyler's equation (Daetwyler *et al.* 2008). Hence, Eq. (10) can be written as:

$$\begin{aligned}
 r_{EGV_{AT}} &= \sqrt{\rho_{A_1}^2 r_{SNP_{A_1}}^2 + \rho_{A_2}^2 r_{SNP_{A_2}}^2} \\
 &= \sqrt{\rho_{A_1}^2 \frac{h_A^2 N_A}{h_A^2 N_A + M_{e_{A_1}}} + \rho_{A_2}^2 \frac{h_A^2 N_A}{h_A^2 N_A + M_{e_{A_2}}}},
 \end{aligned}
 \tag{4.11}$$

where h_A^2 is the heritability of the trait in population A , N_A is the number of individuals from population A in the training population, $M_{e_{A_1}}$ and $M_{e_{A_2}}$ are the effective number of chromosome segments segregating in population A , based on variation in GRM_1 and GRM_2 , respectively. The M_e values represent the effective number of effects that are estimated in the model. The values for M_e can be calculated as the inverse of the variance of within-population GRM (Wientjes *et al.* 2015; Lee *et al.* 2017a).

4.2.1.2 Accuracy of multi-population GP with two separate GRM (MPMG model)

In the case of multi-population GP model with two GRM fitted simultaneously, we assume that individuals from two populations A and B are combined in a training population to estimate the total genomic values for validation candidates from population A (EGV_{A_T}). To estimate the accuracy of EGV_{A_T} from this model, we estimate the accuracy of a selection index in which EGV for the selection candidates are combined from two different models having either population A or B as training population. The first model is a WPMG model using individuals from population A in the training population ($\text{EGV}_{A_{1,A}}$ and $\text{EGV}_{A_{2,A}}$). The second model is an across-population model with two GRM using individuals from population B in the training population ($\text{EGV}_{A_{1,B}}$ and $\text{EGV}_{A_{2,B}}$). The selection index was as follows:

$$\text{EGV}_{A_T} = b_{A_{1,A}}\text{EGV}_{A_{1,A}} + b_{A_{1,B}}\text{EGV}_{A_{1,B}} + b_{A_{2,A}}\text{EGV}_{A_{2,A}} + b_{A_{2,B}}\text{EGV}_{A_{2,B}}. \quad (4.12)$$

The (co)variance matrix of information sources is given as:

$$P = \begin{bmatrix} \text{var}(\text{EGV}_{A_1,A}) & \text{cov}(\text{EGV}_{A_1,A}, \text{EGV}_{A_1,B}) & 0 & 0 \\ \text{cov}(\text{EGV}_{A_1,A}, \text{EGV}_{A_1,B}) & \text{var}(\text{EGV}_{A_1,B}) & 0 & 0 \\ 0 & 0 & \text{var}(\text{EGV}_{A_2,A}) & \text{cov}(\text{EGV}_{A_2,A}, \text{EGV}_{A_2,B}) \\ 0 & 0 & \text{cov}(\text{EGV}_{A_2,A}, \text{EGV}_{A_2,B}) & \text{var}(\text{EGV}_{A_2,B}) \end{bmatrix} \quad (4.13)$$

Again, we assume that the covariances between EGV based on GRM₁ and GRM₂ are zero. Given also that we assume a variance of 1 for TG_{V_A}:

$$\begin{aligned} \text{var}(\text{EGV}_{A_1,A}) &= r_{\text{EGV}_{A_1,A}}^2, \text{var}(\text{EGV}_{A_1,B}) = r_{\text{EGV}_{A_1,B}}^2, \text{var}(\text{EGV}_{A_2,A}) = r_{\text{EGV}_{A_2,A}}^2, \text{ and} \\ \text{var}(\text{EGV}_{A_2,B}) &= r_{\text{EGV}_{A_2,B}}^2. \end{aligned}$$

$\text{cov}(\text{EGV}_{A_1,A}, \text{EGV}_{A_1,B})$ is the covariance of the EGV_{A₁} calculated using SNP effects estimated in the reference populations A and B, respectively. Following the derivation in Appendix A1, we show that:

$$\begin{aligned} \text{cov}(\text{EGV}_{A_1,A}, \text{EGV}_{A_1,B}) &= r_{\text{EGV}_{A_1,A}}^2 r_{\text{EGV}_{A_1,B}}^2 \\ \text{and } \text{cov}(\text{EGV}_{A_2,A}, \text{EGV}_{A_2,B}) &= r_{\text{EGV}_{A_2,A}}^2 r_{\text{EGV}_{A_2,B}}^2. \end{aligned} \quad \text{Therefore,}$$

$$P = \begin{bmatrix} r_{EGV_{A_1,A}}^2 & r_{EGV_{A_1,A}}^2 r_{EGV_{A_1,B}}^2 & 0 & 0 \\ r_{EGV_{A_1,A}}^2 r_{EGV_{A_1,B}}^2 & r_{EGV_{A_1,B}}^2 & r_{EGV_{A_2,A}}^2 & r_{EGV_{A_2,A}}^2 r_{EGV_{A_2,B}}^2 \\ 0 & 0 & r_{EGV_{A_2,A}}^2 & r_{EGV_{A_2,A}}^2 r_{EGV_{A_2,B}}^2 \\ 0 & 0 & r_{EGV_{A_2,A}}^2 r_{EGV_{A_2,B}}^2 & r_{EGV_{A_2,B}}^2 \end{bmatrix}. \quad (4.14)$$

Following Eq. (9), g can be written as:

$$g = \begin{bmatrix} r_{EGV_{A_1,A}}^2 \\ r_{EGV_{A_1,B}}^2 \\ r_{EGV_{A_2,A}}^2 \\ r_{EGV_{A_2,B}}^2 \end{bmatrix}. \quad (4.15)$$

The accuracy of the index, representing r_{g,A_T} can be calculated as:

$$r_{EGV_{A_T}} = \sqrt{\frac{g'P^{-1}g}{var(TGVA)}} = \sqrt{g'P^{-1}g}.$$

With some algebra (see Appendix A2), we show that the equation $\sqrt{g'P^{-1}g}$ for the MPMG model can be represented as:

$$\begin{aligned}
 & \rho_{A_1}^2 \left(\frac{h_A^2}{M_{eA_1}} \left(\frac{h_B^2}{M_{eAB_1}} + \frac{1}{N_B} \right) + r_g^2 \frac{h_B^2}{M_{eAB_1}} \left(\frac{h_A^2}{M_{eA_1}} + \frac{1}{N_A} \right) - 2 \left(\frac{h_A^2}{M_{eA_1}} \right) \left(r_g^2 \frac{h_B^2}{M_{eAB_1}} \right) \right) \\
 & + \rho_{A_2}^2 \left(\frac{h_A^2}{M_{eA_2}} \left(\frac{h_B^2}{M_{eAB_2}} + \frac{1}{N_B} \right) + r_g^2 \frac{h_B^2}{M_{eAB_2}} \left(\frac{h_A^2}{M_{eA_2}} + \frac{1}{N_A} \right) - 2 \left(\frac{h_A^2}{M_{eA_2}} \right) \left(r_g^2 \frac{h_B^2}{M_{eAB_2}} \right) \right) ,
 \end{aligned}
 \tag{4.16}$$

which in matrix form can be represented as:

$$\begin{bmatrix} \rho_{A_1} \sqrt{\frac{h_A^2}{M_{eA_1}}} \\ \rho_{A_1} r_g \sqrt{\frac{h_B^2}{M_{eAB_1}}} \\ \rho_{A_2} \sqrt{\frac{h_A^2}{M_{eA_2}}} \\ \rho_{A_2} r_g \sqrt{\frac{h_B^2}{M_{eAB_2}}} \end{bmatrix}^T \begin{bmatrix} \frac{h^2_A}{M_{eA_1}} + \frac{1}{N_A} & r_g \sqrt{\frac{h_A^2 h_B^2}{M_{eA_1} M_{eAB_1}}} & 0 & 0 \\ r_g \sqrt{\frac{h_A^2 h_B^2}{M_{eA_1} M_{eAB_1}}} & \frac{h^2_A}{M_{eA_2}} + \frac{1}{N_A} & 0 & 0 \\ 0 & 0 & \frac{h^2_B}{M_{eA_2}} + \frac{1}{N_B} & r_g \sqrt{\frac{h_A^2 h_B^2}{M_{eA_2} M_{eAB_2}}} \\ 0 & 0 & r_g \sqrt{\frac{h_A^2 h_B^2}{M_{eA_2} M_{eAB_2}}} & \frac{h^2_B}{M_{eAB_2}} + \frac{1}{N_B} \end{bmatrix}^{-1} \begin{bmatrix} \frac{h_A^2}{M_{eA_1}} \\ \frac{h_B^2}{M_{eAB_1}} \\ \frac{h_A^2}{M_{eA_2}} \\ \frac{h_B^2}{M_{eAB_2}} \end{bmatrix} .
 \tag{4.17}$$

The input parameters are:

h_A^2 = heritability of the trait in population A

h_B^2 = heritability of the trait in population B

N_A = number of individuals from population A in the training population

N_B = number of individuals from population B in the training population

ρ_{A_1} = square root of the proportion of genetic variance explained in the validation population A by GRM₁
 ρ_{A_2} = square root of the proportion of genetic variance explained in the validation population A by GRM₂
 r_g = genetic correlation between populations A and B
 M_{eA_1} = number of effective chromosome segments in population A based on variation in GRM₁
 M_{eA_2} = number of effective chromosome segments in population A based on variation in GRM₂
 M_{eAB_1} = number of effective chromosome segments across populations A and B based on variation in GRM₁
 M_{eAB_2} = number of effective chromosome segments across populations A and B based on variation in GRM₂.

For predicting $r_{\text{EGV}_{A_1}}$, Eq. (4.17) reduces to Eq. (4.11) for within-population GP when r_g between populations A and B is 0. The values for M_{eAB} can be calculated as the inverse of the variance of the across-population block of multi-population GRM (Wientjes *et al.* 2015; Wientjes *et al.* 2016).

4.2.2 Validation of prediction equations using simulations

The aim of this section was to use simulations to validate Eqs. 4.11 (WPMG) and 4.17 (MPMG) in scenarios that differed in the proportion of causal SNPs that are pre-selected and fitted in the models. Consequently, the scenarios differed in the proportion of total genetic variance explained by SNPs in each of the two GRM fitted simultaneously in the models. The scenarios also differed in heritability of the trait. For the MPMG model, the scenarios also differed in the level of genetic correlation between the populations A and B . Genotype data of two existing cattle populations were used in combination with simulated phenotypes. As validation, we compared the empirical accuracies in each simulated scenario to the accuracy obtained using the derived prediction equations.

4.2.2.1 Genotype data

Genotypes for SNPs on the Illumina Bovinesnp50 (Illumina Inc., San Diego, CA, USA) with 48,912 SNPs after quality control, were available on 595 New Zealand Jersey bulls and 5553 Dutch Holstein bulls. These SNPs had at least ten copies of the minor allele in a combined Dutch Holstein and New Zealand Jersey population, with a minor allele frequency (MAF) ranging from 0.009 to 0.5. Hereafter, we will refer to the New Zealand Jerseys as population A and to the Dutch Holsteins as population B.

4.2.2.2 Simulation of phenotypes

Phenotypes for all individuals were simulated using their real genotypes and assuming an additive model. From the 48,912 SNPs, 500 were randomly selected to be causal SNPs in both populations. Allele substitution effect of the causal SNPs (a) were sampled from a bi-variate normal distribution with a mean of 0, variance of 1, and a correlation of 0.8, 0.6 and 0.4 between populations A and B. Since allele substitution effects were sampled independently from their allele frequency, the correlation between allele substitution effect represents the correlation between genomic values of individuals from populations A and B, which is referred to as the genetic correlation between populations (r_g). Within each population, TGV for individual i were calculated as $\sum(x_{i,j} * a_j)$, where $x_{i,j}$ is the genotype of individual i at causal SNP j (coded as 0, 1, 2) and a_j is the allele substitution effect of causal SNP j . The corresponding phenotype was computed as $TGV_i + e_i$, where e_i is the residual effect of individual i , sampled from a standard normal distribution with a mean of 0 and a variance equal to $\sigma_{g_k}^2 * \left(\frac{1}{h^2} - 1\right)$, where $\sigma_{g_k}^2$ is the variance of TGV in population k . For each population, the residual effects were sampled from a separate normal distribution. The heritability (h^2) was set to 0.3 and 0.8 in each population. Simulation of phenotypes was carried out in R (R Core Team 2013) and was replicated 100 times.

4.2.2.3 Genomic prediction

The WPMG (only 476 individuals from population A in the training population) and MPMG (476 individuals from population A and 5,553 individuals from population B in the training population) models were implemented in the software MTG2 (Lee and Van Der Werf 2016). We used three levels for the proportion of causal SNPs that are identified, pre-selected and used to create the first GRM. The number of causal SNPs underlying the simulated trait was always 500. The levels are:

CSNP_125: this level represents a situation in which a quarter of the causal SNPs are identified, pre-selected and used to create the first GRM.

CSNP_250: this level represents a situation in which half of the causal SNPs are identified, pre-selected and used to create the first GRM.

CSNP_500: this level represents the extreme situation in which all 500 causal SNPs are identified, pre-selected and used to create the first GRM.

At all levels, the remaining SNPs that were not used as causal were used to create the second GRM. The level above were evaluated under varying degrees of genetic correlation between populations (0.4, 0.6 and 0.8) and heritability in both populations (0.3 and 0.8). Throughout the study, individuals from population *A* were used as the validation candidates in a 5-fold cross-validation scheme, where individuals from population *A* were randomly split into five sets of 119 individuals each. The GRM fitted in the MPMG model were constructed according to Wientjes *et al.* (Wientjes *et al.* 2017), considering population-specific allele frequencies. The empirical accuracies of prediction at all levels were computed as the correlation between the EGV and the simulated TGV of validation candidates.

To compare how accurate the accuracy of the MPMG model can be predicted compared to a MPSG model, we also fitted a multi-population, single GRM (MPSG) model, with the multi-population GRM made from non-causal SNPs in addition to either the CSNP_125, CSNP_250 or CSNP_500 SNPs. To predict the accuracy of the MPSG model, we used the derived prediction Eq. (4.17) in which $\rho_{A_2}^2$ is set to 0, given that the model has only one GRM. The value used for $\rho_{A_1}^2$ was empirically estimated as the ratio of empirical accuracy of the WPSG model and the predicted accuracy using the formula of Daetwyler *et al.* (2008), assuming all the variance is captured by SNPs. Setting $\rho_{A_2}^2$ to 0 reduces Eq. (17) to the equation derived by Wientjes *et al.* (2016).

4.2.2.4 Values of input parameters for the prediction equations

For the prediction of accuracy using Eqs. (4.11) and (4.17), we used the simulated values as input for the parameters h_A^2 , h_B^2 , and r_g . The values used for $\rho_{A_1}^2$ were 0.25, 0.5 and 1 in the CSNP_125, CSNP_250 and CSNP_500 levels, respectively. This is because a quarter of the causal SNPs explains, on average, a quarter of the total genetic variance for the trait, which was confirmed empirically. An empirical approach was used to determine the appropriate values for the input parameter $\rho_{A_2}^2$. This parameter represents the proportion of total genetic variance of the trait in the

validation population explained by the non-causal SNPs (GRM_2) in the training population. We determined empirically that the non-causal SNPs could only explain 66% of the total genetic variance of the trait in the validation individuals of population A using a within-population model with one GRM including all non-causal SNPs. We did this by taking a ratio of the empirical accuracy obtained from cross-validation and the predicted accuracy using Daetwyler's equation (Daetwyler *et al.* 2008), assuming that 100% of the total genetic variance for the trait is captured by SNPs. Thus, the values for $\rho_{A_2}^2$ as used in the prediction equation were 0.66×0.75 , 0.66×0.5 and 0 , in the CSNP_125, CSNP_250 and CSNP_500 levels, respectively, with the values 0.75 , 0.5 and 0 representing the proportion of total genetic variance unexplained by the causal SNPs (GRM_1) in the CSNP_125, CSNP_250 and CSNP_500 levels, respectively. Throughout the study, M_e within a population was calculated according to Lee *et al.* (Lee *et al.* 2017a) as the inverse of the variance of the within-population GRM, while M_e across populations was calculated as the inverse of the variance of the across-population block terms of the multi-population GRM (Wientjes *et al.* 2015; Wientjes *et al.* 2016).

4.2.3 Potential accuracies of different models in relation to different levels of r_g , ρ_{A_1} , and ME_{AB_2}

The aim of this section was to identify the situations under which the MPMG model will outperform all other models tested in terms of prediction accuracy. We evaluated the potential accuracy of predicting the genomic values of selection candidates from a numerically small population A under three cases. These are hypothetical cases that aim to mimic real life situations in dairy cattle breeding programs.

Case 1

For the first case, we assume that, in addition to individuals from the target population A ($N_A = 476$), individuals from a larger but different population B ($N_B = 5,553$) are available to be included in the training population, mimicking the real sample sizes of the Jersey and Holstein populations used in this study. We investigated the effect of the level of genetic correlation between populations on the accuracy of prediction. The following additional assumptions were made: M_e within population A based on pre-selected SNPs (calculated from real genotype data) = 159; M_e across populations A and B based on pre-selected SNPs (calculated from real genotype data) = 280; M_e within population A based on remaining SNPs (calculated from real genotype data) = 463; M_e across populations A and B based on

remaining SNPs (calculated from real genotype data) = 32,970; heritability of the trait: 0.3 in both populations; ρ^2 due to 500 pre-selected causal SNPs = 0.4, ρ^2 due to all SNPs = 0.8 (assuming the remaining non-causal SNPs capture 66% of the residual genetic variance).

Case 2

Many genome-wide association studies have been carried out in livestock with the aim to identify the causal variants underlying complex traits. The variants that were discovered explain varying proportions of the genetic variance for the traits of interest. Here, we evaluated the potential accuracy of prediction under situations ranging from poor causal SNP discovery (discovered “causal SNPs” explain 0% of genetic variance) to accurate causal SNP discovery (discovered causal SNPs explain 100% of genetic variance for the trait). The following additional assumptions were made: M_e within population *A* based on pre-selected SNPs (calculated from real genotype data) = 159; M_e across populations *A* and *B* based on 500 pre-selected SNPs (calculated from real genotype data) = 280; M_e within population *A* based on 48,412 remaining SNPs (calculated from real genotype data) = 463; M_e across populations *A* and *B* based on 48,412 remaining SNPs (calculated from real genotype data) = 32,970; genetic correlations between populations *A* and *B* = 0.6; heritability of the trait = 0.3 in both populations; proportion of genetic variance explained by all SNPs = 1.

Case 3

In this case, our aim was to explore the impact of the closeness between the validation population and the training populations as measured by the M_e across populations. In most studies, the number of variants identified as causal for complex traits is at most a few hundred. In the context of the MPMG model, the M_e across populations based on the identified potential causal SNPs is usually small, with an upper bound equal to the total number of identified “causal SNPs”. A parameter that is expected to be considerably large, especially with increasing SNP density is the M_e based on the remaining unselected SNPs. Here, we evaluated the effect of M_e across populations based on the remaining unselected SNPs on the potential accuracy of prediction. We varied M_e across populations from 1000 to 50,000. The following additional assumptions were made: M_e within population *A* based on 500 pre-selected SNPs (calculated from real genotype data) = 159; M_e across populations *A* and *B* based on 500 pre-selected SNPs (calculated from real genotype data) = 280; M_e within population *A* based on 48,412 remaining SNPs (calculated from real genotype data) = 463; genetic correlations between populations *A* and *B* = 0.6;

heritability of the trait = 0.3 in both populations; ρ^2 due to 500 pre-selected causal SNPs = 0.4, ρ^2 due to all SNPs = 0.8 (assuming the remaining non-causal SNPs capture 66% of the residual genetic variance).

Case 4

The study of Van den Berg et al. [44], reported that M_e values estimated from the GRM are underestimated by ~80% compared to M_e values back-solved from the empirical accuracy of a GBLUP model. In case 4, we investigated the potential impact of an underestimation of M_e within the predicted population A on the accuracy of prediction. Thus, we used underestimated values of M_e for both GRM_1 and GRM_2 in the prediction equations, with the extent of underestimation ranging from 0 to 90%. The following additional assumptions were made: M_e across populations A and B based on 500 pre-selected SNPs (calculated from real genotype data) = 280; M_e across populations A and B based on 48,412 remaining SNPs (calculated from real genotype data) = 32,970; genetic correlations between populations A and B : = 0.6; heritability of the trait = 0.3 in both populations; proportion of genetic variance explained by SNPs in GRM_1 = 0.5.

In all four cases, we evaluated the potential accuracy of three models using their prediction equations as follows.

Within-population, single-GRM (WPSG) model: To predict the potential accuracy of this model, we used the formula of Daetwyler et al. (Daetwyler *et al.* 2008), which takes the proportion of genetic variance explained by all SNPs (ρ_A^2) into account, as

$$\sqrt{\rho_A^2 \frac{h_A^2 N_A}{h_A^2 N_A + M_{eA_1}}}. \text{ The value for } M_{eA_1} \text{ was calculated based on all SNPs.}$$

Within population, multiple GRM (WPMG) model: to predict the potential accuracy of this model, we used the derived prediction Eq. (4.12).

Multi-population, single GRM (MPSG) model: to predict the potential accuracy of this model, we used the formula of Wientjes *et al.* (2016):

$$\sqrt{\begin{bmatrix} \rho_A \sqrt{\frac{h_A^2}{M_{eA}}} \\ \rho_A r_g \sqrt{\frac{h_B^2}{M_{eAB}}} \end{bmatrix}^T \begin{bmatrix} \frac{h_A^2}{M_{eA_1}} + \frac{1}{N_A} & r_g \frac{\sqrt{h_A^2 h_B^2}}{\sqrt{M_{eA} M_{eAB}}} \\ r_g \frac{\sqrt{h_A^2 h_B^2}}{\sqrt{M_{eA} M_{eAB}}} & \frac{h_B^2}{M_{eAB}} + \frac{1}{N_B} \end{bmatrix}^{-1} \begin{bmatrix} \rho_A \sqrt{\frac{h_A^2}{M_{eA}}} \\ \rho_A r_g \sqrt{\frac{h_B^2}{M_{eAB}}} \end{bmatrix}}.$$

Here also, M_{eA} and M_{eAB} were calculated based on all SNPs.

Multi-population, multiple GRM (MPMG) model: to predict potential accuracy, we used the derived prediction Eq. (4.17).

4.3 Results

4.3.1 Number of independent chromosome segments (M_e) within and across populations

The number of independent chromosome segments per SNP set estimated within population *A* (595 New Zealand Jersey) and across populations *A* and *B* (5553 Dutch Holsteins) are in Table 4.1. The same value of M_e within population *A* (280) was obtained when all 48,912 SNPs and when only the non-causal SNPs were used to construct the GRM. Estimated M_e within population *A* differed markedly between SNP sets only when the number of SNPs used to calculate the GRM was small. At lower SNP densities, M_e across population *A* and *B* were close to the number of SNPs used to calculate the GRM. M_e across populations obtained with higher density SNPs were similar.

Table 4.1 Number of independent chromosome segments (M_e) per SNP set estimated within the target population *A* and across populations *A* and *B*

SNP set (number of SNPs)	M_e within population <i>A</i>	M_e across populations <i>A</i> and <i>B</i>
Non-causal SNPs (48,412)	280	32,970
*CSNPs_500 + Non-causal SNPs (48,912)	280	33,242
*CSNPs_250 + Non-causal SNPs (48,662)	280	33,056
*CSNPs_125 + Non-causal SNPs (48,537)	280	32,984
CSNPs_500 (500)	159	463
CSNPs_250 (250)	107	236
CSNPs_125 (125)	67	124

*Non-causal SNPs combined with the pre-selected SNPs in a single GRM

4.3.2 Empirical versus predicted accuracies of the within-population, multiple GRM (WPMG) model

Empirical and predicted accuracies of the WPMG model (476 individuals in the training population) under the different levels of pre-selection of causal SNPs, with a simulated trait heritability of 0.3 and 0.8 are in Fig. 4.1. For a low heritability trait (0.3), Eq. (4.11) over-predicts the empirical accuracy of the WPMG model, with the extent of over-prediction ranging from 8.4% (CSNP_500) to 11.8% (CSNP_125). For a high heritability trait (0.8), predicted accuracies were close to empirical accuracies, with the predicted accuracies within the standard errors of empirical accuracies (except for CSNP_500).

Figure 4.1 Predicted versus empirical accuracies of the within-population, multiple GRM genomic prediction model (476 individuals from population *A* in the training population), with a simulated trait heritability of 0.3 and 0.8. The standard error bars represent twice the standard deviation of the accuracies across 100 replicates.

4.3.3 Empirical versus predicted accuracies of the multi-population, single GRM (MPSG) model

Empirical and predicted accuracies of the MPSG model under different levels of inclusion of causal SNPs in the GRM and different levels of r_g between populations A and B are shown in Fig. 4.2 (heritability in both populations = 0.3) and Fig. 3 (heritability in both populations = 0.8). As expected, given the large number of M_e across populations, empirical accuracies were not significantly different between scenarios differing in r_g and percentage of causal SNPs included in the GRM. The standard errors of empirical accuracies were higher at low heritability (Fig. 4.2), than at high heritability (Fig. 3), which most likely reflects the higher level of noise in the phenotype at low heritability than at a high heritability. As expected, an increase in heritability resulted in an increase in empirical accuracies. In general, the prediction equation for the accuracy of the MPSG model resulted in an over-prediction of empirical accuracy ranging from ~5 to 10%.

Figure 4.2 Predicted versus empirical accuracies of the multi-population, single GRM genomic prediction model (5553 individuals from population B and 476 individuals

from population *A* in the training population) under the three different levels of inclusion of causal SNPs in the GRM, with genetic correlation (r_g) between populations *A* and *B* of 0.4, 0.6 and 0.8, and a heritability of 0.3 in both populations. The standard error bars represent twice the standard deviation of the accuracies across 100 replicates.

Figure 4.3 Predicted versus empirical accuracies of the multi-population, single GRM genomic prediction model (5553 individuals from population *B* and 476 individuals from population *A* in the training population) under the three different levels of inclusion of causal SNPs in the GRM, with genetic correlation (r_g) between populations *A* and *B* of 0.4, 0.6 and 0.8, and a heritability of 0.8 in both populations. The standard error bars represent twice the standard deviation of the accuracies across 100 replicates.

4.3.4 Empirical versus predicted accuracies of the multi-population, multiple GRM (MPMG) model

The empirical and predicted accuracies of the different levels of pre-selection of causal SNPs, with a simulated genetic correlation between populations *A* and *B* of 0.4, 0.6 and 0.8 and a heritability of 0.3 in both populations, are in Fig. 4.4.

Figure 4.4 Predicted versus empirical accuracies of the multi-population, multiple GRM genomic prediction model (5553 individuals from population *B* and 476 individuals from population *A* in the training population) under the three different levels of pre-selection of causal SNPs, with genetic correlation (r_g) between populations *A* and *B* of 0.4, 0.6 and 0.8, and a heritability of 0.3 in both populations. The standard error bars represent twice the standard deviation of the accuracies across 100 replicates.

Results show that empirical and predicted accuracies increase with an increasing number of pre-selected causal SNPs in the first GRM, and the level of genetic correlation between populations. Using the prediction equation (Eq. 17), predicted accuracies were less than one standard error away from the average empirical

accuracy in seven of the nine scenarios evaluated. We observed over-prediction of accuracies in all the scenarios, but only in the CSNP_125 and the CSNP_250 and with a genetic correlation between populations of 0.4 did the predicted accuracies go outside the standard error of the empirical accuracies.

For a higher heritability trait (0.8), similar patterns of results were observed (Fig. 4.5). Empirical and predicted accuracies increase with increasing number of pre-selected causal SNPs in the first GRM, and the level of genetic correlation between populations. We also observed slight over-predictions of accuracy in the CSNP_125 and CSNP_250 levels, across the three levels of genetic correlation between populations. However, for the CSNP_500 level, we observed a slight under-prediction of accuracies, across the three levels of genetic correlation between populations. The level of over-estimation of empirical accuracy seems to be consistently higher at low heritability (Fig. 4.4) than at high heritability (Fig. 4.5). This also the case for the WPMG model (Fig. 4.1) and for most scenarios of the MPSG model (Figs. 4.2 and 4.3). A possible explanation might be that for a lower heritability trait, the SNPs have more difficulty in explaining all the genetic variance. This can result from a larger environmental effect in the phenotypes, which can be considered as a noise term in the phenotype around the genetic effect. Furthermore, the level of over-estimation of empirical accuracies with the MPMG model (between ~2 to ~10%, Figs. 4.4 and 4.5) is within the range of over-estimation observed with the MPSG model (between ~5 to 10%, Figs 4.2 and 4.3). Hence, the relative advantage of the MPMG model over the MPSG model as assessed by the corresponding deterministic prediction equations should be a good indication of the true advantage of the MPMG over the MPSG model.

Figure 4.5 Predicted versus empirical accuracies of the multi-population, multiple GRM genomic prediction model (5553 individuals from population *B* and 476 individuals from population *A* in the training population) under the three different levels of pre-selection of causal SNPs, with genetic correlation (r_g) between populations *A* and *B* of 0.4, 0.6 and 0.8, and a heritability of 0.8 in both populations. The standard error bars represent twice the standard deviation of the accuracies across 100 replicates.

In general, we observed a positive correlation between the EGV from the two GRM in MPMG, albeit with high standard errors, except in the CSNP_500 for which all the QTL underlying the trait are in one GRM, where the correlation was around zero [see Additional file 1 Table S1].

4.3.5 Potential accuracies of different models in relation to different levels of r_g , ρ_{A_1} , and M_{eAB_2}

The potential accuracies of within-and multi-population GP models, with either single or multiple GRM fitted, in relation to different levels of genetic correlation (r_g ; case 1) between populations are presented in Fig. 4.6. The accuracy of the within-population, single GRM (WPSG) model, is not affected by r_g between populations. The result shows that a 9.6% increase in accuracy is possible by splitting SNPs into separate GRM based on prior information on their causality (WPMG model), when the preselected SNPs explain 40% of the total genetic variance for the trait. Compared to WPSG, the multi-population, single GRM (MPSG) model can result in a small increase in the accuracy of prediction, ranging from 0% ($r_g = 0$) to a maximum of 3.1% ($r_g = 1$). When the multi-population, multiple GRM (MPMG) model is implemented, the increase in accuracy as compared to the WPSG model ranges from 9.6% ($r_g = 0$) to 32% ($r_g = 1$), again, assuming that the preselected SNPs explain 40% of the genetic variance.

Figure 4.6 Potential accuracy of predicting the genomic value of individuals from population A under different models: within-population, single GRM (WPSG), within-population, multiple GRM (WPMG), multi-population, single GRM (MPSG), multi-population, multiple GRM (MPMG), in relation to different values of genetic correlation (r_g) between population A and B. The following assumptions were made: M_e within population A based on 500 pre-selected causal SNPs (calculated from real genotype data) = 159; M_e across populations A and B based on 500 pre-selected causal SNPs (calculated from real genotype data) = 463; M_e within population A based on 48,412 non-causal SNPs (calculated from real genotype data) = 280; M_e across populations A and B based on 48,412 non-causal SNPs (calculated from real genotype data) = 32,970; M_e within population A based on all 48,912 SNPs (calculated from real genotype data) = 280; M_e across populations A and B based on all 48,912 SNPs (calculated from real genotype data) = 33,242; heritability of the trait = 0.3 in both populations; proportion of genetic variance explained by all SNPs = 0.8; proportion of genetic variance explained by 500 pre-selected causal SNPs = 0.4; number of individuals from population A in the training population = 476; number of individuals from population B in the training population = 5553.

The potential accuracies of the different models in relation to different proportions of total genetic variance explained by the pre-selected SNP set (ρ_{A_1} ; case 2) are

presented in Fig. 4.7. When the pre-selected SNP set explains zero proportion of genetic variance, the accuracy is equal for the models with one or two GRM for both within- and multi-population GP. Simply implementing the multi-population single GRM model, instead of the within-population single GRM model results in a negligible ($< 1\%$) increase in the accuracy of prediction. An increasing proportion of genetic variance explained by the pre-selected SNP set results in a linear increase in the accuracy of prediction, up to a maximum of $\sim 18.4\%$ when the WPMG model is implemented instead of the WPSG model. Increase in accuracy, as compared to the WPSG model, reaches a maximum of $\sim 29.2\%$ when the MPMG model is implemented instead.

Figure 4.7 Potential accuracy of predicting the genomic value of individuals from population A under different models: within-population, single GRM (WPSG), within-population, multiple GRM (WPMG), multi-population, single GRM (MPSG), multi-population, multiple GRM (MPMG), in relation to different proportion of genetic variance explained by the causal SNP set. The following assumptions were made: M_e within population A based on 500 pre-selected causal SNPs (calculated from real genotype data) = 159; M_e across populations A and B based on 500 pre-selected causal SNPs (calculated from real genotype data) = 463; M_e within population A based on 48,412 non-causal SNPs (calculated from real genotype data) = 280; M_e across populations A and B based on 48,412 non-causal SNPs (calculated from real genotype data) = 32,970; M_e within population A based on all 48,912 SNPs (calculated from real genotype data) = 280; M_e across populations A and B based on all 48,912 SNPs (calculated from real genotype data) = 33,242; heritability of the trait = 0.3 in both populations; genetic correlation between populations A and B = 0.6; proportion of genetic variance explained by all SNPs = 0.8; number of individuals from population A in the training population = 476; number of individuals from population B in the training population = 5553.

We investigated further the impact of the number of independent chromosome segments between populations based on the non-causal SNPs (M_{eAB} ; case 3) on the accuracy of GP models, the results are presented in Fig. 4.8. For the WPSG and

WPMG models, the parameter M_{eAB} has no relevance. However, in both the MPSG and MPMG models, the accuracy of prediction decreases with increasing M_{eAB} . The rate of decrease in accuracy with increasing M_{eAB} , however, is higher in the MPSG model than in the MPMG model. For example, an increase in M_{eAB} from 1000 to 20,000 resulted in an 11.3% decrease in accuracy of the MPSG model, while this decrease was only 4.8% in the MPMG model. In general, the difference in accuracy between the MPSG and MPMG models increases with increasing M_{eAB} . For M_{eAB} values smaller than 5000, both multi-population models (MPSG and MPMG) have higher accuracies than any within-population GP model. For larger values of M_{eAB} , however, the accuracy of the MPSG model is lower than that of the WBMG model, while the accuracy of the MPMG model tends to flatten above that of the WPMG model.

Figure 4.8 Potential accuracy of predicting the genomic value of individuals from population *A* under different models: within-population, single GRM (WPSG), within-population, multiple GRM (WPMG), multi-population, single GRM (MPSG), multi-population, multiple GRM (MPMG), in relation to different values of M_e across populations due to the non-causal SNP set. The following assumptions were made: M_e within population *A* based on 500 pre-selected causal SNPs (calculated from real genotype data) = 159; M_e across populations *A* and *B* based on 500 pre-selected causal SNPs (calculated from real genotype data) = 463; M_e within population *A* based on 48,412 non-causal SNPs (calculated from real genotype data) = 280; M_e within population *A* based on all 48,912 SNPs (calculated from real genotype data) = 280; heritability of the trait = 0.3 in both populations; genetic correlation between populations *A* and *B* = 0.6; proportion of genetic variance explained by all SNPs = 0.8; proportion of genetic variance explained by 500 pre-selected causal SNPs = 0.4; number of individuals from population *A* in the training population = 476; number of individuals from population *B* in the training population = 5553.

Furthermore, we assessed the potential bias in the predicted accuracies due to the under-estimation of M_e within the predicted population *A* (Fig. 4.9). Across different GP models, an underestimation of M_e within population *A* of ~20% resulted in an inflation of predicted accuracies ranging between 4.5 (MPMG model) to 7.4% (WPSG

model). At level of under-estimation of 90%, inflation of predicted accuracies ranged from 36 (MPMG) to 58% (WPSG)".

Figure 4.9 Potential accuracy of predicting the genomic value of individuals from population A under different models: within-population, single GRM (WPSG), within-population, multiple GRM (WPMG), multi-population, single GRM (MPSG), multi-population, multiple GRM (MPMG), in relation to different levels of underestimation of M_e within the predicted population A (0 – 90%). The following assumptions were made: M_e across populations A and B based on 500 pre-selected SNPs (calculated from real genotype data) = 280; M_e across populations A and B based on 48,412 remaining SNPs (calculated from real genotype data) = 32,970; genetic correlations between populations A and B = 0.6; heritability of the trait = 0.3 in both populations; proportion of genetic variance explained by SNPs in GRM_1 = 0.5; number of individuals from population A in the training population = 476; number of individuals from population B in the training population = 5553.

4.4 Discussion

The objective of this study was to underpin theoretically the advantages and limits of the multi-population, multiple GRM (MPMG) genomic prediction model over the multi-population, single GRM (MPSG) genomic prediction (GP) model, by deriving and validating a deterministic prediction equation for its accuracy. We derived the deterministic prediction equation for the accuracy of the MPMG model using selection index theory and building upon previous works by Daetwyler *et al.* (2008) and Wientjes *et al.* (2016). We showed that, the derived equation can predict the accuracy of the MPMG model under varying levels of genetic correlation between the target population and the additional population in the training population, varying levels of the heritability of the trait and varying levels of the proportions of genetic variance explained by the pre-selected and differentially weighted SNPs. The equation can be used to assess the potential benefit of combining information from different populations, e.g., different breeds or lines for GP in livestock or plants, or different groups of people based on their ethnic background for the prediction of disease risk scores.

To date, in the literature, increase in the accuracy of GP in a multi-population, single-GRM (MPSG) context as compared to within-population GP has been limited as illustrated in Table 2 of the review paper of Lund *et al.* (2014). This is consistent with our result shown in Fig. 4.7, where less than 1% increase in accuracy is projected if the MPSG model is implemented, assuming a genetic correlation of 0.6 between populations, instead of a within-population, single-GRM model. However, Raymond *et al.* (2018) showed that the MPMG model, which differentially weights SNPs based on prior knowledge of potential causality, can yield significant increases in the accuracy of GP as compared to a within-population or multi-population model in which all markers are equally weighted. The prediction equation developed in this study, highlights two parameters: number of independent chromosome segments across populations A and (M_{eAB}) , and the proportion of genetic variance explained by pre-selected SNPs ($\rho_{A_1}^2$), that may underlie the improved performance of the MPMG model as compared to a single GRM model. These parameters and their estimation are discussed below. We also discuss the values for the genetic correlation between populations as used in the prediction equation.

The number of independent chromosome segments across populations (M_{eAB}) is an important parameter that influences the accuracy of multi-population GP, since it determines the effective number of effects that are estimated in the model

(Goddard 2009; Sarup *et al.* 2016). For example, Wientjes *et al.* (2016) showed that when M_{eAB} is large, combining populations together in a multi-population single GRM model is less likely to result in a significant increase in accuracy as compared to single-population GP model. However, the prediction equation developed here shows that the MPMG model is still able to take advantage of information from distantly related populations (large M_{eAB}), mainly by partitioning the parameter M_{eAB} into two components, corresponding to pre-selected and remaining SNPs, respectively. In most cases, the value of M_{eAB} due to the pre-selected SNPs is small, given that, in most cases, only a few hundred SNPs are pre-selected from e.g., a genome-wide association study. The small value for M_{eAB} due to pre-selected SNPs means that the accuracy contributed by the pre-selected SNPs is high, and completely unaffected by the value for M_{eAB} due to the remaining SNPs.

For distantly related populations such as Holstein and Jersey, a good estimate for M_{eAB} based on a few pre-selected SNPs is the number of pre-selected SNPs (Table 4.1), which in any case is the maximum possible value. For M_{eAB} based on the remaining SNPs, it is very difficult to get an estimate without genotype data, as this value depends on the relatedness between populations. Similar to Wientjes *et al.* (2016) our suggestion is to consider genotyping a sample of individuals from each of the populations, e.g. 100 each, from which M_{eAB} can be estimated. With the availability of genotype data on individuals from the target population, the number of independent chromosome segments within population (M_{eA}) can be estimated empirically as the reciprocal of the variance of within-population genomic relationships (Wientjes *et al.* 2015; Lee *et al.* 2017a; Lee *et al.* 2017b). It is also possible to estimate M_{eA} based on population parameters such as the effective population size (Goddard 2009; Hayes *et al.* 2009c; Goddard *et al.* 2011). However, this approach cannot be used to partition M_{eA} into values corresponding to the pre-selected and remaining SNPs, respectively, which are required for the prediction equation.

All the M_e values used in our prediction equation were estimated from the GRM. Van Den Berg *et al.* (2019) argued that M_e values estimated from the GRM are underestimated, and, when used in a prediction equation, result in the over-prediction of empirical accuracy. We observed that an underestimation of M_e within the target population *A* results in a substantial inflation of predicted accuracies (Fig. 4.9). The prediction equation developed in our study, with M_e values calculated from the GRM tend to over-predict the accuracy of the MPMG model, although in most of the

scenarios evaluated, predicted accuracies were still within the standard errors of empirical accuracies. However, the extent of over-prediction of accuracy in Van Den Berg *et al.* (2019) using the deterministic formula of Goddard (2009) with M_e values estimated from the GRM, was much higher than in our study. We cannot pinpoint, with certainty, the underlying reasons for the differences in the extent of over-prediction of empirical accuracies in our study and in Van Den Berg *et al.* (2019). However, there are a few possible explanations. For example, Van Den Berg *et al.* (2019) fitted a polygenic component (pedigree-based relationship matrix) in their model, which was meant to capture the genetic variance for the trait that is not picked up by SNPs. Although the authors did not present the results of variance components estimation, it is likely that the polygenic component picked up some proportion of the total genetic variance for the trait. If that is the case, using a prediction equation with the assumption that SNPs explain 100% of the total genetic variance for the trait, as the authors did, could have resulted in the over-prediction of empirical accuracies. The chances that the polygenic component picks up some proportion of the total genetic variance for the trait are higher when QTL are sampled from sequence variants that have low MAF, and effects that are difficult to estimate or regressed heavily towards zero in the model, than when QTL are sampled from common SNPs (Wientjes *et al.* 2016). Furthermore, in their prediction equation, Van Den Berg *et al.* (2019) corrected for the reduction in prediction error variance as the accuracy of predicted genomic values increases. This correction results in higher predicted accuracies than when the correction is not applied. In our study, however, this correction was not applied. There are other factors, such as the difference in the design of cross-validation schemes to calculate empirical accuracies and structure of the populations analyzed that could potentially underlie the difference in the extent of over-prediction of empirical accuracies between the two studies.

The ability of the MPMG model to take advantage of the small value of M_{eAB} due to pre-selected SNPs depends on the accuracy of SNP pre-selection, which in turn determines the proportion of genetic variance explained by the pre-selected SNPs. Our results (Fig. 4.7) show that for an improvement in accuracy of GP by using the MPMG instead of the MPSG model, it is not sufficient to split randomly SNPs into two classes without accurate prior knowledge on potential causality of the SNPs. Instead, the pre-selected SNPs must explain some proportion of the genetic variance for the trait. Our results agree with those of Sarup *et al.* (2016), who showed that a significant improvement in accuracy by using the “genomic feature” model over the standard GBLUP could be achieved provided the prioritised variants explained more than 10% of the total genetic variance for the trait. In their study, Sarup *et al.* (2016)

fitted a linear mixed model including two random genomic effects, with the genetic effects estimated from “genomic feature” or prioritized variants and the remaining variants, respectively. In general, when the pre-selected SNPs explain zero proportion of genetic variance, the accuracy of the MPMG model is expected to be the same as the accuracy of the MPSG model (Fig. 4.7). Furthermore, if the prior information is accurate, such that the pre-selected SNPs explain some proportion of the genetic variance for the trait, fitting the pre-selected SNPs in a separate GRM in a within-population GP scenario may be more beneficial than combining distantly related populations in a multi-population single-GRM model (Fig. 4.6).

In this study, the proportions of total genetic variance explained in the validation population by the SNPs in the first GRM ($\rho_{A_1}^2$) was determined in a simulation context. There was 100% certainty that the selected SNPs were causal, and that $\rho_{A_1}^2$ did not depend on linkage disequilibrium (LD) between non-causal SNPs and the underlying causal SNPs. In this case, we found that the estimate of genomic heritability (proportion of the genetic variance explained by the SNPs) from the GREML model (Yang *et al.* 2010; Speed *et al.* 2012) was equivalent to $\rho_{A_1}^2$ (results not shown). De Los Campos *et al.* (2015) also showed that when all causal mutations are included in the GREML analysis, the genomic heritability parameter is equivalent to the proportion of genetic variance explained by SNPs (ρ^2). However, in practice, the causal mutations that underlie a trait are not always observed, and pre-selected SNPs will usually come from GWAS studies, in which SNPs can show an association with a trait due to LD with unobserved causal mutations. When SNPs that are proxies for causal mutations are used for the analysis, the variance components estimates depend on the effects of the unobserved causal mutations, the extent of LD between SNPs and the causal mutations, and the extent of LD between the SNPs themselves (De Los Campos *et al.* 2015). Not accounting explicitly for the LD between SNPs and the causal mutations in the model used for analysis can result in the over-estimation of variance components (De Los Campos *et al.* 2015). This means that the estimate of genomic heritability from the GREML model cannot be considered to be equivalent to ρ^2 . Browning and Browning (2011) also showed that population structure can inflate SNP-based heritability estimates. Although genomic heritability from GREML models are higher than ρ^2 , they are still a good indication for ρ^2 , as they have similar trend (Erbe *et al.* 2013). Thus, for comparison of the proportion of genetic variance explained by SNPs in different SNP sets, genomic heritability from a GREML model can be used.

In practice, when two GRM are fitted simultaneously in a bivariate GREML model, two estimates of genetic correlation are obtained, one for each GRM. However, in this study, we used only the simulated value for the genetic correlation between populations for both GRM in the prediction Eq. (4.17). This is because the pre-selected SNPs were randomly sampled from all causal SNPs. In practice, a general expectation is that causal SNPs that are pre-selected from e.g. GWAS have a higher effect on the trait than the remaining unselected SNPs, and are most likely to be more consistent across populations (e.g. Calus *et al.* 2018; Raymond *et al.* 2018). Therefore, the genetic correlation of the preselected SNPs can be higher than that for the remaining SNPs. In the cases that this expectation holds, it is inappropriate to use a single value of genetic correlation for both GRM in the prediction Eq. (4.17). A solution to this issue is to use the estimated genetic correlation. However, one must take into account that the estimated genetic correlation can be biased when the causal and non-causal SNPs used to estimate the genetic correlation do not have similar properties, e.g., similar pattern of allele frequencies (Wientjes *et al.* 2018). In general, it is likely that causal SNPs have lower MAF than non-causal SNPs, which means that estimates of genetic correlation obtained by non-causal SNPs can be underestimating the genetic correlation between populations at the causal SNPs (Raymond *et al.* 2018).

In the derivation of the predicting Eq. (4.17), we assumed that there is no covariance between EGV from different GRM, since the expectation is zero when the GRM are jointly fitted (Yang *et al.* 2011). Because a negative sampling covariance might appear when effects cannot be estimated independently, and the covariance might bias predicted accuracies using the derived equation, we tested this assumption. We calculated the correlations between EGV_{A_1} and EGV_{A_2} in each replicate and averaged these across replicates. In general, we observed a positive correlation between the EGV from the two GRM in MPMG, albeit with high standard errors, except in the CSNP_500 in which case all the causal SNPs underlying the trait are in one GRM, where the correlation was around 0 [see Additional file 1 Table S1]. However, in most cases the correlations were not significantly different from 0. The positive correlation between the EGV mean that individuals that have a high EGV based on the pre-selected SNPs (GRM_1), on average also have a high EGV based on the rest of the genome (GRM_2). To check if the observed correlation between EGV is an artefact of our simulation, we summed the effects of all causal SNPs in GRM_1 as TGV_1 and summed the effect of all causal SNPs not in GRM_2 as TGV_2 . Across 100 replicates, we observed no correlation between the TGV: -0.007 (0.07) for CSNP_125 and 0.01 (0.07) for CSNP_250. The observed correlation between EGV do not seem

to be a result of sampling covariance, since the expected sign of the resulting correlation is negative, and are not an artefact of our simulation, given that the TGV based on the separate sets of causal SNPs were not correlated. We think that the empirical correlation between EGV are a result of an estimation issue, but we are not sure. In any case, given that the level of overestimation of empirical accuracy of the MPMG model is in the range of overestimation observed in a model with only one GRM (MPSG), the impact of any possible covariance between the EGV on the predicted accuracy of the MPMG model is expected to be small.

4.5 Conclusions

In this paper, we presented a deterministic prediction equation for the accuracy of a multi-population, multiple GRM (MPMG) model, which gives insight into the underlying reasons for the superior performance of the MPMG model over the multi-population, single GRM (MPSG) model. With the help of the prediction equation, we showed that an important advantage of the MPMG model is its ability to benefit from the small number of independent chromosome segments (M_e) due to the pre-selected SNPs, both within and across populations, while for the MPSG model, there is only a single value for M_e , calculated based on all SNPs. However, this advantage depends on the condition that the pre-selected SNPs can explain some proportion of the total genetic variance for the trait. The prediction equation developed here can be used as a deterministic tool to assess the potential benefit of combining information from different populations e.g., different breeds or lines for GP in livestock or plants, or different groups of people based on their ethnic background for genomic prediction in humans. This is especially the case, when accurate biological information about the causality of the SNPs is available.

4.6 Appendices

4.6.1 Appendix 1

$$\begin{aligned} cov(\text{EGV}_{A_1, A}, \text{EGV}_{A_1, B}) &= cov\left(\sum_j X_{A_{i,j}} \hat{\beta}_{A_1, j}, r_g \sum_j X_{A_{i,j}} \hat{\beta}_{B_1, j}\right) \\ &= r_g cov(\hat{\beta}_{A_1}, \hat{\beta}_{B_1}), \end{aligned} \quad (\text{A1.1})$$

where $X_{A_{i,j}}$ is the genotype of validation candidate i from population A at locus j , $\hat{\beta}_{A_1, j}$ and $\hat{\beta}_{B_1, j}$ are the estimated SNP effects for locus j in training populations A and B , respectively, and r_g is the genetic correlation between populations A and B .

$$cov(\hat{\beta}_{A_1}, \hat{\beta}_{B_1}) = cor_{\hat{\beta}_{A_1}, \hat{\beta}_{B_1}} \sqrt{var(\hat{\beta}_{A_1}) var(\hat{\beta}_{B_1})},$$

where $cor_{\hat{\beta}_{A_1}, \hat{\beta}_{B_1}}$ is the correlation between the estimated SNP effects of GRM₁ in populations A and B . Using the path coefficient method of Dekkers [24], it can be shown that $cor_{\hat{\beta}_{A_1}, \hat{\beta}_{B_1}} = r_g r_{SNP_{A_1}} r_{SNP_{B_1}}$, where $r_{SNP_{A_1}}$ and $r_{SNP_{B_1}}$ are the accuracies of the estimated SNP effects in populations A and B , respectively. The square root of the variance of estimated SNP effects in populations A and B is equal to the accuracy of the estimated SNP effects, i.e.

$$\sqrt{var(\hat{\beta}_{A_1}) var(\hat{\beta}_{B_1})} = r_{SNP_{A_1}} r_{SNP_{B_1}}.$$

Therefore:

$$\text{cov}(\hat{\beta}_{A_1}, \hat{\beta}_{B_1}) = r_g r_{SNP_{A_1}}^2 r_{SNP_{B_1}}^2 = r_g^2 r_{SNP_{A_1}}^2 r_{SNP_{B_1}}^2,$$

and filling this in Eq. (A1.1) results in:

$$r_g \text{cov}(\hat{\beta}_{A_1}, \hat{\beta}_{B_1}) = r_g^2 r_{SNP_{A_1}}^2 r_{SNP_{B_1}}^2. \quad (\text{A1.2})$$

Assuming that true genomic values are scaled such that they have a variance of 1, Eq. (A1.2) becomes:

$$r_{EGV_{A_1,A}}^2 r_{EGV_{A_1,B}}^2. \quad (\text{A1.3})$$

4.6.1 Appendix 2

Matrix P and vector g are shown below:

$$P = \begin{bmatrix} r_{EGV_{A_1,A}}^2 & r_{EGV_{A_1,A}}^2 r_{EGV_{A_1,B}}^2 & 0 & 0 \\ r_{EGV_{A_1,A}}^2 r_{EGV_{A_1,B}}^2 & r_{EGV_{A_1,B}}^2 & r_{EGV_{A_2,A}}^2 & r_{EGV_{A_2,A}}^2 r_{EGV_{A_2,B}}^2 \\ 0 & 0 & r_{EGV_{A_2,A}}^2 & r_{EGV_{A_2,B}}^2 \\ 0 & 0 & r_{EGV_{A_2,A}}^2 r_{EGV_{A_2,B}}^2 & r_{EGV_{A_2,B}}^2 \end{bmatrix}.$$

$$g = \begin{bmatrix} r_{EGV_{A_1,A}}^2 \\ r_{EGV_{A_1,B}}^2 \\ r_{EGV_{A_2,A}}^2 \\ r_{EGV_{A_2,B}}^2 \end{bmatrix}.$$

Since P is a block diagonal matrix, it can be split into 2x2 sub-matrices for ease of inversion:

$$P_{11} = \begin{bmatrix} r_{EGV_{A_1,A}}^2 & r_{EGV_{A_1,A}}^2 r_{EGV_{A_1,B}}^2 \\ r_{EGV_{A_1,A}}^2 & r_{EGV_{A_1,B}}^2 \end{bmatrix},$$

$$P_{22} = \begin{bmatrix} r_{EGV_{A_2,A}}^2 & r_{EGV_{A_2,A}}^2 r_{EGV_{A_2,B}}^2 \\ r_{EGV_{A_2,A}}^2 & r_{EGV_{A_2,B}}^2 \end{bmatrix}.$$

Then, $P = \begin{bmatrix} P_{11} & 0 \\ 0 & P_{22} \end{bmatrix}$ and $P^{-1} = \begin{bmatrix} P_{11}^{-1} & 0 \\ 0 & P_{22}^{-1} \end{bmatrix}$. The inverse of the two non-zero sub-matrices of P can be calculated as:

$$P_{11}^{-1} = \frac{1}{r_{EGV_{A_1,A}}^2 r_{EGV_{A_1,B}}^2 - r_{EGV_{A_1,A}}^4 r_{EGV_{A_1,B}}^4} \begin{bmatrix} r_{EGV_{A_1,B}}^2 & -r_{EGV_{A_1,A}}^2 r_{EGV_{A_1,B}}^2 \\ -r_{EGV_{A_1,A}}^2 r_{EGV_{A_1,B}}^2 & r_{EGV_{A_1,A}}^2 \end{bmatrix}$$

$$\begin{aligned}
&= \begin{bmatrix} \frac{\gamma_{\hat{B}_{A_1,B}}^2}{\gamma_{\text{EGV}_{A_1,A}}^2 \gamma_{\text{EGV}_{A_1,B}}^2 - \gamma_{\text{EGV}_{A_1,A}}^4 \gamma_{\text{EGV}_{A_1,B}}^4} & \frac{-\gamma_{\hat{B}_{A_1,A}}^2 \gamma_{\hat{B}_{A_1,B}}^2}{\gamma_{\text{EGV}_{A_1,A}}^2 \gamma_{\text{EGV}_{A_1,B}}^2 - \gamma_{\text{EGV}_{A_1,A}}^4 \gamma_{\text{EGV}_{A_1,B}}^4} \\ \frac{-\gamma_{\hat{B}_{A_1,A}}^2 \gamma_{\hat{B}_{A_1,B}}^2}{\gamma_{\text{EGV}_{A_1,A}}^2 \gamma_{\text{EGV}_{A_1,B}}^2 - \gamma_{\text{EGV}_{A_1,A}}^4 \gamma_{\text{EGV}_{A_1,B}}^4} & \frac{\gamma_{\hat{B}_{A_1,A}}^2}{\gamma_{\text{EGV}_{A_1,A}}^2 \gamma_{\text{EGV}_{A_1,B}}^2 - \gamma_{\text{EGV}_{A_1,A}}^4 \gamma_{\text{EGV}_{A_1,B}}^4} \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{\gamma_{\text{EGV}_{A_1,A}}^2 (\gamma_{\text{EGV}_{A_1,A}}^2 \gamma_{\text{EGV}_{A_1,B}}^2)} & \frac{-1}{(1 - \gamma_{\text{EGV}_{A_1,A}}^2 \gamma_{\text{EGV}_{A_1,B}}^2)} \\ \frac{-1}{(1 - \gamma_{\text{EGV}_{A_1,A}}^2 \gamma_{\text{EGV}_{A_1,B}}^2)} & \frac{1}{\gamma_{\text{EGV}_{A_1,B}}^2 (1 - \gamma_{\text{EGV}_{A_1,A}}^2 \gamma_{\text{EGV}_{A_1,B}}^2)} \end{bmatrix}
\end{aligned}$$

Similarly:

$$P_{22}^{-1} = \begin{bmatrix} \frac{1}{\gamma_{\text{EGV}_{A_2,A}}^2 (1 - \gamma_{\text{EGV}_{A_2,A}}^2 \gamma_{\text{EGV}_{A_2,B}}^2)} & \frac{-1}{(1 - \gamma_{\text{EGV}_{A_2,A}}^2 \gamma_{\text{EGV}_{A_2,B}}^2)} \\ \frac{-1}{(1 - \gamma_{\text{EGV}_{A_2,A}}^2 \gamma_{\text{EGV}_{A_2,B}}^2)} & \frac{1}{\gamma_{\text{EGV}_{A_2,B}}^2 (1 - \gamma_{\text{EGV}_{A_2,A}}^2 \gamma_{\text{EGV}_{A_2,B}}^2)} \end{bmatrix}.$$

$$r_{EGV_{AT}}^2 = \sqrt{\begin{bmatrix} r_{EGV_{A1,A}}^2 & r_{EGV_{A1,B}}^2 & r_{EGV_{A2,A}}^2 & r_{EGV_{A2,B}}^2 \end{bmatrix}^T \begin{bmatrix} 1 & -1 & 0 & 0 \\ \frac{r_{EGV_{A1,A}}^2 (r_{EGV_{A1,A}}^2 r_{EGV_{A1,B}}^2)}{(1-r_{EGV_{A1,A}}^2 r_{EGV_{A1,B}}^2)} & \frac{r_{EGV_{A1,B}}^2 (1-r_{EGV_{A1,A}}^2 r_{EGV_{A1,B}}^2)}{(1-r_{EGV_{A1,B}}^2 r_{EGV_{A1,A}}^2 r_{EGV_{A1,B}}^2)} & 0 & 0 \\ -1 & 1 & 0 & 1 \\ 0 & 0 & \frac{r_{EGV_{A2,A}}^2 (1-r_{EGV_{A2,A}}^2 r_{EGV_{A2,B}}^2)}{(1-r_{EGV_{A2,A}}^2 r_{EGV_{A2,B}}^2)} & \frac{r_{EGV_{A2,B}}^2 (1-r_{EGV_{A2,A}}^2 r_{EGV_{A2,B}}^2)}{(1-r_{EGV_{A2,A}}^2 r_{EGV_{A2,B}}^2)} \end{bmatrix} \begin{bmatrix} r_{EGV_{A1,A}}^2 \\ r_{EGV_{A1,B}}^2 \\ r_{EGV_{A2,A}}^2 \\ r_{EGV_{A2,B}}^2 \end{bmatrix}}$$

Simplifying the equation above results in:

$$\sqrt{\begin{aligned} & \frac{r_{EGV_{A1,A}}^2}{(1-r_{EGV_{A1,A}}^2 r_{EGV_{A1,B}}^2)} + \frac{-r_{EGV_{A1,A}}^2 r_{EGV_{A1,B}}^2}{(1-r_{EGV_{A1,A}}^2 r_{EGV_{A1,B}}^2)} + \frac{-r_{EGV_{A1,A}}^2 r_{EGV_{A1,B}}^2}{(1-r_{EGV_{A1,A}}^2 r_{EGV_{A1,B}}^2)} + \frac{r_{EGV_{A1,B}}^2}{(1-r_{EGV_{A1,A}}^2 r_{EGV_{A1,B}}^2)} \\ & + \frac{r_{EGV_{A2,A}}^2}{(1-r_{EGV_{A2,A}}^2 r_{EGV_{A2,B}}^2)} + \frac{-r_{EGV_{A2,A}}^2 r_{EGV_{A2,B}}^2}{(1-r_{EGV_{A2,A}}^2 r_{EGV_{A2,B}}^2)} + \frac{-r_{EGV_{A2,A}}^2 r_{EGV_{A2,B}}^2}{(1-r_{EGV_{A2,A}}^2 r_{EGV_{A2,B}}^2)} + \frac{r_{EGV_{A2,B}}^2}{(1-r_{EGV_{A2,A}}^2 r_{EGV_{A2,B}}^2)} \end{aligned}}$$

$$= \sqrt{\left(\frac{r_{EGV_{A1,A}}^2 + r_{EGV_{A1,B}}^2 - 2r_{EGV_{A1,A}}^2 r_{EGV_{A1,B}}^2}{(1-r_{EGV_{A1,A}}^2 r_{EGV_{A1,B}}^2)} \right) + \left(\frac{r_{EGV_{A2,A}}^2 + r_{EGV_{A2,B}}^2 - 2r_{EGV_{A2,A}}^2 r_{EGV_{A2,B}}^2}{(1-r_{EGV_{A2,A}}^2 r_{EGV_{A2,B}}^2)} \right)} \quad (A2.1)$$

The reliability of estimated genomic values is a product of the reliability of the estimated SNP effects (r_{SNP}^2) and the proportion of genetic variance explained in the validation population by the SNPs (ρ^2). Ignoring ρ^2 , for now, and replacing r_{SNP}^2 with the deterministic equations to predict accuracy of GP [6, 31]:

$$r_{SNP A_1}^2 = \sqrt{\frac{h_A^2 N_A}{h_A^2 N_A + ME_{A_1}}} = \sqrt{\frac{\frac{h_A^2}{ME_{A_1}}}{\frac{h_A^2}{ME_{A_1}} + \frac{1}{N_A}}},$$

$$r_{SNP B_1}^2 = \sqrt{r_g^2 \frac{\frac{h_A^2}{ME_{AB_1}}}{\frac{h_B^2}{ME_{AB_1}} + \frac{1}{N_B}}}, r_{SNP A_2}^2 = \sqrt{\frac{\frac{h_A^2}{ME_{A_2}}}{\frac{h_A^2}{ME_{A_2}} + \frac{1}{N_A}}}, \text{ and } r_{SNP B_2}^2 = \sqrt{r_g^2 \frac{\frac{h_A^2}{ME_{AB_2}}}{\frac{h_B^2}{ME_{AB_2}} + \frac{1}{N_B}}},$$

then Eq. (A2.1) becomes:

$$\sqrt{\left(\frac{\frac{\frac{h_A^2}{ME_{A_1}}}{\frac{h_A^2}{ME_{A_1}} + \frac{1}{N_A}} + r_g^2 \frac{\frac{h_A^2}{ME_{AB_1}}}{\frac{h_B^2}{ME_{AB_1}} + \frac{1}{N_B}} - 2 \left(\frac{\frac{h_A^2}{ME_{A_1}}}{\frac{h_A^2}{ME_{A_1}} + \frac{1}{N_A}} \right) \left(r_g^2 \frac{\frac{h_A^2}{ME_{AB_1}}}{\frac{h_B^2}{ME_{AB_1}} + \frac{1}{N_B}} \right)}{1 - \left(\frac{\frac{h_A^2}{ME_{A_1}}}{\frac{h_A^2}{ME_{A_1}} + \frac{1}{N_A}} \right) \left(r_g^2 \frac{\frac{h_A^2}{ME_{AB_1}}}{\frac{h_B^2}{ME_{AB_1}} + \frac{1}{N_B}} \right)} \right) + \left(\frac{\frac{\frac{h_A^2}{ME_{A_2}}}{\frac{h_A^2}{ME_{A_2}} + \frac{1}{N_A}} + r_g^2 \frac{\frac{h_A^2}{ME_{AB_2}}}{\frac{h_B^2}{ME_{AB_2}} + \frac{1}{N_B}} - 2 \left(\frac{\frac{h_A^2}{ME_{A_2}}}{\frac{h_A^2}{ME_{A_2}} + \frac{1}{N_A}} \right) \left(r_g^2 \frac{\frac{h_A^2}{ME_{AB_2}}}{\frac{h_B^2}{ME_{AB_2}} + \frac{1}{N_B}} \right)}{1 - \left(\frac{\frac{h_A^2}{ME_{A_2}}}{\frac{h_A^2}{ME_{A_2}} + \frac{1}{N_A}} \right) \left(r_g^2 \frac{\frac{h_A^2}{ME_{AB_2}}}{\frac{h_B^2}{ME_{AB_2}} + \frac{1}{N_B}} \right)} \right)} \right) \cdot (A2.2)$$

Multiplying both the numerator and the denominator of the first component of Eq. (A2.2) by $\left(\frac{h_A^2}{ME_{A_1}} + \frac{1}{N_A} \right) \left(\frac{h_B^2}{ME_{AB_1}} + \frac{1}{N_B} \right)$ and the second component by $\left(\frac{h_A^2}{ME_{A_2}} + \frac{1}{N_A} \right) \left(\frac{h_B^2}{ME_{AB_2}} + \frac{1}{N_B} \right)$ results in:

$$\sqrt{\left(\frac{\frac{h_A^2}{ME_{A1}} \left(\frac{h_B^2}{ME_{AB1}} + \frac{1}{N_B} \right) + r_g^2 \frac{h_B^2}{ME_{AB1}} \left(\frac{h_A^2}{ME_{A1}} + \frac{1}{N_A} \right) - 2 \left(\frac{h_A^2}{ME_{A1}} \right) \left(r_g^2 \frac{h_B^2}{ME_{AB1}} \right)}{\left(\frac{h_A^2}{ME_{A1}} + \frac{1}{N_A} \right) \left(\frac{h_B^2}{ME_{AB1}} + \frac{1}{N_B} \right) - \left(\frac{h_A^2}{ME_{A1}} \right) \left(r_g^2 \frac{h_B^2}{ME_{AB1}} \right)} \right) + \left(\frac{\frac{h_A^2}{ME_{A2}} \left(\frac{h_B^2}{ME_{AB2}} + \frac{1}{N_B} \right) + r_g^2 \frac{h_B^2}{ME_{AB2}} \left(\frac{h_A^2}{ME_{A2}} + \frac{1}{N_A} \right) - 2 \left(\frac{h_A^2}{ME_{A2}} \right) \left(r_g^2 \frac{h_B^2}{ME_{AB2}} \right)}{\left(\frac{h_A^2}{ME_{A2}} + \frac{1}{N_A} \right) \left(\frac{h_B^2}{ME_{AB2}} + \frac{1}{N_B} \right) - \left(\frac{h_A^2}{ME_{A2}} \right) \left(r_g^2 \frac{h_B^2}{ME_{AB2}} \right)} \right)} \quad (A2.3)$$

Taking into account the proportion of genetic variance explained by SNPs in GRM₁ ($\rho_{A_1}^2$) and GRM₂ ($\rho_{A_2}^2$), respectively, Eq. (A2.3) becomes:

$$\sqrt{\rho_{A_1}^2 \left(\frac{\frac{h_A^2}{ME_{A1}} \left(\frac{h_B^2}{ME_{AB1}} + \frac{1}{N_B} \right) + r_g^2 \frac{h_B^2}{ME_{AB1}} \left(\frac{h_A^2}{ME_{A1}} + \frac{1}{N_A} \right) - 2 \left(\frac{h_A^2}{ME_{A1}} \right) \left(r_g^2 \frac{h_B^2}{ME_{AB1}} \right)}{\left(\frac{h_A^2}{ME_{A1}} + \frac{1}{N_A} \right) \left(\frac{h_B^2}{ME_{AB1}} + \frac{1}{N_B} \right) - \left(\frac{h_A^2}{ME_{A1}} \right) \left(r_g^2 \frac{h_B^2}{ME_{AB1}} \right)} \right) + \rho_{A_2}^2 \left(\frac{\frac{h_A^2}{ME_{A2}} \left(\frac{h_B^2}{ME_{AB2}} + \frac{1}{N_B} \right) + r_g^2 \frac{h_B^2}{ME_{AB2}} \left(\frac{h_A^2}{ME_{A2}} + \frac{1}{N_A} \right) - 2 \left(\frac{h_A^2}{ME_{A2}} \right) \left(r_g^2 \frac{h_B^2}{ME_{AB2}} \right)}{\left(\frac{h_A^2}{ME_{A2}} + \frac{1}{N_A} \right) \left(\frac{h_B^2}{ME_{AB2}} + \frac{1}{N_B} \right) - \left(\frac{h_A^2}{ME_{A2}} \right) \left(r_g^2 \frac{h_B^2}{ME_{AB2}} \right)} \right)} \quad (A2.4)$$

4.7 Acknowledgements

This study was financially supported by NWO-TTW, the Dutch Ministry of Economic Affairs (TKI Agri & Food project 16022, code BO-22.04-011-001-ASG-LR), and the Breed4Food partners Cobb Europe, CRV, Hendrix Genetics and Topigs Norsvin. The use of the HPC cluster was made possible by CAT-AgroFood (Shared Research Facilities, Wageningen University and Research).

4.8 References

- Brøndum RF, Su G, Janss L, Sahana G, Guldbrandtsen B, Boichard D, et al. Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. *J Dairy Sci.* 2015;98:4107–16.
- Browning SR, Browning BL. Population structure can inflate SNP-based heritability estimates. *Am J Hum Genet.* 2011;89:191–3.
- Calus MPL, Goddard ME, Wientjes YCJ, Bowman PJ, Hayes BJ. Multibreed genomic prediction using multitrait genomic residual maximum likelihood and multitask Bayesian variable selection. *J Dairy Sci.* 2018;101:4279–94.
- Daetwyler HD, Villanueva B, Woolliams JA. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS One.* 2008;3:e3395.
- de los Campos G, Sorensen D, Gianola D. Genomic heritability: What is it? *PLoS Genet.* 2015;11:e1005048.
- de Roos A, Schrooten C, Veerkamp RF, Van Arendonk JA. Effects of genomic selection on genetic improvement, inbreeding, and merit of young versus proven bulls. *J Dairy Sci.* 2011;94:1559–67.
- Dekkers J. Prediction of response to marker-assisted and genomic selection using selection index theory. *J Anim Breed Genet.* 2007;124:331–41.
- Dalton R. No bull: genes for better milk. *Nature News.* 2009;457:369.
- Erbe M, Gredler B, Seefried FR, Bapst B, Simianer H. A function accounting for training set size and marker density to model the average accuracy of genomic prediction. *PLoS One.* 2013;8:e81046.
- Erbe M, Hayes BJ, Matukumalli LK, Goswami S, Bowman PJ, Reich CM, et al. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *J Dairy Sci.* 2012;95:4114–29.
- Falconer D, MacKay T. Introduction to quantitative genetics. Harlow: Longman; 1996.
- Goddard M. Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica.* 2009;136:245–57.
- Goddard ME, Hayes BJ, Meuwissen TH. Using the genomic relationship matrix to predict the accuracy of genomic selection. *J Anim Breed Genet.* 2011;128:409–21.
- Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME. Invited review: genomic selection in dairy cattle: Progress and challenges. *J Dairy Sci.* 2009a;92:433–43.
- Hayes BJ, Bowman PJ, Chamberlain AC, Verbyla K, Goddard ME. Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genet Sel Evol.* 2009b;41:51.
- Hayes BJ, Visscher PM, Goddard ME. Increased accuracy of artificial selection by using the realized relationship matrix. *Genet Res.* 2009c;91:47–60.
- Hazel LN. The genetic basis for constructing selection indexes. *Genetics.* 1943;28:476–90.

- Heffner EL, Jannink JL, Sorrells ME. Genomic selection accuracy using multifamily prediction models in a wheat breeding program. *Plant Genome*. 2011;4:65–75.
- Karoui S, Carabano MJ, Diaz C, Legarra A. Joint genomic evaluation of French dairy cattle breeds using multiple-trait models. *Genet Sel Evol*. 2012;44:39.
- Lee SH, Clark S, van der Werf JH. Estimation of genomic prediction accuracy from reference populations with varying degrees of relationship. *PLoS One*. 2017a;12:e0189775.
- Lee SH, Van der Werf JH. MTG2: an efficient algorithm for multivariate linear mixed model analysis based on genomic information. *Bioinformatics*. 2016;32:1420–2.
- Lee SH, Weerasinghe WSP, Wray NR, Goddard ME, van der Werf JH. Using information of relatives in genomic prediction to apply effective stratified medicine. *Sci Rep*. 2017b;7:42091.
- Lund MS, Su G, Janss L, Guldbrandtsen B, Brodum RF. Genomic evaluation of cattle in a multi-breed context. *Livest Sci*. 2014;166:101–10.
- MacLeod IM, Bowman PJ, Vander Jagt CJ, Haile-Mariam M, Kemper KE, Chamberlain AJ, et al. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. *BMC Genomics*. 2016;17:144.
- Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*. 2001;157:1819–29.
- Olson KM, VanRaden PM, Tooker ME. Multibreed genomic evaluations using purebred Holsteins, Jerseys, and Brown Swiss. *J Dairy Sci*. 2012;95:5378–83.
- Pryce JE, Gredler B, Bolormaa S, Bowman PJ, Egger-Danner C, Fuerst C, et al. Short communication: genomic selection using a multi-breed, across-country reference population. *J Dairy Sci*. 2011;94:2625–30.
- Raymond B, Bouwman AC, Wientjes YC, Schrooten C, Houwing-Duistermaat J, Veerkamp RF. Genomic prediction for numerically small breeds, using models with pre-selected and differentially weighted markers. *Genet Sel Evol*. 2018;50:49.
- Sarup P, Jensen J, Ostensen T, Henryon M, Sørensen P. Increased prediction accuracy using a genomic feature model including prior information on quantitative trait locus regions in purebred Danish Duroc pigs. *BMC Genet*. 2016;17:11.
- Speed D, Hemani G, Johnson MR, Balding DJ. Improved heritability estimation from genome-wide SNPs. *Am J Hum Genet*. 2012;91:1011–21.
- R Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2013.
- Thomasen JR, Guldbrandtsen B, Su G, Brøndum RF, Lund MS. Reliabilities of genomic estimated breeding values in Danish Jersey. *Animal*. 2012;6:789–96.
- van den Berg I, Boichard D, Lund MS. Sequence variants selected from a multi-breed GWAS can improve the reliability of genomic predictions in dairy cattle. *Genet Sel Evol*. 2016;48:83.

- van den Berg I, Meuwissen THE, Macleod IM, Goddard ME. Predicting the effect of reference population on the accuracy of within, across, and multibreed genomic prediction. *J Dairy Sci.* 2019;104:3155–74.
- VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, et al. Invited review: reliability of genomic predictions for North American Holstein bulls. *J Dairy Sci.* 2009;92(92):16–24.
- Wientjes YC, Bijma P, Vandenplas J, Calus MP. Multi-population genomic relationships for estimating current genetic variances within and genetic correlations between populations. *Genetics.* 2017;207:503–15.
- Wientjes YCJ, Calus MPL, Duenk P, Bijma P. Required properties for markers used to calculate unbiased estimates of the genetic correlation between populations. *Genet Sel Evol.* 2018;50:65.
- Wientjes, Y. C., R. F. Veerkamp, P. Bijma, H. Bovenhuis, C. Schrooten *et al.*, 2015 Empirical and deterministic accuracies of across-population genomic prediction. *Genetics Selection Evolution* 47: 5.
- Wientjes YC, Veerkamp RF, Bijma P, Bovenhuis H, Schrooten C, Calus MP. Empirical and deterministic accuracies of across-population genomic prediction. *Genet Sel Evol.* 2015;47:.
- Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet.* 2010;42:565–9.
- Yang, J., T. A. Manolio, L. R. Pasquale, E. Boerwinkle, N. Caporaso *et al.*, 2011 Genome partitioning of genetic variation for complex traits using common SNPs. *Nature genetics* 43: 519.
- Zhong S, Dekkers JC, Fernando RL, Jannink JL. Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. *Genetics.* 2009;192:355–64.
- hou L, Heringstad B, Su G, Guldbrandtsen B, Meuwissen TH, Svendsen M, et al. Genomic predictions based on a joint reference population for the Nordic Red cattle breeds. *J Dairy Sci.* 2014;97:4485–96.

4.9 Supplementary information

Supplementary Table 1: Estimated correlations between the 2 estimated genomic values (EGVs) from the multi-population, multiple genomic relationship matrices (MPMG) model. Standard error of estimates are in parenthesis.

GRM 1*	Correlation (EGV1, EGV2), ($h^2 = 0.3$)	Correlation (EGV1, EGV2) ($h^2 = 0.8$)
$r_g = 0.4$		
CSNPs_125	0.22 (0.14)	0.19 (0.08)
CSNPs_250	0.30 (0.16)	0.26 (0.08)
CSNPs_500	-0.04 (0.40)	0.03 (0.28)
$r_g = 0.6$		
CSNPs_125	0.18 (0.13)	0.16 (0.09)
CSNPs_250	0.23 (0.15)	0.22 (0.09)
CSNPs_500	-0.07 (0.34)	0.03 (0.25)
$r_g = 0.8$		
CSNPs_125	0.13 (0.12)	0.11 (0.09)
CSNPs_250	0.14 (0.15)	0.16 (0.09)
CSNPs_500	-0.07 (0.27)	0.02 (0.20)

*GRM 2 throughout the study was made from ~48k non-causal SNPs.

5

CHAPTER 5.

Using prior information from humans to
prioritize genes and gene-associated variants
for complex traits in livestock

Biaty Raymond^{1,2}, Loic Yengo³, Roy Costilla⁴, Chris Schrooten⁵, Aniek C. Bouwman¹,
Ben J. Hayes⁴, Roel F. Veerkamp¹, and Peter M. Visscher³.

¹Animal Breeding and Genomics, Wageningen University and Research, P.O. Box 338, 6700 AH Wageningen, The Netherlands; ²Biometris, Wageningen University and Research, 6700AA Wageningen, The Netherlands; ³Institute for Molecular Bioscience, University of Queensland, St. Lucia, 4072, Australia; ⁴Queensland Alliance for Agriculture and Food Innovation, University of Queensland, St. Lucia, 4072, Australia; ⁵CRV BV, P.O. Box 454, 6800 AL Arnhem, The Netherlands.

Accepted (PLoS Genetics)

Abstract

Genome-Wide Association Studies (GWAS) in large human cohorts have identified thousands of loci associated with complex traits and diseases. For identifying the genes and gene-associated variants that underlie complex traits in livestock, especially where sample sizes are limiting, it may help to integrate the results of GWAS for equivalent traits in humans as prior information. In this study, we sought to investigate the usefulness of results from a GWAS on human height as prior information for identifying the genes and gene-associated variants that affect stature in cattle, using GWAS summary data on samples sizes of 700,000 and 58,265 for humans and cattle, respectively. Using Fisher's exact test, we observed a significant proportion of cattle stature-associated genes (30/77) that are also associated with human height (odds ratio = 5.1, $p = 3.1\text{e-}10$). Result of randomized sampling tests showed that cattle orthologs of human height-associated genes, hereafter referred to as candidate genes (C-genes), were more enriched for cattle stature GWAS signals than random samples of genes in the cattle genome ($p=0.01$). Randomly sampled SNPs within the C-genes also tend to explain more genetic variance for cattle stature (up to 13.2%) than randomly sampled SNPs within random cattle genes ($p=0.09$). The most significant SNPs from a cattle GWAS for stature within the C-genes did not explain more genetic variance for cattle stature than the most significant SNPs within random cattle genes ($p=0.87$). Altogether, our findings support previous studies that suggest a similarity in the genetic regulation of height across mammalian species. However, with the availability of a powerful GWAS for stature that combined data from 8 cattle breeds, prior information from human-height GWAS does not seem to provide any additional benefit with respect to the identification of genes and gene-associated variants that affect stature in cattle

5.1 Introduction

Genome-wide association studies (GWAS) have become popular and powerful tools for identifying genes involved in complex traits aetiology, and for investigating the genetic architecture of such traits in animal and plant species. In humans for example, GWAS have been used to identify thousands of risk loci for complex diseases such as heart diseases and psychiatric disorders, and have led to the discovery of potential drug targets (Hindorff *et al.* 2009; Visscher *et al.* 2017). In plant and livestock species, the focus of GWAS has primarily been, to identify variants or genomic regions that underlie complex and economically important traits (Brachi *et al.* 2011; Sharma *et al.* 2015). Ultimately, the ambition, with respect to GWAS, is the identification of a subset of genomic variants that together explain the total genetic variance for complex traits.

The ability of GWAS to identify variants that are associated with complex traits depends on sample size. Large sample size is especially important for identifying rare variants with medium-sized effects and common variants with small effects. A classic example of the importance of sample size in GWAS is the trajectory of the number of loci that have been associated with human height over the past decade. That number rose from ~40 with a sample size between 13,000 to ~ 30,000 in 2008 (Gudbjartsson *et al.* 2008; Lettre *et al.* 2008; Weedon *et al.* 2008), to 180 with a sample size of 183,727 in 2010 (Allen *et al.* 2010). By 2014, the number of loci associated with human height had increased to 697 with a sample size of 253,288 (Wood *et al.* 2014), while the latest GWAS for human height from 2018 reported 3,290 near-independently associated loci, with a sample size of ~ 700,000 (Yengo *et al.* 2018). It should be noted that the increase noticed in the past five years also reflects improvements in data analysis methods (Yang *et al.* 2014). It is predicted that as sample sizes continue to increase, the number of loci reliably associated with human height will continue to increase (Visscher *et al.* 2017; Yengo *et al.* 2018).

While GWAS sample sizes have continued to increase in humans, the same is not true for livestock species. While there are millions of animals from livestock species that are regularly genotyped around the world, the data are rarely shared or combined for the purpose of GWAS, given that the data are commercially sensitive and proprietary. Up until recently, for example (e.g. Bouwman *et al.* 2018; Jiang *et al.* 2019; Purfield *et al.* 2019; Xiang *et al.* 2019), most GWAS in livestock have been carried out with cohort level data, with samples sizes varying from a few hundreds to a few thousands (Sharma *et al.* 2015). While GWAS in livestock have resulted in a

number of trait-associated loci, mainly those with large effect sizes, the joint effect of the discovered loci explain only small proportions of the total genetic variance for the traits of interest (Dekkers 2004). Given the sample-size limitation in most GWAS in livestock, it would be of scientific interest to know if the results of GWAS studies for complex traits in humans can be used as prior information to identify the genes and gene-associated variants that affect equivalent traits in livestock species.

The idea to utilize GWAS results for a given trait across species presupposes that the genes underlying the trait are conserved across species. There have been reports of genes and mutations affecting height in humans that are found to affect height in other mammalian species (see Kemper *et al.* 2012 for a review). In their study, Pryce *et al.* (2011) found that variants within 55 cattle genes, whose orthologs are associated with height in humans, were significantly associated with stature in cattle. The result of Pryce *et al.* (2011) is further evidence to suggest that the same genes may be underlying height in humans and cattle. Thus, instead of using a within-species GWAS to identify genes and gene-associated variants for complex traits as was done in Bouwman *et al.* (2018), a GWAS results for the same trait but from a different species could be utilized. With the availability of large sample sizes for GWAS and whole genome sequence data in both humans and cattle, we hypothesised that the evidence for similarity in the biology of height in humans and cattle will become stronger, and that use of GWAS results for height in humans as prior information to identify stature genes in cattle will become feasible. The objective of this study was to test this hypothesis. We therefore investigate the usefulness of GWAS results on human height as a prior information for identifying genes and gene-associated variants that affect stature in cattle. For this, we used GWAS summary data on samples sizes of ~700,000 and 58,265 for humans and cattle, respectively. We find that prior information from human height GWAS is beneficial for identifying genes and gene-associated variants affecting stature in cattle, only when cattle prior information is not available.

5.2 Results

5.2.1 Orthologous genes in humans and cattle are highly correlated in size but not in LD score or gene-effect on height

From the UMD3.1 cattle genome build, we identified 13,742 genes that have orthologs in humans and passed our quality control (QC). Within each species, we estimated the correlation between gene size and gene LD score, defined as the mean LD score of variants in the vicinity of the gene (Methods). We observed no significant ($p < 0.05$) correlation between gene size and gene LD scores across the 13,742 orthologous genes selected within the human genome ($r = 0.0029$, $p = 0.74$), and a significant correlation between gene size and gene LD scores within the cattle genome ($r = 0.028$, $p = 0.001$). We observed a significant positive correlation in gene-size ($r = 0.85$, $p = 2.2 \times 10^{-16}$) but no significant correlation in gene LD scores ($r = 0.012$, $p = 0.16$) between human genes and their cattle orthologs. Furthermore, we estimated the correlation between the effect of orthologous genes on height in humans and cattle, as a high correlation may be an indication that the genes perform similar function across species. To do this, we estimated the aggregated effect of all the SNPs within each orthologous gene on the phenotype (human height and cattle stature), which we refer to as gene-effect. The estimate of correlation between the effect of the orthologous genes on height was low, but significant ($r = 0.12$, $p = 2.2 \times 10^{-16}$).

5.2.2 Height-associated genes in humans and cattle significantly overlap

To address if the same genes underlie height in humans and cattle, we estimated the proportion of orthologous genes in or close to GWAS peaks for human height that are also identified in or close to GWAS peaks for stature in cattle, and performed Fisher's exact test to check if the proportion of overlap is more than can be expected by chance. We identified 77 cattle stature-associated orthologous genes based on the results of Bouwman *et al.* (2018). Out of the 77 cattle orthologous genes, 10 were also found to be associated with human height based on the result of Wood *et al.* (2014), Fig 5.1A. The proportion 10/77 is more than can be observed by chance based on the results of Fisher's exact test (odds ratio = 5.5, $p = 3.7 \times 10^{-5}$). The contingency table for the Fisher's exact test is provided in the S1 table. Details about the 10 overlapping genes are provided in S2 table.

Fig 5.1 Overlap between orthologous genes associated with height in humans based on Wood et al. (2014) that are also associated with stature in cattle based on Bouwman et al. (2018) (Panel A), and overlap between orthologous genes associated with height in humans based on Yengo et al. (2018) that are also associated with stature in cattle based on Bouwman et al. (2018) (Panel B).

The overlap was stronger (30/77) between the 77 cattle genes and human height-associated genes identified based on the results of Yengo *et al.* (2018), Fig 5.1B. The human height-associated genes were identified as those that were overlapping with or nearby the near-independent significant SNPs from the meta-GWAS. The proportion 30/77 is more than expected by chance based on the result of Fisher's exact test (odds ratio = 5.1, p-value = 3.1×10^{-10}). The contingency table for the Fisher's exact test is provided in the S3 table, and details about the 30 overlapping genes are provided in S4 table.

It is worth noting that the 77 orthologous cattle stature-associated genes identified from Bouwman *et al.* (2018), are among the largest in size in both the cattle and human genomes (Fig 5.2), which could inflate the likelihood of overlap. Therefore, we implemented a gene size-stratified Cochran-Mantel-Haenszel test to correct our estimates of overlap for potential bias. Our result show that accounting for gene-size still results in a significant overlap (p-value = 2.4×10^{-9}). The contingency table for the Cochran-Mantel-Haenszel test is provided in S5 table.

Fig 5.2: Distribution of gene-size (in cattle and humans) for the 77 orthologous genes that overlap with or are within 10kb of the lead SNPs from Bouwman *et al.* (2018). Genes were stratified into 10 quantiles based on gene size, with the shortest genes in 1 and the longest genes in 10.

5.2.3 Cattle stature-associated variants are enriched in human height-associated genes

To further check if human height-associated genes also affect stature in cattle, we tested if cattle genes prioritized based on information from a GWAS on human height are more enriched for GWAS signals for cattle stature than random sets of genes from the cattle genome. For this purpose, we utilized two sets of human height-associated genes. The first set comprised of 370 orthologous genes that were identified by Yengo *et al.* (2018) using the summary-based Mendelian Randomization (SMR) analysis (Zhu *et al.* 2016). The SMR analysis combines gene-expression data from multiple tissues and summary level GWAS data to identify genes whose expression level is associated with the trait. We identified the cattle orthologs of the SMR-based human genes that we refer to as SMR candidate (SMR-C) genes. The result of randomized sampling test (Methods) showed that the number of near-independently significant SNPs for cattle stature within and around the 370 SMR-C genes are not more than can be expected within and around random sets of cattle genes matched on size, SNP density and mean LD score (Fig 5.3).

Fig 5.3: Frequency distribution of the number of SNPs with near-independent significant effect on stature ($P < 5 \times 10^{-8}$) in 1,000 control tests for SMR-C genes. In each test, 370 cattle genes were sampled from a total of 13,865 genes that have human orthologs,

and near-independent significant SNPs within and 100kb (Column 1 – panels A,D,G), 50kb (Column 2 – panels B,E,H) or 5kb (Column 3 – panels C,F,I) either sides of the sampled genes were identified. The genes were sampled using stratified random sampling, with the genes stratified based on gene-size (Row 1 – panels A,B,C), number of SNPs in a gene (Row 2 – panels D,E,F) or gene LD score (Row 3 – panels G,H,I). The vertical red lines represents the number of near-independent significant SNPs ($P < 5 \times 10^{-8}$) associated with stature, identified within and around the 370 cattle (SMR-C) candidate genes that have been shown to be associated with height in humans. The p value represents the proportion of samples (out of 1,000) that have equal or higher level of enrichment for GWAS signals than the SMR-C genes.

The second set of human height-associated genes we used comprised of 1,184 orthologous genes that overlap with the 3,290 near-independently significant SNPs affecting human height identified through a conditional and joint effect (COJO) analysis in Yengo *et al.* (2018). The cattle orthologs of the COJO-based human genes were identified and referred to as COJO candidate (COJO-C) genes. We found that the number of near-independent SNPs for stature in cattle within and around the 1,184 COJO-C genes is more than can be observed within and around random sets of cattle genes matched on size, SNP density and mean LD score (Fig 5.4).

Fig 5.4: Frequency distribution of the number of SNPs with near-independent significant effect on stature ($P < 5 \times 10^{-8}$) in 1,000 control tests for COJO-C genes. In each test, 1,184 cattle genes were sampled from a total of 13,865 genes that have human orthologs,

and near-independent significant SNPs within and 100kb (Column 1 – panels A,D,G), 50kb (Column 2 – panels B,E,H) or 5kb (Column 3 – panels C,F,I) either sides of the sampled genes were identified. the genes were sampled using stratified random sampling, with the genes stratified based on gene-size (Row 1 – panels A,B,C), number of SNPs in a gene (Row 2 – panels D,E,F) or gene LD score (Row 3 – panels G,H,I). The vertical red lines represents the number of near-independent significant SNPs ($P < 5 \times 10^{-8}$) associated with stature, identified within and around the 1,184 cattle (COJO-C) candidate genes that have been shown to be associated with height in humans. The p value represent the proportion of samples (out of 1,000) that have equal or higher level of enrichment for GWAS signals than the COJO-C genes.

The difference in sample size (370 vs 1,184) between the SMR-C and the COJO-C genes may result in the observed difference in the significance of enrichment for cattle stature GWAS signals. However, we found that with the same sample size of 370, the COJO-C genes are more enriched for GWAS signals than the SMR-C genes or any random set of 370 genes matched based on gene-size (Fig 5.5). This indicates that the results observed for the SMR-C genes is not just an issue of power (small gene sample size).

Fig 5.5: Frequency distribution of the number of SNPs with near-independent significant effect on stature ($P < 5 \times 10^{-8}$) in 100 control tests. In each test, 370 out of 1,184 cattle (COJO-C) genes were sampled. The number of near-independent significant SNPs within and 100kb either side of the sampled genes were identified and compared to the number obtained within and 100kb either side of the 370 SMR-C genes. The vertical red dashed line represents the number of near-independent significant SNPs ($P < 5 \times 10^{-8}$) associated with stature, identified within and 100kb either side of the 370 SMR-C genes, that have been shown to be associated with height in humans through an SMR analysis. The p value represent the proportion of samples (out of 100) that have equal or lower level of enrichment for GWAS hits than the SMR-C genes.

The SMR analysis uses gene-expression data from multiple tissues. Given that not all tissues are relevant for a given trait, it is critical that gene-expression data from a trait-relevant tissue is used. Hence, we tested if the level of enrichment for GWAS signals in the SMR-C genes will be more than can be expected from random sets of genes if the selection criteria were to be more stringent. In particular, we required that the human ortholog of the gene must show evidence of differential gene expression in three or more tissues. With this selection criterion, we identified a

subset of 121 (SMR-MT-C) genes instead of 370. Our result show that the level of enrichment for GWAS hits nearby the 121 SMR-MT-C genes is more than can be expected from any set of 121 randomly sampled genes, with $p = 0.02$ (Fig 5.6). The p values in the randomized sampling test were calculated as the proportion of randomly sampled gene-sets that had equal or higher level of enrichment for GWAS hits as compared to the candidate gene sets.

Fig 5.6: Frequency distribution of the number of SNPs with near-independent significant effect on stature ($P < 5 \times 10^{-8}$) in 1,000 control tests for SMR-MT-C genes.

In each test, 121 cattle genes were sampled from a total of 13,865 genes that have human orthologs. Stratified random sampling was carried out, with genes stratified based on gene-size. The number of near-independent significant SNPs within and 100k either side of the sampled genes were identified. The vertical dashed line represents the number of near-independent significant SNPs ($P < 5 \times 10^{-8}$) associated with stature, identified within and 100kb either side of the 121 cattle (SMR-MT-C) candidate genes, whose human orthologues show evidence of differential expression in 3 or more tissues. The p value represent the proportion of samples (out of 1,000) that have equal or higher level of enrichment for GWAS hits than the SMR-MT-C genes.

5.2.4 Accuracy of cattle variants selection using within- and across-species prior information

Next, we quantified the proportion of stature variance explained by pre-selected cattle SNPs based on GWAS results for height in humans and stature in cattle. First, we tested the hypothesis that cattle SNPs that are pre-selected based on two information sources 1) genes identified from GWAS results for human height, and 2) Cattle stature GWAS significance level, will explain more genetic variance for stature than SNPs that are pre-selected based on only cattle stature GWAS significance level. To test this hypothesis, we identified the near-independent significant SNPs within 100kb either side of the SMR-C and COJO-C genes, that also segregate in a New Zealand Holstein (NZH) cattle population: (80 SMR-C SNPs, 391 COJO-C SNPs). The NZH population was not included in the cattle meta-GWAS analysis (Bouwman *et al.* 2018) and is therefore ideal as a validation population. It must be noted, however, that some of the bulls in the NZH population are descendants of bulls from the Dutch Holstein population which was included in the meta-GWAS. The result of randomized sampling test showed that the identified candidate SNPs do not explain more genetic variance for stature in NZH population than the most significant SNPs for cattle stature within randomly sampled cattle genes (Fig 5.7).

Fig 5.7: Frequency distribution of the percentage of genetic variance in New Zealand Holstein population, explained by SNPs in 1,000 control tests for SMR-C SNPs. In each test, the most significant SNPs for stature within a sampled gene-set, that had at least 20 copies of the minor allele in the population were used to construct the first GRM fitted in a GREML model. A second GRM was made with randomly sampled 50k SNPs (excluding SNPs in the first GRM) across the genome, to account for population structure. Figure A shows the result of control analysis for the SMR-C SNPs, in which 80 top SNPs in each gene set were used to construct the first GRM, while figure B shows the result of control analysis for the COJO-C SNPs, in which 391 top SNPs within each gene set was used to construct the first GRM. The red vertical lines represents the percentage of genetic variance explained by the candidate SMR-C and COJO-C SNPs, respectively.

Secondly, we tested the hypothesis that cattle SNPs pre-selected using GWAS result for human height will explain more genetic variance for stature, than random set of SNPs from random cattle genes. To test the hypothesis, we randomly sampled 5 SNPs per gene to avoid having a large number of SNP in a set. For the SMR-C genes, this resulted in 1,777 unique SNPs that were segregating in NZH. For the COJO-C genes, the number of unique SNPs extracted was 5,781. Although the results are not statistically significant, we found that in $\sim 90\%$ of time, the candidate SNPs explain more genetic variance for stature in NZH population than randomly sampled SNPs from random sets of genes from the cattle genome (Fig 5.8).

Fig 5.8: Frequency distribution of the percentage of genetic variance in New Zealand Holstein population, explained by SNPs in 1,000 control tests for COJO-C SNPs. In each test, SNPs were sampled randomly (matched by allele frequency to candidate SNPs) within a sampled gene-set (5 SNPs were sampled per gene, within each gene-set). SNPs had at least 20 copies of the minor allele in the validation population. Each sampled SNP set was used to construct the first GRM. A second GRM was made with randomly sampled 50k SNPs (excluding all SNPs in the first GRM) across the genome, to account for population structure. Figure A shows the result of control analysis for the SMR-C SNPs, in which 1,777 random SNPs in each gene set were used to construct the first GRM, while figure B shows the result of control analysis for the COJO-C SNPs, in which 5,781 random SNPs within each gene set was used to construct the first GRM. The red vertical lines represents the percentage of genetic variance explained by the candidate SMR-C and COJO-C SNPs, respectively.

5.3 Discussion

A general expectation with respect to conserved orthologous genes is that they perform the same or very similar functions in different species and thus the functions of the genes in one species could be inferred to another, especially when such functions are critical for evolutionary fitness (Koonin 2005; Descorps-Declère *et al.* 2008; Gabaldón and Koonin 2013). Several experiments in the past have reported a similarity in the function of orthologous genes in different species, although these studies focused on genes causing mendelian disorders, (see Dolinski and Botstein 2007 for a review). For a highly polygenic trait such as height, it is not yet established if the many genes underlying the trait in different species are conserved in their function, and if height-associated genes in one species can be used as prior information to identify the genes affecting height in a different species. In this study, we sought to establish if human height-associated orthologous genes, with mutations affecting height identified through GWAS, can be used as prior information to identify the genes and gene-associated variants that affect stature in cattle.

Due to LD, GWAS does not always result in the identification of the causal variants and/or genes for a given trait. In many cases, variants identified as significantly associated with a trait are proxies of unknown causal variants. One way of validating candidate genes identified from GWAS is to test if the genes are also associated with the same trait in a different species, given that LD pattern and thus the impact of LD on GWAS is different in different species. In our study, we identified a significant proportion (~39%) of cattle stature-associated orthologous genes that are also found to be associated with human height. Given that they appear as candidate genes for height in two species that separated over ~90 million years ago (Murphy *et al.* 2004), the 30 overlapping genes cannot be considered an artifact of cryptic population structure. We found that 6 out of the 30 genes: ADAM12, SIK3, AXL, INSR, SPRED1 and PAPP2 are involved in the Signaling by Receptor Tyrosine Kinases pathway, a pathway that is responsible for signaling of the growth hormone (GH) (Carter-Su *et al.* 2016). Two of the overlapping genes ADAMTS17 and ADAMTS14 participate in the pathway: Defective B3GALT1 causes Peters-plus syndrome (PpS). The Peters-plus syndrome is a recessive disorder that results in, among other things, short stature and delay in growth (Heinonen and Mäki 2009). Altogether, evidence suggests that height in humans and stature in cattle are biologically similar traits and that the orthologous genes associated with height in humans can be considered validated if they are also found to be associated with stature in cattle.

With respect to the identification of genes and gene-associated variants that affect stature in cattle, our results suggest that across-species prior information (results of GWAS for human height) is less informative than within-species prior information (results of GWAS for cattle stature). There are two reasons likely underpinning this finding. Firstly, the cattle (within-species) GWAS (Bouwman *et al.* 2018) was for a trait that is highly heritable. It is easier for GWAS to identify the signals underlying a highly heritable trait such as stature than for a lowly heritable trait e.g. fertility, given that at low heritability, there is larger environmental effect in the phenotype, which acts as noise around the genetic effects. In that sense, the benefit of using prior information from humans to identify genes and gene-associated variants for complex traits in cattle is expected to be higher for lowly heritable cattle traits than for highly heritable traits. For example, Costilla *et al.* (2020) estimated the proportion of additive genetic variance for age at puberty in cattle explained by SNPs in and around cattle orthologous genes that were associated with age at menarche in humans. They showed that the SNPs in and around the candidate genes explained 11.2% of the total additive genetic variance for age at puberty in cattle, which was about twice the variance explained by SNPs in random gene-sets of similar size and SNP number.

Secondly, the within-species GWAS was large in terms of sample size and therefore was powerful enough such that the human prior information does not provide additional information. However, the GWAS study of Bouwman *et al.* (2018) is an exceptional collaboration and data sharing in the livestock genetics community. An important factor for the success of the collaboration was that stature is generally not considered an economic trait, and in that sense, data on stature is less sensitive to share. Studies as large as that of Bouwman *et al.* (2018) for economically important traits, especially those that are difficult or expensive to measure, are less common. For these kinds of traits, GWAS will most likely be underpowered, given that they are carried out at population level, where, in most cases, sample sizes are limiting. Identification of the genes and gene-associated variants that affects such traits will be aided by the integration of the huge resources, namely, GWAS results that are publicly available for complex traits in humans. Furthermore, for traits in cattle where powerful GWAS are not possible for whatever reason, using prior information from human GWAS from similar traits will likely aid the identification of the underlying genes and the associated variants.

The method of determining the genes that underlie height in humans is very important, as this may determine whether or not the orthologs affects height in other species. In this study, we focused on two sets of orthologous genes that were determined to have an effect on human height. The COJO gene set are those that were prioritized as encompassing or close to the top GWAS signals for height in humans from Yengo *et al.* (2018). The cattle orthologs of these genes were more enriched for cattle stature signals and captured more genetic variance for stature than random sample of genes in the cattle genome. The second set of genes were those that were prioritized based on the summary-based mendelian randomization analysis (SMR) in Yengo *et al.* (2018). These are genes whose expression levels and association with height are assumed to be controlled by the same causal variant, and in that sense, assumed to be the most functionally relevant genes for human height. In our analysis, however, the cattle orthologs of the SMR-based human genes were not more enriched for cattle stature GWAS signals than random sets of genes from the cattle genome. There are two possible explanations for this observation. First, some genes are known to show tissue specific gene expression (Nica *et al.* 2011). For their SMR analysis, Yengo *et al.* (2018) used the GTEx-v7 database (Gtex Consortium 2015) that contains expression data of several genes in multiple tissues, not necessarily height related tissues. It is possible, therefore, that some of the genes showed differential level of gene-expression only in specific tissues that are not related to human height. The orthologs of such genes are less likely to affect height in other species. In contrast, genes with consistent differential expression in multiple tissue are more likely to maintain their effect across species. Our result, although only marginally significant ($p=0.02$), suggests that to be the case. Given that the specific tissues that are relevant for specific traits of interest are not always known, the results of SMR analysis that uses gene expression data in only a single tissue, should be taken with caution. The second possible explanation might just be that the cattle meta-GWAS analysis of Bouwman *et al.* (2018) was not powerful enough to detect important cattle stature variants within the cattle orthologs of the SMR-based human genes.

Potentially, GWAS for complex traits in humans could be utilized to aid precision genomic prediction in livestock. For example, knowledge about the position of candidate genes from GWAS in humans or the proportion of genetic variance for the trait that is explained by variants in or around the candidate genes, can be used to specify priors for genomic prediction models in livestock. A non-linear genomic prediction model that can be used for this purpose is the BayesR by segment (BayesRS) model (Brøndum *et al.* 2012). In BayesRS, variant effects are assumed to

belong to a mixture of four normal distributions, with the different distribution reflecting the different proportions of genetic variance explained by different genomic segments. Effectively, the BayesRS model can be specified such that the prior probability for a given variant depends on whether it is within or nearby an important gene identified from a human GWAS. An extended BayesRS model is the BayesRC model (Macleod *et al.* 2016), which also allows for the explicit incorporation of prior knowledge in GP. In BayesRC, information on genes or variants from a human based GWAS can be used to assign variants into two or more classes (C). Other methods of using information from GWAS in humans for genomic prediction in livestock include weighting variants differently based on information from human GWAS in the construction of GRMs (e.g. Su *et al.* 2014; Fragomeni *et al.* 2017), and fitting multiple GRMs in BLUP, with the different GRMs made from different classes of markers (e.g. Speed and Balding 2014; Raymond *et al.* 2018).

Altogether, our results suggest that human height and cattle stature are biologically similar traits. For the purpose of identifying the genes and gene-associated variants that affect stature in cattle, prior information from current human-height GWAS, does not provide more benefit than what a powerful GWAS for cattle-stature already provides. We must state that our findings cannot easily be generalized to other complex traits. This is especially so for traits that have been under strong artificial selection in cattle, given that such traits might have evolved differently and rapidly than in humans. More research is needed to make such a generalization.

5.4 Materials and Methods

5.4.1 Data

The summary level results from meta-GWAS for human height and cattle stature and other data sets used in the study are summarised in Table 5.1.

Table 5.1: A summary of all data sets used in the study

Species	Data type	Sample size	# of SNPs	References
Human	GWAS summary-level result	253,288	2,550,858	(Wood <i>et al.</i> 2014)
Human	GWAS summary-level result	693,529*	2,334,001	(Yengo <i>et al.</i> 2018)
Cattle	GWAS summary-level result	58,265	25.4 million	(Bouwman <i>et al.</i> 2018)
Human	Individual-level genotypes	8,552	16,652,994	(Sonnega <i>et al.</i> 2014)
Cattle	Individual-level genotypes	WGS 1,147	30.3 million	(Daetwyler <i>et al.</i> 2014)
Human	LD scores	10,000	17,613,557	(UK10K Consortium 2015)
Cattle	Individual-level genotypes and phenotype	975	14,341,737	(Raymond <i>et al.</i> 2018)

*Mean sample size. WGS: whole genome sequence

5.4.2 Comparison of gene characteristics in humans and cattle

Many genes are conserved across species, therefore, comparison of traits across humans and cattle can be made at the gene level. For this study, we utilized information on orthologous genes between humans and cattle. Out of the 24,617 cattle genes identified in the UMD3.1 cattle genome build (Elsik *et al.* 2009) and using Biomart Ensemb (Chen *et al.* 2010), we identified 13,742 genes that have orthologs in humans and passed quality control (QC). The QC thresholds used were: identity of the cattle gene to human gene > 70%; cattle gene order conservation score > 70%; cattle gene orthology confidence =1, orthology type: 1 to 1.

Within each species (humans and cattle), we estimated the correlation between gene size and LD score of orthologous genes to see if longer genes have higher or lower level of LD than shorter genes. Furthermore, we estimated the correlation between human gene-size and cattle gene-size for the orthologous genes, and did the same for gene-based LD score. We calculated gene-size by subtracting the start, in base pairs (bp), from the end position of the gene and divided the value by 1,000 to obtain gene size in kilo base pairs (kbp). For humans, we used already calculated

LD scores on UK10K (UK10K Consortium 2015) whole genome sequence variants. The LD score for each SNP was calculated as the sum of LD (r^2) between the SNP and all other SNPs in a 200kb window, with 100kb overlap between two adjacent segments. The mean LD score of a gene was calculated as the mean LD score of all the SNPs within the gene. For cattle, we used GCTA (Yang *et al.* 2011) to calculate the LD scores for sequence variants on individuals from the Run4 of the 1,000 bulls genomes project dataset (Daetwyler *et al.* 2014) that have 3 or more copies of the minor allele. The LD score for cattle SNPs was calculated with the same specifications as was used for human LD score calculations.

A high correlation between the effect of orthologous genes on human height and cattle stature may be an indication that the genes perform similar function across the species. Using the fastBat function in GCTA (Bakshi *et al.* 2016), and GWAS summary level results, we estimated the aggregated effect of all the SNPs within each orthologous gene on the phenotype (human height and cattle stature), which we refer to as gene-effect. We then calculated the correlation between the effect of the genes in both species. For the fastBat analysis to estimate the effect of the orthologous genes on human height, we used summary-level GWAS result from Yengo *et al.* (2018) and individual level imputed genotype data on individuals from the Health and Retirement Study (Sonnega *et al.* 2014) as LD reference. Similarly, for the fastBat analysis to estimate the effect of the orthologous genes on cattle stature, we used the summary-level result from Bouwman *et al.* (2018) and the full sequence data on all individuals from Run4 of the 1,000 Bull Genomes Project (Daetwyler *et al.* 2014) as LD reference. For human data, it was shown previously that the fastBat analysis loses power when SNPs are in very high or perfect LD, hence the recommendation to apply LD threshold of 0.9 (Bakshi *et al.* 2016). We therefore used an LD threshold of 0.9 in the fastBat analysis of human data and set a lower LD threshold of 0.5 for the fastBat analysis of cattle data, given the high average LD between SNPs in the cattle genome. Outputs from the fastBat analysis were gene-based chi-squared (calculated as the sum of chi-squared values for all the SNPs within and 50kb either side of a gene), and the corresponding gene-based p-value that indicate whether or not the effect of a gene on height (or stature in cattle) is significantly different from zero.

5.4.3 Overlap of height-associated genes in humans and cattle

We estimated the proportion of orthologous genes in or close to GWAS peaks for human height that are also identified in or close to GWAS peaks for stature in cattle, and tested if the overlap is more than can be expected by chance. We identified the genes potentially controlling stature in cattle as those that overlap with, or are within 10kbp of the 163 lead SNPs for stature as reported in Bouwman *et al.* (2018). We identified 77 genes that have orthologs in humans. We checked the proportion of the 77 cattle genes that were also identified based on results from Wood *et al.* (2014) and tested if the proportion is more than can be expected by chance. From Wood *et al.* (2014), we identified and used 368 orthologous genes that overlap with, or are within 10kb of the 649 lead SNPs for human height.

Given the larger sample size used by Yengo *et al.* (2018), we wanted to know if there is a stronger overlap between the 77 cattle genes from Bouwman *et al.* (2018), and human height genes from Yengo *et al.* (2018), as against those from Wood *et al.* (2014). To do this, we identified all human orthologous genes that overlap with, or are within 10kb of the 3,290 near independent significant SNPs from Yengo *et al.* (2018), 1,553 genes in total. We then checked the proportion of the 77 cattle genes that are part of the 1,553 identified human orthologous genes and tested if the proportion is more than can be expected by chance. In both cases, we used Fisher's exact test.

There is a chance that genes are enriched for GWAS signals and are identified as important genes for height in both humans and cattle, simply because they are large genes (Mirina *et al.* 2012). Thus, we implemented the Cochran-Mantel-Haenszel test (Cochran 1954; Mantel and Haenszel 1959) to test for the association between cattle stature genes and human height genes, while at the same time, accounting for a possible confounding effect of gene size. We implemented the test only for the overlap between the 77 cattle stature genes identified from Bouwman *et al.* (2018), and human height genes identified from Yengo *et al.* (2018).

5.4.4 Enrichment for cattle stature GWAS signals in and around human height-associated genes

The aim in this section was to quantify the usefulness of information from human height GWAS in the prioritization of stature genes in cattle, an analysis similar to Pryce *et al.* (2011). Specifically, we tested the hypothesis that cattle genes prioritized based on information from a GWAS on human height are more enriched for cattle stature GWAS signals than random sets of genes from the cattle genome. For the

analysis, we used two different sets of human-height genes. First, we used the 610 human height genes identified by Yengo *et al.* (2018) using the Summary-based Mendelian Randomization (SMR) analysis (Zhu *et al.* 2016). The SMR analysis identifies genes whose expression level and association with the trait are assumed to be controlled by the same causal variant. We identified 370 cattle orthologs that will henceforth be referred to as SMR candidate (SMR-C) genes. The number of near-independent significant SNPs for cattle stature within and around the 370 SMR-C genes were then identified. To test if this number is more than can be observed with a random set of 370 cattle genes, we implemented a validation analysis, using randomized permutation test as follows: all 13,742 cattle genes that have human orthologs were stratified into 10 quantiles based on gene-size, number of SNPs within a gene or gene-based LD score, with the genes stratified separately for each criterion. In each quantile, we identified the number of SMR-C genes and subsequently sampled the same number of control genes from the quantile. Using this method, we sampled 370 control genes 1,000 times. We identified the number of near-independent SNPs within and 100kbp, 50kbp or 5kbp either side of the genes in each sampled control gene-set and checked if the number is more, equal or less than was observed within and 100kbp, 50kbp or 5kbp either side of the 370 SMR-C genes, respectively. For the GWAS signal enrichment analysis, near-independent significant SNPs for stature refers to SNPs obtained following a clumping analysis in plink 1.9 (Purcell and Chang 2015), using the summary level result from Bouwman *et al.* (2018) and the fully sequenced individuals from Run4 of the 1000 Bull Genomes Project (Daetwyler *et al.* 2014) as LD reference. The following parameters thresholds were set in the clumping analysis: significance level for index SNPs: 5×10^{-8} ; secondary significance level for clumped SNPs: 5×10^{-8} ; LD threshold for clumping: 0.2; physical distance threshold for clumping: 0.5mbp.

Secondly, we mapped all the 3,290 near-independent significant SNPs affecting human height identified through a conditional and joint effect (COJO) analysis by Yengo *et al.* (2018) to human genes, using Biomart Ensembl and the GRCh37 human reference genome. We identified 1,184 cattle orthologs of the COJO-based human height genes. These 1,184 cattle orthologous genes will henceforth be referred to as COJO candidate (COJO-C) genes. We carried out the same validation analysis as described for the SMR-C genes, with 1,184 control genes sampled 1,000 times.

In addition to possible differences in the usefulness of human prior information used to prioritize the SMR-C and the COJO-C cattle gene sets, respectively, difference in gene sample size (370 vs 1,184) can result in a difference in the level of enrichment

for GWAS signals. To account for the difference in gene sample size, we took a sample of 370 genes from the 1,184 COJO-C genes 100 times, and checked the number of near-independent SNPs within and 100kbp either side of the genes in each sample. We then checked if the number of near independent SNPs for stature in each sample is more or less than was observed within and 100kb either side of the 370 SMR-C genes.

Furthermore, we tested if the level of enrichment for GWAS hits within the SMR-C genes is stronger with more stringent selection criterium, namely, evidence of differential expression in multiple tissues. To do this, we identified 182 out of the 610 human height genes prioritized by Yengo *et al.* (2018) through SMR, that have differential gene expression evidence in 3 or more tissues. We located the cattle orthologs of the 182 genes, with a total of 121 remaining after QC. These 121 cattle orthologs genes will henceforth be referred to as: SMR evidence in multiple tissues candidate (SMR-MT-C) genes. We implemented a validation analysis as earlier described, to test if the number of near-independent SNPs within the 121 SMR-MT-C genes are more than can be observed within and around sampled control genes. In this case, we used stratified random sampling to sample 121 control genes 1,000 times, with genes stratified only based on gene size, as this gives similar results compared to when genes are stratified based on gene-based LD score or SNP density.

5.4.5 Accuracy of cattle variants selection using within- and across-species prior information

The aim in this section was to quantify the accuracy (in terms of the proportion of genetic variance explained for stature), of pre-selecting cattle SNPs based on GWAS results for height in humans. For the analysis, deregressed estimated breeding values (dEBVs) for stature and effective daughter contributions (EDCs), for 975 New Zealand Holstein (NZH) bulls was used. The EDCs were used as weights for the dEBVs in all analyses. To estimate the proportion of genetic variance explained by SNPs in the NZH population, a univariate GREML model with two genomic relationship matrices (GRMs) was fitted using the MTG2 software (Lee and Van Der Werf 2016) follows:

$$y = 1\mu + Wg_1 + Wg_2 + e,$$

where y is a vector containing the phenotypes (dEBVs of the NZH bulls), μ is the mean, W is an incidence matrix linking phenotypes to genetic effects, g_1 and g_2 are vectors of genetic effects corresponding to the two GRMs fitted simultaneously in the model, and e is a vector of residual effects. In this study, the first GRM (GRM₁)

was made using the SNPs of interest and the second GRM (GRM₂) was made from 50k SNPs that were randomly sampled across the genome, excluding all SNPs in GRM₁, and meant to account for possible population structure. Genetic effects in g_1 and g_2 are assumed to be normally distributed as $g_1 \sim N(0, \text{GRM}_1 \sigma_a^2)$ and $g_2 \sim N(0, \text{GRM}_2 \sigma_a^2)$, where σ_a^2 is the genetic variance for stature. Residual effects were assumed to be normally distributed as $e \sim N(0, D \sigma_e^2)$, where σ_e^2 is the residual variance and D is a diagonal matrix that contains the inverse of EDCs, which were used as weights for the dEBVs. SNPs were included in the analysis that had at least 20 copies of the minor allele in NZH population. Thus, MAF threshold was set to $\frac{20}{975} * 2 = 0.04$. We tested two hypotheses as follows:

Hypothesis 1: Cattle SNPs that are pre-selected based on a combination of two information sources: 1) genes identified from GWAS results for human height, and 2) Cattle stature GWAS significance level, will explain more genetic variance for stature than SNPs that are pre-selected based from random cattle genes based on their significance level in a GWAS for cattle stature. To test this hypothesis, we identified the near-independent significant SNPs for stature in and 100kb either side of the SMR-C and COJO-C genes, that segregate in NZH (80 SMR-C SNPs, 391 COJO-C SNPs). We used all the identified candidate SNPs to make GRM₁ in GCTA [40]. For the control analysis, we sampled 1,000 control gene sets, and for each control gene set we ranked the SNPs based on significance level for association with stature and sampled the top 80 (SMR) or 391 (COJO) SNPs. So each set of control SNP set had the same number of SNPs as the candidate SNP set and was used to calculate GRM₁. The proportion of total genetic variance explained by each set of control SNPs were estimated and compared to the genetic variance explained by the candidate SNPs.

Hypothesis 2: Cattle SNPs that are pre-selected based on GWAS result for human height will explain more genetic variance for stature, than random set of SNPs from random cattle genes. In this case, the candidate SNPs were all the SNPs identified within and 100kb either side of the SMR-C and COJO-C genes, respectively. We randomly sampled 5 SNPs per gene to avoid a large number of SNP in high LD with each other, given the long range LD in commercial cattle populations. For the SMR-C genes, this resulted in 1,777 unique SNPs that were segregating in NZH. For the COJO-C genes, the number of unique SNPs extracted was 5,781. Each of these SNP set was used to calculate GRM₁. For the control analysis, the same number of SNPs as the candidate SNPs (1,777 and 5,781 SNPs respectively) were randomly sampled from within and 100kb around the genes in each of the 1,000 control gene-sets and the SNPs were matched for allele frequency to the candidate SNPs.

5.5 References

- Lango Allen H, Estrada K, Lettre G, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*. 2010;467:832-838.
- Bakshi, A., Zhu, Z., Vinkhuyzen, A. *et al.* Fast set-based association analysis using summary data from GWAS identifies novel gene loci for human complex traits. *Sci Rep*. 2016;6:32894.
- Bouwman AC, Daetwyler HD, Chamberlain AJ, Ponce CH, Sargolzaei M, Schenkel FS, et al. Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. *Nat Genet*. 2018;50:362–7.
- Brachi, B., Morris, G.P. & Borevitz, J.O. Genome-wide association studies in plants: the missing heritability is in the field. *Genome Biol*. 2011;12:232.
- Brøndum, R.F., Su, G., Lund, M.S. *et al.* Genome position specific priors for genomic prediction. *BMC Genomics*. 2012;13:543.
- Carter-Su, C., J. Schwartz and L. S. Argetsinger. Growth hormone signaling pathways. *Growth Hormone & IGF Research*. 2016;28: 11-15.
- Chen, Y., F. Cunningham, D. Rios, W. M. McLaren, J. Smith et al. Ensembl variation resources. *BMC Genomics*. 2010;11: 293.
- Cochran, W. G., 1954 Some methods for strengthening the common χ^2 tests. *Biometrics* 10: 417-451.
- Consortium, UK10K. The UK10K project identifies rare variants in health and disease. *Nature*. 2015;526: 82-90.
- Costilla, R., C. Warburton and B. Hayes. Genetic control of fertility traits across species: variance in tropical beef heifers'age at puberty explained by genes controlling age at menarche in women. In: *Proceedings of the 23rd AAABG conference: 27 October - 1st November 2019, Armidale, New South Wales; 2014*, pp. 258-261.
- Daetwyler, H. D., A. Capitan, H. Pausch, P. Stothard, R. Van Binsbergen et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nat. Genet*. 2014 ;46:858.
- Dekkers, J. C. Commercial application of marker-and gene-assisted selection in livestock: strategies and lessons. *Journal of Animal Science*. 2004;82: E313-E328.
- Descorps-Declère, S., F. Lemoine, Q. Sculo, O. Lespinet and B. Labedan. The multiple facets of homology and their use in comparative genomics to study the evolution of genes, genomes, and species. *Biochimie*. 2018;90: 595-608.
- Dolinski, K., and D. Botstein. Orthology and functional conservation in eukaryotes. *Annu. Rev. Genet*. 2007;41: 465-507.
- Elsik, C. G., R. L. Tellam and K. C. Worley. The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science*. 2009;324:522-528.

- Fragomeni, B. O., D. A. Lourenco, Y. Masuda, A. Legarra and I. Misztal. Incorporation of causative quantitative trait nucleotides in single-step GBLUP. *Genetics Selection Evolution*. 2017;49: 59.
- Gabaldón, T., and E. V. Koonin. Functional and evolutionary implications of gene orthology. *Nature Reviews Genetics*. 2013;14:360-366.
- GTEX Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*. 2015;348: 648-660.
- Gudbjartsson, D. F., G. B. Walters, G. Thorleifsson, H. Stefansson, B. V. Halldorsson et al. Many sequence variants affecting diversity of adult human height. *Nature Genetics*. 2008;40: 609.
- Heinonen, T. Y., and M. Mäki. Peters'-plus syndrome is a congenital disorder of glycosylation caused by a defect in the β 1, 3-glucosyltransferase that modifies thrombospondin type 1 repeats. *Annals of Medicine*. 2009;41: 2-10.
- Hindorff, L. A., P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. In: *Proceedings of the National Academy of Sciences*. 2009 ;106:9362-9367.
- Jiang, J., L. Ma, D. Prakapenka, P. M. VanRaden, J. B. Cole et al. A large-scale genome-wide association study in US Holstein cattle. *Frontiers in Genetics*. 2019;10: 412.
- Kemper, K. E., P. M. Visscher and M. E. Goddard. Genetic architecture of body size in mammals. *Genome Biology*. 2012;13: 244.
- Koonin, E. V. Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.* 2005;39:309-338.
- Lee, S. H., and J. H. Van der Werf. MTG2: an efficient algorithm for multivariate linear mixed model analysis based on genomic information. *Bioinformatics*. 2016; 32:1420-1422.
- Lettre, G., A. U. Jackson, C. Gieger, F. R. Schumacher, S. I. Berndt et al. Identification of ten loci associated with height highlights new biological pathways in human growth. *Nature Genetics*. 2008;40: 584.
- MacLeod, I., P. Bowman, C. Vander Jagt, M. Haile-Mariam, K. Kemper et al. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. *BMC Genomics*. 2016;17: 144.
- Mantel, N., and W. Haenszel. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*. 1959;22:719-748.
- Mirina, A., G. Atzmon, K. Ye and A. Bergman Gene size matters. *PloS One*. 2012;7: e49093.
- Murphy, W. J., P. A. Pevzner and S. J. O'Brien. Mammalian phylogenomics comes of age. *TRENDS in Genetics*. 2004;20:631-639.
- Nica, A. C., L. Parts, D. Glass, J. Nisbet, A. Barrett et al. The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genetics*. 2011;7:e1002003.

- Pryce, J. E., B. J. Hayes, S. Bolormaa and M. E. Goddard. Polymorphic regions affecting human height also control stature in cattle. *Genetics*. 2011;187: 981-984.
- Purcell, S., and C. Chang. PLINK 1.9. URL <https://www.cog-genomics.org/plink2>. 2015.
- Purfield, D., R. Evans and D. Berry. Reaffirmation of known major genes and the identification of novel candidate genes associated with carcass-related metrics based on whole genome sequence within a large multi-breed cattle population. *BMC Genomics*. 2019;20: 720.
- Raymond, B., A. C. Bouwman, Y. C. Wientjes, C. Schrooten, J. Houwing-Duistermaat et al. Genomic prediction for numerically small breeds, using models with pre-selected and differentially weighted markers. *Genetics Selection Evolution*. 2018;50:49.
- Sharma, A., J. S. Lee, C. G. Dang, P. Sudrajad, H. C. Kim et al. Stories and challenges of genome wide association studies in livestock—a review. *Asian-Australasian Journal of Animal Sciences*. 2015;28:1371.
- Sonnega, A., J. D. Faul, M. B. Ofstedal, K. M. Langa, J. W. Phillips et al. Cohort profile: the health and retirement study (HRS). *International Journal of Epidemiology*. 2014;43:576-585.
- Speed, D., and D. J. Balding. MultiBLUP: improved SNP-based prediction for complex traits. *Genome Research*. 2014;24:1550-1557.
- Su, G., O. F. Christensen, L. Janss and M. S. Lund. Comparison of genomic predictions using genomic relationship matrices built with different weighting factors to account for locus-specific variances. *Journal of Dairy Science*. 2014;97: 6547-6559.
- Visscher, P. M., N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy et al. 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics*. 2017;101:5-22.
- Weedon, M. N., H. Lango, C. M. Lindgren, C. Wallace, D. M. Evans et al. Genome-wide association analysis identifies 20 loci that influence adult height. *Nature Genetics*. 2008;40:575.
- Wood, A. R., T. Esko, J. Yang, S. Vedantam, T. H. Pers et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics*. 2014;46:1173.
- Xiang, R., I. v. d. Berg, I. M. MacLeod, B. J. Hayes, C. P. Prowse-Wilkins et al. Quantifying the contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits. In: *Proceedings of the National Academy of Sciences*. 2019;116:19398-19408.
- Yang, J., S. H. Lee, M. E. Goddard and P. M. Visscher. GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*. 2011;88: 76-82.
- Yang, J., N. A. Zaitlen, M. E. Goddard, P. M. Visscher and A. L. Price. Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics*. 2014;46:100.

- Yengo, L., J. Sidorenko, K. E. Kemper, Z. Zheng, A. R. Wood et al. Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. *Human Molecular Genetics*. 2018;27: 3641-3649.
- Zhu, Z., F. Zhang, H. Hu, A. Bakshi, M. R. Robinson et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics*. 2016;48: 481.

5.6 Supporting information

S1 Table: Contingency table showing the overlap of cattle stature genes from Bouwman *et al.* (2018) with human height genes from Wood *et al.* (2014).

	In Wood <i>et al.</i> (2014)	Not in Wood <i>et al.</i> (2014)	Marginal totals (rows)
In Bouwman <i>et al.</i> (2018)	10	67	77**
Not In Bouwman <i>et al.</i> (2018)	358	13,307	13,665
Marginal totals (columns)	368***	13,374	13,742*

*Cattle genes with 1 to 1 orthologs in humans (after QC); **Cattle genes within 10kb of the 164 lead SNPs from Bouwman *et al.* (2018) that also have orthologs in humans; ***Human height genes prioritized as those that overlap with, or are within 10kbp either side of the 649 lead SNPs for height in Wood *et al.* (2014), that also have orthologs in cattle. The proportion 10/77 is more than can be expected by chance: Fisher's exact test (odds ratio = 5.5, p-value = 3.7e-05).

S2 Table: Names and coordinates of 10 orthologous genes that are associated with both human height based on results from Wood *et al.* (2014) and cattle stature based on results from Bouwman *et al.* (2018).

Gene name	Chromosome - Cattle	Start (BP) - Cattle	Stop (BP) - cattle	Chromosome - Human	Start (BP) - Human	Stop (BP) - Human
GATAD2A	7	3737212	3838689	19	19385826	19508931
MAN1C1	2	127888646	128008111	1	25617468	25786207
VPS8	1	82706097	82987709	3	184812143	185052614
GNAQ	8	53970972	54280697	9	77716087	78031458
DIS3L2	2	120486066	120794558	2	231961245	232344350
ADAM12	26	45848827	46238138	10	126012381	126388455
INSR	7	17280452	17403052	19	7112255	7294034

GNA12	25	41099259	41171209	7	2728112	2844324
IGF2BP3	4	32077891	32222388	7	23310209	23470467
DNM13A	11	74030425	74059688	2	25227855	25342590

S3 Table: Contingency table showing the overlap of cattle stature genes from Bouwman *et al.* (2018) with human height genes from Yengo *et al.* (2018).

	In Yengo <i>et al.</i> (2018)	Not In Yengo <i>et al.</i> (2018)	Marginal totals (rows)
In Bouwman <i>et al.</i> (2018)	30	47	77**
Not in Bouwman <i>et al.</i> (2018)	1,523	12,142	13,665
Marginal totals (columns)	1,553***	12,189	13,742*

*Cattle genes with 1 to 1 orthologs in humans (after QC); **Cattle genes within 10kb of the 164 lead SNPs from Bouwman *et al.* (2018) that also have orthologs in humans; ***Human height genes prioritized as those that overlap with, or are within 10kbp either side of the 3,290 lead SNPs for height in Yengo *et al.* (2018), that also have orthologs in cattle. The proportion 30/77 is more than can be expected by chance: Fisher's exact test (odds ratio = 5.1, p-value = 3.1e-10).

S4 Table: Names and coordinates of 30 orthologous genes that are associated with both human height based on results from Yengo *et al.* (2018) and cattle stature based on results from Bouwman *et al.* (2018).

Gene name	Chromosome - Cattle	Start (BP) - Cattle	Stop (BP) - cattle	Chromosome - Human	Start (BP) - Human	Stop (BP) - Human
FBN1	10	61877808	62142171	15	48408306	48645849
MAN1C1	2	1.28E+08	1.28E+08	1	25617468	25786207
AXL	18	50677199	50709510	19	41219203	41261766
TBC1D16	19	53181379	53245548	17	79932343	80035848
C13H20orf194	13	52253039	52422570	20	3249305	3407625
ADAMTSL4	3	20187039	20195550	1	1.51E+08	1.51E+08
PLEKHA6	16	1806692	1878608	1	2.04E+08	2.04E+08
PELI2	10	68778347	68974093	14	56117814	56301526
DAAM1	10	71730213	71916529	14	59188646	59371405
NSD2	6	1.1E+08	1.1E+08	4	1871424	1982207
SPSB4	1	1.29E+08	1.29E+08	3	1.41E+08	1.41E+08
HABP4	8	84579583	84615685	9	96450201	96491336
DIS3L2	2	1.2E+08	1.21E+08	2	2.32E+08	2.32E+08
TBC1D22A	5	1.18E+08	1.18E+08	22	46762617	47175699
ADAM12	26	45848827	46238138	10	1.26E+08	1.26E+08
INSR	7	17280452	17403052	19	7112255	7294034
SIK3	15	27942431	28060833	11	1.17E+08	1.17E+08

CCND2	5	1.06E+08	1.06E+08	12	4273762	4405490
GNA12	25	41099259	41171209	7	2728112	2844324
NELFA	6	1.1E+08	1.1E+08	4	1982714	2041903
PAPPA2	16	59132781	59456250	1	1.76E+08	1.77E+08
FNDC3B	1	95891014	96173201	3	1.72E+08	1.72E+08
IGF2BP3	4	32077891	32223388	7	23310209	23470467
ADAMTS17	21	6514360	6924114	15	99971589	1E+08
RAP1GAP2	19	24395974	24452796	17	2755705	3037739
SYN3	5	71475847	71926718	22	32512552	33058372
LAPTM4A	11	78862495	78880461	2	20032650	20052028
NCAPG	6	38765969	38812051	4	17810902	17844862
NHEJ1	2	1.08E+08	1.08E+08	2	2.19E+08	2.19E+08
SPRED1	10	33950024	34069008	15	38252326	38357249

S5 Table: Contingency table showing the overlap of cattle stature genes from Bouwman *et al.* (2018) with human height genes from Yengo *et al.* (2018), while accounting for gene size.

		Gene size	
In Bouwman <i>et al.</i> (2018)	In Yengo <i>et al.</i> (2018)	Short	Long
Yes	Yes	7	23
	No	25	22
No	Yes	736	787
	No	8,850	3,293

Mantel-Haenszel chi-squared (corrected for gene length) = 35.6, p-value < 2.4e-9, common odds ratio = 4.1.

6

CHAPTER 6.

General discussion

6.1 Introduction

The advent of genomic prediction (GP) (Meuwissen *et al.* 2001), a method that uses large scale genotype information to predict an individual's genomic value for a complex trait, has revolutionized animal and plant breeding (e.g. Dalton 2009; Heffner *et al.* 2011), and has the potential to result in personalized medicine in humans (e.g. Wray *et al.* 2007; Abraham and Inouye 2015; Lee *et al.* 2017). Genomic values can be genomically estimated breeding values (GEBVs) of animals or plants for a given trait (Meuwissen *et al.* 2001) or the polygenic scores for complex traits and diseases in humans (Wray *et al.* 2007; Evans *et al.* 2009; Richardson *et al.* 2019). The benefit, and thus the adoption of GP is dependent on its accuracy, which varies across traits and species. For accurate GP, it is essential that the extent of linkage disequilibrium (LD) between markers like single nucleotide polymorphisms (SNPs) and causal variants is high, as the extent of LD determines the proportion of genetic variance for the trait that can be explained by markers (e.g. Calus and Veerkamp 2007; Weir 2008; Hayes *et al.* 2009). To attain a high level of LD between markers and causal variants, high density SNP panels can be used such that every causal variant affecting the trait is tagged by at least one SNP. Given that whole genome sequence (WGS) data is the ultimate high-density panel, the use of WGS instead of low density SNP panels is expected to improve the accuracy of GP. Furthermore, given the expectation that causal variants are included in WGS data, the accuracy of GP may no longer depend on LD between SNPs and causal variants, as the effect of the causal variants themselves can be directly estimated.

Despite promising conclusions from simulations studies (e.g. Meuwissen and Goddard 2010; Macleod *et al.* 2014), studies using real data have reported no added benefit of using WGS for GP instead of marker arrays (e.g. Ober *et al.* 2012; Brøndum *et al.* 2015; Van Binsbergen *et al.* 2015; Heidaritabar *et al.* 2016; Veerkamp *et al.* 2016). In most of these studies, GP was carried out within-population. Results of studies with real data suggest that within population GP accuracy does not benefit from the use of WGS data, possibly due to the long-range LD within population, which makes that all causal variants are already tagged by SNPs on arrays. Given that LD across populations is conserved only over short distances (De Roos *et al.* 2008), WGS holds more promise to increase the accuracy of across- and multi-population GP than within-population GP.

The overall objective of this thesis was to investigate the benefit of WGS data for GP across populations and species. The first sub-objective in the thesis was to investigate strategies for the efficient use of WGS data in across-population, and multi-population GP. In **Chapter 2**, I showed that the accuracy of across-population GP can be improved by using variants that are pre-selected from WGS data based on biological information, in this case the significance level of the variants in a powerful multi-population meta-GWAS (Bouwman *et al.* 2018). In addition, as the third sub-objective, I investigated the benefit of using GWAS result on human height to prioritise genes and gene-associated variants for stature in cattle (**Chapter 5**). I showed that orthologous genes associated with height in humans are enriched for variants that are associated with height in cattle. Randomly sampled SNPs within the prioritised genes also tend to explain more genetic variance for cattle stature than randomly sampled SNPs within random cattle genes. Therefore, in the first part of the general discussion, I discuss the methods for variants pre-selection based on biological information.

While pre-selection of important variants improve the accuracy of GP across populations, results from chapter 2 showed that such variants alone do not capture the total genetic variance for the trait. Therefore, the second sub-objective of the thesis was to implement and test the performance of a multi-population, multiple genomic relationship matrices (GRM) GP model (MPMG) that uses all available variants, but weight pre-selected important variants separately from all other variants in a separate GRM (**Chapter 3**). The predictive performance of the MPMG model was underpinned by deriving and validating a deterministic prediction equation for its accuracy (**Chapter 4**). Using the prediction equation, I showed that the superiority of the MPMG model over single-GRM models is because it benefits from a low number of independent chromosome segments (Me) due to the pre-selected variants, both within and across population. These results show that properly modelling prior biological information can benefit the accuracy of genomic prediction. Therefore, in the second part of the general discussion, I discuss statistical methods for biology-informed GP. In each part, I discuss the current state of knowledge, describe the contribution of this thesis to the current state of knowledge, and discuss possible future directions for research.

6.2 Use of biological information to pre-select variants for GP

In the context of this thesis, I define biological information as information regarding the relationship between changes in genome elements, such as genetic variants, genes, transcripts, regulatory sequences, and changes in observed phenotypes. In an ideal scenario, we would have all the biological information necessary to identify all the causal variants that underlie a given complex trait. The identification of all the causal variants that underlie complex traits, together with the availability of large reference populations and the right model to accurately estimate their effects, should lead to a GP accuracy of 100% (Goddard 2017). While there are developments in animal breeding such as the establishment of the 1,000 bulls genomes project (Daetwyler *et al.* 2014) and the FAANG Consortium (Andersson *et al.* 2015) that are aimed at unravelling biological information coded in the genome, the causal variants that have been identified for any given complex trait are at most a few hundreds. This limitation together with others such as limitations in sample size, genotype by environment interactions, non-additive genetic effects restricts the accuracy of GP to less than 100% (Campos and Toro 2017; Gianola 2017; Goddard 2017). However, while the discovery of all causal variants for any given complex trait is still far from realization, it is still worthwhile to continue to identify likely causal variants or those that are in complete LD with them, as the appropriate modelling of such variants improves the accuracy of GP. For example, results from this thesis showed an improvement in the accuracy of across- and multi-population GP when biological information from a meta-GWAS (Bouwman *et al.* 2018) is used to pre-select variants, which agrees with several other studies in literature (e.g. Brøndum *et al.* 2015; Pérez-Enciso *et al.* 2015; Macleod *et al.* 2016; Fang *et al.* 2017).

In itself, the scientific effort to determine the variants that underlie complex traits is not new, and dates to the time Mendel postulated about “heritable factors” that underlie observed phenotypes. One of the earliest methods used for this purpose was linkage analysis, a statistical approach that examines the co-segregation of variants and trait locus within families and relied mainly on the use of microsatellite markers. Linkage analysis was used to identify the genomic regions that underlie mainly mendelian rather than polygenic traits and had the limitation that it requires multiple generations of pedigree of individuals expressing the mendelian trait. From around 2007 onwards (Wellcome Trust Case Control Consortium 2007), GWAS effectively replaced linkage analysis, mainly due to the availability of dense SNP arrays and the genetics community’s growing interest in complex traits. Henceforth in this section, I will discuss the relevance of GWAS for pre-selection of variants and

the impact on GP prediction accuracy (6.2.1). I will also discuss the use of eQTL analysis to pre-select variants and the potential impact on the accuracy of GP (6.2.2), and other methods for the identification and pre-selection of variants for GP purposes (6.2.3).

6.2.1 GP with variants pre-selected from GWAS

Genome-wide association study was designed to detect associations between genomic variants and complex traits. Both GWAS and GP are based on the same principle, in that both methods rely on the LD between markers and causal variants. The difference between the method is that while GWAS is aimed at identifying variants in the genome that affect complex traits, GP focuses on the prediction of an individual's genomic value based on all variants. My aim here is not to discuss GWAS as a concept, as there are already excellent literatures on GWAS available (e.g. Schmid and Bennewitz 2017; Visscher *et al.* 2017). Rather, my aim is to raise some important discussion points with regards to the use of GWAS results for GP purposes.

Throughout this thesis, I utilized results from a meta-analysis of GWAS results for stature in cattle that combined data on 58,265 individuals from 8 cattle breeds and WGS (Bouwman *et al.* 2018). I pre-selected SNPs from WGS based on their $-\log_{10}(p)$ value from the meta-GWAS and demonstrated the usefulness of the pre-selected variants in different models for GP. Given that such a large meta-GWAS is not common for most economic traits that are typically included in an index for genetic evaluation, it is worth asking what the impact on the accuracy of GP will be if variants are pre-selected from a less powered GWAS, for example, a GWAS using SNP arrays, or a WGS-based GWAS in only a single population. I will first try to address the use of SNP-based GWAS to pre-select variants for GP. Subsequently, I will discuss the use of within-population WGS-based GWAS to pre-select variants for GP.

6.2.1.1 Using QTL from a SNP-Chip based GWAS

When variants are pre-selected from a SNP-chip based GWAS, I do not expect the pre-selected variants to improve the accuracy of GP. That is because SNPs that are on a chip are typically selected because they are common in the population of interest (Matukumalli *et al.* 2009; Groenen *et al.* 2011). Common SNPs are only in high LD with common causal variants (Van Binsbergen *et al.* 2014), and therefore such GWAS will identify mainly QTL that are common in the population and less likely to identify rare QTL, say with minor allele frequencies (MAF) $< 1\%$. Pre-selected common QTL variants from a SNP-chip based GWAS are expected to capture less genetic variance for the trait than using all the SNPs on a SNP-chip. In addition, if the

'rare variants' hypothesis holds (e.g. McClellan *et al.* 2007; Gibson 2012), which asserts that most causal variants for complex traits are rare, the use of SNP arrays for GWAS is less likely to result in the accurate pre-selection of the most relevant variants for complex traits and thus, less likely to yield an improvement in the accuracy of GP. For future GWAS therefore, and if the interest is to use results from the GWAS for the purpose of GP, I recommend to use WGS data instead of SNP arrays. This is given the expectation that WGS-based GWAS is able to identify rare causal/predictive variants. This is also in line with results of GP in this thesis using the variants pre-selected based on a WGS-based GWAS (Bouwman *et al.* 2018), and other results in literature e.g. Brøndum *et al.* 2015; Ober *et al.* 2015; Van Den Berg *et al.* 2016; Al Kalaldehy *et al.* 2019). For example, Al Kalaldehy *et al.* (2019) showed an improvement of between 5% to 9% in the accuracy of multi-breed GP of parasite resistance in Australian sheep, when variants were carefully pre-selected based on their $-\log_{10}(p)$ value from a WGS-based GWAS, as compared to scenarios with no variants pre-selection. Furthermore, they showed that the improvement in accuracy was more pronounced when variants were pre-selected from WGS-based GWAS as compared to when variants were pre-selected from a GWAS based on a high density (HD) SNP panel.

6.2.1.2 Using QTL from a within population GWAS

Although WGS-based GWAS is in principle able to identify more trait associated variants, including those that are rare, using a within-population WGS-based GWAS has one major limitation. Because of the long range LD in highly structured population as is common in animal breeding, a QTL region in a within-population WGS-based GWAS typically contains a large number of variants that are all in strong LD with each other. This makes it difficult to pinpoint the real causal variant(s) in a given QTL region. The use of variants pre-selected from such a GWAS are less likely to improve the accuracy of within-population GP as compared to the scenarios with no variants pre-selection. For example, Veerkamp *et al.* (2016) used WGS data on Holstein-Friesian bulls to perform a GWAS for protein yield, somatic cell score and the interval from first to last insemination, and used results from the GWAS to pre-select variants for GP. Using only the pre-selected variants or separating the pre-selected variants in one GRM and all other variants in a second GRM did not improve the accuracy of GP as compared to the scenarios with no pre-selection. It can be argued that the lack of increase in accuracy as reported in Veerkamp *et al.* (2016) may be due to a limitation in GWAS sample size (5,503). However, I do not expect an increase in the accuracy of prediction if the GWAS were based on larger sample size. With a very large sample size, say a million individuals, the GWAS will be more

powerful and more causal/predictive variants, especially those that are rare, will be identified. However, the Holstein population has a very small effective population size, which means that LD extends over long distances (De Roos *et al.* 2008). As a result, the current SNP arrays are well able to tag the effect of the causal variants, and thus the pre-selection of variants from a GWAS, powerful or not, may not improve within-population GP accuracy.

Furthermore, pre-selecting variants from a within-population WGS-based GWAS for GP in a different population, for example a different breed, may also not work. That is because a variant that has an effect on the trait due to LD in the discovery population, may not have the same effect in the prediction population if the LD pattern between the variant and causal variants is different in the prediction population. An approach to dealing with this issue is to combine multiple populations in meta-GWAS as was done in Bouwman *et al.* (2018), as doing so helps to break down long LD blocks and narrows down the relevant variants to shorter intervals (Karlsson *et al.* 2007; Bovine Hapmap Consortium 2009; Hayes *et al.* 2009). Still, a possible trade-off is that estimated variants effects in a multi-population GWAS can be skewed towards those in the largest populations in the meta-GWAS. For example, if a multi-population meta-GWAS is dominated by large number of Holsteins as is commonly the case in cattle, the estimated variants effects will most likely reflect those in a Holstein population and not the other numerically small populations included in the same meta-GWAS. The use of result from such a meta-GWAS to pre-select variants for GP in a numerically small Jersey population for example, should be done with caution. This is especially the case when there are population specific causal variants or when some causal variants are fixed in some populations while they still segregate in others. In addition to combining multiple-populations in a GWAS, an option to narrow down the most relevant variants in a given GWAS peak is to use methods such as the conditional and joint analysis (Yang *et al.* 2012) as was implemented in Chapter 2 and 3 of this thesis, which results in the selection of only the variants that independently contributes to explaining the genetic variance for the trait of interest.

6.2.2 Use of eQTL analysis to pre-select variants for GP

In GWAS, some QTL are found to be in or very close to genes that are the likely causal genes. However, most of the QTL detected in GWAS are intergenic. The biological mechanism underpinning the detection of intergenic QTL in GWAS is unclear. A possible explanation for the detection of QTL in non-coding regions of the genome is that they may be underlying the expression of some genes (Nica and Dermitzakis

2013). If a genetic variant is an expression quantitative trait locus (eQTL), that is, it influences the gene-expression level of a given gene, individuals carrying different genotypes of the variant will show differences in their level of gene-expression. Such an eQTL can be detected by performing a simple association test between genetic variants and gene-expression data measured on several individuals. Provided that the expression level of the gene influences the phenotype, then individuals carrying different genotypes of the eQTL will also show differences in their phenotype. In this case, the causal variant indirectly affect the phenotype through gene-expression. Another case that may exist is that the causal variant exerts a direct pleiotropic effect on both gene-expression and the phenotype (Zhu *et al.* 2016), and detection of such an eQTL will require the integration of eQTL analysis and GWAS results. The integration of eQTL analysis with GWAS helps to further prioritise the most functionally relevant variants in a GWAS peak and to identify likely causal genes. There are two possible approaches to utilizing eQTL analysis for the pre-selection of variants for GP.

The first approach is to use eQTL analysis to identify the functionally relevant variants in a given GWAS peak and to ultimately identify the candidate causal genes. This is given the expectation that not all the variants in a GWAS peak are functionally relevant, as some of the variants are significant in GWAS only due to LD with a causal loci. There are several examples where eQTL analysis resulted in the identification of target genes for complex traits following GWAS in humans (e.g. Nica *et al.* 2010; Hernandez *et al.* 2012; Farh *et al.* 2015; Gtex Consortium 2015), cattle (e.g. Brand *et al.* 2016; Littlejohn *et al.* 2016; Lopdell *et al.* 2017), pigs (e.g. Liaubet *et al.* 2011; Ponsuksili *et al.* 2014; González-Prendes *et al.* 2017) and poultry (e.g. Blum *et al.* 2011; Johnsson *et al.* 2016). While this approach may lead to the discovery of some causal genes and give us more insight into the biology of complex traits, I do not expect that, in the short term, the use of variants prioritised as functionally relevant following GWAS and eQTL analyses will improve GP accuracy as compared to the use of variants pre-selected from only GWAS.

The reason is that, if a genetic variant is causal for the expression of a given gene and the expression of the gene has an effect on a phenotype of interest, such a causal variant will be among the significant results from a GWAS analysis. In this case, an eQTL analysis can only narrow down the number of significant GWAS variants, but not identify additional variants. Furthermore, if a given variant is causal for the expression of a gene, and the gene does not affect the trait of interest, GWAS will not pick up the variant. However, that does not matter, as the variant is not relevant

for the trait of interest. This is usually the case when gene-expression measurements are made in tissues that are not relevant for the trait of interest. While there are genes whose expression levels are consistent in multiple tissues from the same individual, other genes have been shown to have tissue specific expression patterns (Hernandez *et al.* 2012; GTEx Consortium 2015). The implication is that a positive result for an eQTL will be uninformative if the tissue used is not relevant for the trait. Studies have shown that QTLs and eQTL overlap strongly when the tissue is relevant for the trait (e.g. Brown *et al.* 2013; Torres *et al.* 2014). An eQTL analysis using gene-expression data in tissues that are not relevant for the trait may explain why in some cases (e.g. Van Den Berg *et al.* 2019), there is little overlap between eQTL and QTL. Thus, prioritising only variants that are both an eQTL and a QTL will rule out many possible causal or predictive variants that have a direct effect on the trait or an indirect effect through some other mechanism that is not gene-expression. Consequently, GP accuracy will not improve as compared to using variants that are directly selected from GWAS alone.

The second approach to utilizing eQTL analysis results is in the identification of additional causal/predictive variants for complex traits, that is, in addition to variants that are already identified through GWAS. If a GWAS has a low power, for example in the case of a low number of discovery sample, some causal/ predictive variants may not be statistically significant and as such will not be pre-selected for GP based on GWAS alone. However, evidence that such variants affect gene expression in an eQTL analysis can be used to pre-select them for GP. It is not clear, however, if pre-selected variants using this approach can improve the accuracy of GP, especially within-population, as compared to the variants pre-selected directly from only GWAS.

With regards to variants pre-selection using eQTL analysis, it is important to keep in mind that while gene-expression is largely regulated by genetic factors, non-genetic factors such as nutrition and stress can also trigger gene-expression through epigenetic mechanisms (Choi and Friso 2010; Duraimani *et al.* 2015). Thus, an association between a genetic variant and gene-expression is not necessarily indicative of the causality of the variant for gene-expression. This additional layer of complexity could result in a mismatch between eQTL and QTL from GWAS, further reducing the utility of eQTL analysis as a variants pre-selection strategy for GP.

For research into the biology underlying complex traits, I expect that eQTL analysis will continue to be relevant as it has the potential to shed more light on the

mechanism through which variants exert their effect on complex traits. In the short term however, I do not expect to see an adoption of eQTL analysis as a variant pre-selection strategy for GP, at least not to the same extent as GWAS will be. In addition, the data required for GWAS is already routinely collected, while generating gene expression data for different tissues requires additional costs. For breeding organizations, any benefit of eQTL analysis may therefore not outweigh the additional cost.

6.2.3 Pre-selection of variants using other methods

In addition to GWAS and eQTL analysis, other methods exist that can be used to prioritise or weight variants in the genome for GP. For example, to aid the identification of causal variants, especially those with large effect on complex traits, Derks *et al.* (in preparation) proposed and designed a pipeline that relies on input from GWAS, gene-expression, the combined annotation dependent depletion (CADD) score (Rentzsch *et al.* 2018; Groß *et al.* 2020), DNA methylation, gene orthology, etc. A major advantage of the pipeline is its ability to distinguish between a causal variant and all other variants in a GWAS peak that are all in strong LD. Using the pipeline, (Derks *et al.* in preparation) identified causal variants that are already known in literature to affect complex traits in pigs. For example, they found a variant in the MC4R gene that is already known to affect production traits in pigs (Kim *et al.* 2000), a causal variant in the VRTN gene affecting the number of teats in pigs (Van Son *et al.* 2019), and a causal variant in the PRKAG3 gene affecting meat quality (Milan *et al.* 2000). In addition to enhancing our understanding of the biology of complex traits, the pipeline can be used to pre-select likely causal variants that can continuously be added on to the array for GP.

Another type of information that can be used to pre-select variants for GP is the functional and evolutionary trait heritability (FAETH) score proposed by (Xiang *et al.* 2019). The FAETH score is used to rank all variants in the genome based on their functionality and predicted heritability. In their study, (Xiang *et al.* 2019) showed that variants with high FAETH score significantly improved the proportion of total genetic variance explained for complex traits and improved genomic prediction accuracy than variants with low FAETH scores. It would be interesting however to investigate if the variants with the high FAETH score can capture more genetic variance and improve the accuracy of GP for complex traits than the standard 50k SNP-chip. The FAETH scores are now publicly available and can be used to pre-selected relevant variants for GP of complex traits.

6.3 Models for biology-informed GP

In the previous sections, I focused on the different methods for variants pre-selection from WGS data based on biological information. There are several means to include good biological information in breeding programs. For example, if GWAS or functional studies such as eQTL analysis results in the discovery of a major gene for a given trait, a program can be designed to, in the long term, increase the frequency of the gene in the population. This strategy has been used in the past, for example, to increase the frequency of double muscling gene in some beef cattle breeds (Dunner *et al.* 2003), increase the frequency of the polledness gene in dairy cattle breeds (Windig *et al.* 2015), or to eliminate deleterious alleles of certain genes in a population (Upperman *et al.* 2019). This strategy is used mainly for monogenic traits. However, most traits are complex, in that they are controlled by many genes and many causal variants with moderate to small effects. Thus, in this section, I will discuss the different ways of modelling prior biological information in GP. First, I will focus on the so-called GBLUP model as a representative of linear GP models. Secondly, I discuss the use of non-linear Bayesian models that explicitly incorporates prior biological information for GP. Finally, I will discuss some advantages and potential disadvantages of explicit modelling of prior biological information for GP.

6.3.1 Modelling prior biological information with GBLUP

The GBLUP model as originally proposed (Vanraden 2008; Strandén and Garrick 2009) does not account for variations in the effect of genomic variants on a given trait. Rather, an assumption of the model is all variants have the same infinitesimally small effect on the trait. With the increasing availability of WGS data and the increasing availability of biological knowledge on genomic variants, making the infinitesimal model assumption in GBLUP is sub-optimal. The GBLUP model can be modified to allow for differential weighting of variants based on prior biological knowledge. One way to do that is to allow the differential weighting of variants based on prior biological knowledge in the construction of genomic relationship matrix (GRM). Several studies have demonstrated that accurate weighting of variants that are used to build the GRM resulted in a higher predictive ability of the GBLUP model than a scenario in which all variants are equally weighted in a GRM. For example, Zhang *et al.* (2010) used the variance explained by variants as weights to construct trait-specific GRMs. They observed an increase in accuracy from 0.73 when all variants had the same weight in the construction of GRM to 0.79 when variants were differentially weighted. Su *et al.* (2014) used results from Bayesian model for estimating SNP effects and GWAS (posterior variances of SNP effects, SNP effects

and $-\log_{10}(p)$ values) as weights in the construction of GRMs, with up to 2% increase in the reliability of predicted genomic values as compared to the use of unweighted GRM. These results are similar to others in literature (e.g. De Los Campos *et al.* 2013; Zhou *et al.* 2014; Tiezzi and Maltecca 2015; Fragomeni *et al.* 2017).

In general, weighting of variants in the construction of GRM is an attractive approach when marker density is not very high and in cases where almost all the variants are expected to have non-zero effect on the trait, either directly or because they are in LD with unobserved causal variants. When marker density is very high, for example at WGS level, differential weighting of all the variants in the construction of GRM may not be the best approach. That is because many or a majority of the variants will have a zero-effect on the trait. While assigning a weight of zero to such variants will effectively exclude them from any contribution to prediction, leaving them in the model anyways increases model complexity and possibly computing time and resources. An alternative to using weighted GRMs for GP with WGS is to use the classical GBLUP model with GRM built using only a set of variants that are pre-selected based on prior-biological information. In this case, the variants that are expected to have zero effect on the trait are left out of the model a priori. In Chapters 2 of this thesis for example, I pre-selected SNPs based on their $-\log_{10}(p)$ value from the stature meta-GWAS of Bouwman *et al.* (2018), and used only the pre-selected variants to build a GRM for use in a GBLUP model. For GP across different breeds of cattle, using a GRM built from only 133 top SNPs resulted in an accuracy of 0.23, which is more than double the accuracy obtained from using a GRM built from ~ 14 million WGS variants (0.08). In Chapter 2, the 133 top SNP pre-selected from the meta-GWAS were assigned the same weight in the construction of the GRM. It is still left to be seen if differential weighting of the 133 SNPs either based on their $-\log_{10}(p)$ values or their estimated effects will result in the an improvement in the accuracy of prediction. Given that the number of pre-selected variants (here 133) is far less than the size of the reference population, my expectation is that the GBLUP model is able to accurately estimate the effect of the SNPs and that prior weighting of the SNPs in the GRM will not improve the accuracy any further. As the number of pre-selected variants approach or exceed the reference population size, however, it may be beneficial to differentially weight them in the construction of the GRM, with the weights set to reflect the variation in their effect on the trait.

While pre-selection of variants for GP isolates the effect of the pre-selected variants from noise, there is a trade-off in the proportion of genetic variance for the trait that can be explained. In chapter 2 for example, the 133 pre-selected top SNPs explained

between ~40 to ~60% of the total genetic variance for stature in three different cattle populations, while the ~14 million WGS variants explained between ~78 to ~98% of the total genetic variance for stature. An approach to circumvent this trade-off is to fit multiple GRMs in GBLUP. For example, one GRM can be made from the pre-selected variants and another from SNPs on a standard 50k SNP-chip. In Chapter 3, I showed that the GBLUP model with multiple GRMs resulted in higher accuracies of within-, across- and multi-population GP than a GBLUP model using only pre-selected SNPs in one GRM. The prediction equation developed in Chapter 4 helped to underpin the reasons for the improvement in accuracy, which are well discussed in Chapter 4. Mainly, in addition to the fact that the pre-selected variants in the first GRM are isolated from the noise effect of other variants (as such their effects are accurately estimated), the number of independent chromosomal segment due to the pre-selected variants is low, given that only a few hundred variants are typically pre-selected. This is only the case however, when the pre-selected variants explain some proportion of the total genetic variants for the trait. The results of Chapter 3 agree with results in literature, obtained using the so-called genomic feature BLUP models (e.g. Sørensen *et al.* 2013; Speed and Balding 2014; Sarup *et al.* 2016; Fang *et al.* 2017). The Genomics feature models are linear GBLUP models that fit two genomic relationship matrices, one computed using informative variants pre-selected based on some form of biological information and the other based on neutral or uninformative SNPs. At the moment, identifying all the causal variants for a given complex trait is still not possible. To that effect and based on the results from Chapter 3 and 4 of thesis, it is always recommended to fit a separate GRM, in addition to the GRM made from pre-selected variants in GP model. In the worst case scenario, the accuracy of GP remain the same as a model that uses only a single GRM, but not lower. For future research, it would be interesting to investigate the performance of the multiple-GRMs model in a multi-trait context, especially given that most breeding organization use an index made up of several component traits for genomic selection decisions.

6.3.2 Modelling prior biological information with Bayesian models

Non-linear Bayesian GP models can also be optimised to allow for the explicit modelling of prior biological information. One such non-linear model is the BayesR by segment (BayesRS) model, proposed by Brøndum *et al.* (2012). Similar to the bayesR model (Erbe *et al.* 2012), variant effects in BayesRS are assumed to belong to a mixture of four normal distributions. While in BayesR, the mixture proportion (π) are assigned a uniform Dirichlet priors as $\pi \sim \text{Dirichlet}(1,1,1,1)$, they are assigned a non-uniform prior as $\pi_s \sim \text{Dirichlet}(\mathbf{k}_s)$ in BayesRS, where $\mathbf{k}_s$ is a vector of priors

for the distributions in a given genomic segment s . In (Brøndum *et al.* 2012), the priors ($\mathbf{k}_s$) were set to reflect the genetic variance explained by the genomic segments from a previously fitted BayesR model. Effectively, In BayesRS, the probability of the effect of a given variant to belong to any of the four mixture distributions depends on the genomic segment (s) to which the variant belongs. In that way, the BayesRS model estimates the mixture proportion based on both the prior information, namely, prior knowledge on the variance explained by the genomic segments to which the variants belong, and evidence in the data. Brøndum *et al.* (2012) showed an increase in the accuracy of across-breed GP of up to 3.5% by using the BayesRS model as against the BayesR model. The BayesRS model can use other forms of prior biological information, such as those discussed in section 6.2 of this Chapter. While in principle the BayesRS model should result in improved prediction accuracy than BayesR in the presence of prior biological information, it has one potential drawback. That is, the model is not able to update the mixture proportions of a given variant if it is wrongly placed a priori even with the availability of large data, which may hinder its performance in cases where the prior biological information used is inaccurate.

Another model called BayesRC was proposed by Macleod *et al.* (2016), and allows for the explicit incorporation of prior-biological information in GP. In BayesRC, prior-biological information, such as those discussed in Section 6.2 of this Chapter, are used to assign variants to two or more classes (C). The effect of variants within a given class, say class 1, are assumed to be drawn from a mixture of four normal distributions, while the effect of variants in a different class, say class 2, are assumed to be drawn from a different mixture of four normal distributions. Furthermore, the BayesRC model allows the iterative updating of the distribution of a given marker effect within a class, if the data supports it. For example, it is possible that at iteration number i , the effect of marker j belongs to the mixture proportion π , while at iteration number $i + 1$, the mixture proportion for the marker changes to π_1 . A major difference between BayesRC and BayesRS is that in BayesRC, a uniform Dirichlet priors, $\pi \sim \text{Dirichlet}(1,1,1,1)$, is used for the four normal distributions in a given class. Thus, within a given class, BayesRC estimates the mixture proportions only based on evidence in the data. Effectively, the BayesRC model is a BayesR model with multiple, instead of one class. Results obtained by Macleod *et al.* (2016) showed that the BayesRC model yields equal or higher accuracies of GP and QTL detection than both BayesR and the classical GBLUP models. They also showed that the greatest benefit of the BayesRC model over other tested models was observed when

the reference and validation population were distantly related, indicating that the model will be very useful for GP across populations and generations.

The Bayesian mixture models shrink the parameter space by assigning variants into several distributions instead of one. That is analogous to splitting variants into several GRMs in the GBLUP model as implemented in Chapters 3 and 4. By doing so, the Bayesian mixture models are able to estimate the variants effects within each distribution more accurately than if variants were all in the same distribution, and in that sense are able to improve the accuracy of prediction. While the incorporation of prior biological information in non-linear GP such as BayesRS and BayesRC improves the accuracy of GP as compared to classical Bayesian models, it is debatable if such models will have any predictive advantage over linear models like GBLUP that also explicitly incorporate prior biological knowledge. This is still an area that needs more research.

6.3.3 Advantages and disadvantages of modelling prior biological information in GP

The use of prior biological knowledge to prioritise variants in GP has a number of advantages and disadvantages that are highlighted in this section. One of the main advantage of variants prioritization based on biological information is that it allows GP to be driven by the most relevant variants for a given trait (Zhang *et al.* 2010). For example, when prior biological information is used to pre-select variants or used as weights for variants in the construction of a GRM, the noise effect of irrelevant variants are either significantly reduced or completely removed, with resultant increase in the accuracy of GP. Furthermore, variants prioritization based on accurate prior biological information means that the pre-selected variants are either themselves the causal variants or they are close to the causal variants. Thus GP across generations or populations using the prioritised variants is expected to be more accurate than using neutral variants, given that the difference in LD pattern across generations or populations is less likely to influence the estimated effect of the prioritised variants. For example, training the effects of pre-selected variants in a Holstein population and using the estimated effects to estimate the genomic values of Jersey bulls can be possible as shown in Chapter 2 of this thesis. The closer we are to the causal variants for a given trait, the less GP depends on the similarity in the pattern of LD between the reference and validation populations, and GP will become possible even across distantly related populations.

There are however several potential disadvantages of using prior biological information to prioritise variants for GP which must be kept in mind. When genomic selection decision for a given trait is based on a few variants that are pre-selected based on biological information, allele frequencies of such variants in the genome can be drastically shifted with potential negative consequences on other vital traits. This is especially so for causal or predictive variants that seem to be under natural balancing selection. It is possible that increasing the frequency of the variants of interest is deleterious or negatively impacts fitness, and this must be kept in mind.

Furthermore, in GP, weighted GRMs have the disadvantage that they are trait-specific, given that the prior biological information used as weighting factors are mostly trait specific. That means, for example, that a GRM weighted using the results of a GWAS for milk production cannot be used to predict genomic values for stature, except if milk production and stature are highly correlated traits. Creating trait specific GRMs for all traits is not practical for a breeding organization, given that most breeding organization use the same GRM for all the traits included in their index. A potential solution, which to my knowledge has not yet been investigated is as follows. First, for each trait in the selection index, prior biological information such as GWAS results can be used to pre-select the most relevant variants. Secondly, all the pre-selected variants for all the traits can be combined into a single set and used to create a GRM. Thirdly, a second set of common SNPs can be used to create a second GRM and both GRMs can be fitted simultaneously in the typically multi-trait analysis used for genetic evaluations. Assume for simplicity that there are two traits, A and B in the index and assume that 300 variants were pre-selected for trait A and the same number of variants were pre-selected for trait B. In the unlikely case that the traits are uncorrelated, the 300 variants for trait A becomes random noise for trait B and vice versa. As shown in Chapter 2, if the number of random noise variants for a given trait is not very high, they have limited negative impact on the accuracy of prediction for the trait. My expectation is that this approach will improve the accuracy of GP for all traits in the index as compared to current methods employed for genetic evaluation. In addition to the suggested approach, there is also the possibility to use trait-independent weights in the construction of a GRM, such as the CADD score (Rentzsch *et al.* 2018; Groß *et al.* 2020) that are derived mainly from genome annotation information. However, the impact of using trait-independent weighting factors such as CADD on the accuracy of GP is something that still needs to be investigated.

In non-linear but also linear models, biological information that is used as priors have the tendency to exert a strong influence on model posterior estimations and predictions, especially where the data is limited or when evidence in the data is simply not there to counter the effect of the priors. This can be a problem in situations where the prior biological information is inaccurately obtained. However, a large enough reference population is able to counter any negative effect of inaccurately specified priors except in models such as BayesRS where the prior distribution of variants are fixed and cannot be updated based on evidence in the data.

Finally, when prior biological information are obtained using data from a given population, use of the information to prioritise variants and the subsequent use of the prioritised variants for GP in the same population will likely lead to an over-estimation of variants effects and the accuracy of prediction.

6.4 Concluding remarks

Increasing the accuracy of GP has been of major interest in the past decade, mainly in animal and plant breeding, but also in human genetics. Results from this thesis have contributed to our knowledge of how that can be achieved. Evidence in this thesis shows that WGS data can be used to improve the accuracy of GP mainly across, but not within populations. It seems, however, that the strategy of simply increasing the density of markers up to WGS level will not increase accuracy as compared to accurate identification and pre-selection of variants in WGS based on biological information. I have discussed in this chapter the different methods that can be used to pre-select variants in WGS data for the purpose of GP. I concluded that GWAS is the most relevant and practical method for variants pre-selection. The first reason underpinning this conclusion is that GWAS is able to identify both the variants that have a direct effect on complex traits and other variants that have an indirect effect on the trait through biological mechanisms such as gene-expression. In that sense, I do not expect functional studies such as eQTL analysis to result in the pre-selection of more relevant variants for GP than GWAS, but may aid in pinpointing the candidate causal variants. The second reason is that the data required to implement GWAS is already routinely collected by breeding organizations, while generating the data needed for functional studies will require further investment. I also discussed that the pre-selection of variants based on biological information is in itself only the first step towards improving the accuracy of GP. It is important that the pre-selected variants are properly modelled, for example by fitting pre-selected variants in one

GRM and other variants in a different GRM as demonstrated in this thesis. I expect the next phase of GP in breeding organization to be the utilization of GWAS data to pre-select variants and the implementation of prediction models that differentially weight the pre-selected variants.

6.5 References

- Abraham, G., and M. Inouye, 2015 Genomic risk prediction of complex human disease and its clinical application. *Current opinion in genetics & development* 33: 10-16.
- Al Kalaldehy, M., J. Gibson, N. Duijvesteijn, H. D. Daetwyler, I. MacLeod *et al.*, 2019 Using imputed whole-genome sequence data to improve the accuracy of genomic prediction for parasite resistance in Australian sheep. *Genetics Selection Evolution* 51: 32.
- Andersson, L., A. L. Archibald, C. D. Bottema, R. Brauning, S. C. Burgess *et al.*, 2015 Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome biology* 16: 57.
- Blum, Y., G. Le Mignon, D. Causeur, O. Filangi, C. Désert *et al.*, 2011 Complex trait subtypes identification using transcriptome profiling reveals an interaction between two QTL affecting adiposity in chicken. *BMC genomics* 12: 567.
- Bouwman, A. C., H. D. Daetwyler, A. J. Chamberlain, C. H. Ponce, M. Sargolzaei *et al.*, 2018 Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. *Nature genetics* 50: 362.
- Bovine HapMap Consortium, 2009 Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science* 324: 528-532.
- Brand, B., M. O. Scheinhardt, J. Friedrich, D. Zimmer, N. Reinsch *et al.*, 2016 Adrenal cortex expression quantitative trait loci in a German Holstein× Charolais cross. *BMC genetics* 17: 135.
- Brøndum, R. F., G. Su, L. Janss, G. Sahana, B. Guldbrandtsen *et al.*, 2015 Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. *Journal of dairy science* 98: 4107-4116.
- 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 genomics* 13: 543.
- Brown, C. D., L. M. Mangravite and B. E. Engelhardt, 2013 Integrative modeling of eQTLs and cis-regulatory elements suggests mechanisms underlying cell type specificity of eQTLs. *PLoS genetics* 9.
- Calus, M., and R. Veerkamp, 2007 Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. *Journal of Animal Breeding and Genetics* 124: 362-368.
- Campos, G., and M. Toro, 2017 Can we make genomic selection 100% accurate? *Journal of Animal Breeding and Genetics* 6: 437-437.
- Choi, S.-W., and S. Friso, 2010 Epigenetics: a new bridge between nutrition and health. *Advances in nutrition* 1: 8-16.
- 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. *Nature genetics* 46: 858.

- 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 genetics* 9.
- De Roos, A., B. J. Hayes, R. Spelman and M. E. Goddard, 2008 Linkage disequilibrium and persistence of phase in Holstein–Friesian, Jersey and Angus cattle. *Genetics* 179: 1503-1512.
- Dunner, S., M. E. Miranda, Y. Amigues, J. Cañón, M. Georges *et al.*, 2003 Haplotype diversity of the myostatin gene among beef cattlebreeds. *Genetics Selection Evolution* 35: 103.
- Duraimani, S., R. H. Schneider, O. S. Randall, S. I. Nidich, S. Xu *et al.*, 2015 Effects of lifestyle modification on telomerase gene expression in hypertensive patients: a pilot trial of stress reduction and health education programs in African Americans. *PLoS One* 10: e0142689.
- e.g. Dalton, R., 2009 No bull: genes for better milk. *Nature* 457: 369-370.
- Erbe, M., B. Hayes, L. Matukumalli, S. Goswami, P. Bowman *et al.*, 2012 Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *Journal of dairy science* 95: 4114-4129.
- Evans, D. M., P. M. Visscher and N. R. Wray, 2009 Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Human molecular genetics* 18: 3525-3531.
- Fang, L., G. Sahana, P. Ma, G. Su, Y. Yu *et al.*, 2017 Use of biological priors enhances understanding of genetic architecture and genomic prediction of complex traits within and between dairy cattle breeds. *BMC genomics* 18: 604.
- Farh, K. K.-H., A. Marson, J. Zhu, M. Kleinewietfeld, W. J. Housley *et al.*, 2015 Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* 518: 337-343.
- Fragomeni, B. O., D. A. Lourenco, Y. Masuda, A. Legarra and I. Misztal, 2017 Incorporation of causative quantitative trait nucleotides in single-step GBLUP. *Genetics Selection Evolution* 49: 59.
- Gianola, D., 2017 It is unlikely that genomic selection will ever be 100% accurate. *Journal of animal breeding and genetics= Zeitschrift fur Tierzuchtung und Zuchtungsbiologie* 134: 438.
- Gibson, G., 2012 Rare and common variants: twenty arguments. *Nature Reviews Genetics* 13: 135-145.
- Goddard, M. E., 2017 Can we make genomic selection 100% accurate? *Journal of Animal Breeding and Genetics* 134: 287-288.
- González-Prendes, R., R. Quintanilla, A. Cánovas, A. Manunza, T. F. Cardoso *et al.*, 2017 Joint QTL mapping and gene expression analysis identify positional candidate genes influencing pork quality traits. *Scientific reports* 7: 1-9.
- Groenen, M. A., H.-J. Megens, Y. Zare, W. C. Warren, L. W. Hillier *et al.*, 2011 The development and characterization of a 60K SNP chip for chicken. *BMC genomics* 12: 274.

- Groß, C., M. Derks, H.-J. Megens, M. Bosse, M. A. Groenen *et al.*, 2020 pCADD: SNV prioritisation in *Sus scrofa*. *Genetics Selection Evolution* 52: 4.
- GTEx Consortium, 2015 The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 348: 648-660.
- Hayes, B. J., P. J. Bowman, A. Chamberlain and M. Goddard, 2009 Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of dairy science* 92: 433-443.
- Heffner, E. L., J.-L. Jannink and M. E. Sorrells, 2011 Genomic selection accuracy using multifamily prediction models in a wheat breeding program. *The Plant Genome* 4: 65-75.
- Heidaritabar, M., M. P. Calus, H. J. Megens, A. Vereijken, M. A. Groenen *et al.*, 2016 Accuracy of genomic prediction using imputed whole-genome sequence data in white layers. *Journal of Animal Breeding and Genetics* 133: 167-179.
- Hernandez, D. G., M. A. Nalls, M. Moore, S. Chong, A. Dillman *et al.*, 2012 Integration of GWAS SNPs and tissue specific expression profiling reveal discrete eQTLs for human traits in blood and brain. *Neurobiology of disease* 47: 20-28.
- Johansson, M., K. B. Jonsson, L. Andersson, P. Jensen and D. Wright, 2016 Quantitative trait locus and genetical genomics analysis identifies putatively causal genes for fecundity and brooding in the chicken. *G3: Genes, Genomes, Genetics* 6: 311-319.
- Karlsson, E. K., I. Baranowska, C. M. Wade, N. H. S. Hillbertz, M. C. Zody *et al.*, 2007 Efficient mapping of mendelian traits in dogs through genome-wide association. *Nature genetics* 39: 1321.
- Kim, K. S., N. Larsen, T. Short, G. Plastow and M. F. Rothschild, 2000 A missense variant of the porcine melanocortin-4 receptor (MC4R) gene is associated with fatness, growth, and feed intake traits. *Mammalian genome* 11: 131-135.
- Lee, S. H., W. S. P. Weerasinghe, N. R. Wray, M. E. Goddard and J. H. Van Der Werf, 2017 Using information of relatives in genomic prediction to apply effective stratified medicine. *Scientific reports* 7: 42091.
- Liaubet, L., V. Lobjois, T. Faraut, A. Tircazes, F. Benne *et al.*, 2011 Genetic variability of transcript abundance in pig peri-mortem skeletal muscle: eQTL localized genes involved in stress response, cell death, muscle disorders and metabolism. *BMC genomics* 12: 548.
- Littlejohn, M. D., K. Tiplady, T. A. Fink, K. Lehnert, T. Lopdell *et al.*, 2016 Sequence-based association analysis reveals an MGST1 eQTL with pleiotropic effects on bovine milk composition. *Scientific reports* 6: 25376.
- Lopdell, T. J., K. Tiplady, M. Struchalin, T. J. Johnson, M. Keehan *et al.*, 2017 DNA and RNA-sequence based GWAS highlights membrane-transport genes as key modulators of milk lactose content. *BMC genomics* 18: 968.
- MacLeod, I., P. Bowman, C. Vander Jagt, M. Haile-Mariam, K. Kemper *et al.*, 2016 Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. *BMC genomics* 17: 144.

- MacLeod, I. M., B. J. Hayes and M. E. Goddard, 2014 The effects of demography and long-term selection on the accuracy of genomic prediction with sequence data. *Genetics* 198: 1671-1684.
- 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.
- McClellan, J. M., E. Susser and M.-C. King, 2007 Schizophrenia: a common disease caused by multiple rare alleles. *The British Journal of Psychiatry* 190: 194-199.
- Meuwissen, T., and M. Goddard, 2010 Accurate prediction of genetic values for complex traits by whole-genome resequencing. *Genetics* 185: 623-631.
- 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.
- Milan, D., J.-T. Jeon, C. Looft, V. Amarger, A. Robic *et al.*, 2000 A mutation in PRKAG3 associated with excess glycogen content in pig skeletal muscle. *Science* 288: 1248-1251.
- Nica, A. C., and E. T. Dermitzakis, 2013 Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences* 368: 20120362.
- Nica, A. C., S. B. Montgomery, A. S. Dimas, B. E. Stranger, C. Beazley *et al.*, 2010 Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet* 6: e1000895.
- 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 genetics* 8.
- Ober, U., W. Huang, M. Magwire, M. Schlather, H. Simianer *et al.*, 2015 Accounting for genetic architecture improves sequence based genomic prediction for a *Drosophila* fitness trait. *PLoS one* 10.
- Pérez-Enciso, M., J. C. Rincón and A. Legarra, 2015 Sequence-vs. chip-assisted genomic selection: accurate biological information is advised. *Genetics Selection Evolution* 47: 43.
- Ponsuksili, S., E. Murani, N. Trakooljul, M. Schwerin and K. Wimmers, 2014 Discovery of candidate genes for muscle traits based on GWAS supported by eQTL-analysis. *International journal of biological sciences* 10: 327.
- Rentzsch, P., D. Witten, G. M. Cooper, J. Shendure and M. Kircher, 2018 CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research* 47: D886-D894.
- Richardson, T. G., S. Harrison, G. Hemani and G. D. Smith, 2019 An atlas of polygenic risk score associations to highlight putative causal relationships across the human phenome. *Elife* 8: e43657.
- Sarup, P., J. Jensen, T. Ostensen, M. Henryon and P. Sørensen, 2016 Increased prediction accuracy using a genomic feature model including prior information on quantitative trait locus regions in purebred Danish Duroc pigs. *BMC genetics* 17: 11.

- Schmid, M., and J. Bennewitz, 2017 Invited review: Genome-wide association analysis for quantitative traits in livestock—a selective review of statistical models and experimental designs. *Archiv fuer Tierzucht* 60: 335.
- Sørensen, P., S. M. Edwards, P. Madsen, P. Jensen, I. F. Sørensen *et al.*, 2013 Genomic feature models, pp. in *Conference on Genomics of Common Diseases*.
- Speed, D., and D. J. Balding, 2014 MultiBLUP: improved SNP-based prediction for complex traits. *Genome research* 24: 1550-1557.
- Strandén, I., and D. Garrick, 2009 Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. *Journal of dairy science* 92: 2971-2975.
- Su, G., O. F. Christensen, L. Janss and M. S. Lund, 2014 Comparison of genomic predictions using genomic relationship matrices built with different weighting factors to account for locus-specific variances. *Journal of dairy science* 97: 6547-6559.
- Tiezzi, F., and C. Maltecca, 2015 Accounting for trait architecture in genomic predictions of US Holstein cattle using a weighted realized relationship matrix. *Genetics Selection Evolution* 47: 24.
- Torres, J. M., E. R. Gamazon, E. J. Parra, J. E. Below, A. Valladares-Salgado *et al.*, 2014 Cross-tissue and tissue-specific eQTLs: partitioning the heritability of a complex trait. *The American Journal of Human Genetics* 95: 521-534.
- Upperman, L. R., B. P. Kinghorn, M. D. MacNeil and A. L. Van Eenennaam, 2019 Management of lethal recessive alleles in beef cattle through the use of mate selection software. *Genetics Selection Evolution* 51: 36.
- van Binsbergen, R., M. C. Bink, M. P. Calus, F. A. van Eeuwijk, B. J. Hayes *et al.*, 2014 Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. *Genetics Selection Evolution* 46: 41.
- Van Binsbergen, R., M. P. Calus, M. C. Bink, F. A. van Eeuwijk, C. Schrooten *et al.*, 2015 Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. *Genetics Selection Evolution* 47: 71.
- Van den Berg, I., D. Boichard and M. S. Lund, 2016 Sequence variants selected from a multi-breed GWAS can improve the reliability of genomic predictions in dairy cattle. *Genetics Selection Evolution* 48: 83.
- Van Den Berg, I., B. Hayes, A. Chamberlain and M. Goddard, 2019 Overlap between eQTL and QTL associated with production traits and fertility in dairy cattle. *BMC genomics* 20: 291.
- Van Son, M., M. S. Lopes, H. J. Martell, M. F. Derks, L. E. Gangsei *et al.*, 2019 A QTL for number of teats shows breed specific effects on number of vertebrae in pigs: Bridging the gap between molecular and quantitative genetics. *Frontiers in Genetics* 10: 272.
- VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. *Journal of dairy science* 91: 4414-4423.
- Veerkamp, R. F., A. C. Bouwman, C. Schrooten and M. P. Calus, 2016 Genomic prediction using preselected DNA variants from a GWAS with whole-

- genome sequence data in Holstein–Friesian cattle. *Genetics Selection Evolution* 48: 95.
- Visscher, P. M., N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy *et al.*, 2017 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics* 101: 5-22.
- Weir, B., 2008 Linkage disequilibrium and association mapping. *Annu. Rev. Genomics Hum. Genet.* 9: 129-142.
- Wellcome Trust Case Control Consortium, 2007 Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447: 661.
- Windig, J. J., R. A. Hoving-Bolink and R. F. Veerkamp, 2015 Breeding for polledness in Holstein cattle. *Livestock Science* 179: 96-101.
- Wray, N. R., M. E. Goddard and P. M. Visscher, 2007 Prediction of individual genetic risk to disease from genome-wide association studies. *Genome research* 17: 1520-1528.
- Xiang, R., I. v. d. Berg, I. M. MacLeod, B. J. Hayes, C. P. Prowse-Wilkins *et al.*, 2019 Quantifying the contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits. *Proceedings of the National Academy of Sciences* 116: 19398-19408.
- Yang, J., T. Ferreira, A. P. Morris, S. E. Medland, P. A. Madden *et al.*, 2012 Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature genetics* 44: 369.
- Zhang, Z., J. Liu, X. Ding, P. Bijma, D.-J. de Koning *et al.*, 2010 Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. *PloS one* 5.
- Zhou, L., M. S. Lund, Y. Wang and G. Su, 2014 Genomic predictions across Nordic Holstein and Nordic Red using the genomic best linear unbiased prediction model with different genomic relationship matrices. *Journal of animal breeding and genetics* 131: 249-257.
- Zhu, Z., F. Zhang, H. Hu, A. Bakshi, M. R. Robinson *et al.*, 2016 Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature genetics* 48: 481.

Summary

Summary

Whole genome sequence (WGS) data is becoming increasingly available in animal and plant breeding. Theory suggests that WGS data can be utilized to improve the accuracy of genomic prediction (GP), as the data is expected to contain the causal variants that underlie complex traits. However, several studies have shown that WGS data have little to no impact on accuracy when GP is carried out within populations. In this thesis, I explored the potential of WGS data to increase the accuracy of genomic prediction across-populations and species. In this thesis, I present a number of methods utilizing WGS that can be used to improve the accuracy of GP across populations, and discussed the implications for breeding.

In **chapter 2**, I investigated the effect of using WGS data instead of SNP arrays on the accuracy of GP for stature across different breeds of cattle. I showed that the accuracy of GP across breeds does not improve by simply increasing the density of variants up to WGS level. Instead, I showed that accuracy can be significantly improved when variants are pre-selected from WGS data based on biological information such as the effect of variants on the trait in a genome-wide association study (GWAS). This is because the pre-selection helps to isolate the effect of important variants from the noise effect of the many neutral variants in WGS data. The results indicate, however, that there is a trade-off with respect to the pre-selection and use of only a few important variants for GP, namely, the pre-selected variants alone cannot explain the total genetic variance for the trait of interest.

In **chapter 3**, I implemented a biology-informed GP model, when multiple populations, for example multiple breeds or lines are combined in the reference population. In the model, instead of fitting a single GRM made from important pre-selected variants, SNPs on a standard 50k SNP-chip are used to create a second GRM, and both GRMs are fitted simultaneously. I showed that the multi-population, multiple GRMs (MPMG) model outperforms GP models in which only a single GRM, made using all variants, is used. In addition to the fact that the pre-select variants are isolated from the noise effect of other neutral variants, I showed that the model uses the 50k SNPs to capture the residual genetic variance that cannot be explained by the pre-selected variants. Furthermore, the model weighs information from all individuals in the reference population by their genetic correlation with the validation individuals. Thus, individuals in the reference population can contribute towards prediction only if they have a genetic correlation with the validation individuals that is different from zero.

In **chapter 4**, I used selection index theory to derive a prediction equation for the accuracy of the MPMG model. The model was used to underpin the advantages and limitations of the MPMG model over single-GRM models. Using the equation, I showed that the advantage of the MPMG model over other tested models is that in most cases, only a few hundred variants are pre-selected based on biological information and as such, the number of independent chromosomal segments is low. However, I showed that the advantages holds only when the biological information used to pre-select the variants is accurate, such that the pre-selected variant explain some proportion of the total genetic variance for the trait. The prediction equation derived in chapter 4 can be used to optimize the design of biology-informed genomic breeding programs.

The results of chapters 2, 3 and 4 all show the importance of identifying and pre-selecting important variants for complex traits to improve the accuracy of GP. Therefore, in **chapter 5**, I investigated the usefulness of summary-level result from the largest human height GWAS to date as prior information for identifying genes and gene-associated variants that affect stature in cattle. I found a strong evidence for a similarity in the biology of height across humans and cattle, but also that prior information from humans is not more useful for identifying cattle stature genes and variants than within-species prior information (cattle stature GWAS). In the absence of cattle stature GWAS, however, the result suggests that human height GWAS results can be useful in identifying cattle stature genes and associated variants.

Curriculum Vitae

About the author

Biaty Raymond was born on the 29th of July 1988 in Pella, Adamawa State, Nigeria. For his secondary school education, he attended the General Murtala Mohammed College in Yola, Nigeria, where he graduated in 2005. In 2007, he gained admission into the Federal University of Technology, Yola, where he studied for five years and graduated with a Bachelor of Technology (B.Tech.) in Animal Science and Range Management. Between July 2010 and June 2011, he studied at the Kinneret Academic College, Israel, as part of a student exchange program. After graduating from college in 2012, he participated in the compulsory Nigerian National Youth Service program for one year as a secondary school teacher in Kano State, Nigeria. In 2014, Biaty obtained an Erasmus Mundus scholarship to follow the European Master in Animal Breeding and Genetics (EMABG) program. In the first year of the program, he studied at Wageningen University and did a major research thesis investigating the use of genomic information in estimating direct and indirect genetic effects for survival time in crossbred laying hens. For the second year of the program, he studied at the Swedish University of Agricultural Sciences in Uppsala, Sweden, where he did a genome-wide association study for bone strength in laying hens as a second major thesis. Immediately after graduating from the EMABG program in 2016, Biaty started his PhD about the use of whole genome sequence data for genomic prediction at the Animal Breeding and Genomics Group and Biometris. The results of his PhD research are described in this thesis. During his PhD, Biaty stayed for 6 months at the Institute for Molecular Bioscience, Program in complex trait genomics, University of Queensland, Australia. Since May 2020, Biaty Works as a research scientist for Wageningen Livestock Research.

Peer-reviewed publications

Brinker T., **Raymond B.**, Bijma P., Vereijken A., Ellen E.D., Estimation of total genetic effects for survival time in crossbred laying hens showing cannibalism, using pedigree or genomic information, *Journal of Animal Breeding and Genetics*. 134 (2017) 60-68.

Raymond B., Bouwman A.C., Schrooten C., Houwing-Duistermaat J., Veerkamp R.F., Utility of whole-genome sequence data for across-breed genomic prediction, *Genetics Selection Evolution*. 50 (2018) 27.

Raymond B., Johansson A.M., McCormack H.A., Fleming R.H., Schmutz M., Dunn I.C., De Koning D.J., Genome-wide association study for bone strength in laying hens, *Journal of Animal Science*. 96 (2018) 2525-2535.

Raymond B., Bouwman A.C., Wientjes Y.C., Schrooten C., Houwing-Duistermaat J., Veerkamp R.F., Genomic prediction for numerically small breeds, using models with pre-selected and differentially weighted markers, *Genetics Selection Evolution*. 50 (2018) 49.

Raymond B., Bouwman A.C., Wientjes Y.C., Schrooten C., Veerkamp R.F., A deterministic equation to predict the accuracy of multi-population genomic prediction with multiple genomic relationship matrices, *Genetics Selection Evolution*. 52 (2020) 22.

Manuscript accepted

Raymond B., Yengo L., Costilla R., Bouwman A.C., Schrooten C., Hayes B.J., Veerkamp R.F., Visscher P.M., Using prior information from humans to prioritize genes and gene-associated variants for complex traits in livestock. *bioRxiv*. doi: <https://doi.org/10.1101/2020.04.16.044685>.

Conference proceedings

Raymond B., Bouwman A.C., Schrooten C., Houwing-Duistermaat J., Veerkamp R.F. (2017). Utility of whole genome sequence data for across breed genomic prediction. In: book of Abstracts of the 68th Annual Meeting of the European Federation of Animal Science (2018), Tallinn, Estonia, ISBN: 978-90-8686-312-9.

Raymond B., Bouwman A.C., Schrooten C., Houwing-Duistermaat J., Veerkamp R.F., Genomic prediction across populations, using pre-selected markers and differential weight models, Proceedings of the 11th World Congress on Genetics Applied to Livestock Production (2018), Auckland, New Zealand.

Raymond B., Bouwman, A.C., Wientjes Y.C.J., Schrooten C., Houwing-Duistermaat J. Veerkamp R.F. An equation to predict accuracy of multi-breed genomic prediction model with multiple random effects. In: book of Abstracts of the 69th Annual Meeting of the European Federation of Animal Science (2018), Dubrovnik, Croatia, ISBN: 9789086863235 .

Raymond B., Yengo L., Costilla R., Schrooten C., Bouwman A.C., Veerkamp R.F., Hayes B., Visscher P.M. Common genomic regions underlie height in humans and stature in cattle. In: book of Abstracts of the 70th Annual Meeting of the European Federation of Animal Science (2019), Ghent, Belgium.

Training and education

The Basic Package (3.0 credits)	
WIAS Introduction Day	2016
Course on philosophy of science and/or ethics	2018
Course on essential skills	2017
Disciplinary Competences (16 credits)	
Quantitative Genetics Discussion Group (QDG)	2016-2020
Design of Breeding Programs with Genomic Selection (Wageningen)	2017
Statistical Methods for Genome-Enabled Selection (Wrocław, Poland)	2018
Writing of project proposal	2016
Introduction to Graphical Models with Applications to Quantitative Genetics and Genomics (Armidale, Australia)	2019
EGSABG Summer school "emerging technologies in Animal Breeding" (Wageningen)	2017
Professional Competences (7.4 credits)	
Techniques for Writing and Presenting a Scientific Paper	2018
Brain Training	2018
Organization of the DuPont Symposium in Wageningen - Breeding Data: Statistical Advances in Modern Plant Breeding	2018
Organization of WIAS Science day - Beyond sustainability	2017
Project and time management	2019
The Final Touch: Writing the General Introduction and Discussion	2019
Career Perspectives	2019
Presentation Skills (4.0 credits)	
Oral presentation at EAAP (Tallinn, Estonia)	2017
Oral presentation at EAAP (Dubrovnik, Croatia)	2018
Oral presentation at EAAP (Ghent, Belgium)	2019
Oral presentation at WCGALP (Auckland, New Zealand)	2018
Teaching competences (4.0 credits)	
Supervision of MSc Student	2018
Assisting in BSc course: Animal Breeding and Genetics	2017
Assisting in the MSc Course: Genetic Improvement of Livestock	2017
Total credits	34.4

Acknowledgement

My PhD journey has now ended. While I am delighted to have completed this journey, I must acknowledge that I did not make it on my own. Along every step of the way were wonderful people who supported and cheered me on, to whom I would like to express my deepest gratitude.

First and foremost, I would like to thank my core supervisory team: Roel, Aniek, Yvonne and Chris, without whom this journey would not have been possible.

Roel, you once said something to me during one of our many meetings that have stuck with me: “when results from an analysis do not align with underlying theory, it is not the theory that is wrong”. To me, this statement characterizes you. You always encouraged me to look deeper than what the numbers say. You hardly ever lose sight of the big picture and I am very grateful that you have trained me to have this mindset. I believe that I am a better scientist today because of it. Thank you for always making yourself available and for always showing interest in my work. Thank for frequently saying the words “well done Biaty”. Hearing these words from you meant so much to me and kept me motivated. It has been an honor to have worked with and learnt from you.

Aniek, despite being the supervisor and having far more experience than me, you always listened to what I had to say. On those difficult days when my analyses were failing, when things didn’t make sense, you encouraged me in your usual calm demeanor, to keep at it and to learn from and enjoy the process. You encouraged me to own my research and to be proud of it. From helping to write some lines of code to visiting and supporting me and my wife when we had a baby, you were always there and I could always count on you. For these and many more than I can innumerate here, I am thankful. I look forward to continue to work and learn from you in the new job.

Yvonne, almost halfway through my PhD, you joined my supervisory team, and I am very thankful today that you did. You joined because you found my research interesting, that really encouraged me. I have learnt so much from the several hours that we spent in meeting rooms, working on equations, planning analyses and discussing results. You are a very smart and principled scientist and I am glad that I had the chance to work with you. No doubt, you have positively impacted my way of thinking and doing research, thank you.

Chris, thank you for the time you have invested into my research. Thank you for the many times you travelled from Arnhem to Wageningen to attend our meetings, for preparing and sending data when we make a request, for bringing the industry perspective to our discussions, for reading and commenting on manuscripts and many more. I am glad that I got to know and worked with you. I would also like to thank all other colleagues at CRV B.V. who always took interest in my research and supported me in various ways.

To my wonderful colleagues at ABG, thank you for creating a stimulating environment where young researchers such as myself thrive. I enjoyed the lively discussions at QDG, TLM, genomic prediction meetings, focus group meetings, coffee corner chats, social events and many more. While I cannot mention all of you by name, I would like to mention a few that deserve special thanks. Mario, Piter, Han and Jeremie, thank you so much that I could always and confidently approach you with challenges in my research and for always creating the time to attend to me even when I show up unannounced in your offices. Harmen, Maria, Zhou, Langqing, Chiara, thank you for creating the best office environment for a PhD candidate at ABG ☺. I have enjoyed every joke, every memory and story shared, every birthday celebration, every moment. I wish I had moved into the office much earlier. You guys are the best! Ibrahim, thank you for being a reliable friend and brother. Lisanne, I always looked forward to the days you worked at ABG, thank you for your friendship and for being my paranymph. Sanne, thank you for the wonderful times we shared in the topbreed project. Even outside ABG, you remain one of my closest friends, and someone I can always confide in, thank you, also for being my paranymph. Shuwen, you are a wonderful friend, thank you for the nice walks around campus during lunch and coffee breaks and for the great discussions we had. Pascal, thank you for always being ready to share and discuss research ideas and for the great times we had in Brisbane and Armidale. Lisette, thank you for your great smiles and your readiness to help with administrative requests. Thank you for helping to submit my PhD thesis and for your helpful suggestion on the layout. I wish you so much success at the WIAS office. Wilma, thank you for always asking how I am doing.

I would like to thank my colleagues at Biometris who supported me along my PhD Journey. Dinie, thank you for always helping with administrative task concerning my PhD. Fred, thank you for coordinating the topbreed project and for always approving my requests to attend courses and conferences abroad. Henry, thank you for the great discussions we had about the topbreed project.

I would like to thank everyone at PCTG, University of Queensland, for a wonderful six months in Australia. Loic, thank you for helping me to settle into the group and for your very smart suggestions for analyses. Thank you for always telling me that I am doing well, even when I felt that I was falling short. Roy, I am grateful to you and Emma for making me and my family feel at home in Brisbane. Thank you for the so many picnics, dinners and trips. I have found a wonderful friend in you. Ben, thank you for always creating the time to meet and discuss when I needed it. Working with you was truly inspirational. Peter, I consider the opportunity to work with you as one of the highlights of my PhD. I am sometimes shocked at how humble and down to earth you are. Thank you for always creating the time to discuss my research and for making me feel so comfortable in the group. I hope I have the opportunity to work with you again in the future. Naomi, thank you for always asking how I and my family are doing and for inviting us to your home. It meant a great deal to us.

I would like to express my thanks to my family of faith, the Amazing Grace Parish Church in Wageningen for providing a spiritual home, where I felt loved and cared for. To pastors Farai and Busi, and every member of the Church, thank you for everything. I have come to rely on your love and spiritual guidance especially for navigating those difficult times when I experienced overwhelming family loss. God bless you and continue to strengthen you for the work of love you do in Wageningen.

While the PhD research took four years to complete, to me, the journey began much earlier, thanks to my parents. To my mum and my late daddy, thank you for planting the idea of going all the way to a PhD level in my head from when I was little. Thank you for loving me, believing in and supporting me with everything you had. To you and all my siblings, I am eternally grateful. I love you.

To my wife, Homtapwa, words fail me. It is amazing how strong and supporting you have been throughout my PhD journey. Thank you for being my love, my best friend, my rock. Thank you for listening to my complaints of being tired even when you yourself are exhausted from taking care of the family. Thank you for being my audience when I had to practice for my presentations. Thank you for being a wonderful mother to our two amazing daughters. I love you with my whole being.

To Chayil and Charissa, my adorable daughters, I am thankful to God that you are in my life. I hope that through my PhD journey, I have shown you that anything is possible if you truly set your mind to it. Daddy loves you so much.

Colophon

This research was financially supported by the Netherlands Organization for Scientific Research (NWO) and the Breed4Food consortium partners Cobb Europe, CRV, Hendrix Genetics, and Topigs Norsvin, through the Topbreed project. The research visit of Biaty Raymond to the University of Queensland, Brisbane (Australia) was funded by the Wageningen Institute for Animal Sciences (WIAS).

The cover of this thesis was designed by Fenna Schaap.

Printed by: DigiForce | Proefschriftmaken.nl