ACCOUNTING FOR POPULATION ADMIXTURE IN GENOMIC EVALUATIONS

DOCTORAL THESIS

MAHLAKO L. MAKGAHLELA

ACADEMIC DISSERTATION

To be presented, with the permission of the Faculty of Agriculture and Forestry of the University of Helsinki, for public examination in Infokeskus Korona, Lecture Hall 2, Auditorium 235 at Viikki

Viikinkaari 11, Helsinki, on Friday, February 7th, 2014, at 12 o'clock noon.

Helsinki 2014

DEPARTMENT OF AGRICULTURAL SCIENCES | PUBLICATIONS | 30

Custos: **Professor Pekka Uimari**
Department of Agricultural Sciences
Box 28, FIN-00014, University of Helsinki, Finland

Supervisors: **Professor Esa Mäntysaari**
MTT Agrifood Research Finland
Biotechnology and Food Research, Biometrical Genetics
Myllytie 1, FIN-31600 Jokioinen, Finland

Adjunct Professor Jarmo Juga
Department of Agricultural Sciences
Box 27, FIN-00014, University of Helsinki, Finland

Docent Ismo Strandén
MTT Agrifood Research Finland
Biotechnology and Food Research, Biometrical Genetics
Myllytie 1, FIN-31600 Jokioinen, Finland

Professor Mikko J. Sillanpää
Departments of Mathematical Sciences, Biology and Biocenter Oulu,
Box 3000, FIN-90014, University of Oulu, Finland

Reviewers: **Professor Freddy Fikse**
Swedish University of Agricultural Sciences
Inst för HGEN
Box 7023, Gerda Nilssons väg 2, 750 07 Uppsala, Sweden

Professor Nicolas Gengler
University of Liège – Gembloux Agro-Bio Tech (GxABT)
Agricultural Sciences Department
Passage des Déportés 2, B-5030 Gembloux, Belgium

Opponent **Professor Theodorus Meuwissen**
Department of Animal and Aquacultural Sciences
Norwegian University of Life Sciences
Box 5003, 1432 Ås, Norway

Cover Illustration: © Viking Genetics

ISBN 978-952-10-8889-6 (Paperback)
ISBN 978-952-10-8890-2 (PDF)
Electronic publication at http://ethesis.helsinki.fi

Unigrafia
Helsinki 2014

To my daughter, Tumisang, the source of my inspiration

TABLE OF CONTENTS

LIST OF ORIGINAL PUBLICATIONS

This thesis is based on the following publications, which have been reprinted with the kind permission of their copyright holders:

I. **Makgahlela M. L.**, E. A. Mäntysaari, I. Strandén, M. Koivula, U. S. Nielsen, M. J. Sillanpää and J. Juga. 2013. Across breed multi-trait random regression genomic predictions in the Nordic Red dairy cattle. Journal of Animal Breeding and Genetics. 130:10-19.

II. **Makgahlela M. L.**, I. Strandén, U. S. Nielsen, M. J. Sillanpää and E. A. Mäntysaari. 2013. The estimation of genomic relationships using breedwise allele frequencies among animals in multibreed populations. Journal of Dairy Science. 96:5364-5375.

III. **Makgahlela M. L.**, I. Strandén, U. S. Nielsen, M. J. Sillanpää and E. A. Mäntysaari. 2013. Using the unified relationship matrix adjusted by breed-wise allele frequencies in genomic evaluation of a multibreed population. Journal of Dairy Science. DOI: 10.3168/jds.2013-7167.

The publications are referred to in the text by their Roman numerals.

The author participated in: 1) planning of studies I-III 2) data preparations for analyses 3) method developments and statistical analyses 4) interpretation of results 5) dissemination of research outcomes in journals as the main author.

ABBREVIATIONS

SNP	Single Nucleotide Polymorphism
QTL	Quantitative Trait Loci
LD	Linkage Disequilibrium
GS	Genomic Selection
AF	Allele Frequency
BP	Breed Proportion
EBV	Estimated Breeding Value
DRP	De-Regressed Estimated Breeding Value
IDD	Individual Daughter Deviations
EDC	Effective Daughter Contribution
DGV	Direct Estimated Genomic Value
GEBV	Genomic Enhanced Breeding Value
MME	Mixed Model Equations
BLUP	Best Linear Unbiased Prediction
GBLUP	Genomic Best Linear Unbiased Prediction
RDC	Red Dairy Cattle

ABSTRACT

Genomic evaluations of animals in multi-breed and admixed populations tend to ignore the population structure and assume that these populations are homogeneous, which may lead to limited success in the application of this technology. The objective of this Ph.D. thesis was to develop approaches for accounting for the admixed structure of the Nordic Red dairy cattle (RDC) and furthermore, investigate the predictive ability of these methods in the estimation of genomic enhanced breeding values. The Nordic RDC population is a composite of the Finnish Ayrshire (FAY), Swedish Red (SRB), Norwegian Red (NRF), Danish Red (RDM), and their crosses with other breeds. The study was carried out using individual breed proportions derived from the pedigree to define the base breeds, dense marker genotypes and phenotypes of progeny tested bulls with reliabilities from traditional evaluations close to one.

Two approaches were developed: (1) the multi-trait random regression model, which accounts for the interactions between marker effects and base breed origin of alleles, (2) the adjusted genomic relationship matrices by allele frequencies (AF) estimated within breeds versus across breeds, estimated from the currently genotyped versus the base (founding) population. Then, the predictive ability of genomic relationships accounted for breed composition was investigated in genomic evaluations with GBLUP of genotyped animals only, and GBLUP of both genotyped and ungenotyped animals (single-step GBLUP). Information in all evaluation models were weighted by the reliability of the phenotype (i.e., bull or cow deregressed breeding value). The validation of genomic evaluations for all models was assessed as the regression of phenotype on direct estimated genomic values or genomic enhanced breeding values.

Gains in validation reliabilities were 2 and 3% for milk and protein, respectively, and -1% using the multi-trait random regression model in comparison to GBLUP model that assumed a homogeneous population. The use of AF within breeds greatly reduced differences in additive genomic relationship coefficients between populations, when assessed both across and within sub-populations. This was more evident and closer to pedigree relationships when breed-wise AF were estimated from the base population. Whereas the use of AF across breeds increased genomic relationships, especially for individuals that were originating from populations that were further from the mean population AF across breeds. Accounting for the population structure with breed-wise AF also, relaxed assumptions when incorporating pedigree-based relationships for single-step GBLUP. This advantage however, was not achieved in genomic evaluations. The validation reliabilities between GBLUP with breed-wise AF and GBLUP with AF across breed were generally similar at 33% for milk and protein and 43% for fat. The validation reliabilities increased to 37%, 40% and 47% for milk, protein and fat, respectively, but were similar irrespective of AF used to compute genomic relationships in single-step GBLUP. The improvement in at least 5% for all traits with single-step GBLUP shows the benefit of utilizing all the available information into genomic evaluations.

From the methods developed, it was concluded that accounting for the population structure overall had marginal advantage in the predictive ability of genomic evaluations. However, as genomic selection is becoming a dominant tool, biased evaluations in multi-breeds from ignoring differences between breeds is clearly to be feared. Therefore, a more reasonable and cautious approach for integrating genomic information in multi-breeds would be from single-step evaluations that utilize cow performance record as phenotype and genomic relationships accounted for varying AF between the breeds’ founder populations.

1 OVERVIEW

1.1 INTRODUCTION OF GENETIC EVALUATIONS

Genetic improvement in livestock populations through the application of animal breeding techniques has been undoubtedly successful for many decades. Animal breeding has achieved its gains by estimating the genetic merit of selection candidates based on phenotype and pedigree information (Henderson, 1984). The genetic information is further used to make selection decisions. The high cost and time taken to identify animals of high genetic merit (i.e., breeding animals) has remained an impediment for even faster genetic progress (Schaeffer, 2006). More recently, developments in high-throughput genotyping platforms have allowed scientists and breeders to extend their tools to accommodate the new generated data, for long-term gain at a reduced cost and time (Meuwissen et al., 2001; Schaeffer, 2006). In dairy cattle, optimal use of all phenotypic, pedigree and genomic information currently plays a crucial role in genetic evaluations (Hayes et al., 2009a; Kearney et al., 2009; Reinhardt et al., 2009; Su et al., 2010, Aguilar et al., 2010).

1.2 TRADITIONAL EVALUATIONS

In traditional genetic evaluations, knowledge of individual phenotypic measurements and pedigree information is used to estimate breeding values (EBV) most often using best linear unbiased prediction (BLUP; Henderson, 1984) models. BLUP models often assume the infinitesimal model, which states that trait variation is determined by infinitely many unlinked genes, each of infinitesimally small additive effect (Falconer and Mackay, 1996). The simple additive model of genetic effects has been sufficient for the estimation of EBV for individuals in single breeds. Following the breeder’s interest in crossbreeding, BLUP models in multi-breed and admixed evaluations were easily extended to account for both intrabreed and interbreed additive effects, and non-additive genetic effects such as heterosis (Lo et al., 1993; Pollak and Quaas, 1998; García-Cortés and Toro, 2006).

Artificial insemination (AI) has been a method of choice for most dairy farmers globally (~80%), as a result, obtaining sire proofs through progeny testing is of utmost importance for widespread use. With large amount of data, the prediction reliability for such elite bulls for most economic traits can approach 100%. The EBVs of young unproven bulls however, remain mid-parent values, until their measured and tested daughters (i.e., after 5 to 6 years) are available. Then, an actual estimate of the bull's Mendelian segregation term, which is due to sampling of gametes from parents, is obtained. The reliability would generally be less (~80%) and gradually increase with increasing information from effective daughters and relatives.

1.3 GENOMIC EVALUATIONS

Over the last decade, genetic evaluations have been gradually extended to integrate DNA markers; the latest in this development is called genomic selection (GS). Genomic selection (also known as genomic evaluation or genomic prediction) utilizes whole-genome high-density single nucleotide polymorphism (SNP) markers or haplotype segments of these markers in the estimation of animal breeding values (Meuwissen et al., 2001; Goddard, 2009). In its most basic implementation, prediction equations are trained using older individuals with genotypes and phenotypes. Predictions are then applied to genotypes of young individuals assumed to have no phenotypes. Commonly used terms for these two sets of individuals are training set for older animals and the validation set for younger animals. The main advantage of GS is the reduction in generation interval by being able to predict the genetic merit (i.e., including Mendelian sampling term) of juvenile individuals without performance records. This increases the genetic gain through early selection. In principle, selection could be done as soon as the DNA is available (Pryce and Daetwyler, 2012) but in practice bull-calves are selected between 1 to 2 months of age. Reduced genotyping costs facilitated the application of GS in livestock (see for example Hayes et al., 2009a, Daetwyler et al., 2012, Chen et al., 2011; Forni et al., 2011) and plant (Resende et al., 2012a; 2012b) species.

1.3.1 Methodologies for genomic evaluations

One of the key issues in GS is to define the variance of the quantitative trait loci (QTL) explained by SNP markers, which is determined by the extent of linkage disequilibrium (LD) (i.e., a phenomenon in which two alleles at a locus do not occur independently in a population) between the QTL and SNP markers (Meuwissen et al., 2001). The QTL variance can be explained using either single SNP genotypes or haplotype segment of several markers (Calus et al., 2008; Hayes et al., 2009a; de Roos et al., 2011). Analytical methods have been mainly categorized into linear BLUP models, which assume SNP effects are drawn from a normal distribution with constant variance, and Bayesian models (i.e., Bayesian "a*lphabets*"), which may assume prior knowledge of unequal distribution of SNP effects and variances (Meuwissen et al., 2001; VanRaden, 2008; Gianola et al., 2009; Goddard, 2009; Hayes and Goddard, 2010). The performances of BLUP and Bayesian approaches tend to be comparable although Bayesian models perform better when the genetic architecture of the trait deviates from the infinitesimal model (Moser et al., 2009; Clark et al., 2011; Daetwyler et al., 2010). However, linear BLUP models have been most commonly used in practice due to straightforward implementation into existing evaluation tools and inexpensive computational demands.

Developments in genomic BLUP estimation of breeding values have been reviewed (e.g., Hayes et al., 2009a; Goddard and Hayes, 2010; de los Campos et al., 2013). Genomic evaluations are commonly implemented in a multi-step procedure. Firstly, EBV from traditional evaluations has to be deregressed and used as pseudo-data for GS (Garrick et al., 2009). This is done because the true genetic merit of the animal is unknown and also, as the phenotypic daughter yield deviations are not reported. The training population, which contains individuals with marker genotypes and pseudo-data, is then used to estimate SNP effects. Next, the estimated effects are summed over all markers to predict direct estimated genomic values (DGV) for selection candidates without phenotypes (i.e., SNPBLUP). Alternatively, DGV can be predicted using a genomic relationship matrix (**G**) in place of the numerator relationship matrix (**A**) within the mixed model equations (i.e., GBLUP) (Strandén and Garrick, 2009). Finally, genomic enhanced breeding values (GEBV) could be predicted by blending DGV and EBV using selection index procedure, to account for ancestral information from the EBV (VanRaden et al., 2009). Due to inconsistencies in accurate use of data between studies (e.g., response variables, weighting of phenotypes), Garrick et al. (2009) demonstrated an approach of deregressing breeding values, which pools different data sources while avoiding bias by weighting phenotypes. Several studies later examined this approach and noted that deregressed breeding values as phenotypes were more appropriate than EBV (Guo et al., 2010; Ostersten et al., 2011; Gao et al., 2013).

In GBLUP, the construction of genomic relationship matrix (**G**) from dense marker data plays a crucial role (Nejati-Javaremi et al., 1997; Habier et al., 2007). In contrast to the expected relationships in **A**, coefficients in **G** are based on the actual sharing of chromosome segments between individuals, which tend to deviate from expected relationships for closely related individuals. Furthermore, **G** matrix includes information on genes identical by state and also, captures unrecorded pedigrees (Powell et al., 2010). Several ways of deriving **G** within a population have been demonstrated (VanRaden, 2008; Yang et al., 2010). In their methods, each genotype is a deviation from marker specific population mean, which is calculated with population level AF. The construction of **G** in multi-breeds is currently carried out using observed AF across breeds (Hayes et al., 2009b), which may bias the derivation of **G** due to differences in AF between breeds (Harris and Johnson, 2010; Simeone et al., 2011).

Empirical application of multi-step evaluations heightened concerns such as loss of information and numerous assumptions, which in turn may limit the model performance. To address these issues and more, a single-step approach was developed by constructing and using a unified relationship matrix that combined genomic and pedigree information, for the estimation of GEBV for genotyped and extending the estimation of GEBV to ungenotyped individuals (Misztal et al., 2009; Aguilar et al., 2010; Christensen and Lund, 2010). Single-step evaluations, although requiring a little more computational time, provide a unified framework because the only change to conventional evaluations is to include genomic information (Aguilar et al., 2010). The accurate construction of **G** and optimal blending of **G** and **A** relationship matrices is the cornerstone for single-step evaluations (Forni et al., 2011; Meuwissen et al., 2011; Christensen et al., 2012).

1.3.2 Accuracy (reliability) of genomic evaluations

The accuracy (r) of GS is measured as the correlation between the estimated and true BV and has a linear relationship with response to selection (Meuwissen et al., 2001; Daetwyler et al., 2008). With empirical data, the true genetic merit of the animal is unknown and therefore, validation reliability (r^2), which has a similar function, is often used to test predictors (Mäntysaari et al., 2010). In simulation experiments, the accuracy of linear models for selection candidates range from 60 to 85% (Meuwissen et al., 2001; VanRaden, 2008; Vitezica et al., 2011; Daetwyler et al., 2013). The validation reliabilities for yield traits in breeds such as Holstein range from 50 to 67% and are over twice as high as those from parental average (Hayes et al., 2009a; Su et al., 2012a). Validation reliabilities for yield traits are generally 2 to 4% higher with single-step than multi-step evaluations (Vitezica et al., 2011; Gao et al., 2012; Koivula et al., 2012).

While prediction ability of GS is clearly better than that of the parental average, other challenges have emerged. The performance of GS appears to be limited in small populations (Thomasen et al., 2012; Brøndum et al., 2011). It was pointed out that one way to overcome the small training set is to combine data from multiple populations (de Roos et al., 2009; Hayes et al., 2009b; Brøndum et al., 2011). This strategy improved the validation reliabilities; however, the observed reliability in multi-breed and admixed populations is lower compared to homogeneous populations with large training set (Hayes et al., 2009a; Hayes et al., 2009b; Kizilkaya et al., 2010).

1.3.2.1 Factors affecting accuracy of genomic evaluations

Although the genetic mechanism is currently unclear, several factors underlie the prediction accuracy of GS. The key finding from simulations by Daetwyler et al. (2008) is that the accuracy of GS depends primarily on, 1) the amount of marker-QTL LD, which is a function of effective population size (i.e., breeding animals in an ideal population in which the effects on random drift and inbreeding would be similar to the actual population) and the number of markers 2) the size and structure of the training population (also known as the reference population) 3) heritability (i.e., proportion of variance due to additive genetic variance), and 4) the number of QTL and distribution of their effects.

1.3.2.2 Accuracy of genomic evaluations in multi-breed populations

Generally, multi-breed and admixed populations do not have either or both of the first two factors above required for improved accuracy. This is because population admixture constitutes a systematic differences in AF and LD phases between breeds due to differences in genetic background (Ewens and Spielman, 1995; Deng, 2001), which overall lowers the marker-QTL LD and hence the accuracy (de Roos et al., 2009; Hayes et al., 2009b). More so, SNP effects estimated from one breed would not accurately predict DGV for other breeds (Hayes et al., 2009b). In practice, however, evaluations ignore population structures and model common effects, assuming that multi-breeds are homogenous populations (Hayes et al., 2009b; Brøndum et al., 2011; Pryce et al., 2012).

Simulation studies indicated that the accuracy in admixed populations could be improved by increasing the marker density for the marker-QTL LD to persist across breeds (Ibánez-Escriche et al., 2009; de Roos et al., 2009). For such cases, there would be no need to account for breed-specific effects (Ibánez-Escriche et al., 2009). But this strategy may not hold because it addresses the artifact LD due to admixture as pointed out by Ewens and Spielman (1995), which might not reflect the actual LD within breeds and also, for more genetically isolated populations. Genomic selection in multi-breeds must be carried out using multi-breed procedures to account for all the genetic effects within and across breeds, as typically with conventional evaluations.

2 AIMS OF THE STUDY

The general aim of this study was to develop methods for accounting for the population structure in the estimation of genomic breeding values in the admixed Nordic RDC population. The specific aims (the order follows the list of articles) were:

I. To evaluate the predictive ability of a multi-trait random regression model that accounts for interactions between marker effects and breed of origin in the estimation of direct estimated genomic values in the Nordic RDC population.

II. To investigate whether the use of estimated breed-wise allele frequencies in the calculation of genomic relationships would provide a more accurate estimation of genomic relationships than using allele frequencies across breeds, and to determine the effect on genomic relationships when allele frequencies are estimated from the base population versus the currently genotyped population.

III. To investigate if accounting for breed origin of alleles in the calculation of genomic relationships derived with either currently genotyped or base population allele frequencies would improve the reliability of genomic enhanced breeding values using single-step GBLUP model.

3 MATERIALS AND METHODS

Materials and methods described in the original publications are referred to here with the Roman numerals I-III.

3.1 MATERIALS

3.1.1 DATA (I-III)

Data were published EBV for milk, protein and fat indices obtained from March 2010 routine evaluations of the Nordic Cattle Genetic Evaluation (NAV) (Interbull, 2008). The genomic information for 6,145 bulls generated using the Illumina BovineSNP50 BeadChip (Illumina Inc., 2005) was provided by the Nordic Genomic Selection project. Genotyped bulls were born between 1971 and 2006. The full RDC pedigree file contained 4,624,453 animals.

3.2 METHODS

3.2.1 POPULATION STRUCTURE (I-III)

The structure of the Nordic RDC population, which was used in Studies I- III, is an admixture of mainly the Danish Red, Swedish Red and the Finnish Ayrshire populations. These sub-populations are categorized by the country of birth or registration of the animal being Denmark (DNK), Sweden (SWE) and Finland (FIN). The full RDC pedigree was used to calculate the individual breed proportions (BP) for 16,010 bulls as shown by Lidauer et al. (2006). The information from BP revealed 13 known base breeds in the gene pool of the RDC. The names of the breeds identified have been given in paper I. Figures 1, 2 and 3 in paper I, illustrate trends in average BP between the years 1980 and 2006 for the Danish, Swedish and Finnish registered bulls, respectively. The average BP for most breeds in the data were however too small. Only 3 breeds contributed 10% or more to the gene pool. Therefore, breeds for Studies I-III as presented in Table 1, were defined as the Swedish Red (SRB), Finnish Ayrshire (FAY), Norwegian Red (NRF) and the remaining breeds with proportions less than 10% were combined in to breed "Other". In paper I, further information about the breakdown of BP percentage share by the 4 defined breeds has been provided.

3.2.2 GENOTYPES AND PHENOTYPES

The original genomic data were edited to remove uninformative SNP markers (I-III), for example, those with poor quality score or call rates, missing genotypes on more than 20% of the population and low minor allele frequencies. Markers with missing genotypes on at most 20% of the population were imputed using fastPHASE software (Scheet and Stephens, 2006). After the above edits, the final genotype data available for analyses in studies I-III were as presented in Table 1.

The original data included the EBV, their reliabilities and effective daughter contribution (EDC) for genotyped bulls (I) and cows (II-III). NAV models for evaluation of EBV account for heterosis among the base breeds, genetic groups and also, are corrected for heterogeneous variances among sub-populations (Lidauer et al., 2010). The EDC were calculated in *ApaX99* software following the approach described by Interbull (2004). For cows with records (II-III), the calculation of EDC was modified to exclude information provided by the dam, and the EDC indicated the amount of information in an individual cow. Deregression of EBV used an iterative procedure of Jairath et al. (1998) and Schaeffer (2001), implemented in *MiX99* software package (Lidauer and Strandén, 1999). Deregressed estimated breeding values (DRP) for the index traits were calculated by using *DeRegress* option (Strandén and Mäntysaari, 2010) with pedigree of bulls (I) and full animal model pedigree (II-III). Deregression models were weighted by EDC to account for differences in the information content between the individuals' EBV. An individual's reliability of DRP was calculated as $\mathrm{r}^2_{\mathrm{DRP}_i} = \mathrm{EDC}_i/(\mathrm{EDC}_i + \lambda)$, where $\lambda = (4 - h^2)/h^2$ (I) and $\lambda = (1 - h^2)/h^2$ (II-III). Thus, deregression of bull EBV included all bulls in the pedigree and used a sire model (I) while cow DRP were computed using an animal model (II-III). The genetic parameters and variance ratios used in deregression were obtained from NAV routine evaluations (Table 1). For each trait (I-III), the DRP with reliability less than 20% were removed from the data.

In paper II, individual daughter deviations (IDD), which are cow performances adjusted for fixed effects, non-genetic random effects and genetic effects of the cow's dam (Mrode and Swanson, 2004), were computed from deregressed cow EBV using an animal model from 305 day combined EBV (Mäntysaari et al., 2011). Thus, IDD are meta-EBV obtained by fitting animal model using cow DRP, an intermediate step in the calculation of daughter yield deviations. The difference between IDD versus cow DRP as data is that IDD account for the mates of the dams in the evaluation of genotyped bulls only but this information is excluded with cow DRP.

After merging different data, 4,142 genotyped bulls also had phenotype and BP information. As shown in Table 1, genotyped bulls were divided into the reference population, which were evaluated for the first time before 2005 NAV routine evaluations and young validation bulls that were not evaluated in 2005.

Table 1 Description of different data and trait parameters used for analyses in Studies I-III

					Trait parameters[4]
Study	Breeds[1]	No. of	Genotyped	No. of	In order of the traits
	% mean BP	markers	bulls[2]	records[3]	milk, protein, fat
I	SRB (20 %)	37,995	3,330[a]	Bull DRP	h^2 = 0.39, 0.31, 0.36
	FAY (46 %)		812[b]	3,330	$R^2_{\mathrm{DRP}_{\mathrm{ref}}}$ = 0.99, 0.98, 0.98[a]
	NRF (12 %)				$R^2_{\mathrm{DRP}_{\mathrm{val}}}$ = 0.94, 0.94, 0.92[b]
	OTHER (22 %)				
II	SRB (20 %)	38,194	3,300[a]	Cow IDD	h^2 = 0.40, 0.28, 0.32
	FAY (46 %)		806[b]	1,995,606	$R^2_{\mathrm{DRP}_{\mathrm{ref}}}$ = 0.96, 0.95, 0.95[a]
	NRF (12 %)				$R^2_{\mathrm{DRP}_{\mathrm{val}}}$ = 0.95, 0.93, 0.94[b]
	OTHER (22 %)				
III	SRB (20 %)	38,194	3,300[a]	Cow DRP	h^2 = 0.40, 0.28, 0.32
	FAY (46 %)		806[b]	2,816,745	$R^2_{\mathrm{DRP}_{\mathrm{ref}}}$ = 0.96, 0.95, 0.95[a]
	NRF (12 %)				$R^2_{\mathrm{DRP}_{\mathrm{val}}}$ = 0.95, 0.93, 0.94[b]
	OTHER (22 %)				

[1] Breeds defined in the data by % mean breed proportions (BP) = Swedish red (SRB), Finnish Ayrshire (FAY), Norwegian red (NRF), Combined breeds (OTHER); [2] Genotyped bulls were split into the reference population[a] and validation bulls[b]; [3] Pseudo phenotypes = deregressed estimated breeding values (DRP), individual daughter deviations (IDD), [4] heritabilities (h^2) used in the deregression of breeding values, and average reliabilities of DRP in the reference ($R^2_{\mathrm{DRP}_{\mathrm{ref}}}$) and validation ($R^2_{\mathrm{DRP}_{\mathrm{val}}}$) data sets.

3.2.3 ESTIMATION OF PEDIGREE AND GENOMIC RELATIONSHIPS

Pedigree relationships for all animals were estimated from the full RDC pedigree using *RelaX2* computer program (Strandén and Vuori, 2006). The genomic relationships in papers I-III (shown in *Appendix A*) were constructed following methods demonstrated by VanRaden (2008) and Yang et al. (2010). The effect of AF on **G** were examined by estimating AF for use in the construction of **G** in different approaches: 1) simple AF across breeds in the observed genotyped population (I-III) 2) AF across breeds estimated from the base (founder) population (II, III) 3) AF within breeds in the observed genotyped population and 4) AF within breeds estimated from the base population (II, III). Allele frequencies within breeds were estimated using either a linear (see the *Appendix A*) or binomial regression of gene content (i.e., number of copies of one allele in a genotype) on BP. Allele frequencies from the base population were estimated using an algorithm proposed by Gengler et al. (2007) (shown in *Appendix A*), which uses classical BLUP to impute genotypes for ungenotyped base animals and subsequently generate an estimate of selection and drift of AF.

In paper II, various approaches of estimating AF and their use in the construction of **G** are demonstrated. The original relationship matrices were computed following method 1 (**Gorg**) and 2 (**Gorg2**) of VanRaden (2008). The adjusted relationship matrices were calculated by modifying method 1 (**Gadj**) and 2 (**Gadj2**) of VanRaden (2008). Both methods were examined because method 1 within breeds is limited by scaling coefficients with the expected marker variances summed across the genome, which was achieved using method 2. Note that the labeling of different genomic relationship matrices in II and III was different but referring to the same methods. Accordingly, **Gorg** in II is the same as $\mathbf{G_{AB}}$ in III. Also, **Gadj2** in II is the same as $\mathbf{G_{BW}}$ in III.

The unified relationship matrices, which combined pedigree and genomic information, were derived following approaches by Aguilar et al. (2010) and Christensen and Lund (2010)

(III). In this study, the pedigree-based relationship matrix **A**, which included both genotyped and ungenotyped animals, was combined with different genomic relationship matrices **G**. The differences in **G** were based on AF used, where $\mathbf{G_{AB}}$ was computed with AF across breeds, and $\mathbf{G_{BW}}$ was derived with AF within breeds (II-III). Firstly, all elements in $\mathbf{G_{AB}}$ were scaled with factor $r = \frac{\text{trace}(\mathbf{A_{11}})}{\text{trace }(\mathbf{G})}$, where $\mathbf{A}_{11}$ is a sub-matrix of genotyped bulls, so that diagonals of $r\mathbf{G_{AB}}$ and $\mathbf{A}_{11}$ on average are equal. This is because coefficients in **A** and **G** are typically expressed differently. The correction factor *r* was not used for $\mathbf{G_{BW}}$ because the modification with breed-wise AF was expected to scale $\mathbf{G_{BW}}$ and **A** to the same level. Also, genomic predictions tested using $\mathbf{G_{BW}}$ with or without factor *r* converged similarly. Finally, each relationship matrix (i.e., $\mathbf{G_{AB}}$ or $\mathbf{G_{BW}}$) was combined with **A** for all pedigreed animals. Detailed illustration of incorporating **A** and **G** into a unified relationship matrix (**H**) is presented in III.

3.2.4 VARIANCE COMPONENTS ESTIMATION AND GENOMIC EVALUATIONS

A multi-trait random regression model (shown in *Appendix B*), which accounts for interactions between marker effects and breeds from which they originate, was developed to estimate breed-wise genetic variances for each trait (I). This model can be considered as an approximation of the multi-breed variance approach proposed by Lo et al. (1993) and García-Cortés and Toro (2006). Lo et al. (1993) described rules to estimate the additive genetic covariance between relatives in multibreed, which includes individual breed proportions and segregation variances. The covariance matrix can then be used with standard BLUP models however, the estimation of genetic variance tend to be challenging. The model by García-Cortés and Toro (2006) splits the EBV into breed-specific components and segregation terms, and allow the estimation of genetic variance but numerically expensive in practice. Both the above methods may not easily be adapted to genomic evaluations. The multi-trait random regression model in paper I estimates breed-wise variance components and DGV by fitting individual BP as fixed regression effects of the breed and also as random regression effects of the sire however, it does not account for the segregations terms. Strandén and Mäntysaari (2013) used a small example to demonstrated that the EBV were comparable (correlation=0.987) between the multi-trait random regression model (i.e., including segregation deviations) and multi-breed variance approach by García-Cortés and Toro (2006). The analyses of variance components in I and II were carried out using *ASReml 3.0* (Gilmour et al., 2009).

Pedigree-based EBVs were estimated using animal model (I, III). The predictions of DGV and GEBV were carried out using phenotypes of the reference population in *MiX99* software (I-III). In GBLUP analyses, the prediction of DGV for genotyped bulls were obtained by replacing **A** with **G** within the mixed model equations (MME) and fitting only the general mean in the model (I, II). In single-step GBLUP analyses, the prediction of GEBV for all animals in the pedigree were obtained by replacing **A** with unified relationship matrices **H**, within the MME (III). Differences between GBLUP evaluations (II) were based on whether **G** was derived accounting for breed origin of alleles or assuming single population and also, whether AF were estimated from the currently genotyped or from the base breed populations. Similarly, single-step GBLUP evaluations differed in the unified **H** matrix (III), where the **G** in **H** was either computed with breed-wise or across breed AF and whether AF were estimated from the currently genotyped versus the base breed population. All analytical models used the reliability of the phenotype as weight, defined as the EDC, to account for level of accuracy in the phenotypes as these were not the true breeding values of the animals (I-III).

3.2.5 VALIDATION OF GENOMIC EVALUATIONS (I-III)

The validation of DGV and GEBV generally followed the protocol for the Interbull validation test for genomic evaluations (Mäntysaari et al., 2010). Briefly, a linear regression model of DRP on DGV or GEBV, weighted by R^2_{DRP} of the bull was fitted in the validation population. Coefficient of determination (R^2) of the validation model was then used to address the accuracy of the DGV and GEBV, and the regression coefficient (b_1) was used to assess the biasedness in the prediction of DGV and GEBV.

4 RESULTS AND DISCUSSION

The primary objective of this study was to develop methods for accounting for the admixed structure of the Nordic RDC and furthermore, investigate their predictive ability in the estimation of genomic breeding values. We developed and validated the multi-trait (breed) random regression model (I), accounted for breed composition in the construction of genomic relationships (II) and assessed the performance of the modified genomic relationships in GBLUP (II) and single-step GBLUP (III).

4.1 BREED PROPORTIONS AND THE POPULATION STRUCTURE (I-III)

In paper I the RDC population structure as described by base breed proportions, has been shown to constitute 98% of individuals that are composite of at least 2 base breeds. Breed proportions by sub-population showed that the genetic constitution of the Swedish and Finnish populations comprises of 4 base breeds: SRB, FAY, NRF and the Canadian Ayrshire (CAY). Moreover, the amount of base breed crosses during the years 1980 and 1994 was smaller in SWE (~30%) and FIN (~20%) as demonstrated by trends in average BP (Figures 2 and 3, respectively, in Publication I). On the other hand, the genetic composition of the Danish population was more admixed with BP from at least 7 different breeds represented (Figure 1 in Publication I). In DNK, trend in average BP from the Danish Red breed dropped drastically between 1980 and 1991 while trend in average BP from the American Brown Swiss increased at nearly the same rate. After this period, genes from more breeds were also introduced, resulting in the DNK population being the most admixed of the 3 sub-populations constituting the Nordic RDC (Figure 1 in Publication I).

Breed proportions provide information on the level of base breed crosses in a population as recorded in pedigrees. One typical reason for crossbreeding is due to an increase in the level of inbreeding, which is associated with depression in performance of the animals (e.g., Thompson et al., 2000a; 2000b). Thus, the increased level of base breed crosses or number of breeds represented in DNK was partly a breeding program decision to control an increase in the rate of inbreeding that might have been observed, for example, prior to 1980 when the genetic constitution of the DNK population was over 80% from RDM. Increased inbreeding levels are especially common in bulls entering the AI progeny testing programs as the dairy industry rely heavily on few selected elite sires for breeding purposes and consequently, having an impact on the genetics of the breed or population (Thompson et al., 2000a; 2000b). On the contrary, importation of genetic materials into SWE and FIN was mainly driven by the expectation of extra genetic gain from elite bulls.

The accuracy of breed proportions depends greatly on the pedigree depth and completeness (Sørensen et al., 2008). In the Nordic RDC, most bulls have pedigree tracing back to the years 1950 and 1960, which would have the pedigree depth to 6 or 7 generations. In addition, some of the elite NRF bulls used heavily in SWE (SRB) and FIN (FAY) have pedigree tracing back to 1910-1920. However, pedigree information content was limited for a few bulls in DNK, which could influence the estimation of their BP. The equivalent complete generations, which measures the number of generations separating the individual from its furthest known ancestor (Maignel et al., 1996), was on average 4.8 in the entire RDC pedigree. Therefore, the RDC pedigree used in this study was generally considered to be deep and complete for accurate estimation of individual genetic contributions.

Previous studies on genomic analyses of the Nordic RDC have defined sub-populations by country of registration of individuals (i.e., DNK, SWE and FIN) (Schulman et al., 2009; Brondum et al., 2011; Rius-Vilarrasa et al., 2011). However, having characterized this population at the genetic level with individual breed composition, it is clear that the sub-populations defined by registration country are also admixed. Therefore, a more ideal approach to define sub-groups would be according to BP because breed fractions characterizes the sub-groups by the genetic constitution instead of their registration country.

Several methods of inferring breed composition or population structure have been developed (see review Price et al. 2010). These methods (e.g., principal component, structured association and cryptic relatedness) infer breed composition at the population level, and have been widely used in many fields. More appealing, algorithms have been developed to estimate the actual local ancestry at typed loci (Tang et al., 2006; Kuehn et al., 2011; Frkonja et al., 2012). Using locus-specific BP may be more informative versus pedigree-based BP, which are expected values and tend to assume that the contributions from all ancestors of a generation are equivalent (Sölkner et al., 2010). Our limitation in estimating locus-specific BP was the unavailability of pure base breed animals because methods that infer local ancestry along the chromosome initially estimate AF within the base breeds. In populations with pure base breeds and their crosses, it may be beneficial to consider actual estimates of chromosomal segments originating from a particular breed.

4.2 PEDIGREE AND GENOMIC RELATIONSHIPS (I-III)

4.2.1 Statistics of relationship coefficients

By examining the diagonal elements from different genomic relationship matrices in comparison to diagonal elements in **A**, it was found that coefficients in **G** had wider range (0.773-1.450) than **A** (1.000-1.135) (Table 3 in Publication II). Similarly, the variability of diagonal elements as measured by standard deviations was greater for **G** matrices compared to **A**. These observations were consistent when diagonal elements were examined across populations and within sub-populations (i.e., DNK, SWE and FIN). The differences in scale between pedigree-based and genomic relationship coefficients were unsurprising because the **A** matrix contains expected genome sharing between individuals given pedigree data, whereas **G** measures actual sharing between individuals at genotyped loci. Because **G** accounts for more variation among individuals (i.e., including Mendelian sampling deviations) than **A**, particularly for closely related individuals (e.g., full-sibs or half-sibs), it would characterize more adequately genome sharing than achieved through pedigree-based expectations only. More so, in cases were pedigree information is lacking or incomplete. In II, demonstration of our results focused on diagonal elements between methods however, both diagonal and off-diagonal elements were assessed. It was found that methods behaved similarly on the estimation of both diagonal and off-diagonal elements.

4.2.2 Effect of allele frequencies on genomic relationship coefficients (II)

With marker-derived relationships widely used in genomic evaluations, it remained important to address the precision of assuming multi-breed populations as homogeneous, which is currently done using AF across breeds to compute **G** (Hayes et al., 2009b; Koivula et al., 2012; Pryce et al., 2012). Indeed, the use of simple genotyped AF across breeds in **G** was found to scale genomic relationship coefficients unevenly between sub-populations. In paper II, Table 3 presents descriptive statistics of diagonal elements from different genomic relationship matrices. The means and standard deviations of diagonal elements were generally smaller when accounting for breed origin of alleles in **Gadj** and **Gadj2** (i.e., using AF within breeds) compared to **Gorg**, which ignored the population structure (i.e., using AF across breeds). Yang et al. (2010) proposed a different scaling of diagonal elements in **G** than presented here, which was also tested in this data, and resulted in smaller variation in diagonal elements.

Diagonal elements of **G** within sub-populations had smaller averages but slightly larger standard deviations in SWE and FIN using AF within breeds than across breeds. Of particular interest, the averages of pedigree diagonals were smaller in DNK (1.007) and greater in FIN (1.016) however; these averages were reversed for DNK (1.136) and FIN (0.979) in **Gorg** (Table 3 in II). These results imply that diagonal elements in **Gorg** increased for DNK registered animals and decreased for animals born in FIN when genomic relationships were computed with AF across breeds. This was contrary to earlier findings (e.g., Brøndum et al., 2011) and trends in BP (I) that the DNK population was more admixed than SWE and FIN and hence, exhibit low inbreeding levels in **A**. Thus, because genomic relationships are expressed as deviations from the mean population AF, DNK animals were further from the mean AF across breeds, which made their genotypes appear more related to each other than in reality. The mean AF across breeds was influenced significantly by animals registered in SWE and FIN. This was expected because firstly, they are genetically more related but are both distantly related to DNK animals (I). Secondly, these populations were well represented in the combined population while DNK had the least number of animals, as observed elsewhere (Toro et al., 2011; Simeone et al., 2011). This confirms thoughts noted earlier that diagonal elements in multi-breed could be distorted if breed means and variances are not accounted for in **G** (Harris and Johnson, 2010). On the other hand, such differences in coefficients between populations were clearly avoided in the current study by using AF estimated within breeds (II), as pointed out by Toro et al. (2011) that pooled data need clear definition of AF. In all cases, it is critical that the pedigree information is deep and complete because pedigree completeness influences the estimation of BP (Sørensen et al., 2008) and subsequently, AF within breed. An incomplete pedigree will also result in an imprecise estimation of **A** relationship matrix. The pedigree relationship matrix in our study accounted for common ancestry shared among the base breeds animals. Thus, ignoring differences in genetic level among these breeds may not approximate well the estimation of **A** for multi-breed populations.

4.2.3 Effect of base population definition on genomic relationship coefficients (II)

Pedigree coefficients, which are twice the expected average identity by descent (IBD) of Malécot (1948), are classically expressed relative to the base or founding population. The founder animals have no known parents; often assumed to be unselected and unrelated. In the genomic context, relationships are widely expressed relative to the current base generation defined by scaling coefficients with AF of the observed genotypes (e.g., VanRaden, 2008; Powell et al., 2010; Yang et al., 2010; Goddard et al., 2011). Although rarely used in practice, the base population of **G** could also be defined in previous base generations by scaling coefficients with AF estimated for ungenotyped base animals from the pedigree data (Gengler et al., 2007; VanRaden, 2008; VanRaden et al., 2009).

The distributions in diagonal elements from different **G** built assuming the observed genotyped population to be the founder generations have been presented in Figure 1. Similarly, these distributions have been presented in Figure 2 but assuming the founder population in the past generation. Averages of diagonal elements from **G** using AF within breeds and from the base population were close but less than 1.0, for an unknown reason (Table 4 in II). An uneven tendency of using AF across breed in the genotyped population is clearly illustrated by two peaks in **Gorg** (Figure 1). The distribution of off-diagonal elements for **Gorg** also had 2 peaks across populations. In sub-populations, **Gorg** had two peaks for both diagonal and off-diagonal elements in DNK but not in SWE and FIN. The peak smoothed slightly when AF were estimated from the base population (Figure 2). This unevenness was avoided in both methods that utilized AF within breeds. The advantage of using AF from the base population of each breed was observed in Figure 2 where the spread of the distribution was further reduced. Thus, pedigree information accounted for selection and drift in AF over time thereby adjusting coefficients, especially for genetically distant individuals; with their respective breed means and variances that may have been imprecise in the currently genotyped generation. Moreover, correlations between diagonal elements of **G** and **A** were all close to zero with the current base generation but increased to 0.16 and 0.38 for **Gorg** and **Gadj2**, respectively, with the past base generation (Paper II). In the estimation of base-breed AF, our study only defined the base breeds as SRB, FAY, NRF and breed "Other", which combined small breeds with average BP <10% in the population. Alternatively, further division of breed "Other" into many smaller base breeds might yield different estimates of genomic relationships. As mentioned above, it is critical that the pedigree quality is good as subsequent analyses depend on its depth and completeness.

The observed correlations between diagonal elements of **A** and **G** were comparable to those of Aquilar et al. (2010) but smaller than estimates reported by VanRaden (2008), Toro et al. (2011) and VanRaden et al. (2011). These differences may be attributed to varying population structures of the analyzed data. However, the agreement is that the **G** matrix derived with AF from the base population is more correlated to **A** (VanRaden, 2008), which is logical because **G** and **A** would be somewhat expressed relative to a similar base generation. Furthermore, using base population AF within breeds to some extent yielded improved values in **Gadj2** relative to **A**, which simplified the blending of these information sources into a unified relationship matrix **H**. In ssGBLUP, scaling of **G** before combining it with **A** tends to be complex due to strong assumptions but is currently used in evaluations (Chen et al., 2011; Forni et al., 2011; Meuwissen et al., 2011; Christensen et al., 2012). This scaling had no effect on ssGBLUP evaluations after modifying **Gadj2** with AF within breeds (Paper III).

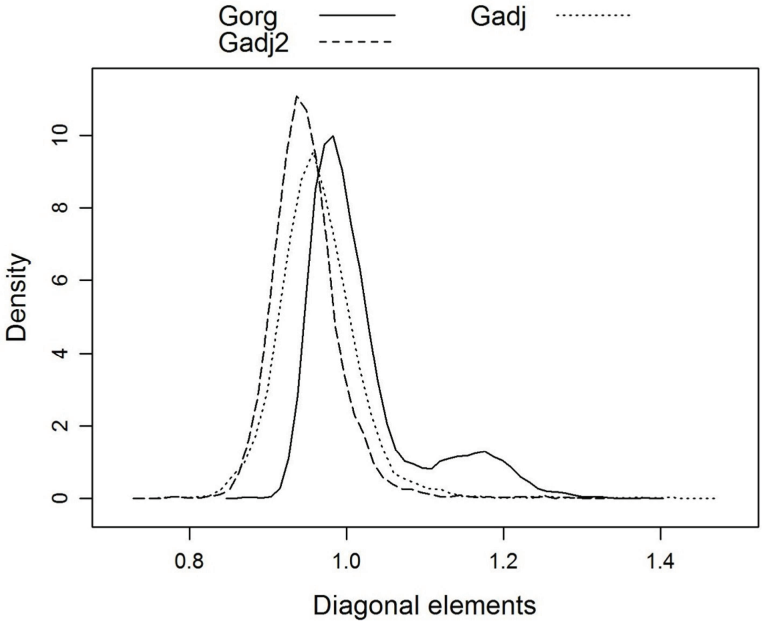

Figure 1 Distributions of diagonal elements from genomic relationship matrices with allele frequencies (AF) from the observed population. **Gorg** (**G_{AB}** in III) was built using the original method 1 of VanRaden (2008) and AF across breeds; **Gadj** and **Gadj2** (**G_{BW}** in III) were built adjusting method 1 and 2, respectively, of VanRaden (2008) and AF within breeds.

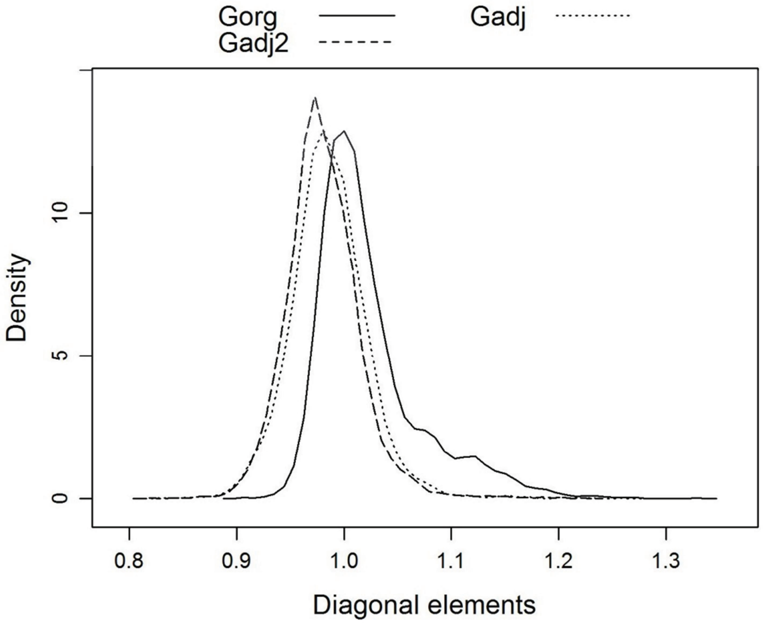

Figure 2 Distributions of diagonal elements from genomic relationship matrices with allele frequencies (AF) from the base population. **Gorg** (**G_{AB}** in III) was built using the original method 1 of VanRaden (2008) and AF across breeds; **Gadj** and **Gadj2** (**G_{BW}** in III) were built adjusting method 1 and 2, respectively, of VanRaden (2008) and AF within breeds.

4.3 ESTIMATED VARIANCE COMPONENTS: EFFECT OF DATA AND MODELS

Breed-specific sire variances and their averages for each trait estimated with bull DRP as data are presented in Tables 2 and 3, respectively, in paper I. Sire genetic variances were not greatly different between breeds, except they were higher in NRF, which may have been influenced by the smaller average BP in the data. Averages of sire variances were close to 100 for all traits in the DRP scale from NAV, which is due to standardization of EBV and depends on the accuracy of EBV. However, using bull DRP greatly inflated the estimated residual variances, which led to twice as high variance ratios compared to traditional evaluations. Because the same residual variances were estimated with both GBLUP and multi-trait random regression model, bull DRP as data for genomic evaluations may have limitations. Estimated additive genetic and especially residual variances were more logical when IDD or cow DRP were used as data (Table 2). The benefits and drawbacks between different response variables will be discussed later.

Our multi-trait random regression model allowed easier estimation of breed-wise sire variances, which has been numerically expensive in earlier studies (Lo et al., 1993; García-Cortés and Toro, 2006). The estimation of breed-wise residual variances and covariance between breeds remained computationally challenging (I). Covariance between random regression terms was not accounted for in models of García-Cortés and Toro (2006) and Strandén and Mäntysaari (2013), most likely because it's included in the segregation variance. The segregation variance results from differences in allelic frequencies between pure breeds, and is derived as the difference in additive variances between breed groups (Lo et al., 1993). Segregation deviations however, were not accounted for in our model. As the multi-trait random regression model assumed different marker effects between breeds, it can be thought that covariance information would have being an indication of breed-wise marker differences (I). Although our model may have suffered from the current admixed structure, the same model was later shown to be more efficient in multi-breeds with distinct base breeds and their crosses (Olson et al., 2012).

The observed bias in sire and residual variances with bull DRP may be due to sampling of heavily selected individuals in the reference population. Using single-step GBLUP and raw phenotypes, Forni et al. (2011) noted that additive genetic variances for litter size were sensitive to a method used to construct **G** when most individuals in **A** are genotyped. This appears to concur with our findings that a subsample of genotyped data could yield imprecise variance estimates. The authors suggested that a reason for biased estimates could be the differences in scale between **G** and **A** relationship matrices. However, our estimates were not significantly different between methods used to construct **G** (Table 2). The underlying reason for the dependency of variance components on the data is unclear but regardless of the cause, biased variances or heritabilities further influences the predictive power. As Hill (2010) said "BLUP is the best in the sense of minimum variance among linear predictors, but only if population parameters are well estimated."

For the models tested, genomic measures that correspond to heritability (i.e., the ratio of additive genetic variance to total variance) were less than those traditionally estimated with pedigree information (I, II). This agrees with the general consensus among studies that genomic measurements of heritability tend to be lower than traditional evaluations (Visscher et al., 2008; Rolf et al., 2010; Yang et al., 2010; Jensen et al., 2012). This appears to be true irrespective of the population structure and has been associated with incomplete marker-QTL LD due to lower minor allele frequency of the causal variants than in available commercial SNP marker data (Yang et al., 2010). Nonetheless, comparing estimates from classical BLUP and GBLUP may be unreasonable because BLUP is based on the infinitesimal model and GBLUP utilizes only a finite number of SNP markers (Daetwyler et al., 2012; de los Campos et al., 2012). Secondly, in addition to having a few genotyped animals, the expression of additive genetic variation is different in both models due to differences in the definition of founder populations in their covariance relationship matrices (Study II). Single-step evaluations, on the other hand, were found to estimate the additive genetic variances that were more stable and comparable to pedigree estimates, irrespective of the choice of **G**, when analysis include all genotyped and ungenotyped animals (Forni et al., 2011). In study III, genetic parameters from traditional evaluations were used directly in single-step GBLUP. Thus, single-step evaluations of all animals in the pedigree would be an ideal strategy to avoid possible biases in the estimation of additive genetic and residual variances. This assumes that pedigree and genomic data are weighted optimally and in study III, we have showed an easier integration of these information sources for multi-breed populations.

Table 2 The estimated additive genetic variance (σ_a^2) and residual variance (σ_e^2) by trait

Method[1]	Milk		Protein		Fat	
	$\hat{\sigma}_a^2$	$\hat{\sigma}_e^2$	$\hat{\sigma}_a^2$	$\hat{\sigma}_e^2$	$\hat{\sigma}_a^2$	$\hat{\sigma}_e^2$
Observed AF						
Gorg	31.27	293.60	33.58	408.04	28.47	382.06
Gadj	32.66	293.61	34.67	408.05	29.58	382.07
Gadj2	30.53	293.61	32.84	408.05	27.98	382.06
Base population AF						
Gorg	31.55	293.603	33.91	408.04	28.78	382.06
Gadj	39.70	293.61	35.02	408.05	29.60	382.07
Gadj2	31.37	293.61	33.75	408.05	28.07	382.07

[1] **Gorg** (**G$_{\mathbf{AB}}$** in III) was built using the original method 1 of VanRaden (2008) and allele frequencies (AF) across breeds; **Gadj** and **Gadj2** (**G$_{\mathbf{BW}}$** in III) were built adjusting method 1 and 2, respectively, of VanRaden (2008) and AF within breeds.

4.4 THE VALIDATION RESULTS

The accuracy and unbiasedness of the predictions in Studies I-III as measured by regression coefficients and reliabilities from the validation models are presented in Table 3. The validation results are presented for the EBV (I, III), DGV (I, II) and GEBV (III) of selection candidates or validation bulls for milk, protein and fat.

4.4.1 Validation regression coefficients

Regression coefficients in the validation analyses were generally higher from genomic evaluations than from pedigree-based animal model (I, III). In paper I, the validation regression coefficients for milk and protein were slightly higher at 0.06 and 0.03 units, respectively, when accounting for breed-specific effects in the model compared to assuming a homogeneous population. However, regression coefficients were similar between models for fat. This means that the level of bias was slightly reduced for milk and protein but not for fat when accounting for breed-specific SNP effects than modeling these effects similarly across breeds. In study II, the b_1 regression coefficients were in general similar across traits, regardless of whether the covariance matrix in GBLUP (i.e., **G** matrix) accounted for breed composition of the individuals by using AF within breeds or ignoring the population's admixed structure and using AF across breeds. The b_1 regression coefficients in single-step GBLUP (III) were slightly higher when **G** was computed using AF across breeds compared to AF within breeds. In addition, regression coefficients were slightly higher when genomic relationship matrices used AF from the currently genotyped versus the base population. Thus, although AF significantly influenced the estimation of **G** coefficients in II and III, there was little improvement if any in reducing the bias in GS when using the modified relationship matrices in both GBLUP and single-step GBLUP.

The validation regression coefficients b_1 in I-III were in agreement with the literature reports for single (Aguilar et al., 2010; Vitezica et al., 2011; Christensen et al., 2012; Gao et al., 2012) and multi-breed (Koivula et al., 2012; Su et al., 2012a; Harris et al., 2012) populations. The observed regression coefficients however, were reported to be less than the expected value of one, which suggests that genomic evaluations (i.e., DGV or GEBV) tend to be inflated or biased, hence overestimate the phenotypes (i.e., DYD, DRP or performance measurements) for validation bulls (Mäntysaari et al., 2010). Inflation of DGV and GEBV has been a widely reported concern for all models utilized in GS and the source is currently unclear (Olson et al., 2011; Vitezica et al., 2011; Forni et al., 2011). Olson et al. (2011) noted that pre-selection of validation bulls based on EBV or DRP when genotyping could reduce the validation regression coefficients from its expectation. But in the current study, this could not have been the case because the population analyzed in I-III included all bulls in almost all the birth years to reduce the possibility of selective genotyping. Furthermore, the inflation was also found in the validation of pedigree-based parental averages (I, III). Inflation of parental averages is associated with preferential treatment to the bull-dams (Olson et al., 2011). Information from bull-dams is often excluded in genomic evaluations, and hence, the source of bias or inflation of DGV and GEBV remains unknown, and would need to be investigated.

Simulating traits with different heritabilities, Vitezica et al., (2011) examined the cause of bias as measured by the validation regression coefficients, prediction error variance and mean square error between GBLUP and single-step methods. They found negligible differences between the b_1 terms at 0.01-0.03 units but in favour of single-step. The differences increased and still in favour of single-step for the remaining two measurements of bias depending on the simulated heritability and criteria of selection for breeding purposes. This tells that levels of bias found were slightly better with single-step GBLUP. However, more efforts are needed to reduce this inflation to a level close to zero.

4.4.2 Validation reliabilities

The gain in validation reliabilities when accounting for breed-specific effects (i.e., multi-trait random regression models) over GBLUP was 2% and 3% for milk and protein, respectively, using bull DRP as data (I). Here, the validation reliabilities from both the multi-trait random regression and GBLUP models were twice of those from pedigree-based evaluations.

Reliabilities for GBLUP seemed slightly higher for milk and protein using cow IDD (II) versus bull DRP (I) as data. However, it should be emphasized that cow IDD were used for convenience and were not expected to contain any additional information. But because we earlier noticed that direct use of cow DRP in GBLUP excludes information from the mates and therefore, yielded lower validation reliabilities. Although cow IDD and DRP as data for genomic evaluations resulted in higher validation reliabilities, the validation regression coefficients from these evaluations were surprisingly smaller than found for bull DRP. A possible explanation could be that the EBV of the cow is typically less reliable than that of the bull hence; there was smaller variance in the DGV estimated with bull DRP compared to cow IDD or DRP. In study I and II, the validation reliabilities for fat were similar between methods that accounted for or ignored the population structure. The validation reliabilities from pedigree evaluations were higher in III than I (Table 3). This increase in reliabilities was due to more information in III as evaluations included genotyped and ungenotyped animals while evaluations included only genotyped bulls and their pedigree (I).

Ideally, the true animal genetic merit should be used as phenotype for GS but this is unknown. In the absence, daughter yield deviations (DYD), which measure actual deviation of performance of the daughters, and DRP have been shown to be reliable indicators of genetic information (VanRaden , 2008; Garrick et al., 2009; Guo et al., 2010; Ostersten et al., 2011). These analogue variables were derived after EBV, which are easily accessible, were found to shrink genomic breeding values thereby changing their scale and also, tend to double–count information from relatives (Guo et al., 2010). These issues would not matter with DYD. However, DYD are not readily available from the routine evaluation databases. As a result, EBVs are typically deregressed (i.e., DRP) to be similar to DYD (Garrick et al., 2009; Strandén and Mäntysaari, 2010). Alternatively, in a recent study of Vandenplas and Gengler (2012), Bayesian procedures were improved simulating dairy cattle set-up, to integrate different sources of data while avoiding double-counting of information from relatives. Although it only attends to the issue of double counting, computational demands were also found to increase as double-counting was avoided.

Accounting for breed composition of an individual in the construction of **G** unexpectedly, resulted in no gain in the validation reliability (II, III). Reliabilities were all similar (II) and in some cases 1-2% higher (III) when AF were obtained across breeds compared to those estimated within the base breeds, and also, when AF were estimated from the currently genotyped individuals as opposed to AF from the base population. As mentioned earlier, this indicates that coefficients in **G** were sensitive to AF used. However, the predicted individual genetic values were unaffected. The tendency of **G** being sensitive to AF used but generating similar genomic values was earlier noted for single breeds with GBLUP (VanRaden, 2008) and single-step evaluations (Forni et al., 2011). In multi-breeds, Harris et al. (2012) used single-step with performance records to evaluate purebred Holstein and Jersey, and their crossbreds. In agreement to our results, they found small differences between validation reliabilities when **G** was adjusted to account for the population structure.

While the validation reliabilities from multi-step GBLUP ranged from 30-33% for milk and protein, and 42-43% for fat, the corresponding ranges increased to 37%-40% for milk and protein and 46-47% for fat using single-step GBLUP. Our results fall within the reported range (21-57%) for GBLUP evaluation of production traits in multiple populations (Harris and Johnson, 2010; Hayes et al., 2009b; Pryce et al., 2011; Koivula et al., 2012). Bayesian models generally achieve 0-3% higher reliabilities than GBLUP (Moser et al., 2009; Pryce et al., 2011; Gao et al., 2013). Our ranges however, were smaller than 53-67% for GBLUP in single breed evaluations (Hayes et al., 2009a; Kearney et al., 2009; Reinhardt et al., 2009; Su et al., 2010). Results from single-step GBLUP were comparable to those by Gao et al. (2012) in Holstein population but smaller compared to Harris et al. (2012) in crossbreds of Holstein and Jersey breeds. These results clearly show the added advantage of including all pedigreed individuals in genomic evaluations, regardless of their genotypic status. Despite this fact, also, highlighting a critical gap between the reliability of GS in single and multiple or admixed populations, which needs to be addressed through further research.

Table 3 The validation regression coefficients (b_1) and reliabilities (R^2) of pedigree-based estimated breeding values (EBV) (I, III), direct estimated genomic values (DGV) (I, II) and genomically enhanced breeding values (GEBV) (III) by trait

Study	Method[1]	Regression coefficient (b_1)			Validation reliability (R^2)		
		Milk	Protein	Fat	Milk	Protein	Fat
I	PED	0.74	0.73	0.88	0.15	0.15	0.23
	GBLUP	0.78	0.82	0.94	0.30	0.29	0.43
	mt-RRBLUP	0.84	0.85	0.94	0.32	0.32	0.42
II	$\mathrm{GBLUP}_{\mathrm{AB}_{\mathrm{obs}}}$	0.71	0.75	0.81	0.32	0.33	0.43
	$\mathrm{GBLUP}_{\mathrm{BW}_{\mathrm{obs}}}$	0.71	0.75	0.80	0.32	0.33	0.42
	$\mathrm{GBLUP2}_{\mathrm{BW}_{\mathrm{obs}}}$	0.72	0.76	0.82	0.33	0.33	0.43
	$\mathrm{GBLUP}_{\mathrm{AB}_{\mathrm{base}}}$	0.71	0.75	0.81	0.32	0.33	0.43
	$\mathrm{GBLUP}_{\mathrm{BW}_{\mathrm{base}}}$	0.71	0.75	0.80	0.32	0.33	0.42
	$\mathrm{GBLUP2}_{\mathrm{BW}_{\mathrm{base}}}$	0.72	0.76	0.82	0.33	0.33	0.43
III	PED	0.72	0.89	0.81	0.24	0.25	0.28
	$\mathrm{ssGBLUP}_{\mathrm{AB}_{\mathrm{obs}}}$	0.77	0.90	0.85	0.37	0.40	0.47
	$\mathrm{ssGBLUP2}_{\mathrm{BW}_{\mathrm{obs}}}$	0.75	0.88	0.84	0.36	0.39	0.47
	$\mathrm{ssGBLUP}_{\mathrm{AB}_{\mathrm{base}}}$	0.76	0.86	0.82	0.37	0.40	0.47
	$\mathrm{ssGBLUP2}_{\mathrm{BW}_{\mathrm{base}}}$	0.72	0.78	0.80	0.36	0.38	0.46

[1] Pedigre-based animal model (PED); multi-trait random regression model (mt-RRBLUP); genomic best linear unbiased prediction (GBLUP) with the genomic relationship matrix (**G**) computed using: 1) observed allele frequencies (AF) across breeds (GBLUP, $\mathrm{GBLUP}_{\mathrm{AB}_{\mathrm{obs}}}$), 2) observed breed-wise AF ($\mathrm{GBLUP}_{\mathrm{BW}_{\mathrm{obs}}}$and $\mathrm{GBLUP2}_{\mathrm{BW}_{\mathrm{obs}}}$), 3) base population AF across breeds ($\mathrm{GBLUP}_{\mathrm{AB}_{\mathrm{base}}}$) or breed-wise ($\mathrm{GBLUP}_{\mathrm{BW}_{\mathrm{base}}}$); **G** in II were built using method 1 ($\mathrm{GBLUP}_{\mathrm{AB}}$) or adjusting methods 1 and 2 ($\mathrm{GBLUP}_{\mathrm{BW}}$ and $\mathrm{GBLUP2}_{\mathrm{BW}}$) of VanRaden (2008); single-step GBLUP (ssGBLUP) analyses with **G** computed as described in II

4.4.3 Why the low validation reliability in multi-breed populations?

Most evaluations in admixed and multi-breed populations ignore breed composition and assume that these populations are homogenous (e.g., de Roos et al., 2009; Hayes et al., 2009b; Pryce et al., 2011). Firstly, because genomic selection exploits LD, where the assumption is that marker effects are the same across the population given sufficient marker-QTL LD (Meuwissen et al., 2001). As we have earlier mentioned, this LD is an artifact in multi-breeds and, hence, this assumption implies that the genetic backgrounds within breeds are not accounted for, that marker effects across breeds are similar and residuals follow a single normal distribution. Secondly, modelling breed composition has been ignored because simulations showed no gain in the accuracy when fitting breed-specific effects (Ibanez-Escriche et al., 2009; Zeng et al., 2013). This finding was somewhat different from reports by Hayes et al. (2009b) and Kizilkaya et al. (2009), who found that marker effects from one breed do not accurately predict genomic values when applied to other breeds, hence, the need to account for differences in LD phase between breeds. However, in support to earlier findings, our multi-trait random regression model, which defines vectors for breed-specific effects as well as animal genomic values, achieved negligible gain when applied in this population, and elsewhere (Olson et al., 2012). Furthermore, there was no gain when accounting for varying allele means and variances between breeds by adjusting genotypes with AF estimated within breeds in this study, and elsewhere (Harris et al., 2012).

On the other hand, theoretical (de Roos et al., 2009) and empirical (Hayes et al., 2009b) arguments indicated that reliabilities could be improved by increasing the marker density such that the marker-QTL LD persist across breeds, particularly for distantly related populations. The feasibility of imputation software's like fastPHASE and Beagle, amongst others (Scheet and Stephens, 2006; Browning and Browning, 2009, respectively) in imputing available markers to higher densities were then examined (Hayes et al., 2012; Brøndum et al., 2012). However, the validation reliabilities improved by about 5% using higher density data (i.e., ~800K) over 50K in single and multiple breeds (Harris et al., 2011; Su et al., 2012b). Note that although validation reliabilities are low for marketing and breeding purposes, these reliabilities are more than twice of those from parental averages. Admixture is not only affecting the predictive ability of GS but also, genome-wide association studies have been equally reporting spurious associations and inflation problems (see for example reviews by Astle and Balding, 2009 and Price et al. 2010). Similarly, Janss et al. (2012) and Sul and Eskin (2013) noted minimal differences between models with or without population correction factors.

With these issues, the simple answer to the above question is uncertain. However, while the ultimate goal for genomic evaluations is to generate individual genomic breeding values that validate accurately or reliably, it may be beneficial to achieve this without imposing strong assumptions. The effect of sufficient marker-QTL LD on the accuracy is clear, but improving LD by increasing data instead of modeling inconsistencies in marker-QTL LD between breeds may well improve the accuracy in the short-term. However, the long-term consequences in breeding programmes may well become a prospective challenge. In Zeng et al. (2013), the response to selection was generally higher with breed-specific over additive models but they argued that the superiority of breed-specific over additive models may be due to dominance effects versus differences in marker-QTL LD between breeds.

4.5 Future considerations

In spite of progress in fundamental aspects such as, analytical approaches, development of various marker panels, strategic genotyping by imputation techniques, and generating reference populations, for multi-breeds, several unresolved issues would need to be addressed in the future. The gap of progress between multi-breeds and Holstein populations is widening very rapidly. Because the structure of Holstein populations tend to be more suited for genomic evaluations, for example, large reference populations, small effective population sizes and hence sufficient marker-QTL LD. This gap will be more noticeable for novel and new traits (e.g., feed efficiency, health, fertility and milk composition), where genomic evaluations are expected to offer the most benefit.

There is paucity of information about the underlying confounding factors due to admixture. This limits our understanding of the true source of confounding, to be accounted for or reduced in methods development. Although this is not an easy undertaking, studies like Deng (2001), which investigate factors or the role of population admixture itself, as the potential cause for hampering analyses, are encouraged. Disentangling admixture would ensure that prediction models are not negatively affected and hence, maintain long-term genetic improvement.

The low marker-QTL LD may be improved by constructing haplotype segments of markers instead of individual markers. Because haplotype segments include several markers, they typically originate from common ancestry thereby associating with unique alleles. Several methods of constructing haplotypes have been described, and found to be more reliable than individual markers (Hayes *et al*., 2007; Calus *et al*., 2008; de Roos *et al*., 2011). The availability of high marker density or sequence data may even enable the construction of haplotype segments surrounding causal mutations.

Smaller number of reference populations relative to the Holstein is the other limitation. Size of the reference population is a key issue because it has a linear relationship with the prediction accuracy. The reference population can be increased by genotyping all available cows. Including at least 2000 cows in the reference population has been shown to increase the accuracy by 10% (Calus et al., 2011). Genotyping cows would also benefit in the evaluation of new traits where proven bulls may not have reliable EBV as their daughters may not have measurements (Buch et al., 2012). In North America, over 50,000 cows have been already genotyped with marker panels of various densities. If costs are limited, an effective strategy would be to: i) genotype randomly across families with high density, ii) genotype remaining animals with lower density and iii) perform imputations to higher densities.

5 CONCLUSIONS

Genomic selection has indeed offered animal breeders new tools for evaluating young individuals without performance information more accurately, which will subsequently lead to much faster genetic progress at a reduced time and cost. The success has been more evident for breeds with population structures that are suitable for the application of this technology. In this Ph.D. thesis, two approaches have been developed to explore the prospects of genomic selection in multi-breed and admixed populations when accounting for the population structure, using information on breed composition.

Firstly, when the multi-trait random regression model, which accounts for the interactions between marker effects and base breed origin of alleles, was used, we found that gains in validation reliabilities were 2 and 3% for milk and protein, respectively, and -1% for fat in comparison to a model that assumed a homogeneous population. This model could be more beneficial for evaluations in multi-breed populations with many base breed crosses but also, including a reasonable number of pure base breed individuals.

Secondly, our results evidently showed the crucial role played by allele frequencies in the estimation of genomic relationships as we observed that relationship coefficients were sensitive and varied greatly with allele frequencies utilized. Genomic relationships increased and were more variable when ignoring the structure by using allele frequencies across breeds. Furthermore, coefficients for individuals from populations that were genetically distant from the mean population allele frequency across breeds appeared to be even higher than expected when compared to pedigree-based relationships. These problems were avoided (i.e., both across and within sub-populations) when accounting for breed composition by using allele frequencies within breeds. In addition, genomic relationships were lower, less variable and more comparable to pedigree-based relationships when the estimation utilized allele frequencies from the base population versus the currently genotyped individuals. The use of allele frequencies from the base population of each breed subsequently, made easier the incorporation of genomic and pedigree information for single-step GBLUP. Thus, to avoid possible short term errors in genomic relationships, and long term consequence in breeding programs, it may be advisable to estimate genomic relationships accounting for varying allele frequencies from the base (founder) population of every breed.

The effect of accounting for breed composition in genomic relationships was however, not as evident for genomic evaluations. The validation reliabilities when accounting for or ignoring the population structure were generally similar across models at 33% for milk and protein and 43% for fat with GBLUP models of genotyped individuals only, and increased to 37%, 40% and 47% for milk, protein and fat, respectively, with single-step GBLUP of both genotyped and ungenotyped individuals. This gain of at least 5% in single-step validation reliabilities indicates the benefit of utilizing all available data in to genomic evaluations. In study I and II, it was found that the estimation of variance components with cow compared to bull information as phenotype appeared to be more desirable. Overall, accounting for the population structure achieved marginal advantage in the predictive ability of genomic evaluations. However, to incorporate genomic information into existing breeding programs for multi-breeds cautiously, single-step evaluations that utilize cow performance record as phenotype and genomic relationships accounted for varying allele frequencies between the breeds' founder populations could be a reasonable approach for long term genetic improvement.

6 REFERENCES

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

Astle, W. and D. J. Balding. 2009. Population structure and cryptic relatedness in genetic association studies. Statist. Sci. 43:415-471.

Browning, B. L. and S. R. Browning. 2008. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am. J. Hum. Genet. 84:210-223.

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.

Brøndum, R. F., E. Rius-Vilarrasa, I. Strandén, G. Su, B. Guldbrandtsen, W. F. Fikse, and M. S. Lund. 2011. Reliabilities of genomic prediction using combined reference data of the Nordic Red dairy cattle populations. J. Dairy Sci. 94:4700-4707.

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

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

Calus, M. P. L., Y. de Haas, M. Pszczola, and R. F. Veerkamp. 2011. Predicted response of genomic selection for new traits using combined cow and bull reference populations. Interbull Bull. 44:231-7234.

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

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

Christensen, O. F., P. Madsen, B. Nielsen, T. Ostersen, and G. Su. 2012. Single-step methods for genomic evaluation in pigs. Animal 6:1565-1571.

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

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

Daetwyler, H. D., J. M. Hickey, J. M. Henshall, S. Dominik, B. Gredler, J. H. J. Van Der Werf and B. J. Hayes. 2010. Accuracy of estimated genomic breeding values for wool and meat traits in a multi-breed sheep population. Anim. Prod. Sci. 50:1004-1010.

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

de los Campos, G., J. M. Hickey, R. Pong-Wong, H. D. Daetwyler and M. P. L. Calus. 2013. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics. 193:327-345.

Deng, H. W. 2001. Population admixture may appear to mask, change or reverse genetic effects of genes underlying complex traits. Genetics 159:1319-1323.

de Roos, A. P. W., B. J. Hayes, and M. E. Goddard. 2009. Reliability of genomic predictions across multiple populations. Genetics 183:1545-1553.

de Roos, A. P. W., C. Schrooten, and T. Druet. 2011. Genomic breeding value estimation using genetic markers, inferred ancestral haplotypes, and the genomic relationship matrix. J. Dairy Sci. 94:4708-4714.

Ewens W.J., Spielman R.S. (1995) The transmission/disequilibrium test: history, subdivision, and admixture. Am. J. Hum. Genet., 57:455-464.

Falconer, D. S., and T. F. C Mackay. 1996. Introduction to quantitative genetics. 4th Edition. Harlow, Longmans Green.

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

Frkonja, A., B. Gredler, U. Schnyder, I. Curik, and J. Sölkner. 2012. Prediction of breed composition in an admixed cattle population. Anim. Genet. 43:696-703.

Gao, H., O. F. Christensen, P. Madsen, U. S. Nielsen, Y. Zhang, M. S. Lund, and G. Su. 2012. Comparison on genomic predictions using three GBLUP methods and two single-step blending methods in the Nordic Holstein population. Genet. Sel. Evol. 44:8.

Gao, H., M. S. Lund, Y. Zhang, and G. Su. 2013. Accuracy of genomic prediction using different models and response variables in the Nordic Red cattle population. J. Anim. Breed. Genet. In press.

García-Cortés, L. A., and M. Á. Toro. 2006. Multibreed analysis by splitting the breeding values. Genet. Sel. Evol., 38, 601-615.

Garrick, D. J., J. F. Taylor, and R. L. Fernando. 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet. Sel. Evol. 41:55.

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

Gilmour, A. R., B. J. Gogel, B. R. Cullis and R. Thompson. 2009. ASREML User Guide Release 3.0. VSN International Ltd, Hemel Hempstead, UK.

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

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

Guo, G., M. S. Lund, Y. Zhang, and G. Su. 2010. Comparison between genomic predictions using daughter yield deviation and conventional estimated breeding value as response variables. J. Anim. Breed. Genet., 127:423-432.

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

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

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

Harris, B. L., A. M. Winkelman, and D. L. Johnson. 2012. Large-scale single-step genomic evaluation for milk production traits. Interbull Bull. 46:20-24.

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

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

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

Hayes, B. and M. Goddard. 2010. Genome-wide association and genomic selection in animal breeding. Genome. 53:876-883.

Hayes, B. J., P. J. Bowman, H. D. Daetwyler, J. W. Kijas and J. H. J. Van Der Werf. 2012. Accuracy of genotype imputation in sheep breeds. Anim. Genet. 43:72-80.

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

Hill, W. G. 2010. Understanding and using quantitative genetic variation. Phil. Trans. R. Soc. B. 365:73-85.

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

Illumina Inc. 2005. Illumina GenCall Data Analysis Software—Gen-Call software algorithms for clustering, calling, and scoring genotypes. Illumina. Pub. No. 370-2004-009. Illumina Inc., San Diego, CA.

Interbull. 2004. Code of Practice April 27th 2004 < http://www.interbull.se/>

Interbull. (2008) National genetic evaluation system. Accessed 02 June 2010. http://www-interbull.slu.se/national_ges_info2/framesida-ges.htm.

Jairath, L., J. C. M. Dekkers, L. R. Schaeffer, Z. Liu, E. B. Burnside, and B. Kolstad. 1998. Genetic evaluation for herd life in Canada. J. Dairy Sci. 81:550-562.

Janss, L., G. de los Campos, N. Sheehan and D. Sorensen. 2012. Inferences from genomic models in stratified populations. Genetics 192:693-704.

Jensen, J., G. Su and P. Madsen. 2012. Partitioning additive genetic variance into genomic and remaining polygenic components for complex traits in dairy cattle. BMC Genet. 13:44.

Kearney, F., A. Cromie, and D. P. Berry. 2009. Implementation and uptake of genomic evaluations in Ireland. Interbull Bull. 40:227-230.

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

Koivula, M., I. Strandén, G. Su, and E. A. Mäntysaari. 2012. Different methods to calculate genomic predictions-Comparisons of BLUP at the single nucleotide polymorphism level

(SNP-BLUP), BLUP at the individual level (G-BLUP), and the one-step approach (H-BLUP). J. Dairy Sci. 95:4065-4073.

Kuehn, L. A., J. W. Keele, G. L. Bennett, T. G. McDaneld, T. P. L. Smith, W. M. Snelling, T. S. Sonstegard, and R. M. Thallman. 2011. Predicting breed composition using breed frequencies of 50,000 markers from the US Meat Animal Research Center 2,000 bull project. J. Anim. Sci. 89:1742-1750.

Lidauer, M., and I. Strandén. 1999. Fast and flexible program for genetic evaluation in dairy cattle. Interbull Bull. 20:20.

Lidauer, M., E. A. Mäntysaari, I. Strandén, J. Pösö, J. Pedersen, U. S. Nielsen, K. Johansson, J-Å. Eriksson, P. Madsen, and G. P. Aamand. 2006. Random heterosis and recombination loss effects in a multibreed evaluation for Nordic Red dairy cattle. Communication 24-02 in Proc. 8th World Congr. Genet. Appl. Livest. Prod., Belo Horizonte, Brazil. Instituto Prociência, Minas Gerais. Brazil.

Lidauer, M. H., E. A. Mäntysaari, J. Pösö, J.-Å. Eriksson, U. S. Nielsen, and G. P. Aamand. 2010. Heterogeneous variance adjustment in across-country genetic evaluation with country-specific heritabilities. In: Proceedings of the 9th World Congress on Genetics Applied to Livestock Production, 1-6th August 2010, Leipzig, Germany.

Lo, L. L, R. L. Fernando, and M. Grossman. 1993. Covariance between relatives in multibreed populations. Theor. Appl. Genet., 87:423-430.

Maignel, L., D. Boichard, and E. Verrier. 1996. Genetic variability of French dairy breeds estimated from pedigree information. Interbull Bull. 14: 49–54.

Malécot, G. 1948. Les Mathématiques de L'hérédité. Masion. Paris.

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

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

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

Moser, G., B. Tier, R. Crump, M. Khatkar and H. Raadsma. 2009. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genet. Sel. Evol. 41:56.

Mrode, R. A. and G. J. T. Swanson. 2004. Calculating cow and daughter yield deviations and partitioning of genetic evaluations under a random regression model. Livest. Prod. Sci. 86:253-260.

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

Mäntysaari, E. A., M. Koivula, I. Strandén, J. Pösö, and G. P. Aamand. 2011. Estimation of GEBV using deregressed individual cow breeding values. Interbull Bull. 44:26-29.

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

Olson, K. M., P. M. VanRaden, M. E. Tooker, and T. A. Cooper. 2011. Differences among methods to validate genomic evaluations for dairy cattle. J. Dairy Sci. 94:2613-2620.

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

Ostersen, T., O. F. Christensen, M. Henryon, B. Nielsen, G. Su, and P. Madsen. 2011. Deregressed EBV as the response variable yield more reliable genomic predictions than traditional EBV in pure-bred pigs. Genet. Sel. Evol. 43:38.

Price, A. L., N. A. Zaitlen, D. Reich, and N. Patterson. 2010. New approaches to population stratification in genome-wide association studies. Nat. Rev. Genet. 11:459-463.

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

Pryce, J. E., B. Gredler, S. Bolormaa, P. J. Bowman, C. Egger-Danner, C. Fuerst, R. Emmerling, J. Sölkner, M. E. Goddard, and B. J. Hayes. 2011. Short communication: Genomic selection using a multi-breed, across-country reference population. J. Dairy Sci. 94:2625-2630.

Pollak E. J., and R. L. Quaas. 1998. Multibreed genetic evaluations in beef cattle. In: Proceedings of 6th World Congress Applied to Livestock Production, vol. 23. Armidale, Australia, pp. 81–88.

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

Reinhardt, F., Z. Liu, F. Seefried, and G. Thaller. 2009. Implementation of genomic evaluation in German Holsteins. Interbull Bull. 40:219-226.

Resende, Jr., M. F. R., P. Munoz, M. D. V. Resende, D. J. Garrick, Fernando, R. L., Davis, J. M., Jokela, E. J., Martin, T. A., Peter, G. F., M. Kirst. 2012a Accuracy of genomic selection methods in a standard data set of loblolly pine (Pinus taeda L.). Genetics. 190: 1503–1510.

Resende, M. D. V., M. F. R. Resende, Jr., Sansaloni, C. P., Petroli, C. D., Missiagia, A. A., Aquiar, A. M., Abad, J. M., Takahashi, E. K., Rosado, A. M., Faria, D.A, Pappas Jr, G. J., Kilian, A and D. Grattapaglia. 2012b. Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. New Phytol. 194:116-128.

Rius-Vilarrasa, E., T. Iso-Touru, I. Stranden, N. Schulman, B. Guldbrandtsen, E. Strandberg, M. S. Lund, J. Vilkki, and W. F. Fikse. 2011. Characterization of linkage disequilibrium in a Danish, Swedish and Finnish Red Breed cattle population. In: Proc. 62nd Annu. Mtg. Eur. Fed. Anim. Sci., Stavanger, Norway. Wageningen Academic Publishers, Wageningen, the Netherlands. p. 177.

Rolf, M. M., J. F. Taylor, R. D. Schnabel, S. D. McKay, M. C. McClure, S. L. Northcutt, M. S. Kerley and R. L. Weaber. 2010. Impact of reduced marker set estimation of genomic relationship matrices on genomic selection for feed efficiency in angus cattle. BMC Genetics. 11:24.

Schaeffer, L. R. 2001. Multiple trait international bull comparisons. Livest. Prod. Sci. 69:145-153.

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

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

Schulman, N. F., G. Sahana, T. Iso-Touru, M. S. Lund, L. Andersson-Eklund, S. M. Viitala, S. Värv, H. Viinalass and J. H. Vilkki. 2009. Fine mapping of quantitative trait loci for mastitis resistance on bovine chromosome 11. Anim. Genet. 40:509-515.

Simeone, R., I. Misztal, I. Aguilar and A. Legarra. 2011. Evaluation of the utility of diagonal elements of the genomic relationship matrix as a diagnostic tool to detect mislabelled genotyped animals in a broiler chicken population. J. Anim. Breed. Genet. 128:386-393.

Strandén, I., M. Lidauer, E. A. Mäntysaari, and J. Pösö. 2001. Calculation of Interbull weighting factors for the Finnish test day model. Interbull Bull. 26:78-79.

Strandén, I., and K. Vuori. 2006. RelaX2: pedigree analysis program. In: Proc. 8th World Congr. Genet. Appl. Livest. Prod. Belo Horizonte, Brazil.

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

Strandén, I., and E. A. Mäntysaari. 2010. A recipe for multiple trait deregression. Interbull Bull. 42:21-24.

Strandén, I., and O. F. Christensen. 2011. Allele coding in genomic evaluation. Genet. Sel. Evol. 43:25.

Strandén I., and E.A. Mäntysaari. 2012. Use of random regression model as an alternative for multibreed relationship matrix. J. Anim. Breed. 130:4-9.

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

Su, G., P. Madsen, U. S. Nielsen, E. A. Mäntysaari, G. P. Aamand, O. F. Christensen, and M. S. Lund. 2012a. Genomic prediction for Nordic Red Cattle using one-step and selection index blending. J. Dairy Sci. 95:909-917.

Su, G., R. F. Brøndum, P. Ma, B. Guldbrandtsen, G. P. Aamand, and M. S. Lund. 2012b. Comparison of genomic predictions using medium-density (~54,000) and high-density (~777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations. J. Dairy Sci. 95:4657-4665.

Sul, J. H., and E. Eskin. 2013. Mixed models can correct for population structure for genomic regions under selection. Nat. Rev. Genet. 14:300.

Sullivan, P. G., and P. M. VanRaden. 2009. Development of genomic GMACE. Interbull Bull. 40:157-161.

Sölkner J., A. Frkonja, H. W. R. Raadsma, E. Jonas, G. Thaller, E. Egger-Danner, and B. Gredler. 2010. Estimation of individual levels of admixture in crossbred populations from SNP chip data: examples with sheep and cattle populations. Interbull Bull. 42:62-66.

Sørensen, M. K., A. C. Sørensen, R. Baumung, S. Borchersen and P. Berg. 2008. Optimal genetic contribution selection in Danish Holstein depends on pedigree quality. Livestock Science. 118:212-222.

Tang, H., M. Coram, P. Wang, X. Zhu, and N. Risch. 2006. Reconstructing genetic ancestry blocks in admixed individuals. Am. J. Hum. Genet. 79:1-12.

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

Thompson, J.R., R.W. Everett, and C.W. Wolfe. 2000a. Effects of inbreeding on production and survival in Jerseys. J. Dairy Sci. 83:2131-2138.

Thompson, J.R., R.W. Everett, and N.L. Hammerschmidt.2000b. Effects of inbreeding on production and survival in Holsteins. J. Dairy Sci. 83:1856-1864.

Toro, M. A., T. H. E. Meuwissen, J. Fernández, I. Shaat and A. Mäki-Tanila. 2011. Assessing the genetic diversity in small farm animal populations. Animal 5:1-15.

Vandenplas, J. and N. Gengler. 2012. Comparison and improvements of different bayesian procedures to integrate external information into genetic evaluations. J. Dairy Sci. 95:1513-1526.

Van Doormaal, B. J., G. J. Kistemaker, P. G. Sullivan, M. Sargolzaei, and F. S. Schenkel. 2009. Canadian implementation of genomic evaluations. Interbull Bull. 40:214-217.

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

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

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

Visscher, P. M, W. G. Hill, and N. R. Wray. 2008. Heritability in the genomics era – concepts and misconceptions. Nat. Rev. Genet. 9:255-266.

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

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

Zeng, J., A. Toosi, R. L. Fernando, J. C. M. Dekkers and D. J. Garrick. 2013. Genomic selection of purebred animals forcrossbred performance in the presence ofdominant gene action. Genet. Sel. Evol.45:11.

7 APPENDICES

7.1 APPENDIX A: The construction of genomic relationship matrices (G)

The original G – Derived using allele frequency across breeds

Let there be *n* individuals that have been genotyped for *m* markers. Let u_{ij} denote the genotype j of animal i. The genotype in u_{ij} has value 0, 1 or 2 if animal i is homozygote 11, heterozygote 12 or homozygote 22, respectively, at locus j. Let the frequency of the 2nd allele at locus j be p_j. Then, the original **G** matrices as proposed in method 1 and 2 of VanRaden (2008) can be defined as:

$$\mathbf{Gorg} = \mathbf{ZZ}'/k\,, \tag{A.1}$$

$$\mathbf{Gorg2} = \mathbf{Z}^{*}\mathbf{Z}^{*\prime}/m, \tag{A.2}$$

where in method 1 (A.1), the matrix **Z** contains centred genotypes (i.e., centred by the expected mean allele frequency $2p_j$) with the element of animal i for marker j in **Z** being u_{ij}-$2p_j$; the scaling factor is $k = 2\sum_j p_j(1-p_j)$, which is the expected variance of marker j. In method 2 (A.2), for each marker column j in the **Z** matrix denoted $\mathbf{Z}_j$, the coefficients were further standardized as:

$$\mathbf{Z}_j^{*} = \mathbf{Z}_j/\sqrt{2\mathrm{p}_j(1-\mathrm{p}_j)}$$

Equations A.1 and A.2 were also calculated using the base population allele frequencies p_j. See equation A.4 for the estimation of allele frequencies from the base population.

Adjusted G – Derived using allele frequency within breeds

Allele frequencies (AF) within breeds were estimated by solving a simple multiple regression vector ($\boldsymbol{\beta}$) of genotypes (**y**) on breed proportions (**X**) for every marker (equation A.3). There were 4 defined breeds, therefore, **X** has dimension $n \times 4$. Following the gene content algorithm of Gengler et al. (2007), AF from the base population were solved by extending equation A.3 by including a design matrix **Q** associating animal genetic effects **g** with vector **y** (Equation A.4). Briefly, in A.4 the assumption is that the covariance between gene contents (i.e., number of copies of one allele) is proportional to the additive relationships between animals. Pedigree relationship matrix is used to estimate the expected gene contents for ungenotyped ancestors.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \tag{A.3}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Q}\mathbf{g} + \mathbf{e}, \tag{A.4}$$

where we assumed that $\mathbf{e} \sim N(0, \mathbf{I}\sigma_{\mathrm{e}}^2)$ with σ_{e}^2 being the residual variance component, set to 0.01. It was assumed that $\mathbf{g} \sim N(0, \mathbf{A}\sigma_{\mathrm{a}}^2)$, where **A** is the pedigree relationship matrix and σ_{a}^2 is the additive genetic variance, assumed to be 1.0. The expectation of AF for marker j in A.3 and A.4 is given by $\widehat{\mathrm{P}}_j = \mathbf{X}\widehat{\boldsymbol{\beta}}_j$, where AF is in $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1, \dots, \widehat{\boldsymbol{\beta}}_4)$ for the four breeds. The AF from the base population across breed were solved in A.4.

Now, equations A.1 and A.2 were adjusted using $\widehat{\mathrm{p}}_{ij}$ as follows:

$$\mathbf{Gadj} = \mathbf{M}\mathbf{M}'/k\,, \tag{A.5}$$

$$\mathbf{Gadj2} = \mathbf{M}^{*}\mathbf{M}^{*\prime}/m, \tag{A.6}$$

where with the same notation as in **Z**, element $\mathbf{M}_{ij}$ is $u_{ij} - 2p_{ij}$ where p_{ij} is the expected AF for marker j when accounting for the breed background of animal i. In A.6, each element in **M** was further scaled by the standard deviation of the expected marker effects:

$$\mathbf{M}^* = \mathbf{M}_{ij} / \sqrt{2\mathrm{p}_{ij}(1 - \mathrm{p}_{ij})}$$

7.2 APPENDIX B: The Multi-trait Random Regression model

Breed-specific variances and breeding values can be obtained by model:

$$y_i = \mu + \sum_{k=1}^{4} \mathrm{c}_{ik}\,\mathrm{b}_k + \sum_{k=1}^{4} \sqrt{\mathrm{c}}_{ik}\,\mathrm{a}_{ik} + \mathrm{e}_i\,, \tag{A.7}$$

where y_i is a vector of phenotype for bull i; μ is the overall mean; b_k is the fixed regression effect of breed k ($k=1,...,4$); c_{ik} is the breed proportion of bull i for breed k, so that $\sum_k c_{ik} = 1$ for all i. For purebreds, t: $c_{ik}=1$ and $c_{it}=0$ for all $t \neq k$. Here $\sqrt{\mathrm{c}}_{ik}$was used to equalize the proportion of sire variance accounted for by breeds and avoid high variation between purebred and crossbred sire variances when fitting c_{ik}. The $\boldsymbol{a} = (a_{ik})$ is a vector of genomic breeding values with length of 4 times n, so that bull i has a sub-vector with 4 breed specific breeding values (a_{i1}, a_{i2}, a_{i3}, a_{i4}). It was assumed that $\mathbf{a} \sim N(\mathbf{0}, \mathbf{G} \otimes \mathbf{G}_0)$, where $\mathbf{0}$ is a vector of zeros of length $4n$; **G** is the genomic relationship matrix of dimension $n \times n$ and $\mathbf{G}_0$ is a 4×4 diagonal matrix of breed specific sire variances. Assumption for the random residuals common across breeds e_i is assumed as $e_i \sim N(0, \sigma_e^2/w_i)$, where weight w_i is the reliability of the phenotype scaled by $\lambda = \frac{4-h^2}{h^2}$ and heritabilities are given in Table 1.

Model A.7 in matrix notation is:

$$\mathbf{y} = \mathbf{1}\mu + [\mathbf{C}_1\mathbf{1} \cdots \mathbf{C}_4\mathbf{1}]\begin{bmatrix} \mathrm{b}_1 \\ \vdots \\ \mathrm{b}_4 \end{bmatrix} + [\mathbf{S}_1\mathbf{Z}_\mathrm{a} \cdots \mathbf{S}_4\mathbf{W}_\mathrm{a}]\begin{bmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_4 \end{bmatrix} + \mathbf{e},$$

where $\boldsymbol{y}$ is a n x 1 vector of phenotype; μ is the general mean; $\mathbf{1}$ is a unit vector; $\mathbf{C}_i$ is an n x n diagonal matrix with BP for all bulls in breed i on the diagonal and $\mathbf{S}_i$ is square root of $\mathbf{C}_i$; $\mathbf{b}$ is a 4×1 vector of fixed breed effects; $\mathbf{W}_\mathrm{a}$ is an $n \times n$ incidence matrix associating random breed specific genetic effects to the records; here $\mathbf{W}_\mathrm{a} = \mathbf{I}$ when all the individuals included in the data had a record ($n = n$); $\boldsymbol{a}$ is a vector of random breed specific animal genetic effects ordered by animals within breed, and $\mathbf{e}$ is an n x 1 vector of random residual terms common across breeds.

8 ACKNOWLEDGEMENTS

I would like to express my gratitude to institutions and individuals who have contributed to my professional and personal life.

Phenotypic data, including records, breed proportions and pedigrees for the research were obtained from the Nordic Cattle Genetic Evaluation Ltd (NAV). Genotypes were obtained from the Nordic Genomic Selection project financed by Viking Genetics, NAV, FABA Finland, Svensk Mjölk (Växa Sverige) and the Dannish Cattle Federation. I gratefully acknowledge financial support from the Suomen NaudanJalostussäätiö and the Finnish Ministry of Agriculture and Forestry. Also, myriad of thanks to the University of Helsinki and Maa- ja elintarviketalouden tutkimuskeskus (MTT), for providing work resources.

Words seem inadequate to express my profound gratitude for invaluable assistance, sound advices and guidance from my Professors, Esa Mäntysaari, Jarmo Juga, Ismo Strandén and Mikko Sillanpää. It has been a great pleasure working with you and thank you for sharing your vast scientific knowledge. Esa, your enthusiasm and patience when explaining concepts has afforded me a better understanding. Most important for my career, you showed me that research is fun and that a good researcher sees hurdles as opportunities. Jarmo, your great leadership in the department, and the independence you afford your graduate students has taught me to become a responsible researcher. With many responsibilities, you still ensured my work-related and practical necessities were in order. Ismo, your simple approach when solving my technical questions always left me wondering “why didn’t I think of that?” which taught me to be critical. Mikko, our discussions have been inspiring to me; I was always looking forward to them while amazed by your sharp research ideas. My sincere gratitude to Professors Nicolas Gengler and Freddy Fikse, for examining my thesis, which greatly improved its content, Professor Theodorus Meuwissen, for allowing me the rare opportunity of being my opponent and Professor Pekka Uimari for an outstanding task as my custos.

Most important to my life, my family, you believed in me, sacrificed where you had to, supported and encouraged me throughout the journey. To my beloved daughter, Tumisang, you are my pillar of strength. Thank you for your practical and emotional support as I added or abandoned the role of a mother, to the competing demands of work, study and personal development. To my parents, a big thank you for creating an environment that made learning seem so natural, and providing the necessary tools for studying. A special feeling of gratitude to my brothers Nasa and Wilson, and their families, for the support whenever I needed help, and my playful little sister, Mosima, you always make me laugh ☺

In addition, I would like to thank the ARC in South Africa, for acquainting me with the necessary research skills through the PDP mentorship programme, and allowing me room for growth. It is with immense gratitude that I acknowledge my mentors and dear friends, Dr. Banga and Professor Norris, for believing in me. Without your motivation and encouragement, I would not have considered a graduate career. Lastly, though by no means at all the least, a big thank you to all my research colleagues, friends and extended family for their support and encouragement. Especially, Timo Pitkänen and Alban Bouquet, for helping with codes and eliminate BUGS from my programs/scripts, Minna Koivula, for helping with data and analysis at every point, Marjatta Säisä and Ria Kuokkanen for helping with practical matters, the Bouquet family, for the much needed friendship as we were both finding our way in Helsinki in what could have otherwise been unexciting and my special friend Natsuha Yamaga for all the good times we have shared ☺