Chapter 1. Introduction

The genetic improvement of livestock has historically been based on phenotypic selection, where animals have been selected based on conformation, visual appeal and performance. The introduction of quantitative genetics theory has altered many breeding programs by establishing the concept of estimated breeding values (EBV), using Best Linear Unbiased Prediction (BLUP) (Henderson, 1950). Selection using BLUP EBV's is based on information regarding phenotypic performance and the performance of an animal's relatives. In BLUP a linear mixed model is used to correct phenotypic information for common environmental effects such as birth type and year of birth. Furthermore, the accuracy in which the breeding value is estimated increases as more information becomes available; for example, a breeding value will have a low accuracy if the only available information is on an animal's parents, accuracy increases when that animal has a phenotypic measurement and finally the breeding value becomes highly accurate when that animal has many progeny. Furthermore, the mixed model that produces BLUP breeding values is also used to estimate of the amount of additive genetic variation in the population (Mrode, 2005).

The discovery of genetic markers such as restriction fragment length polymorphisms (RFLP), microsatellites and single nucleotide polymorphisms (SNP) allowed for research to begin investigating the underlying causes of genetic variation. Initially, genetic markers were used to identify regions of the genome that cause genetic variation in quantitative traits, called quantitative trait loci (QTL) (Haley & Knott, 1992; Meuwissen et al., 2002; Grapes et al., 2004). Original expectations were that QTL's would be widely used in animal breeding and genetic

evaluation, however, this has not eventuated for several reasons: 1) early marker densities were very low and marker maps were sparse so the QTL could not be fine mapped; 2) the effects of QTL tended to be overestimated and unable to be validated in different populations; 3) the experiments to find QTL were also often underpowered and were unable to detect QTL with moderate to small effects (Goring et al., 2001); 4) there were fewer QTL with large effects than what was initially expected and 4) finding and mapping QTL was very time consuming and expensive. Although there were many limitations for QTL detection and mapping, some QTL and markers for QTL were still used in a limited number of breeding programs in marker assisted selection (Guillaume et al., 2008).

To attempt to avoid the limitations associated with QTL detection, Meuwissen et al. (2001) examined the possibility of fitting markers from across the entire genome simultaneously in a statistical model to predict breeding value. These markers had no verified association with QTL; rather it was assumed that each QTL would be in linkage disequilibrium with at least one marker. The initial simulation results of this 'genomic selection' were very promising as accuracies of greater than 0.8 for animals with no phenotypic information were achieved. The methods used for genomic selection ranged from a BLUP method that fits all SNP as random effects (ridge regression BLUP or RR-BLUP), to Bayes B, which is a variable selection method that allows some SNP to have an effect and others are assumed to have no effect. The effects of each SNP can be combined, depending on the number of alleles an animal carries, into a genomic estimated breeding value (GEBV). Furthermore, a group of animals with phenotypic measurements and genotypes called the reference population or dataset are used to estimate these

SNP effects and GEBV's are predicted for animals with only genotypic measurements (Goddard & Hayes, 2009).

Markers can also be used to make a genomic relationship matrix (GRM) (Villanueva et al., 2005; VanRaden, 2008; Yang et al., 2010) and combined into BLUP, by replacing the pedigree derived numerator relationship matrix with the GRM. This method is called genomic BLUP or gBLUP and has been shown to be equivalent to RR-BLUP (Habier et al., 2007). Instead of estimating SNP effects directly, gBLUP uses information from relatives in the reference population to make a direct prediction of breeding value. The precise makeup of the reference dataset for each method is relatively unknown; Habier et al. (2007) showed that regardless of the method used, the relationship between animals in these populations can affect the accuracy of genomic selection.

The sequencing of many livestock genomes has resulted in the identification of a large number of SNP markers which can be placed on microchips for fast and more cost effective genotyping of individuals. Initially a chip with 1000 markers was regarded as a dense chip, now there is a possibility of using all SNPs present on the genome (millions of SNPS or SNP sequence) (Meuwissen & Goddard, 2010). The availability of high density marker chips has allowed for genomic selection to be validated in many livestock populations. The accuracies reported in real data have generally been dependent on the type of trait, the method used to predict breeding value and the population structure (Habier et al., 2010). Given this, the accuracies that have been achieved in dairy cattle using 50,000 markers have been high and generally range between 0.5

and 0.8 (VanRaden et al., 2007; Hayes et al., 2009). These high accuracies have resulted in many dairy cattle breeding schemes routinely reporting GEBVs to breeders.

Given the high accuracies of GEBV's for young animals, Schaeffer (2006) noted that genomic selection could double the rate of genetic gain in the dairy industry and change the structure of entire breeding program. Furthermore breeding programs may also be impacted due to the effect of the use of GEBV's on inbreeding (Daetwyler et al., 2007). In simulation, Sonesson et al. (2010) showed that GEBV's could increase merit and genomic relationship information could be used for the management of inbreeding in optimal selection. Although there have been many positive outcomes for the use of genomic information in dairy breeding programs, in species, such as beef cattle and sheep, the accuracy of breeding value estimation has been much more variable (Daetwyler et al., 2010). Furthermore the accuracies of GEBV's in beef cattle and sheep still seem to be dependent on the type of trait, the method used to predict breeding value and the population structure i.e. the size and structure of the reference dataset and the large number of breeds used in these industries.

The aim of this thesis was to examine the effect of genomic information on the genetic evaluation of livestock. This thesis represents many aspects of this rapid changing field of research, and each chapter provides insight into the impact of genomic information on genetic evaluation.

Chapter 2 is a review of literature that discusses the methods used for genomic evaluation in detail. It also gives a current review of genomic evaluations that are occurring in many livestock populations and identifies some interesting aspects of genomic selection.

Chapter 3 examines how different methods are affected by the underlying model of genetic variation, giving insight into why some traits are more accurately predicted than others. It also examines the effect of different markers densities and SNP sequence on genomic selection given the different model of variation. Furthermore, it discusses how the relationships between animals can still affect the accuracy of genomic selection under the alternative methods and models of variation.

Chapter 4 built on the concept of how the relationships between animals can affect the accuracy of genomic selection. It compared the accuracy of an animal's breeding value that has a strong pedigree relationship with a reference data set with that of an animal that is essentially unrelated to the reference data set, and discussed the effect of these relationships on the design of reference data sets used in genomic selection breeding schemes.

Chapter 5 examines the use of genomic information in optimal selection strategies where both genetic merit and inbreeding need to be balanced. Primarily we discuss the use of genomic relationship information to manage inbreeding and how this can be utilised in a breeding program.

Chapter 6 examines the properties of the genomic relationship matrix and the impact of building the GRM using identical by state (IBS) or identical by decent (IBD) information. We use real data from the Australia sheep CRC to observe the effect these alternative relationship matrices have on breeding values and variance component estimates.

Chapter 7 is the general discussion of this thesis. It discusses five main topics arising from this work. These are: 1) the number of markers used in genomic evaluation 2) The makeup and construction of reference populations in livestock breeding 3) across breed and family prediction 4) the value of genomic information to the breeding program and 5) using simulation and real data to validate genomic selection methods.

Chapter 2. Review of Literature

The success of a breeding program is measured by response to selection. Response to selection can be increased by: 1) increasing the accuracy in which animals are selected, 2) by reducing the generation interval i.e. breeding from animals at a younger age and 3) increasing selection intensity. The measurements of phenotypes on individuals and their relatives can increase accuracy, however in many cases this may increase generation interval (Falconer & Mackay 1996; Lynch & Walsh 1998). The discovery of regions on the genome that are associated with variation in quantitative traits (quantitative trait loci (QTL)) enabled the possibility of using this genomic information in selection. Initially there were large projects designed to detect and map these QTL and find genetic markers that were also highly associated with these QTL so that this information could be incorporated into the breeding program through marker assisted selection (MAS).

Marker assisted selection offered an increase in response to selection, especially for traits that were measured late in life, had a low heritability or were difficult to measure, because if there was knowledge about an animal's genome, this information could be used to accurately select an animal at a young age. However, the process of finding QTL was costly and only few QTL were actually found and validated in separate populations. To reduce the need for QTL mapping Meuwissen et al. (2001) fitted many markers simultaneously in a statistical model to give an estimate of breeding value. The sequencing of many livestock genomes and the resulting high-density genotyping with genetic markers has led to the widespread interest in using this 'genomic

selection' in livestock breeding programs. Genomic selection offers the same benefits as MAS of accurate selection at a young age and is predicted to increase the rate of genetic improvement in livestock breeding schemes (Schaeffer 2006; Dalton 2009).

This literature review will focus on the use of DNA information in breeding programs. It will briefly discuss the era of the detection and mapping of QTL, it will discuss the variations in DNA that are commonly used as genetic markers. Then a more detailed review will be on using dense markers information in genomic predictions and the methods used to gain an understanding into what is actually being predicted in genomic selection.

2.1. Genetic markers

Different types of variations in DNA can be used as markers for QTL in animal breeding programs. There are two main types of genetic variants used as markers in genetic evaluation and genome wide association studies. These are microsatellites and single nucleotide polymorphisms (SNP). Microsatellites are segments of DNA that consist of a variable number of tandem repeats (VNTR) of a simple sequence of nucleotides. For example, the base pairs AC are repeated 12 times in succession. Microsatellites markers have multiple alleles with a variable number of repeats. With many alleles most individuals are heterozygous, which makes each marker more informative in tracking the inheritance of chromosome segments. This gives the power to detect associations between the marker alleles and the performance of progeny, thus inferring the inheritance of specific QTL alleles (Walsh & Henderson, 2004). Microsatellites are large in size, highly polymorphic and therefore highly informative which makes them useful genetic markers.

However, microsatellites are hard to genotype and have a limited density on the genome and have been replaced by SNP markers in genomic evaluation systems.

Single nucleotide polymorphisms are the most abundant genetic marker on the genome and are now widely used in genomic evaluation and genome wide association studies (GWAS). These SNPs are the result of a single base pair change in the DNA structure and may alter the function of the transcribed protein (Moon et al., 2007; Womack, 2005). However, not all SNP result in this functional change, they can still be used as markers for QTL. Since the sequencing of many livestock genomes, high densities of SNP have been placed on microchips for large scale genotyping. In sheep, the current density of the SNP chip used in genomic evaluation is 48,640 markers (Daetwyler et al., 2010) and in humans 1.3 million markers are now being used for GWAS (Kwee et al., 2012). In GWAS, SNPs are used to find QTL and regions that cause variation in phenotypes. If a SNP is highly associated with variation in phenotype it may be used to develop a marker for that QTL (Pereira et al., 2005; Womack, 2005) and also, may be a useful tool in mapping the QTL within these regions (Pereira et al., 2005). Other variations in DNA can include: deletions, substitutions, inversions and copy number variants (CNV), however these are not commonly used in large scale genomic evaluations. All classes of genetic variants can be used as markers for QTL and specific genes. Markers can be direct markers and represent a change in the function transcribed protein (Dekkers, 2004; van der Werf, 2000). They can also be in linkage equilibrium with QTL, this is when a marker and QTL are associated yet the effect of the markers varies within different families (Dekkers, 2004; Farnir et al., 2000; Notter, 2004). Finally, markers can also be in population-wide linkage disequilibrium (LD) with the QTL or

functional mutation such that selection on the markers will result in a change in the frequency of the QTL.

Linkage disequilibrium is an important factor in understanding the effectiveness of genomic evaluation (Dekkers, 2004; Khatkar et al., 2006; Moore et al., 2003). When a marker is in linkage disequilibrium with a QTL there is a higher probability that if a particular allele is found for a gene marker, a prediction can be made about the QTL allele (Farnir et al., 2000). The extent of LD depends on the distance between the marker and the QTL (van der Werf, 2000). If the two loci are tightly linked then recombination does not occur rapidly and therefore LD is maintained for a long period. The extent of LD in the population is also highly depended on the effective population size (*Ne*). If this is small, LD can be present for large distances; therefore fewer markers are needed to make an accurate prediction of breeding value. In contrast if *Ne* is large more markers are needed such that every QTL is in LD with at least one marker. If the extent of LD is large, there is the potential for a locus to be segregating and in LD with a QTL across an entire population or even a whole species (Dekkers, 2004; Farnir et al., 2000, Goddard, 2009).

2.2. Genome wide association studies and QTL mapping

In livestock and human genetics, genome wide association studies (GWAS) and QTL mapping are used to detect locations of the genome that have an effect on important quantitative traits and disease susceptibility. In GWAS, each SNP is regressed on the phenotypes of a group of individuals (single SNP regression) to attempt to find associations between the marker and causative loci. Following the detection of highly significant SNP, often these areas are fine

mapped and candidate genes identified which may provide insight into biological pathways that are responsible for the variation in phenotype.

Initially the primary method to detect and map QTL was interval mapping using linkage analysis within crosses or sibling groups (Hayley & Knott, 1992). This method used the probability of a QTL genotype conditional on the genotypes of adjacent markers to provide likelihood calculations and the logarithm of the odds favouring linkage score (LOD score) that assessed the probability that a QTL was located in a specific region or not. Linkage disequilibrium information can also be used to map QTL. In pure LD mapping, no specific family structure is required to obtain associations between marker alleles and QTL (Hayes et al., 2006). However, it can be more efficient to use both LD and linkage information simultaneously to map QTL (Meuwissen et al., 2002). This LDLA method uses both linkage and LD information to define the probability that a QTL is present at a particular site on the genome. The extent of across population LD is typically much less than within family linkage (Farnir et al, 2000); therefore the LDLA method requires a denser marker map than interval mapping in order to gain information from population-wide LD.

In both GWAS and QTL mapping experiments factors such as; population size, the size of the QTL effects and marker density can influence the detection of QTL (Grapes et al., 2004; Zhao et al., 2007). Large populations of individuals with phenotypic information are needed for the detection of QTL so that; all genotypes are observed and that there are a large number of animals in each genotype group. Furthermore, sufficient phenotypic information is also needed so that

there is enough statistical power to accurately detect QTL, so that the error around this detection is low (Moghaddar & van der Werf, 2007). This 'power' relates to the size of the QTL effect that is able to be detected. For example a larger amount of data is required to find smaller QTL effects. The marker density required to detect QTL also varies for the size of the QTL effect. Larger QTL effects can be detected with lower marker densities; however the exact number of markers that is needed is dependent on the structure of the population.

Initially, it was thought that QTL and markers for QTL would be very useful in livestock genetic evaluation, to be used in marker assisted selection (MAS). However, their impact has been limited due to various reasons. Firstly, in early genomic research the marker densities used were very low and marker maps were sparse so the QTL could not be fine mapped. Secondly, the size of the populations used to find QTL were often too small and therefore the effects of QTL tended to be overestimated and unable to be validated in different populations. Moreover, the inadequate numbers of phenotypes also resulted in the experiments to find QTL often being underpowered and were therefore only able to detect a few large QTL (Goring et al., 2001). After many GWAS and mapping experiments it is now suggested that only a few large QTL actually exist and that most of the genetic variation is due to many small QTL effects (Yang et al., 2010).

Given these limitations, GWAS and QTL mapping have still been used to discover many significant markers, QTL and causal mutations. However, in many cases the areas identified only explain a small amount of the cumulative genetic variation. For example, in human genetics the 30 genetic markers identified for Crohn's disease account for less than 10% of the additive

genetic variance similarly for the trait height, less than 5% of genetic variation is explained by the 44 identified loci (Visscher, 2008). This has led to discussion regarding this 'missing heritability' and many theories have been suggested to explain why all genetic variation is not captured by these significant regions (Maher, 2008, Yang et al., 2010). Fearnhead et al. (2004) suggested that the inconsistencies that exist between high estimates of heritability and the small proportion of total genetic variance explained by QTL may be due to rare variants that are difficult to detect using QTL mapping and GWAS. However, Yang et al. (2010) observed that a large proportion of genetic variance can be explained by common variants, although many of these variants have small effects and they are often too small to be detected. Furthermore, they suggested that some of the missing heritability observed in GWAS may be due to markers that are not in complete LD with QTL and therefore all of the genetic variance at the QTL level is not being captured by markers. The debate about the missing heritability has also created discussion about the patterns in which QTL effects follow. These range from models that have few QTL with large effects like those used by Meuwissen et al. (2001) to infinitesimal like models, very many loci explain some variation, each of them having a small effect (Fisher, 1918). Moreover, non-additive models such as the epistatic model where all QTL interact with each other have also been discussed as possible models of genetic variation (Zuk et al., 2012). The assumptions made about the model of variation may therefore affect the success of genetic evaluation using genomic information.

2.3. Genetic evaluation using genomic information

To avoid the limitations associated with QTL detection, Meuwissen et al. (2001) examined the possibility of fitting all markers simultaneously in a statistical model to predict breeding value.

Unlike traditional marker assisted selection, the markers used in this genome wide evaluation had no verified association with QTL; rather it was assumed that each QTL would be in linkage disequilibrium with at least one marker. This assumption allows for a reduced cost of using genomic information in breeding value prediction as it negates the need for QTL detection and mapping (Hayes & Goddard, 2009). In simulations the accuracy of the estimated breeding values using this genome wide evaluation are often very high, however they are dependent on the assumed model of variation used in the simulation (Meuwissen et al., 2001; Muir, 2007). Similar to QTL mapping, genomic selection relies on a group of individuals that have phenotypic and genotypic information recorded on them. Information from this 'reference population' is used to predict animals that only have genotypic information recorded.

The size and make-up of the reference population is important for a number of reasons. 1) As in QTL mapping and GWAS the reference population needs to be large enough such that markers effects are accurately estimated. 2) The relationships between animals in the reference population and the selection candidates can also have an impact on the prediction of breeding value (Habier et al., 2007, Habier et al., 2010). If marker effects are representative of QTL effects based on LD information, these effects may persist over many generations. However, if genomic predictions are reliant on information from close relatives then these predictions may erode within a few generations (Habier et al., 2007, Habier et al., 2010). Therefore the makeup of the reference may also govern how often marker effects need to be re-estimated, such that genomic predictions remain accurate.

2.4. Methods used for genomic evaluation

There have been a number of methods proposed for genomic evaluation. These range from methods that assume all markers have an effect and share the same variance such as ridge regression best linear unbiased prediction (RR-BLUP) and genomic BLUP (gBLUP) to variable selection methods such as Bayes A and B that assume only few markers have large effects and many have small or zero effects and each marker has its own variance.

2.4.1. Ridge Regression Best Linear Unbiased Prediction (RR-BLUP)

Genetic markers can be incorporated into the mixed model equations as random effects. This method has been termed RR (ridge regression or random regression) BLUP (Habier et al., 2010) or SNP BLUP. This method has been extensively examined by Meuwissen et al. (2001) and Habier et al. (2007) and assumes the model;

$$y = 1_n\mu + \sum_i \boldsymbol{W} q_i + e$$

where μ is the mean and q_i is the effect of each SNP. The elements in the matrix $\boldsymbol{W}$ are formed by subtracting $2p_j$ (where p_j is the minor allele frequency of marker j) from the genotype code (0, 1 or 2) such that the sum of the coefficients in each column is zero. Here the marker effects are treated as random and these effects are summed over all SNP to give an estimate of breeding value. Each marker effect is treated as random because it is difficult to gain a fixed estimate of marker effects when there are large numbers of effects to be estimated from a limited number of phenotypic records. This may lead to some over fitting of the data and result in inaccurate estimation of marker effects, which in turn results in poor breeding value estimates. Meuwissen

et al. (2001) showed that this approach results in an inaccurate prediction of breeding values when they fitted SNP effects in a least squares analysis. Treating marker effects as random allows for an effect to be estimated for each marker provided the variance of markers effects is assumed to be known. When all markers share the same variance then the variance at each locus is small and large marker effects are often strongly shrunk towards zero. The genetic variance explained by the SNP effects is given by $\boldsymbol{WW'}\sigma_q^2$ and the residual variance is $\mathbf{I}\sigma_e^2$, and the variance-covariance matrix among observations is therefore $\boldsymbol{WW'}\sigma_q^2 + \mathbf{I}\sigma_e^2$, Rather than assuming the variance for each SNP to be equal, this variance can also be assumed different at each locus, in which case it has to be estimated. This is only possible if prior assumptions are made about the variance and the distributions of SNP effects. Bayesian methods have been proposed to achieve this (Meuwissen et al., 2001; Habier et al., 2010; Habier et al., 2011).

2.4.2. Bayes A and B

Meuwissen et al. (2001) proposed two Bayesian methods for the genomic prediction of additive genetic effects. These were called Bayes A and Bayes B. In the Bayes A method all markers are assumed to have an effect, like RR-BLUP (above) and in Bayes B only a proportion of markers have an effect. The general model assumed by these methods is;

$$y_i = 1\mu + \sum_{j=1}^{k} X_{ij}\beta_j\delta_j + e_i$$

where y_i is the phenotype of animal i, μ is the overall mean, k is the number of marker loci, X_{ij} is a matrix that contains the marker genotype at locus j which is coded as 0, 1, or 2 and is the number of copies of the SNP allele that individual (i) carries, β_j is the allele substitution effect at

locus j, δ_j is a 0/1 indicator variable and determines whether the marker is included in the model. In Bayes B the probability (π) defines how many markers are assumed to have an effect and in Bayes A this value is assumed to be 0, such that all markers have an effect. The final term e_i is the random residual effect (Meuwissen et al., 2001; Habier et al., 2007).

In Bayes A and B parameters are estimated using a Gibbs sampling procedure. The Gibbs sampling algorithm draws parameter values (μ, e and β) from conditional distributions (e.g. μ~uniform, e ~ N(0, σ_e^2) and β ~ N (0, σ_β^2)) where marker effects and the residual are normally distributed. A prior distribution of the variance of the residual and of the marker effects needs to be given. In Meuwissen et al. (2001) an inverted chi-squared distribution was assumed, i.e. $\sigma_\beta^2 \sim x^{-2}(v, S)$, using hyper-parameters v (degrees of freedom (df)) and S (scale) to define the shape of the prior distribution. The scale and df parameters control the shape of the conditional distributions, hence controlling the ability of the model to differentiate between markers with large effect to those with small effects. After running the Gibbs chain for a sufficient number of iterations the posterior mean will stabilize according to the marginal (posterior) distribution. The initial estimates of marker effects can be biased towards the starting value and therefore may be important to discard a number of initial cycles as burn in, so that unbiased estimates can be obtained. Estimates of the parameters and SNP effects are average values from the final cycles and these mean SNP effects are aggregated to give an estimate of breeding value.

In the Bayes A and Bayes B methods it is difficult to observe the impact of the prior on the final estimates of marker effects and variance estimates. One criticism of the Bayesian models used by

Meuwissen et al. (2001) is that Bayes A and B are dependent on the prior and they don't allow for Bayesian learning to occur (Gianola et al., 2009). This means that the hyper-parameters (the scale and df of the prior distribution) that are assigned to the variance may have a strong influence on the extent of shrinkage on marker effects, regardless of the amount of data used. Ideally Bayesian learning should be such that the shrinkage value should tend to 0 as more data becomes available, therefore reducing the dependency of the model on the prior and letting the data be more influential (Gianola et al., 2009).

Bayes B uses the same model as Bayes A, however another variable indicating the absence (with probability π) or presence (with probability $1 - \pi$) of locus *j* in the model is included (represented by δ_j in the above equation). This allows for further differentiation between markers with large effects from those with no effect (Meuwissen et al., 2001). Given that some markers are now assumed to explain zero variance, it is difficult to use Gibbs sampling to sample the variance from the prior distribution, as it no longer moves through the entire sample space, because the variance due to the prior is greater than zero. To achieve marker variance estimates, the Bayes B algorithm adds another step such that variance estimates are obtained by running a Metropolis Hastings (MH) algorithm; the MH accepts a new sample of the marker variance as drawn from the prior distribution with a probability determined by the likelihood of the data given this variance. The MH step in Bayes B can be run for many cycles to obtain an estimate of marker variance, instead of simply sampling it from the prior distribution as in Bayes A. Similar methods such as Bayes C π use this principle of some markers having zero effect, however the proportion of markers that have an effect (π) is estimated based on the data (e.g. see Habier et al., 2011).

2.4.3. Genomic Best Linear Unbiased Prediction (gBLUP)

All methods used by Meuwissen et al. (2001) relied on estimating marker effects from the data. However markers can also be included in genomic evaluations through the use of a marker derived relationship matrix, known as the genomic relationship matrix (GRM) (Nejati Javaremi et al., 1997; Villanueva, 2004; VanRaden, 2008).

The traditional BLUP methodology relies on pedigree information to define the covariance between known relatives, based on expected additive genetic relationships. This covariance can also be defined by the genotypes of a large number of DNA markers that are spread across the entire genome. Combining the information from genetic markers into a relationship matrix was first suggested by Nejati- Javaremi et al. (1997). Similarly, Villanueva et al. (2004) examined the inclusion of a marker derived relationship matrix in genomic evaluation. They suggested that BLUP using a GRM can be used to produce higher accuracy estimates than pedigree based BLUP by representing additive relationships between individuals based on information using shared DNA markers. Relationship estimates in the GRM can deviate from the expected relationship given in the numerator relationship matrix (A). For example variation in the relationship between 2 full siblings may range from 0.4 to 0.6 instead of the expectation of 0.5 given in A (Visscher et al., 2006; Hill & Weir, 2011). Moreover, relationships between more distantly related individuals (often related beyond the existing pedigree) can also be used. This exploitation of the variation in relationships is what makes the GRM a useful tool in genomic evaluations.

The gBLUP method commonly used in many evaluations was introduced by VanRaden (2008) and Habier et al. (2007). In practice the model used to implement gBLUP is;

$$y = \boldsymbol{X}b + \boldsymbol{Z}g + e$$

where y is a vector of phenotypes, **X** is a design matrix relating the fixed effects to each animal, b is a vector of fixed effects, **Z** is a design matrix allocating records to genetic values, g is a vector of additive genetic effects for an individual and e is a vector of random normal deviates with variance σ_e^2. Furthermore var(g) = $\mathbf{G}\sigma_g^2$ where **G** is the genomic relationship matrix, and σ_g^2 is the genetic variance for this model.

Habier et al. (2007) showed that gBLUP and RR BLUP are actually equivalent models. Given that the variance of y in the gBLUP model is given by $\left(\mathbf{ZGZ}'\sigma_g^2 + \mathbf{I}\sigma_e^2\right)$ where σ_e^2 is the residual variance and σ_g^2 is the genetic variance (See above) is equal to the variance of y in the RR-BLUP model: var (y) = $\mathbf{WW}'\sigma_q^2 + \mathrm{I}\sigma_e^2$ (assuming equal variance for each SNP) with σ_q^2 being the genetic variance and $\sigma_q^2 = \sigma_g^2$ (see above). Although these methods are equivalent, gBLUP has three important features that makes it more desirable to use than RR-BLUP: (1) the dimensions of the genetic effects in the mixed model equations is reduced from m x m (where m is the number of markers) in RR BLUP to n x n (where n is the number of individuals in the population) in gBLUP, which is computationally more efficient; (2) the accuracy of an individual's genomic estimated breeding value (GEBV) can be calculated in the same way as in pedigree based BLUP. And (3) gBLUP information can be incorporated with pedigree information in a single step method (e.g. see Misztal et al., (2009)).

The use of gBLUP also has some properties that are not ideal. In pedigree based evaluations, the NRM can be incorporated into the mixed model equations and inverted using specific rules based on pedigree information. This allows for fast and efficient genetic evaluations for large numbers of animals. The GRM cannot be inverted using these rules and is often hard to invert because; in many cases the GRM is not positive definite or diagonal dominant. Furthermore, the GRM is very dense and therefore inversion becomes very slow when the number of animals included in the evaluation is large.

2.4.1. The genomic relationship matrix (GRM)

Estimates of the genomic relationship matrix can be formed using different methods and various ways to make the GRM have been proposed (VanRaden, 2008; Yang et al., 2010; Goddard et al., 2011). Some of these will be presented in this section. Firstly, the method presented by VanRaden (2008) develops an incidence matrix **M**, coded as -1, 0, 1 that specifies which alleles each individual has inherited. The minor allele frequency at locus *i* is p_i, and the matrix **P** contains the allele frequencies expressed as a difference from 0.5 and multiplied by 2, such that column i of **P** is $2(p_i - 0.5)$. Subtraction of **P** from **M** gives exactly ***W*** (termed *matrix **Z*** in VanRaden (2008)). The minor allele frequency correction forces the sum of coefficients across animals to be zero for each marker. It also gives more weighting to rare alleles than to common alleles when calculating genomic relationships (VanRaden 2008). The genomic relationship matrix is calculated as $G = \boldsymbol{WW'}/[2\sum p_i(1 - p_i)]$. The division by 2∑pi (1 - pi) places G on a similar scale as the numerator relationship matrix (A) which is used widely in livestock breeding

(VanRaden 2008), however this may only be if the allele frequencies used to scale G are referring to the same base population as used in A.

Other genomic matrices have been proposed, such as the one used by Yang et al. (2010). Here they combined the information on all SNPs (i) (M is now coded 0, 1, 2) to calculate the relationship between individuals j and k into the GRM (G_{ijk}). Weighting the off-diagonal and diagonal elements differently, when j≠k then;

$$G_{jk} = \frac{1}{N}\sum_i G_{ijk} = \frac{1}{N}\sum_i \frac{(m_{ij} - 2p_i)(m_{ik} - 2p_i)}{2p_i(1-p_i)}$$

When j is equal to k (i.e. the relationship of an individual to itself), then;

$$G_{jk} = \frac{1}{N}\sum_i G_{ijk} = 1 + \frac{1}{N}\sum_i \frac{m_{ij}^2 - (1-2p_i)m_{ij} + 2p_i^2}{2p_i(1-p_i)}$$

where m_{ij} is the element of $\boldsymbol{M}$ pertaining to marker i and individual j and N is the number of markers. These estimates of relationship are all relative to a base population in which the average relationship between individuals is zero (all individuals are completely unrelated). Yang et al. (2010) used the individuals in the sample as the base so that the average relationship between all pairs of individuals is 0 and the average relationship of an individual with him- or herself is 1.

The matrix G can also be regressed towards the numerator relationship matrix (A), which contains the expected numerator relationships as derived from the pedigree information. Then, $\hat{G}$ can be calculated as;

$$\hat{G} = [\, A + b(G - A)]$$

Where σ_a^2 is the additive genetic variance and σ_g^2 is the variance of each of the marker effects, $b = \sigma_g^2/\sigma_a^2$. The regression of $\hat{G}$ back towards A is said to now remove some of the error associated with estimating genomic relationships from a finite number of markers, therefore acknowledging that $\hat{G}$ is an estimate of the true genomic relationship (Goddard et al., 2011).

2.5. Results using genomic evaluation methods in real data

The use of genome wide evaluation methods has been examined in many research articles and has been shown to obtain as accurate or more accurate breeding values in livestock breeding programs than pedigree based BLUP. It has been used extensively in dairy cattle breeding with many countries now reporting genomic breeding values to breeders (VanRaden et al., 2009). Genome wide evaluations have also been examined in beef cattle (Garrick et al., 2011), sheep (Daetwyler et al., 2010), Chickens (Wolc et al., 2011) and also for some plant populations (Jannick et al., 2010). In some cases genomic evaluation has been shown to increase genetic gain, for example in layer hens, Wolc et al. (2011) illustrated that the genomic selection could double the rate of genetic gain for early stage selection on egg weight through the increased accuracy of selection.

In real data, the accuracy of genomic breeding values has been estimated using the correlation between GEBV's and accurate progeny test breeding values (PT BV) or by cross validation. Cross validation involves the sampling groups of animals to be used as the reference and predicting the phenotype (corrected for fixed effects) of animals in the validation group. Often re-sampling of the reference and validation groups occurs many times to gain a more accurate estimate of accuracy.

In dairy cattle, VanRaden et al. (2009) used the correlation between GEBV and PT BV to measure accuracy and reported increases in breeding value accuracies of 20-50%. Similarly Harris et al. (2009) used the gBLUP method to generate genomic predictions on 4,500 dairy cattle and found that reliabilities were 16 to 33 % higher than the breeding values based on parent average information for milk production traits. Other studies have used cross validation to estimate accuracy. For example in beef cattle, accuracies generally range between 0.4 and 0.73 depending on the trait, and at this stage are limited to within breed predictions (Garrick et al., 2011). In sheep, Daetwyler et al. (2010) reported average accuracies of 0.2 to 0.8 and found that the accuracy of genomic breeding values was dependent on the trait, its heritability, the breed of animal and the number of animals with phenotypes in the reference population. These accuracies also are dependent on the relationships between animals; Habier et al. (2010) showed that these predictions quickly erode when the relationship between individuals with phenotypic information and those being predicted reduces.

There has generally been very little difference between using gBLUP and the non-linear models to estimate breeding values in real data (i.e. Bayes A, B) (Moser et al., 2009). This has generated some discussion regarding the model of variation. In many cases we observe that variation is controlled by a finite number of genes and genetic variation appears to be controlled by many small QTL in a model like Fisher's (1918) infinitesimal model. However, there are also traits that have mutations that explain significant amounts of genetic variation (i.e. the DGAT1 mutation explains ~20% of the variation in milk yield and ~50% of the variation in milk fat % in dairy cattle) (Spellman, 2002; Schennink et al., 2004). This means that it is important to understand how different methods perform under alternative models of variation so that the correct method can be used for each trait.

There have been many positive results from the use of genomic evaluation in livestock breeding. However, there are still many interesting questions in which this thesis aims to examine. This thesis studies the selection of methods to be used in genomic selection and how these methods are affected by the underlying model of genetic variation. Moreover we examine the impact of marker density on genomic evaluations, the makeup of reference populations, and the use of genomic information in breeding programs where diversity (inbreeding) needs to be managed.

Different models of genetic variation and their effect on genomic evaluation

Samuel A Clark [1,2§], John M Hickey [1], Julius HJ van der Werf [1,2]

[1]School of Environmental and Rural Science, University of New England, Armidale, NSW, 2351, Australia

[2]Cooperative Research Centre for Sheep Industry Innovation, Armidale, NSW, 2351, Australia

[§]Corresponding author

Email addresses:

SAC: sclark9@une.edu.au

JMH: john.hickey@une.edu.au

JHJW: jvanderw@une.edu.au

Published in Genetics Selection and Evolution, May 2011

Chapter 3. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes

Samuel A. Clark[1,2§], John M. Hickey[1], Hans D. Daetwyler[2,3] and Julius H.J. van der Werf[1,2]

[1] University of New England, Armidale, NSW 2351, Australia

[2] CRC for Sheep Industry Innovation, University of New England, Armidale, NSW 2351, Australia

[3] Biosciences Research Division, Department of Primary Industries, 1 Park Drive, Bundoora, Vic. 3083, Australia

[§]Corresponding author

Email addresses:

SAC: sclark9@une.edu.au

JMH: john.hickey@une.edu.au

HDD: hans.daetwyler@dpi.vic.gov.au

JHJW: jvanderw@une.edu.au

Published in Genetics Selection and Evolution, February 2012

Chapter 4. The effect of genomic information on optimal selection in livestock breeding programs

S.A. Clark[1 2], B.P. Kinghorn[2], J.M. Hickey [2], J. H. J. van der Werf [1 2]

[1] University of New England, Armidale, NSW 2351, Australia

[2] CRC for Sheep Industry Innovation, University of New England, Armidale, NSW 2351, Australia

Corresponding Author: Sam Clark; email: sclark9@une.edu.au

Email addresses:

SAC: sclark9@une.edu.au

BPK: bkinghor@une.edu.au

JMH: john.hickey@une.edu.au

JHJW: jvanderw@une.edu.au

Prepared for Genetics Selection and Evolution

Chapter 5. Comparisons of identical by state and identical by descent relationship matrices derived from SNP Markers in genomic evaluation.

S.A. Clark[1 2], B.P. Kinghorn[2], J.M. Hickey [2], J. H. J. van der Werf [1 2]

[1] University of New England, Armidale, NSW 2351, Australia

[2] CRC for Sheep Industry Innovation, University of New England, Armidale, NSW 2351, Australia

Corresponding Author: Sam Clark; email: sclark9@une.edu.au

Email addresses:

SAC: sclark9@une.edu.au

BPK: bkinghor@une.edu.au

JMH: john.hickey@une.edu.au

JHJW: jvanderw@une.edu.au

Prepared for Genetics Selection and Evolution

Chapter 6. General Discussion

This thesis has aimed to improve the understanding of genetic evaluation using genomic data in livestock. Genetic evaluation systems in many breeding programs have undergone significant changes due to the inclusion of genomic information. The use of genomic information in animal breeding provides the opportunity to increase response to selection by increasing the accuracy of breeding values at a young age and therefore reducing the generation interval. It has also created an opportunity to reduce costs, especially for industries that rely on progeny testing for sire selection. Genomic evaluation has now started to be implemented in many livestock breeding sectors. The dairy cattle industry has started reporting genomic estimated breeding values (GEBV) for young bulls and this information is now being used in the selection of breeding animals (Wiggans et al., 2012). In other species such as pigs and poultry some companies are establishing pipelines to enable future large scale genomic evaluation and GEBVs are being predicted for some selection candidates (Cleveland et al., 2010; Wolc et al., 2011). In beef cattle and sheep, research into genomic evaluation has commenced, and some promising results have been reported (Garrick, 2011; Daetwyler et al., 2010b). Given this interest in, and use of genomic technologies, it is important to understand what is actually being predicted in genomic evaluations because of its impact on accuracy of breeding values, the design of reference populations and management of inbreeding in the breeding program.

This thesis has focused on various different aspects of genomic evaluations that range from the choice of methods (Chapter 3 and 6) to the design of reference populations (chapter 4) and use of

genomic information in a structured breeding program (Chapter 5). The third chapter of this thesis focuses on the selection of appropriate methods used for genomic evaluation. This study showed that the performance of Bayes B and gBLUP is influenced by the underlying patterns that control genetic variation, the density of marker information and the relationship between animals in the reference and test populations. It showed that the gBLUP method was robust against the changes to the assumed model of variation; however Bayes B was more accurate when larger QTL effects controlled variation. Under these assumptions, it also showed that Bayes B would be more accurate than gBLUP when the animals in the reference dataset were more unrelated to those in the validation set. Furthermore, even under an infinitesimal like model Bayes B performed as well as gBLUP as it used marker effects to parameterise the relationships between animals. When the marker density was increased to whole genome SNP sequence, accuracy of genomic predictions using Bayes B increased, however this was dependent on the presence of few large QTL effects. When sequence information was used to predict breeding values under infinitesimal assumptions it had not effect on the accuracy of predictions. Chapter 4 built on the notion, that the relationships between animals in the reference and test datasets are important to the accuracy of breeding value estimation. Animals that had close relatives in the reference gained the highest accuracy of the breeding value. However; we show that even when animals are thought to be unrelated, gBLUP can use small relationship coefficients to give some 'baseline' accuracy. This baseline accuracy is dependent on the size and makeup of the reference population.

Chapter 5 examined the concept of using genomic relationship information in optimal selection so that inbreeding could be controlled and merit maximised. It showed that GEBV's could be

used to increase merit because they were more accurate than pedigree based breeding values. It also showed that GEBV's are highly dependent on information from close relatives for high accuracy however, some of the accuracy achieved is due to within family sources of variation (Mendelian sampling) not accounted for by pedigree based methods. The potential to exploit variation in relationship was also assessed i.e. optimal contribution selection where merit and inbreeding are balanced. In half sib families there was no gain from using genomic information to restrict inbreeding because there was already some differentiation between family members using pedigree based methods. However, using genomic information to manage inbreeding increased gains when selection candidates where from large full sib families because genomic relationships allowed for some differentiation between members of the same family. This allowed for the selection of high performing family members that actually share a lower relationship to the selected animal. This benefit occurred despite the fact that animals that are less related to the top animal are more likely to have a lower breeding value because of a high correlation between the relationship to the best sibling and breeding values.

The sixth chapter of this thesis concentrated on the differentiation between IBD and IBS information for estimation of genomic relationships and the subsequent prediction of breeding values. Here we showed that differentiating between the different marker states has little effects on accuracy and there was a high correlation between the breeding values estimates of the different methods used. It also gave insights into the properties of genomic relationship matrices and how differences between relatives in these matrices can affect breeding values estimations and variance component analysis. Given the findings of this thesis, this general discussion will focus on five main points. These are; 1) the validation of genomic selection methods using

simulated and real data, 2) the number of markers used in genomic evaluation, 3) the construction of reference populations in livestock breeding, 4) across breed evaluations and using information from crossbred individuals, and 5) the value of genomic information to the breeding program.

6.1.1. The validation of genomic selection methods using simulated and real data

There are many applications for simulated data in testing genomic evaluation methods. Simulated data is often used to test important questions that cannot be achieved in real data because of population structure and a lack of knowledge regarding the true underlying models. However there is often debate regarding the validity of the methods and models assumed to simulate genomic data. Various methods range from forward simulation methods that involve random mating of individuals until an appropriate LD structure is found to methods that use coalescent theory that simulates genotypic information based on past ancestral gene flow. However, both methods have many limitations and many are not relevant to real livestock populations (Woolliams & Corbin, 2012). Initially, simulations were used because there was a limit of genotypic data from real livestock populations, now there is a large amount of genotypic information that has enabled the wide scale validation of genomic evaluation in many species. This large amount of data from real populations is often not structured for testing specific hypotheses, as can be done using simulation examples, and real data results are often much harder to interpret. There are two main difficult associated with genomic evaluations in real populations 1) the ways to validate genomic evaluations and 2) the population structure of livestock populations.

In simulation, accuracy is measured by the correlation between true (simulated) breeding value and estimated breeding value. In real populations, true breeding values are unknown and therefore a common method used to validate genomic predictions is the use of progeny tested, accurate breeding values in place of a 'true' breeding value (VanRaden et al., 2009). Often there are too few animals with highly accurate breeding values (0.99) and this threshold is reduced. This reduction in the accuracy threshold can mean that the accurate breeding is now also subject to error and is no longer a good estimate of true breeding value. This results in giving a lower accuracy for the estimated breeding values (EBV). Often comparisons are made between pedigree based parent average EBVs and GEBVs, however there is often a correlation between pedigree based and PT breeding values and this can inflate accuracies for parent average breeding values. Furthermore, progeny test BVs can be subject to selection and therefore may bias the empirical correlation. Using a correlation as a tool for validation can also overlook possible differences in scale and dispersion of various breeding values and therefore regressions statistics should also be calculated.

Another method of testing genomic predictions is through cross validation of GEBV on a phenotype corrected for fixed effects (Daetwyler et al., 2010). Cross validation is when the population is split into a reference and validation. For example the population is split into 5 groups, where 4 groups are combined as the reference dataset and used to predict breeding values of the fifth group (Daetwyler et al., 2010). This is undertaken such that all animals in each group receive an estimate of breeding value based on no phenotypic information. Given that the relationships between animals in the reference and validation populations affects the accuracy of genomic selection (Habier et al., 2007; Habier et al., 2010; Hayes et al., 2009 and chapter 4 of

this thesis), the way in which the dataset is dissected will also have a large impact on predictions. For example if animals in the reference are only from one family and information from these animals is used to predict another entire family accuracies will be lower than if all families are represented in each population. Furthermore the corrected phenotype used to assess accuracy may not be a good estimate of true breeding value.

Estimates of accuracy can be made using the prediction error variance (PEV) of the mixed model equations and have been shown to be similar to those estimated from the correlation between GEBV and progeny test BVs. One advantage of estimating accuracy using the PEV is that animals that share different relationships to the reference are given different accuracies. However this is only useful for the gBLUP method of genomic evaluation as no methods have been defined for variable selection methods that can assess the accuracy of prediction. However it is important to note that this accuracy may be influenced by selection i.e. the Bulmer effect. Bijma (2012) noted that the correlation between true and estimated breeding values will reduce due to selection, however the standard error (use in the calculation of the accuracy using the PEV) remains the same, hence selection is not taken into account. While this is the case, the PEV is still a good measure of how much a breeding value is expected to change and therefore can be used to evaluate accuracy. However, to use this measure of accuracy when calculating response, some adjustments may be needed to account for selection. The structure and management of real populations also affects the success of genomic predictions in livestock populations.

The definition of genetic groups can also have an effect on the accuracy of genomic predictions. For example if breed effects are not fitted, SNP effects will be biased by breed. Furthermore, species and breeds that are heterogeneous and/or contain sub structures within a breed (e.g. sub-strains of Merino sheep) also need to be accounted for in the model. If these genetic groups are not fitted it can cause misleading results as often differences between breeds may be well predicted and within breed predictions may be inaccurate. Furthermore, often pedigree and genomic based breeding values are blended in some evaluations therefore it is important that both pedigree based and genomic based breeding values are estimated using the same model and account for the same genetic group structure. All of these factors need careful consideration when using and validating genomic information in the genetic evaluation of livestock. Understanding the method used to validate genomic predictions is important when comparing the success of genomic evaluations in different livestock populations.

6.1.2. The number of markers used in GS and the promise of sequence data

The theory of genomic selection relies on the assumption that markers are in linkage disequilibrium (LD) with QTL. This allows for the use of marker information in genetic evaluation and the prediction of breeding values [10]. The number of markers used in genomic evaluation is expected to have an impact on the accuracy of genomic breeding values. The marker density needs to be high enough such that all QTL are in LD with at least one marker, and therefore allowing for accurate predictions of QTL effects from the marker information (Goddard, 2009; Daetwyler et al., 2010a; Goddard & Hayes, 2009). There have been proposals of using markers that range in number from 384 (Garrick, 2011) to whole genome SNP sequence (Meuwissen & Goddard, 2010; Clark et al., 2011) for genomic evaluation.

The number of SNP needed depends on the distance in which LD extends over the genome. If this distance is large, i.e. for across breed prediction, then there is a need for a high density of markers such each QTL is in LD with at least one marker. If SNP are too far apart then they may not capture all of the variance at the QTL level and breeding value predictions will not be accurate (Yang et al., 2010). For example, de Roos et al. (2009) found that there was LD conserved over the distance of 10kb in *Bos taurus* cattle breeds and therefore approximately 300,000 markers would be needed for some prediction of merit across breed. In contrast, the number of markers needed for prediction of merit within a breed is substantially less (Weigel et al., 2009). Vazquez et al. (2010) showed that accurate genomic predictions could be made from 2,000 markers in Holstein cattle. Furthermore, Garrick (2011) also showed that predictions of merit could be made from as little as 600 specially selected markers using variable selection methods. The numbers of markers used for within breed predictions is also dependent on the structure of the population. Factors such as how related the animals in the reference and test dataset are (i.e. predictions of offspring or predictions of unrelated animals), the effective population size and the number and size of QTL effects to be estimated can all impact how many markers are needed for genomic prediction.

The relationships between animals in the reference and test datasets will have an impact on the number of markers needed to accurately predict genetic merit. If there is a close relationship between animals in the two populations fewer markers will be needed to make accurate predictions. However because of this also means that the predictions are less reliant on LD

information, the effects estimated from these populations may not be relevant in different unrelated subsets of the population. To make accurate predictions of merit from LD information, the reference population size needs to be large and high densities of markers are needed such that predictions can be made across families (provided a range of families are represented in the reference). The amount of LD within the population is controlled by the effective population size (*Ne*) and the structure of that population.

The effective size of a population is a major driver of LD in the population, for example if the *Ne* is low the LD between markers and QTL will be high and accurate predictions can be made from a small number of markers. This may explain why the accuracy of genomic predictions in Holstein and Jersey cattle have not increased when increasing the number of SNP from 50K to 770K (Erbe et al., 2012; Su et al., 2012) as these breeds of cattle have small effective population sizes. Where *Ne* is large, as in humans, a large number of SNP (> 1,000,000) are needed to detect marker effects (Yang et al., 2010). The benefit of increasing the number of markers used in genomic evaluation is also affected by the number and size of the QTL that are to be estimated. If few QTL with large effects are responsible for genetic variation, then moving to higher SNP densities will increase accuracy (Goddard, 2009; Goddard & Hayes, 2009). Furthermore whole genome SNP sequence provides the opportunity to have enough markers to find causative loci. In simulation, where few QTL control genetic variation, the use of sequence data enabled accuracies of close to 1.0 to be achieved (Meuwissen & Goddard, 2010). To gain the advantages associated with the use of high density genotypes and SNP sequence variable selection methods such as Bayes A and B need to be used to predict breeding values (Erbe et al., 2012). As shown in chapter 3, the use of sequence data may have no effect on the prediction of breeding values

when using gBLUP however, for variable selection methods, accuracy can increase. Furthermore, as shown in chapter 3, if there are many QTL with small effects, somewhat like Fisher's (1918) infinitesimal model, increasing marker density alone may have very little impact on the accuracy of predictions.

To gain all of the benefits that have been suggested regarding the use of SNP sequence a large number of phenotypic records also need to be obtained to make accurate estimates of marker effects. Although some theoretical formula have been suggested for the number of phenotypes needed (Goddard, 2009; Hayes et al., 2009; Goddard & Hayes, 2009), the exact number of phenotypic records required is still largely unknown. In a broad context, to gain a precise estimate of a markers effect there needs to be sufficient 'power' in the experiment. A large number of animals is needed to detect significant associations when the size of QTL effects is small and fewer animals are needed if QTL effects are large (Goddard & Hayes, 2009).

The need for phenotypes may well be the major limiting factor in the use of high density marker information. In some industries, the availability of smaller, cheaper SNP chips may enable the genotyping of a large number of animals that have phenotypic measurements. For example in the US dairy industry approximately 40,000 cows with phenotypic records have been genotyped with a 3K chip (Wiggans et al., 2012). If there is a relationship between animals with low density marker genotypes with those that have high density or sequence information then many more SNP can be imputed such that animals with low density information now have high density genomic information (Scheet & Stephens, 2006). This allows for a large increase in the number

of phenotypes used in genomic evaluations. Given the high reliance on phenotypic information, it is important that reference populations (or animals with phenotypic records) are constructed carefully such that all advantages associated with genomic selection can be achieved.

6.1.3. The makeup of reference populations for genomic selection

The reference population is the group of animals that have accurate genotypic and phenotypic measurements taken on them. These phenotypes then inform us about selection candidates, as these candidates share QTL and/or they share a direct relationship with animals in the reference. The makeup of the reference population depends on the animals that are to be predicted. Original expectations were that genomic evaluation predicted the effects of markers in LD with QTL and these effects would be consistent across families and breeds. However, many results show that the accuracy of genomic evaluations tend to be impacted heavily by the relationships between animals in the reference and test populations (Habier et al., 2007; Habier et al., 2010; Clark et al., 2012). This reliance on family information shows that to achieve highly accurate predictions, animals that are to be predicted from the reference need to be closely related to those in the reference (as shown in Chapter 4). Interestingly, this may also imply that the current genomic predictions rely heavily linkage information rather than population wide LD to make accurate predictions of merit. Animals that are relatively unrelated to those in the reference will still receive some accuracy, which we termed the baseline accuracy in chapter 4. This is information from unrelated individuals is more likely to be based on LD information alone. If the reference population is small, the base line accuracy would be expected to be low. By increasing the number of animals in the reference this base line accuracy may increase due to selection candidates sharing more small relationships that would contribute to the estimate of the EBV.

There is often a high degree of similarity between the phenotypes of relatives and this means that each phenotypic observation is not independent. In a reference population it is important to have enough family members, such that accurate phenotypes can be observed. Furthermore it is important that the reference consists of many unrelated families, rather than one family alone. This also implies that the relationships between animals within the reference population are also important, especially for the prediction of breeding values across different families or even breeds. For example, if there are limitations on the number of animals to be used in this reference, animals within the reference should share a high relationship to the selection candidates and a low relationship to other animals in the reference (Pszczola et al., 2012).

In many species, such as dairy cattle, the reference population consists of all animals with genotypic and phenotypic information (Wiggans et al., 2012). If phenotypic records are not available it can also be replaced by accurate breeding values (i.e. phenotypic information on progeny). Given that genomic selection has the greatest effect on traits that are hard to measure, measured late in life, sex limited or of low heritability, there is the opportunity to specially construct reference populations (van der Werf et al., 2010). In the Australian sheep industry the concept of a centralised nucleus for measuring these traits has been set up so that genomic information can be used to increase genetic gains for these important traits. Ideally the reference population would be continually updated such that all genotypic and phenotypic information is recorded on all animals (as in dairy) (Wiggans et al., 2012). This would also automatically indicate that selection candidates would have a close relationship to the reference. This

continually updating system is in fact similar to the current pedigree based evaluation system. Such a large scale updating reference population is not possible for centralised nucleus systems and therefore animals need to be specially selected to enter the reference. One way to construct a reference would be to include a large number of unrelated families (sires), and measure all important traits on a limited number of the offspring of these sires. The information gained from these animals is then used to predict animals in commercial populations. To construct a reference population in this way there is a need for a large amount of initial recording of genotypes and phenotypes, yet in later generations updating can be reduced to key young sires that are not highly related to those already in the reference. Given that there still has to be a high relationship to the reference to achieve high accuracies it presents some difficulties associated with using information from one breed to predict the merit of another. It may be that such problems become smaller if sequence data is used, as predictions may be from the causal DNA variants and may be valid across more generations and even across breeds (Meuwissen & Goddard, 2010). The impact of sequence data on genomic predictions will become clearer in the near future as more sequence data is now being analysed.

6.1.4. Across breed prediction

In species such as beef cattle and sheep, it may be advantageous to be able to predict the merit of animals across different breeds. Especially for breeds that do not have enough phenotypes to form their own reference population i.e. small breeds of sheep and beef cattle. As previously noted, the number of markers used in these predictions needs to be sufficient so that there are enough markers to be in complete LD with the QTL in both breeds (de Roos et al., 2009; Goddard & Hayes, 2009). Furthermore, to be able to capture all QTL in different breeds there

must be assumptions made that QTL will have the same or similar effects in different breeds and that new QTL have not formed since the breeds diverged from a common ancestor. Additionally, if variation is controlled by many small QTL, across breed predictions may be difficult because different breeds share smaller amounts of their genome than animals of the same breed origin. These issues are difficult to overcome and may result in across breed predictions being of lower accuracy than within breed predictions.

Given the restrictions of across breed predictions it is still plausible that animals in different breed can share QTL that have been conserved over time, however to use this information specific methods have to be used. Similarly to getting the value out of a high density of markers to predict across breeds, variable selection methods may be more suitable to gain some information from across breed predictions (Erbe et al., 2012). However the predictions may still be affected by the reference population makeup and the breed composition of the reference. For example marker effects may be biased towards the major breed in the reference population as the phenotypes used to train these predictions will be dominated by the larger breed. In some instances accuracy can actually be reduced by adding information from another breed, this watering down is mainly evident when there are few animals in the reference of the target breed. When using gBLUP to predict across different breeds it may be important to the GRM based on breed specific allele frequencies. However, to date, non-variable selection methods, such as gBLUP, have only been able to predict merit with low or no accuracy using across breed information (Erbe et al., 2012).

An alternative to using groups of purebred individuals to predict merit is the use of information from crossbreds to inform purebred selection candidates (Toosi et al., 2010). These crossbred animals carry alleles from both breeds and therefore may contain some information about each breed. Wei and van der Werf (1995) suggested the use of information for both crossbred and purebred animals in a combined crossbred and purebred selection (CCPS) and found that this could increase genetic gain in pedigree based breeding programs. This is often difficult to implement in large scale situations because crossbred individuals are not routinely recorded (Dekkers, 2007). The value of crossbred information to genomic evaluation still remains largely unknown and how to deal with this information presents some complications and opportunities (Ibánez-Escriche et al., 2009). One such complication may be due to non-additive sources of variation in cross bred animals, as these are less likely to occur in purebreds. This will therefore make the phenotypes that are being trained on different to those that need to be predicted. However this also provides opportunities for research into understanding some non-additive variation such as dominance and epistasis. Using genomic information may allow for better exploitation of these factors in the development of optimal breed crosses.

In cross breeding programs, it may also be valuable to be able to estimate the breed proportions of high performing crosses. Current marker densities allow for the identification of breed crosses through methods such as principle component analysis and denser marker chips may be expected to increase the accuracy of this identification. Furthermore, as more breeds within species have genotypic information recorded on them, breed specific haplotype libraries may be possible for the identification of breed crosses. This may have important implications for systems in which knowledge of ancestry is limited and also where the performance of different breed crosses is

dependent on the environment (such as in subsistent or large scale farming systems). Similar to crossbreeding breeding programs, there are also opportunities for using genomic information in more traditional breeding programs.

6.1.5. The value of genomic information to a structured breeding program

In traditional breeding programs, the accuracy of an animal's breeding value will increase as more information becomes available. Young animals have no information of their own therefore all of the accuracy achieved for breeding values of these animals is from parents and other relatives. As the animal ages it may have a phenotypic measurement recorded on it which increases accuracy. Once animals are selected they may also have progeny where accuracy is further increased. In many species, phenotypic measures are recorded after sexual maturity and gestation lengths are high such that it takes many years to obtain accurate measures of merit to make selection decisions. If genomic selection can offer accuracies of a phenotypic measurement then more accurate selection can occur at a younger age. In some species selection can occur in young animals after birth and can be used to determine which animals enter progeny test program, in a two stage selection (i.e. pre-selection of dairy bulls) (VanRaden et al., 2009).

In the breeding program, there is also an opportunity to use genomic information in conjunction with artificial breeding technologies. Artificial insemination and multiple ovulation reproduction technologies are currently used to increase genetic gain in some livestock species i.e. dairy and beef cattle. Genomic selection could be used in conjunction these technologies, especially female reproductive technologies, as pre-selection could occur at the embryonic level such that higher

value embryos could be selected. The use of juvenile females as dams could also drastically reduce generation intervals in many species. However, it would be important to maintain genetic diversity such that long term gains could be achieved.

Genomic information may also be useful in a breeding program to both control inbreeding and to increase genetic gain. In Chapter 5 we showed that inbreeding could be controlled by using the genomic relationship matrix as the measure of co-ancestry. Many methods used to constraint inbreeding limit the use of animals from the same family and are often limited by the amount of variation that is attributed to Mendelian sampling variation. As in Chapter 5, variation in breeding values is such that V(a) = ¼ V(sire) + ¼ V(dam) + ½ V (MS) + e. Information about mendelian sampling can only be obtained from the measurement of own performance, progeny or through a DNA test. With some of the Mendelian sampling variation explained, selection of more than one animal from a high performing family is possible (Avendano et al., 2004). When breeding values are predicted using parent information, full sibling share a high intra-class correlation between their breeding values and no information about the Mendelian sampling variation is known. Therefore with the use of genomic information the gains are expected to be the highest in breeding programs that select from large full sib families i.e. pig breeding programs. In other breeding programs where less related animals are to be selected, genomic information is expected to perform as well as pedigree based methods. Using genomic methods to control inbreeding may also be advantageous when pedigree is unavailable or incomplete (for example in conservation breeding programs). There may also be the opportunities to use genomic information to manage recessive genetic disorders whilst balancing merit. The culling of all carrier animals can eradicate the disorder however; some carrier individuals are often of

high merit. Genomic information could be used to manage the use of carrier individuals so that all animals can be effectively used in the breeding program. Managing the diversity of the population may also be possible by reducing the number of opposing homozygotes on the genome of an individual. As often many genetic disorders are recessive and are only observed when animals become more inbred. There are also other opportunities and challenges for the use of genomic data in the prediction of breeding values in real data.

6.2. Conclusions

The use of genomic information in the genetic evaluation of livestock has provided many interesting developments in animal breeding programs. This thesis has shown that the success of genomic prediction is dependent on the methods used to predict breeding values. We have shown that IBS or specific IBD based gBLUP methods and variable selection methods such as Bayes B can all be used to predict breeding value. The distribution of QTL effects resulting from the underlying genetic model can also affect breeding value estimation such that some traits may be more suited to be analyses with one method and another may be analysed with the other. We have clearly shown that the relationship between animals in the reference and test populations is important to the accuracy of genomic selection. Furthermore we have shown that estimates of genomic relationships can be used to manage inbreeding in an optimised breeding program and to increase genetic gain. We have discussed these issues in the context of both real and simulated populations and identified some areas for future research, which includes the use of sequence data in genomic evaluations and utilizing crossbred performance in genomic evaluations.

Chapter 7. Consolidated Reference List

1. Avendano, S., Woolliams, J.A. & Villanueva, B. (2004). Mendelian sampling terms as a selective advantage in optimum breeding schemes with restrictions on the rate of inbreeding. *Genet. Res.* 83: 55-64.
2. AWI, MLA (2010). Gain from genetics. *Australian Wool and Innovation, Meat and Livestock Australia.* [http://www.makingmorefromsheep.com.au/gain-from-genetics/index.htm].
3. Bijma, P. (2012). Accuracies of estimated breeding values from ordinary genetic evaluations do not reflect the correlation between true and estimated breeding values in selected populations. *J. Anim. Breed. Genet.* 129: 345-358.
4. Browning, B.L. & Browning, S.R. (2011). A Fast, Powerful Method for Detecting Identity by Descent. *Am. J. Hum. Genet.* 88: 173-182.
5. Bulmer, M.G. (1971). The effect of selection on genetic variability. *American Naturalist.* 105: 201-211.
6. Calus, M.P.L., Meuwissen, T.H.E., de Roos, A.P.W., & Veerkamp R.F. (2008). Accuracy of genomic selection using different methods to define haplotypes. *Genetics* 178: 553-561.
7. Chen, G.K., Marjoram P. & Wall J.D. (2009). Fast and flexible simulation of DNA sequence data. *Genome Res.* 19: 136-142.
8. Clark, S.A., Hickey, J.M, Daetwyler, H.D. & van der Werf, J.H.J. (2012). The importance of information on relatives for the prediction of genomic breeding values

and the implications for the makeup of reference data sets in livestock breeding schemes. *Genet. Sel. Evol.* 44: 4

9. Clark, S.A., Hickey, J.M. & van der Werf, J.H.J. (2011). Different models of genetic variation and their effect on genomic evaluation. *Genet. Sel. Evol.* 43: 18.
10. Cleveland, M.A., Forni, S., Garrick, D.J. & Deeb, N. (2010). Prediction of genomic breeding values in a commercial pig population. *In Proc. 9th World Congr. Genet. Appl. Livest. Prod. Leipzig, Germany*
11. Daetwyler, H.D., Hickey, J.M., Henshall, J.M., Dominik, S., Gredler, B. et al., (2010). Accuracy of estimated genomic breeding values for wool and meat traits in a multi-breed sheep population. *Anim. Prod. Sci.* 50: 1004-1010.
12. Daetwyler, H.D., Pong-Wong, R., Villanueva, B. & Woolliams, J.A. (2010). The impact of genetic architecture on genome-wide evaluation methods. *Genetics* 185: 1021-1031.
13. Daetwyler, H.D., Villanueva, B., Bijma, P. & Woolliams, J.A. (2007). Inbreeding in genome-wide selection, *J. Anim. Breed. Genet.* 124: 369-376
14. Dalton, R., (2009). No bull: genes for better milk. *Nature* 457(7228):369
15. de Roos, A.P.W., Hayes, B.J. & Goddard, M.E. (2009). Reliability of genomic breeding values across multiple populations. *Genetics* 183: 1545-1553.
16. Dekkers, J.C.M. (2007). Marker-assisted selection for commercial crossbred performance. *J. Anim. Sci.* 85: 2104-2114
17. Dekkers, J. (2004). Commercial application of marker- and gene-assisted selection in livestock: Strategies and lessons. *J. Anim. Sci.* 82: 313-328.

18. Donnelly, K.P. (1983). The probability that related individuals share some section of the genome identical by descent. *Theor. Pop. Biol.* 23: 34-64

19. Erbe, M., Hayes, B.J., Matukumalli, L.K., Goswami, S., Bowman, P.J. et al., (2012). Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *J. Dairy Sci.* 95(7): 4114-29

20. Falconer, D.S. & MacKay, T.F.C. (1996). Introduction to quantitative genetics. *4th edition. Longman Scientific & Technical,* Burnt Mill, Harlow, United Kingdom.

21. Farnir, F., Coppieters, W., Arranz, J., Berzi, P., Cambisano, N. et al., (2000). Extensive Genome-wide Linkage Disequilibrium in Cattle. *Genome Res.* 10: 220-227.

22. Fearnhead, N.S., Wilding, J.L., Winney, B., Tonks, S., Bartlett, S. et al., (2004). Multiple rare variants in different genes account for multifactorial inherited susceptibility to colorectal adenomas. *Proc. Natl. Acad. Sci. USA.* 101: 15992-15997.

23. Fisher, R.A. (1918). The correlation between relatives on the supposition of mendelian inheritance. *Trans. R. Soc. Edin.* 52: 399-433.

24. Forni, S., Aguilar, I. & Misztal, I. (2011). Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. *Genet. Sel. Evol.* 43:1-8

25. Garrick, D.J. (2011). The nature, scope and impact of genomic prediction in beef cattle in the United States. *Genet. Sel. Evol*. 43:17

26. Gianola, D., de los Campos, G., Hill, W.G., Manfredi, E. & Fernando, R. (2009). Additive Genetic Variability and the Bayesian Alphabet. *Genetics* 183: 347-363.

27. Gilmour, A.R., Gogel, B.J., Cullis, B.R. & Thompson, R. (2009). ASReml User Guide Release 3.0. *Hemel Hempstead:* VSN International Ltd.

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

29. Goddard, M.E., Hayes, B.J., McPartlan, H. & Chamberlain, A.J. (2006). Can the same genetic markers be used in multiple breeds? *In Proc. 9th World Congr. Genet. Appl. Livest. Prod. Brazil.* 22-16.

30. Goddard, M.E., Hayes, B.J. & Meuwissen T.H.E. (2011). Using the genomic relationship matrix to predict the accuracy of genomic selection. *J. Anim. Breed Genet.* doi: 10.1111/j.1439-0388.2011.00964.x.

31. Goddard, M.E., Hayes, B.J. & Meuwissen, T.H.E. (2010). Genomic selection in livestock populations. *Genet. Res.* 92: 413-421.

32. Goddard, M.E. & Hayes, B.J. (2007). Genomic Selection. *J Anim Breed Genet* 124: 323-330.

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

34. Goring, H., Terwilliger, J.D. & Blangero, J. (2001). Large upward bias in estimation of locus-specific effects from genome wide scans. *Am. J. Hum. Gen.* 69: 1357-1369.

35. Grapes, L., Dekkers, J., Rothschild, M. & Fernando, R. (2004). Comparing Linkage Disequilibrium-Based Methods for Fine Mapping Quantitative Trait Loci. *Genetics* 166, 1561-1570.

36. Guillaume, F., Fritz, S., Boichard, D. & Druet, T. (2008). Short communication: Correlations of marker-assisted breeding values with progeny-test breeding values for eight hundred ninety-nine French Holstein bulls. *J. Dairy Sci.* 91: 2520-2522.

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

38. Habier, D., Fernando, R.L., Kizilkaya, K. & Garrick, D.J. (2011). Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics* 12:186

39. Habier, D., Fernando, R.L., Kizilkaya, K. & Garrick, D.J. (2010). Extension of the Bayesian Alphabet for Genomic Selection. *In Proceedings of the 9th Congress on Genetics Applied to Livestock Production: 1-6 August 2010; Leipzig. 2010*

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

41. Haley, C.S. & Knott, S.A. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity 69*: 315-24.

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

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

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

45. Hayes, B.J., Chamberlain, A. & Goddard, M.E. (2006). Use of linkage markers in linkage disequilibrium with QTL in breeding programs. In *Proc. 8th World Congr. Genet. Appl. Livest. Prod. Belo Horizonte, Brazil,* 8: 30-06.

46. Henderson, C.R. (1950). Estimation of genetic parameters. *Ann. Math. Stat.* 9: 309

47. Henderson, C.R. (1975). Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics* 31 (2): 423–447. doi:10.2307/2529430. PMID 1174616.

48. Hickey, J.M. & Tier, B. (2009). AlphaBayes: user manual. *UNE, Australia.*

49. Hickey, J.M., Kinghorn B.P., Tier, B., Clark, S.A., van der Werf, J.H.J., & Gorjanc, G. (2012). Genomic evaluations using similarity between haplotypes. *J Anim Breed Genet*

50. Hill, W.G. & Weir, B.S. (2011). Variation in actual relationship as a consequence of Mendelian sampling and linkage. *Genet. Res.* 93(1): 47-64.

51. Ibánez-Escriche, N., Fernando, R.L., Toosi, A. & Dekkers, J.C. (2009). Genomic selection of purebreds for crossbred performance. *Genet Sel Evol* 41:12.

52. Jannink, J.L. (2010). Dynamics of long-term genomic selection. *Genet. Sel. Evol.* 42:35.

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

54. Khatkar, M., Collins, A., Cavanagh, J., Hawken, R., Hobbs, M. et al., (2006). A First-Generation Metric Linkage Disequilibrium Map of Bovine Chromosome 6. *Genetics* 174, 79-85.

55. Kijas, J.W., Townley, D., Dalrymple, B.P., Heaton, M.P., Maddox, J.F. et al., (2009). A genome wide survey of SNP variation reveals the genetic structure of sheep breeds. *PLoS One* 4(3): e4668. doi:10.1371/journal.pone.0004668.

56. Kinghorn, B.P. (2012). A simple method to calculate ibd probabilities. *In the Proc of the Inter Conf of Quant Genet (ICQG)*, Edinburgh June 2012

57. Kwee, I., Rinaldi, A., Rancoita, P., Rossi, D., Capello, D. et al., (2012). Integrated DNA copy number and methylation profiling of lymphoid neoplasms using a single array. *Brit. J. of Haem.*156(3): 354-357,

58. Luan, T., Woolliams, J.A., Ødegård, J., Dolezal, M., Roman-Ponce, S.I. et al., (2012). The importance of identity-by-state information for the accuracy of genomic selection. *Genet. Sel. Evol* 44:28

59. Lynch, M. & Walsh, B. (1998). Genetics and the analysis of quantitative traits. *Sinauer Associates Inc.,* Sunderland, MA.

60. Maher, B. (2008). Personal genomes: the case of the missing heritability. *Nature* 456:18-21.

61. Maki-Tanila, A. & Kennedy, B.W. (1986). Mixed model methodology under genetic models with a small number of additive and non-additive loci. *Proc. 3rd World Congr. Genet. Appl. Livest. Prod. 16-22 July 1986; Lincoln.*

62. Meuwissen, T.H.E. & Goddard, M.E. (2010). Accurate prediction of genetic values for complex traits by whole-genome resequencing. *Genetics* 185: 623-31.

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

64. Meuwissen, T.H.E, Karlsen, A., Lien, S., Olsaker, I. & Goddard, M.E. (2002): Fine mapping of quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. *Genetics* 161: 373-379.

65. Meuwissen, T.H.E. (1997): Maximizing the response of selection with a predetermined rate of inbreeding *J. Anim. Sci.* 75: 934-940

66. Misztal, I., Legarra, A. & Aguilar, I. (2009). Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. *J. Dairy Sci.* 92(9): 4648-4655.

67. Moghaddar, N. & van der Werf, J. (2007). Effect of Accuracy of QTL Parameter Estimation on Response to Genotype-Assisted selection. *In the 3rd Intern. Conf. of Quant. Genet., Hangzhu, China*.

68. Moon, S., Shin, H., Cheong, H., Cho, H., Namgoong, S. et al. (2007). BcSNPdb: Bovine Coding Region Single Nucleotide Polymorphisms Located Proximal to Quantitative Trait Loci. *J. Biochem. and Mol. Biol.* 40(1): 95-99.

69. Moore, S., Li, C., Basarab, J., Snelling, W., Kneeland, J. et al. (2003). Fine mapping of quantitative trait loci and assessment of positional candidate genes for backfat on bovine chromosome 14 in a commercial line of Bos taurus. *J. Anim. Sci.* 81: 1919-1925.

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

71. Mrode, R.A. (2005). Linear Models for the Prediction of Animal Breeding Values. *CAB International, Oxon, UK.*

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

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

74. Nielsen, H.M., Sonesson, A.K. & Meuwissen, T.H.E. (2011). Optimum contribution selection using traditional best linear unbiased prediction and genomic breeding values in aquaculture breeding schemes. *J. Anim. Sci*. 89(3): 630-638.

75. Notter, D. (2004). Multiple-Trait Selection in a Single-Gene World. *Ann. Res. Sym.and An. Meet. Beef Imp. Fed.,* 20-25.

76. Pereira, A., Alencar, M., Oliveira, H. & Regitano, L. (2005). Association of GH and IGF-1 polymorphisms with growth traits in a synthetic beef cattle breed. *Genet. Mol. Biol.* 28(2): 230-236.

77. Pong-Wong, R. & Woolliams, J.A. (2007). Optimisation of contribution of candidate parents to maximise genetic gain and restricting inbreeding using semidefinite programming. *Genet. Sel. Evol.* 39: 3-25.

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

79. Price, K. & Storn, R. (2004). Differential evolution. *Dr. Dobb's Journal* 78:18-24

80. Pryce, J.E. & Daetwyler, H.D. (2012): Designing dairy cattle breeding schemes under genomic selection: a review of international research. *Anim. Prod. Sci.* 52(2-3): 107-114.

81. Pryce, J.E., Hayes, B.J. & Goddard, M.E. (2012): Novel strategies to minimize progeny inbreeding while maximizing genetic gain using genomic information. *J. Dairy Sci.* 95(1): 377-388.

82. Pszczola, M., Strabel, T., Mulder, H.A. & Calus, M.P. (2012). Reliability of direct genomic values for animals with different relationships within and to the reference population. *J Dairy Sci* 95(1): 389-400.

83. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A. et al.,(2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81:559-575

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

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

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

87. Schennink, A., Stoop, W.M., Visker, M.H.P.W., Heck, J.M.L., Bovenhuis, H. et al., (2007). DGAT1 underlies large genetic variation in milk-fat composition of dairy cows. *Anim. Genet.* 38(5): 467-473

88. Schierenbeck, S., Pimentel, E.C.G., Tietze, M., Koerte, J., Reents, R. et al., (2011). Controlling inbreeding and maximizing genetic gain using semi-definite programming with pedigree-based and genomic relationships. *J. Dairy Sci.* 94(12): 6143-6152.

89. Sonesson, A.K., Woolliams, J.A. & Meuwissen, T.H.E. (2010). Maximising genetic gain whilst controlling rates of genomic inbreeding using genomic optimum contribution selection, *In Proc. 9th World Congr. Genet. Appl. Livest. Prod, Leipzig, Germany (2010)* Abstract 0892

90. Spelman, R.J., Ford, C.A., McElhinney, P., Gregory, G.C. & Snell, R.G. (2002). Characterization of the DGAT1 gene in the New Zealand dairy population. *J. Dairy Sci.* 85: 3514-7.

91. Su, G., Brøndum, R.F., Ma, P., Guldbrandtsen, B., Aamand, G.P. et al., (2012).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(8):4657-65.

92. Thompson, E.A. & Basu, S. (2003). Genome sharing in large pedigrees: multiple imputation of ibd for linkage detection. *Human Heredity* 56:119-125.

93. Thompson, E.A. (2008). The IBD process along four chromosomes. *Theor. Pop. Biol.* 73: 369-373.

94. Toosi, A., Fernando, R.L. & Dekkers, J.C.M. (2010). Genomic selection in admixed and crossbred populations. *J. Anim. Sci.* 88: 32-46

95. van der Werf, J.H.J., Kinghorn, B.P. & Banks, R.G. (2010): Design and role of an information nucleus in sheep breeding programs. *Anim. Prod. Sci.* 50: 998-1003.

96. van der Werf, J. (2000). Basics of Linkage and Gene Mapping. *Ident. Incorp. Genet. Mark. Maj. Genes Anim. Breed. Prog.*, 45-54.

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

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

99. Vazquez, A.I., Rosa, G.J.M., Weigel, K.A., de los Campos, G., Gianola, D. et al., (2010). Predictive ability of subsets of SNP with and of parent average for several traits in US Holstein. *J. Dairy Sci.* 93:5942-5949.

100. Villa-Angulo, R., Matukumalli, L.K., Gill, C.A., Choi, J., Van Tassell, C.P. et al., (2009). High-resolution haplotype block structure in the cattle genome. *BMC Genetics* 10:19.

101. Villanueva, B., Pong-Wong, R., Fernandez, J. & Toro, M.A. (2005). Benefits from marker-assisted selection under an additive polygenic genetic model. *J. Anim. Sci.* 83: 1747-1752.

102. Visscher, P.M., Medland, S.E., Ferreira, M.A., Morley, K.I., Zhu, G. et al., (2006). Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genet.* 2:41

103. Visscher, P.M. (2008). Sizing up human height variation. *Nat. Genet.* 40(5): 489-490.

104. Walsh, B. & Henderson, D. (2004). Microarrays and beyond: What potential do current and future genomics tools have for breeders? *Amer. Soc. Anim. Sci.* 82: 292-299.

105. Wei, M. & van der Werf, J.H.J. (1995). Genetic correlation and heritabilities for purebred and crossbred performance in poultry egg production traits. *J. Anim. Sci.* 73: 2220-2226.

106. Weigel, K.A., de los Campos, G., González-Recio, O., Naya, H., Wu, X.L. et al., (2009). Predictive ability of genomic breeding values estimated from selected subsets of single nucleotide polymorphism markers for lifetime net merit in Holstein cattle. *J. Dairy Sci*. 92:5248-5257.

107. Wiggans, G.R., Cooper, T.A., VanRaden, P.M., Olson, K.M. & Tooker, M.E. (2012). Use of the Illumina Bovine3K BeadChip in dairy genomic evaluation. *J. Dairy Sci*. 95:1552-1558

108. Wolc, A., Stricker, C., Arango, J., Settar, P., Fulton, N.E. et al., (2011). Breeding value prediction for production traits in layer chickens using pedigree or genomic relationships in a reduced animal model. *Genet. Sel. Evol*. 43:5

109. Womack, J. (2005). Advances in livestock genomics: opening the barn door. *Genom. Res.* 15(1699-1705).

110. Woolliams, J. & Corbin, L. (2012). Coalescence theory in livestock breeding. *J. Anim. Breed. Genet*. 129: 255-256

111. Wray, N.R. & Goddard, M.E. (1994). Increasing long-term response to selection. *Genet. Sel. Evol.* 26: 431-451

112. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S.D., Henders, A.K. et al., (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* 42:565-571.

113. Zhao, H., Fernando, R. & Dekkers, J. (2007). Power and Precision of Alternate Methods for Linkage Disequilibrium Mapping of Quantitative Trait Loci. *Genetics* 175: 1975-1986.

114. Zuka, O., Hechtera, E., Shamil, R., Sunyaeva, B., Lander, E. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *PNAS* doi/10.1073/pnas.1119675109