Components of the accuracy of genomic prediction in a multi-breed sheep population[1]

H. D. Daetwyler,*†[2] K. E. Kemper,*‡ J. H. J. van der Werf,†§ and B. J. Hayes*†#

*Biosciences Research Division, Department of Primary Industries, Bundoora, VIC 3083, Australia; †Cooperative Research Centre for Sheep Industry Innovation, Armidale, NSW 2351, Australia; ‡Faculty of Land and Environment, University of Melbourne, Parkville, VIC 3010, Australia; §School of Environmental and Rural Science, University of New England, Armidale, NSW 2351, Australia; and #La Trobe University, Bundoora, VIC 3086, Australia

ABSTRACT: In genome-wide association studies, failure to remove variation due to population structure results in spurious associations. In contrast, for predictions of future phenotypes or estimated breeding values from dense SNP data, exploiting population structure arising from relatedness can actually increase the accuracy of prediction in some cases, for example, when the selection candidates are offspring of the reference population where the prediction equation was derived. In populations with large effective population size or with multiple breeds and strains, it has not been demonstrated whether and when accounting for or removing variation due to population structure will affect the accuracy of genomic prediction. Our aim in this study was to determine whether accounting for population structure would increase the accuracy of genomic predictions, both within and across breeds. First, we have attempted to decompose the accuracy of genomic prediction into contributions from population structure or linkage disequilibrium (LD) between markers and QTL using a diverse multi-breed sheep (*Ovis aries*) data set, genotyped for 48,640 SNP. We demonstrate that SNP from a single chromosome can achieve up to 86% of the accuracy for genomic predictions using all SNP. This result suggests that most of the prediction accuracy is due to population structure, because a single chromosome is expected to capture relationships but is unlikely to contain all QTL. We then explored principal component analysis (PCA) as an approach to disentangle the respective contributions of population structure and LD between SNP and QTL to the accuracy of genomic predictions. Results showed that fitting an increasing number of principle components (PC; as covariates) decreased within breed accuracy until a lower plateau was reached. We speculate that this plateau is a measure of the accuracy due to LD. In conclusion, a large proportion of the accuracy for genomic predictions in our data was due to variation associated with population structure. Surprisingly, accounting for this structure generally decreased the accuracy of across breed genomic predictions.

Key words: genomic prediction, genomic selection, principal component analysis, population structure, relationship, sheep

J. Anim. Sci. 2012.90:3375–3384
doi:10.2527/jas2011-4557

INTRODUCTION

Genome-wide association studies (**GWAS**) aim to find QTL which are associated with the same marker across a population. Therefore, GWAS need to filter out signals due to population structure, whether they are due to linkage or spurious linkage disequilibrium (**LD**), to avoid false positives (e.g., Freedman et al., 2004; Marchini et al., 2004). In contrast to GWAS, the aim of genomic prediction is to predict breeding values from dense SNP data (Meuwissen et al., 2001; Goddard and Hayes, 2009).

[1]We thank Daniel Brown for providing validation ram ASBV, Andrew Swan for compiling a pedigree, Mike Goddard and John Hickey for discussions, Ken Geenty and Klint Gore for managing the phenotypic information, Cedric Gondro for performing part of the genotype quality control and the farm staff from the Sheep CRC Information Nucleus Flock and the Falkiner Memorial Field Station for their assistance in data collection and processing. We also thank the Cooperative Research Centre for Sheep Industry Innovation and the Sheep Genomics project, including Meat and Livestock Australia and Australian Wool Innovation Limited, for use of data and for funding.

[2]Corresponding author: hans.daetwyler@dpi.vic.gov.au

Received August 5, 2011.
Accepted March 29, 2012.

Population structure contributes significantly to the accuracy of genomic predictions when the selection candidates are closely related to the reference population (Habier et al., 2007, 2010). Thus the accuracy of genomic predictions could be expected to have 2 main contributors: i) prediction based on genomic relationships arising from population structure, and ii) prediction based on LD between markers and QTL. The 2 contributors are likely correlated, because closer relatedness between subgroups (i.e., more recent divergence) increases LD between subgroups and close relationships within a subgroup increases linkage. The extent to which the 2 sources contribute to prediction accuracy is currently unclear. However, the distinction is important because the accuracy due to LD is more likely to persist across generations and breeds than the accuracy due to relationships (Meuwissen et al., 2001; Habier et al., 2007; De Roos et al., 2009).

This reasoning leads us to propose the following hypothesis: if population structure is different in the reference and validation population, the accuracy of genomic breeding values will be reduced if population structure is not accounted for. This is analogous to the reasoning in GWAS, where failure to account for population structure can lead to spurious associations. However, this does not imply that all population structure is spurious, as, for example, the phase between QTL and markers may differ between subpopulations.

We attempt to decompose the accuracy of genomic prediction into the contributions from population structure and SNP in LD with QTL, in a large multi-breed sheep population. We then investigate whether accounting for population structure is beneficial in genomic prediction both within and across breeds to test this hypothesis.

METHODS

Phenotype and Genotype Data

Two phenotypic traits were investigated in sheep: yearling greasy fleece weight (mean 3.3kg; SD 0.96kg) and ultrasound scanned eye muscle depth (mean 23.5 mm; SD 4.9 mm), which have heritability estimates of 0.37 and 0.23, respectively (Safari et al., 2005; Mortimer et al., 2010). Datasets from the Cooperative Research Centre for Sheep Industry Innovation (van der Werf et al., 2010) and Falkiner Memorial Field Station (Oddy et al., 2005) were combined to increase the size of the reference population. This resulted in 3,341 and 7,431 animals with phenotypes and genotypes available for greasy fleece weight and eye muscle depth, respectively. The breed content of the reference populations of the 2 traits is shown in Figure 1. Both reference populations had a significant proportion of Merino individuals, and only this breed had a substantial proportion of purebred animals. Most other animals were crossbreds of meat breeds and Merinos. Breed group size ranged from 3,792 animals for purebred Merino sheep to 5 for a Border Leicester/East Friesian/Polled Dorset cross. A total of 196 rams sired our reference population, and the size of the resulting half-sib families ranged from 385 to 1. Thus, there was great diversity not only in breed-crosses but also in half-sib family size. Notably, the size of the ram half-sib families was often larger than the number of animals in the respective breed-cross groups.

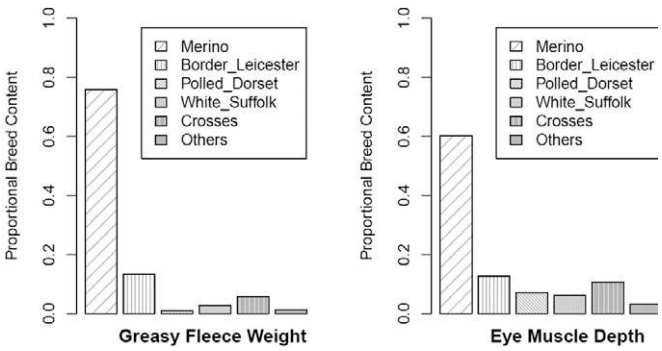

Figure 1. Proportional breed content of crossbred animals in reference populations for greasy fleece weight (n = 3341) and eye muscle depth (n = 7431). Animals in category "Crosses" have a complex multiple breed makeup

All animals were genotyped using the Illumina 50K ovine SNP chip (Illumina Inc., San Diego, CA), which reacts to 54,977 SNP. The following quality control measures were applied to the SNP data: SNP were removed if they had a call rate of <95%, an Illumina Gentrain (**GC**) score of <0.6, a minor allele frequency of <0.01, were out of Hardy-Weinberg equilibrium (a *P*-value cutoff of 1^{-15}), had no genome location or were in >0.99 r^2 with another SNP on the chip. After these measures were applied, 48,640 SNP were used. Data for genotyped animals were removed if their genotype call rate was <0.9 or if their mean heterozygosity was >0.5, which would indicate sample contamination. The genotype database was built over a period of time; early genotypes were imputed using fastPHASE (Scheet and Stephens, 2006), and more recent genotypes were imputed for missing genotypes using Beagle (Browning and Browning, 2009), when this program became available.

Genomic Prediction Methods

Genomic BLUP (**GBLUP**) was used for most analyses by fitting the following model in ASReml (Gilmour et al., 2009):

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zg} + \mathbf{e}$$

where **y** is a vector of phenotypic records, **X** is a design matrix relating the fixed effects to the animal, **b** is a vector of fixed effects, **Z** is a design matrix relating animal effects to phenotypes, **g** is a vector of additive genetic

effects and **e** is the vector of residuals. The following distributions were assumed: $\mathbf{g} \sim N(0, \sigma_g^2 \mathbf{G})$ and $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$, where **G** was a genomic relationship matrix, calculated using the method of Yang et al. (2010), describing the genomic relationships among all pairs of individuals in both the reference and validation populations. The base model included the following fixed effects: sex, birth type, rearing type, contemporary group (birth year × site × management group) and age at trait recording, and for eye muscle depth and weight (kg) at scanning.

Bayesian stochastic search variable selection (**BayesSSVS**; Verbyla et al., 2009) was used to investigate whether it would improve across breed genomic prediction when compared with GBLUP. The main difference between BayesSSVS and GBLUP is that BayesSSVS has 2 Normal distributions of SNP, 1 with very small variance (effectively 0) and another with a larger variance, whereas GBLUP assumes 1 Normal distribution for all SNP. BayesSSVS applies the following model:

$$\mathbf{y} = \mathbf{Xb} + \sum_{j=1}^{NSNP} Z_j \beta_j + \mathbf{e},$$

where z_j is an indicator variable for SNP j, β_j is the allele substitution effect associated with SNP j, and NSNP is the number of SNP. The prior distribution for the SNP effects was a mixture of 2 student-t distributions, with variances sampled from inverse scaled χ^2 distributions, one with parameters as described by Meuwissen et al. (2001), and the second with mean 0 and variance as in Meuwissen et al. (2001) divided by 100 (Verbyla et al., 2009). The indicator variable determined which distribution SNP j was sampled from and its prior was sampled from a Bernoulli distribution with $p_j = 0.1$ (Verbyla et al., 2009). Fixed effects and covariates were fitted as for GBLUP. Ten replicate Markov Chain Monte Carlo chains with 50,000 iterations (10,000 burn-in) were run for each trait. Genomic breeding values were calculated as the sum of SNP allele substitution effects for all validation animals averaged over the 10 replicates.

The accuracy of genomic predictions was evaluated in a population of purebred rams which had been widely used in the general Australian sheep population and subsequently had high accuracy Australian sheep breeding values (**ASBV**) based on progeny records. Phenotypic records used in the genomic prediction analysis were not used to calculate the ram ASBV. Genomic prediction accuracies were calculated as the Pearson correlation of genomic breeding values and ASBV within the following breeds: Merino (n = 181), Border Leicester (n = 35), Polled Dorset (n = 69), and White Suffolk (n = 67). Breed means of ASBV accuracy were all above 0.83, except for the White Suffolk and Polled Dorset industry rams which were low for greasy fleece weight (0.62 and 0.64, respectively), because wool traits are not often measured in progeny of these terminal breeds. Thus, results were not reported in greasy fleece weight for these 2 breeds. The number of validation sires was smaller in the breeds other than Merino; therefore, small changes in accuracy may have been difficult to detect because they are likely to be within the sampling variance.

Principal Component Analysis

Principal component analysis (**PCA**) was performed on **G** using the R function eigen (R Core Developent Team, 2010). This resulted in a matrix of eigenvectors (with dimensions equal to the number of animals) ordered by descending eigenvalues. Note that we refer to these eigenvectors of **G** as principal component (**PC**), where the PC1 had the largest eigenvalue. The PC were plotted and annotated with breed composition of the animals to investigate whether PCA was able to cluster the various breed combinations together. Incidence vectors were coded by membership to breed or family group. For example, for the Merino breed, this would be a vector with 1 for all purebred Merino sheep and a 0 for all other animals. These incidence vectors were then correlated with the first 200 PC with the expectation that correlations would be high at PC where a breed-cross group was differentiated from the remaining animals. The same process was repeated for all ram half-sib families.

Accounting for Population Structure and Exploring its Impact on Accuracy

We assessed the contribution of population structure to accuracy of genomic prediction by fitting chromosome specific **G** matrices. Chromosome specific **G** matrices were calculated using only the SNP on a particular chromosome, denoted $\mathbf{G_1}$, $\mathbf{G_2}$, … $\mathbf{G_{26}}$, for chromosomes 1 through 26. Each chromosome was then fit on its own instead of the genome-wide **G**. Fitting $\mathbf{G_1}$ and $\mathbf{G_{26}}$ was done with including no PC or 200 PC in the base model. Sire and dam breed was also fitted simultaneously with $\mathbf{G_1}$ or $\mathbf{G_{26}}$. Accuracies of chromosomal and genome-wide **G**s were then compared. A measure of the proportion of the total genetic variance explained by each chromosome can be attained by fitting all 26 individual chromosomal **G** matrices simultaneously. It is expected that the proportion of the total genetic variance explained is correlated to the length of the chromosome. Due to memory limitations in the ASReml software, we fit 3 chromosome-wide **G** matrices together with a **G** matrix consisting of all SNP on the remaining chromosomes, and repeated these analyses until all chromosomes had been fitted. Yang et al. (2011) demonstrated that population structure could be quantified by regressing the difference in phenotypic variance explained by individually and simultaneously

fitted chromosomal **G**s on chromosome length. The expectation was that the intercept of this regression ($\boldsymbol{b_{0d}}$) will indicate the degree of relatedness in the population. We calculated the proportion of the genetic variance explained by population structure by division of b_{0d} with the intercept from regressing the variance explained by individually fitted chromosomes on chromosome length ($\boldsymbol{b_{0i}}$). In contrast to Yang et al. (2011), our analysis explicitly includes close relatives.

Attempts to account for population structure included fitting PC, sire and dam breed effects or polygenic effects. A range of PC from 1 to 200 was fitted cumulatively as fixed covariates in GBLUP analysis of greasy fleece weight and eye muscle depth in addition to the base model. Sire and dam breed were also fitted with the base model as fixed effects to evaluate their effect on accuracy. A random polygenic effect was fitted in GBLUP with the following model:

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

where terms are as defined above and **u** was a vector of random polygenic animal effects. Over-fitting was avoided by subtracting pedigree relationships from **G** using the following approach: the numerator relationship matrix (**A**) was calculated using a pedigree including all animals in **G** and their ancestors. Pair-wise pedigree relationships for animals in **G** were then extracted from **A** to form a new matrix of expected relationships among individuals that were in the **G** matrix, $\mathbf{A_{22}}$. Off-diagonals of $\mathbf{A_{22}}$ were converted to the **G** scale by subtracting the difference in mean off-diagonals of $\mathbf{A_{22}}$ and **G** from the off-diagonals of $\mathbf{A_{22}}$. Diagonals were scaled similarly. Over-fitting was avoided by subtracting $\mathbf{A_{22}}$ from **G**, and the two random terms were fitted in GBLUP for each animal with distributions $\mathbf{g} \sim N(0, \sigma_g^2 \mathbf{G\text{-}A_{22}})$ and $\mathbf{u} \sim N(0, \sigma_a^2 \mathbf{A_{22}})$, respectively.

The accuracy achieved when predicting across breeds is a measure of accuracy due to LD between markers and QTL, because the impact of population structure is expected to be minimal due to very low relatedness between breeds. We investigated accuracy of across-breed prediction in 2 ways. First, a Merino-only subset of the reference population was used to predict genomic breeding values of Border Leicester, Polled Dorset, and White Suffolk validation rams for each trait. Second, animals with Border Leicester, Polled Dorset, and White Suffolk ancestry were excluded, one at a time, from the multibreed reference population, and the accuracy of genomic breeding values for the excluded breed were assessed. The LD important to across-breed genomic prediction is expected to be similar to the LD relevant to a GWAS. Hence, accounting for population structure with PC may remove spurious associations and, in turn, increase accuracy of prediction across breed. This was evaluated by fitting a varied number of PC (0, 10, 25, 50, 100, 150, and 200 PC) in GBLUP and BayesSSVS analysis.

RESULTS AND DISCUSSION

We investigated the impact of population structure on the accuracy of genomic prediction in a diverse sample of the Australian sheep population. We first tested the hypothesis that the accuracy of genomic prediction was derived only from SNP in LD with QTL. Most results in real sheep data point toward models of quantitative trait genetic architecture with a very large number of QTL with small effects (e.g., Daetwyler et al., 2010; Kemper et al., 2010). If the only contribution of the SNP markers to the accuracy of genomic prediction was through LD with QTL, then a genomic relationship matrix constructed from only the SNP on 1 chromosome should capture genetic variation in proportion to its length, assuming there is no population stratification. The genomic matrix from a single chromosome captured up to 86% of the total accuracy achieved with the full genomic relationship matrix (Table 1), compared with the expectation of 2 and 11% for chromosomes 26 and 1, respectively. It was therefore clear that a large proportion of the accu-

Table 1. Accuracy of genomic prediction using genomic best linear unbiased prediction (GBLUP) in greasy fleece weight and eye muscle depth when using a genomic relationship matrix calculated genome-wide [All chromosomes (Ch)] or only with SNP on chromosomes 1 (Ch 1) or 26 (Ch 26). Analyses fitted no principal components (PC), 200 PC, breeds or pedigree

	Accuracy					
Breed and model	Greasy fleece weight			Eye muscle depth		
	All Ch	Ch 1	Ch 26	All Ch	Ch 1	Ch 26
Merino						
No PC	0.72	0.62	0.49	0.56	0.46	0.41
Breed	0.72	0.64	0.52	0.53	0.43	0.29
200 PC	0.15	0.08	0.09	0.06	0.00	0.11
Pedigree	0.72	–	–	0.39	–	–
Border Leicester						
No PC	0.43	0.24	0.24	0.31	0.15	0.35
Breed	0.44	0.26	0.29	0.27	0.15	0.34
200 PC	0.21	0.03	0.17	0.08	0.17	0.08
Pedigree	0.46	–	–	0.37	–	–
Polled Dorset						
No PC	–	–	–	0.48	0.41	0.14
Breed	–	–	–	0.46	0.39	0.08
200 PC	–	–	–	0.18	0.13	0.19
Pedigree	–	–	–	0.45	–	–
White Suffolk						
No PC	–	–	–	0.39	0.48	0.20
Breed	–	–	–	0.30	0.42	0.21
200 PC	–	–	–	0.09	0.40	0.08
Pedigree	–	–	–	0.36	–	–

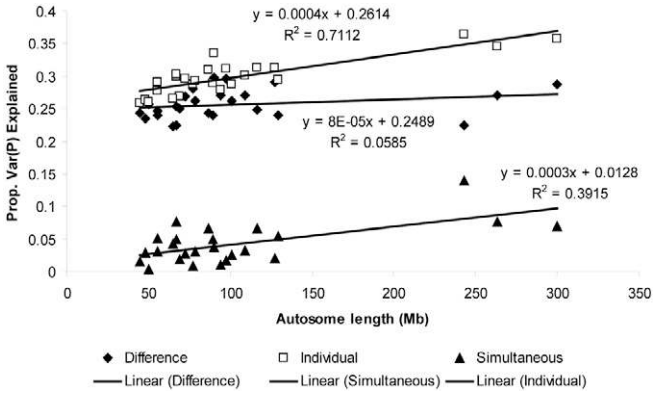

Figure 2. Proportion (Prop.) of phenotypic variance [Var(P)] explained per chromosome. Chromosomes fitted individually (top regression) or simultaneously (bottom regression). Middle regression results from plotting the difference between top and bottom regressions.

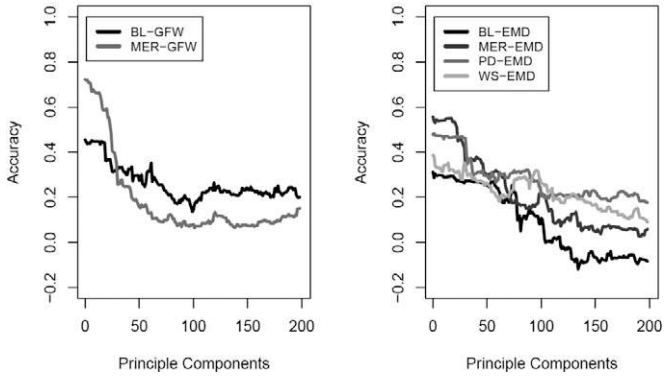

Figure 3. Accuracy of genomic prediction in greasy fleece weight (GFW) and eye muscle depth (EMD) when an increasing number of principle components are fitted in addition to the base model, where BL is Border Leicester, MER is Merino, PD is Polled Dorset, and WS is White Suffolk.

racy of genomic prediction in sheep is due to population structure at the current SNP density. In other words, only a small proportion of the accuracy was due to LD between markers and QTL alone. The trend for reduced accuracy when considering chromosome 26 versus chromosome 1 is likely because there are many fewer SNP on chromosome 26. We also tested the hypothesis that fitting of all chromosome-wide matrices simultaneously would result in each chromosome explaining a fraction of the total genetic variance proportional to its length. This expectation was generally confirmed in both traits (Figure 2). However, the regression lines had only a moderate R^2 of 0.39 in greasy fleece weight and 0.32 in eye muscle depth and the slopes were slightly flatter at 0.0003 than their expectation of 0.0004. The slightly lower slope may indicate that the accuracy due to population structure was evenly distributed among the chromosomes. We also applied the approach by Yang et al. (2011) and found that 95 and 89% of the genetic variance explained was due to relatedness in greasy fleece weight and eye muscle depth, respectively (Figure 2, $b_{0d}/b_{0i} = 0.249/0.261 = 0.95$).

In human populations, fitting PC of the genomic relationship matrix among individuals derived from SNP data has become a widely accepted way to account for population structure in GWAS (Patterson et al., 2006; Price et al., 2006). In our data we fitted up to 200 PC with genome or chromosome-wide relationship matrices in an attempt to correct for population structure in the genomic predictions (Figure 3). However, it was unclear in this population at which point the different aspects of population structure (i.e., breed or family groups) would contribute to the PC. Plots of PC annotated by breed-cross type showed a trend for purebreds to cluster in the respective breed groups (e.g., Merino, Border Leicester; Figure 4). However, crossbreds failed to form clear clusters. Labeling breed crosses by first sire breed showed that they do form broad clusters which nevertheless were smeared between the locations of the purebreds in the cross, likely a reflection of variable proportions of the genome derived from each purebred (results not shown).

We coded incidence vectors indicating membership of a breed or family, and correlated them with the PC. The incidence vectors showed the greatest correlation for Merino sheep (or not) at **PC1**, which accounted for 3.5% of variation in the genomic relationship matrix. This indicated that the breed was distinguished from the remainder of the animals by this PC (Figure 5). The same trend is seen in Figure 4, where Merino sheep form a clear cluster separated from the other groups. As breed groups became smaller, the PC number at which they were differentiated increased. For sire families, the incidence vector correlation also accurately pinpointed which PC were most important for a particular family (Figure 5). The largest 2 families had the greatest correlation at **PC3** (1.1% of variation), which coincided with them being differentiated at PC3 in Figure 6. Again, the smaller the half-sib family, the greater the PC at which they were distinguished, and for families, this group size effect was stronger and more reliable than for the breed-groups. This illustrates that the PC at which a group is differentiated is heavily influenced by its size, thus making it difficult to use PCA to only correct for breed effects and leave structure due to families intact. It also confirms analytical and simulated results by McVean (2009), which show PC to be sensitive to group size. Considering the results in this study, it seems that the general practice of fitting only the first few PC would be inadequate to account for population structure due to breed in diverse datasets.

It has been noted that PCA may also not sufficiently account for family structures (Price et al., 2010). An approach which corrects for both stratification and family structure based on *P*-value adjustments has been proposed (Won et al., 2009). In genomic prediction, it may be more useful to be able to account only for spurious population structure (e.g., structure from admixture) rather than adjust for family structures, as doing so

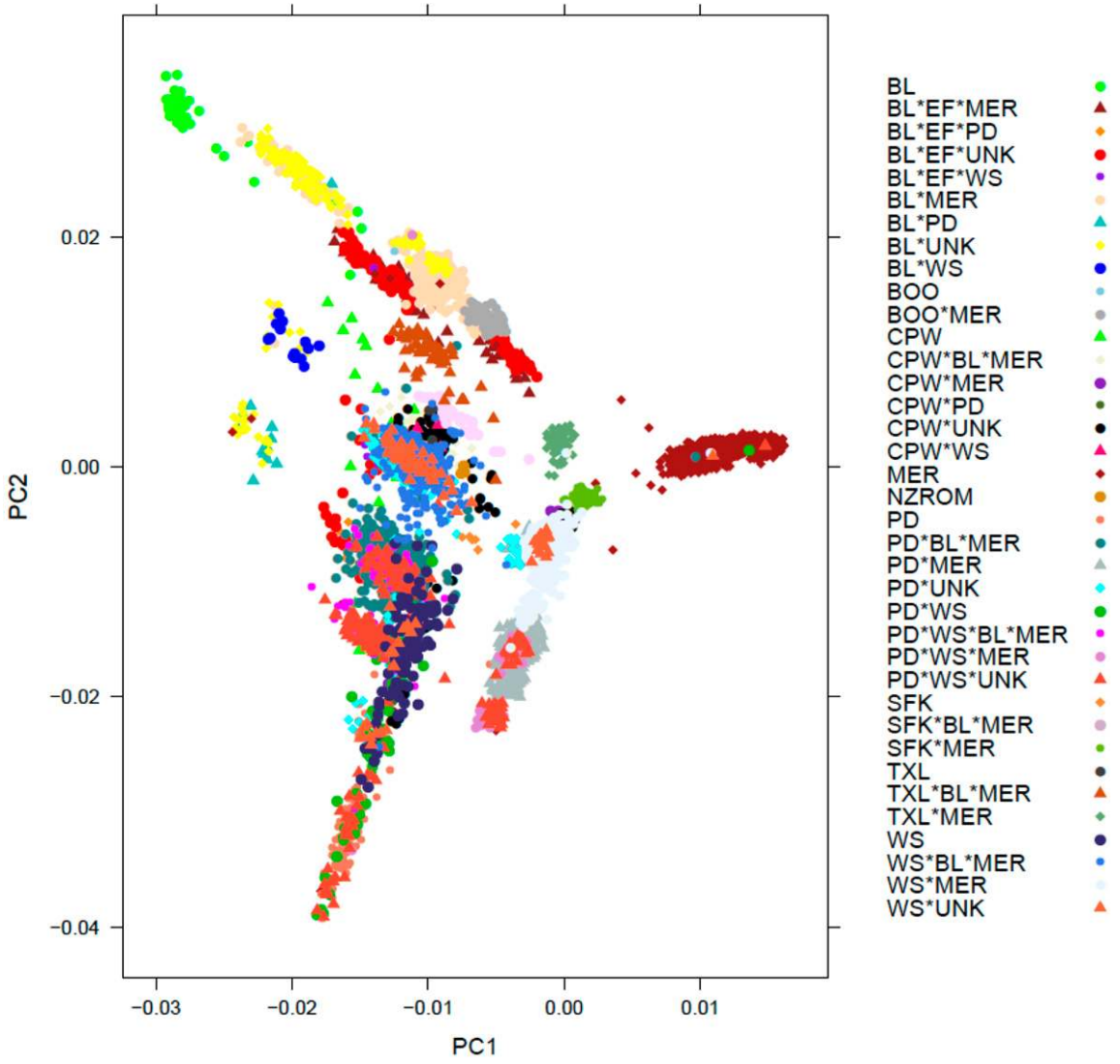

Figure 4. Plot of principal components (PC) 1 and 2 for all genotyped animals in reference and validation sets. Colored by breed combination, where BL is Border Leicester, EF is East Friesian, MER is Merino, PD is Polled Dorset, WS is White Suffolk, UNK is unknown, BOO is Booroola, CPW is Coopworth, NZROM is New Zealand Romney, SFK is Suffolk, and TXL is Texel. See online version for figure in color.

would remove predictive effects within a breed. However accounting for spurious population structure without accounting for relatedness is not trivial in this case.

Accounting for population structure could increase accuracy if the population structure in the reference differs from the validation population or by removing spurious associations. When the phenotypes were corrected for the PC, for both greasy fleece weight and eye muscle depth, the accuracy of genomic prediction clearly declined as an increasing number of PC was fitted (Figure 3). Most breeds eventually reached a lower plateau of accuracy. For greasy fleece weight, the Merino group reached the lower plateau at approximately PC 50, whereas Border Leicester sheep reached this plateau at approximately PC 80. We speculate that these lower plateaus correspond to the accuracy due to LD of markers and QTL as the majority of the effect of population structure is accounted for. The relatively greater accuracy observed in Border Leicester than Merino sheep may correspond to a much lower effective population size (N_e) in Border Leicester (J. Kijas, CSIRO, Brisbane, Australia, personal communication). This would lead to relatively greater LD in this breed than Merino and, in turn, this may have translated into a greater accuracy due to LD.

For eye muscle depth, a similar trend of decaying accuracy as more PC are fitted can be seen in Figure 3 and for this trait the various breeds were more equally represented in the reference population. All 4 validation breeds reached lower plateaus between PC110 and 130. Although initially Merino sheep had the greatest accuracy at low PC, the Polled Dorset and White Suffolk breeds had greater accuracies once the lower plateau was reached. This may again be due to the lower N_e of these terminal breeds when compared with Merino sheep. However, this trend did not seem to hold true for Border Leicester sheep in eye muscle depth. Note that the accuracy plateaus cannot continue indefinitely, as eventually the PC will remove variation due to LD.

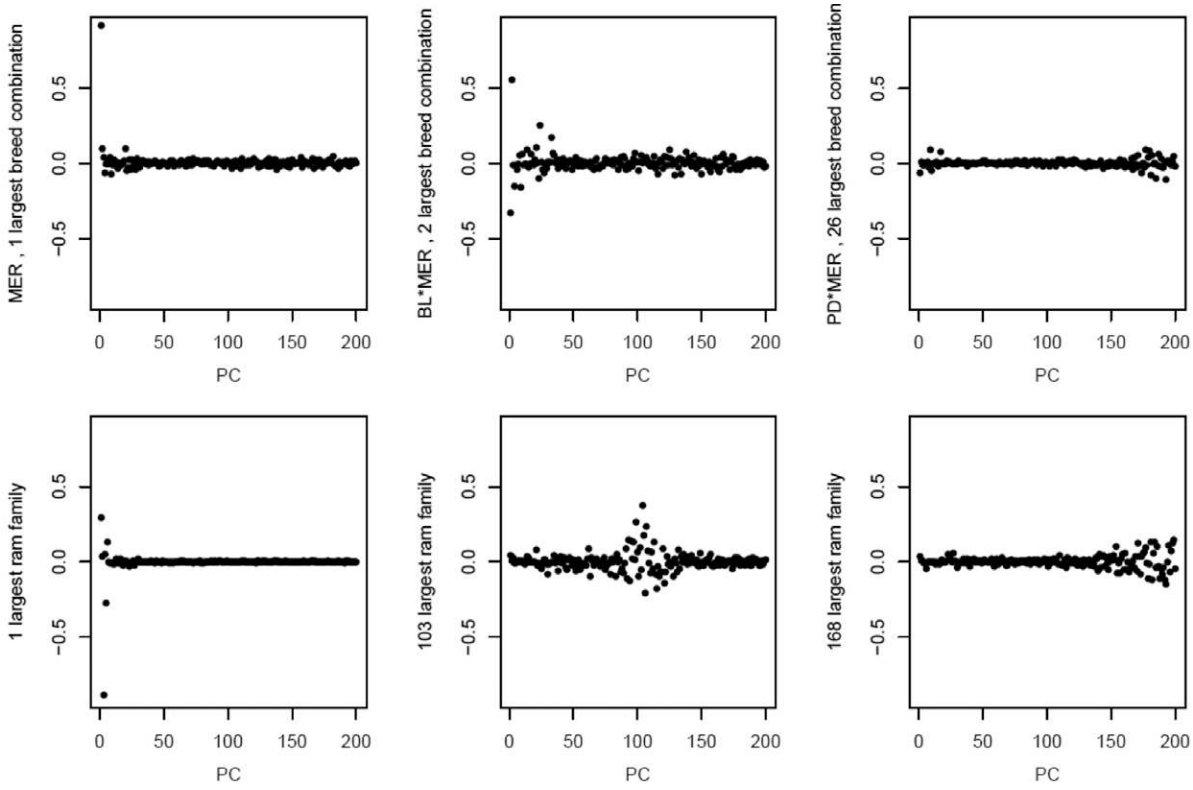

Figure 5. Correlations of principal components (PC) with incidence vectors for 3 breed composite groups (top) and 3 ram half-sib families (bottom). BL is Border Leicester, MER is Merino, and PD is Polled Dorset.

An alternative approach to PCA to control for population structure is to capture the structure through a pedigree-derived relationship matrix, as well as breed effects (e.g., Daetwyler et al., 2008; MacLeod et al., 2010; Liu et al., 2011). In our data, fitting a random polygenic effect through a pedigree relationship matrix within the GBLUP model had only a limited impact on the accuracy of genomic prediction (Table 1). In most breeds, there was no notable difference to the base model. There were 2 exceptions in eye muscle depth, where the accuracy was reduced in Merino sheep and increased in the Border Leicester breeds. However, no consistent change in accuracy was observed by fitting a polygenic effect. These findings could in part be due to an incomplete pedigree in our sheep dataset. Indeed, the proportion of the phenotypic variance explained by the polygenic effect versus the genomic relationship matrix, which was corrected for the polygenic effect, was low (Table 2). The impact of a polygenic effect should be reevaluated in data where more complete pedigree records are available. Fitting sire and dam breed, which has been used as a way to account for population structure in the literature, had a small effect on accuracy (Table 1).

The final measure used to assess the accuracy due to LD between QTL and SNP, which is perhaps more conservative than the previous approaches, is the accuracy achieved when predicting breeding values across breeds (Table 3). This was investigated in 2 ways. First, Border Leicester, Polled Dorset, and White Suffolk rams were predicted from a reference of pure Merino sheep. Second, these 3 breeds were predicted from references excluding the breed to be predicted (e.g., predicting Polled Dorset sheep from reference without Polled Dorset). In addition to the base model, we analyzed the data with BayesSSVS. This method had been shown to increase across breed prediction (Hayes et al., 2009), possibly because it assigns SNP to either a distribution with very small variance (near 0) or 1 with a larger variance in the prediction model, unlike GBLUP which assumes that all SNP have the same variance. The accuracy achieved for across breed prediction was always less than when the breed to be predicted was included in the reference population (compare Tables 1 and 3). For greasy fleece weight, Border Leicester sheep had an increased accuracy when a small number of PC were included as covariates (Figure 7, Table 3). This increase in accuracy, compared with fitting no PC, was greater in both traits when predicted with BayesSSVS than with GBLUP. For eye muscle depth, a different trend was observed (Figure 7, Table 3). The accuracy for Polled Dorset and White Suffolk sheep was reduced by fitting an increasing number of PC and the BayesSSVS method always had a slightly less accuracy than GBLUP.

A possible explanation for the large disparity between accuracy due to population structure and accuracy due to LD is that the sheep SNP chip is only of medium density, at approximately 50,000 SNP, relative to N_e in our population. This would limit LD between SNP and

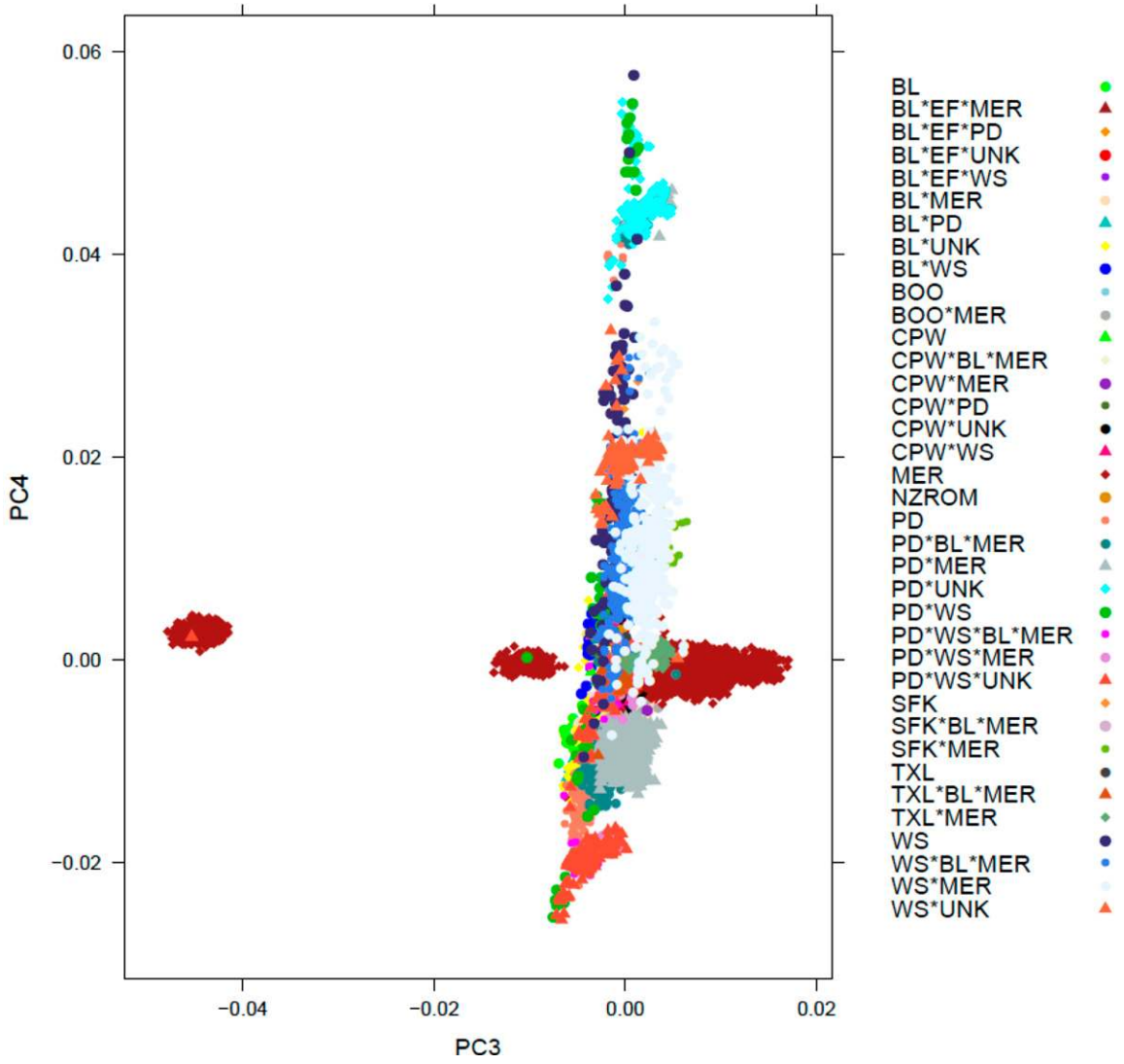

Figure 6. Plot of Principle Components (PC) 3 and 4 for all genotyped animals. Colored by breed combination, where BL is Border Leicester, EF is East Friesian, MER is Merino, PD is Polled Dorset, WS is White Suffolk, UNK is unknown, BOO is Booroola, CPW is Coopworth, NZROM is New Zealand Romney, SFK is Suffolk, and TXL is Texel. See online version for figure in color.

QTL thereby limiting the accuracy of this component. This effect would be especially pronounced in Merino sheep which have a N_e that is larger than the other 3 breeds. Indications of this can be seen in the lower accuracy plateaus observed in Merino sheep (Figure 2). In the longer-term increasing the number of SNP used in the analysis or even moving to full sequence would be expected to increase the accuracy due to LD, leading to greater capability to make genomic predictions across breeds. Meuwissen (2009) proposed that the number of SNP needed to predict unrelated individuals is equal to $10N_eL$, where $\boldsymbol{L}$ is the length of the genome in Morgans. The N_e of Merino sheep is approximately 800 (J. Kijas, CSIRO, Brisbane, Australia, personal communication) and L is approximately 27 Morgans and across breed prediction in sheep would require at least 216,000 SNP. It is clear that 50,000 SNP in sheep is not sufficient to allow for accurate across breed prediction.

Table 2. Proportion of the phenotypic variance explained (i.e., heritability) by matrices **G**, **G-A$_{22}$**, and **A$_{22}$** from GBLUP analysis. Matrix **G** was fitted by itself but matrices **G-A$_{22}$**, and **A$_{22}$** were fitted jointly

	Proportion of Var(P)[1]	
Matrix	GFW	EMD
G	0.65	0.34
G-A$_{22}$	0.36	0.21
A$_{22}$	0.11	0.15

[1]GFW = greasy fleece weight; EMD = eye muscle depth.

In GWAS, accounting for population structure is justified because the aim is to isolate causal variants for traits. Here we have evaluated the value of population structure adjustments when the aim is to predict genomic estimated breeding values. The main attraction of combining potentially diverse populations, such as our multibreed sheep population, is to increase the reference population and attempt to take advantage of prediction

Table 3. Summary of the best accuracies for across breed genomic predictions using genomic best linear unbiased prediction (GBLUP) and Bayesian stochastic search variable selection BayesSSVS (SSVS) for greasy fleece weight and eye muscle depth in Border Leicester (BL), Polled Dorset (PD) and White Suffolk (WS) breeds. Values presented are the accuracy achieved with either the Merino only reference set or with the reference set which excluded the respective target breed, whichever was greater. Max. refers to maximum accuracy achieved at any of 0, 10, 25, 50, 100, 150, and 200 PC. All values are shown in Figure 7[1]

	Accuracy							
	Greasy fleece weight				Eye muscle depth			
	No PC	200 PC	Maximum across breed		No PC	200 PC	Maximum across breed	
Breed	GBLUP	GBLUP	GBLUP	SSVS	GBLUP	GBLUP	GBLUP	SSVS
BL	0.12	0.05	0.24	0.44	0.08	−0.04	0.08	0.21
PD	–	–	–	–	0.33	0.14	0.33	0.20
WS	–	–	–	–	0.26	0.04	0.26	0.18

[1]PC = principal components.

across subgroups. We have shown that the impact of population structure (i.e., relatedness) on the accuracy of genomic prediction is very large at the current SNP density and that most of the prediction accuracy is because the reference population is highly related to the predicted animals. Accounting for population structure with PCA or other methods only decreased accuracy for most breeds in scenarios where across-breed prediction was evaluated. Hence, we conclude that there is no benefit to accounting for population structure in genomic prediction at the current SNP densities in sheep. Increasing the number of SNP is expected to shift the influence of population structure on accuracy toward accuracy due to LD between SNP and QTL.

Population structure is an important component of genetic variation. In a sample of individuals, population structure may be encountered because some individuals are more related than others or because a sample may contain animals from different subgroups (e.g., breeds, ethnicities, admixture). Genomic prediction across closely related individuals is based on linkage whereas predicting distantly related individuals requires LD between QTL and markers. Linkage is a form of LD, but associations only occur within groups of closely related individuals, and the LD may stretch over relatively long segments of the genome. In contrast, LD is consistency of phase between SNP and QTL, at least within a whole population, but possibly even across populations. The more distant the relationship between individuals, the shorter the genomic distance over which phase will be consistent. The time scales which reduce linkage to LD are very long, involving many generations. An example is the very short stretches of the genome that different breeds may share, resulting from common ancestors predating breed divergence. Rather than concentrate on LD between SNP and QTL alone, genomic prediction can use both linkage and LD. It seems pragmatic to not restrict either component and maximize the accuracy of genomic prediction and realize that, at current

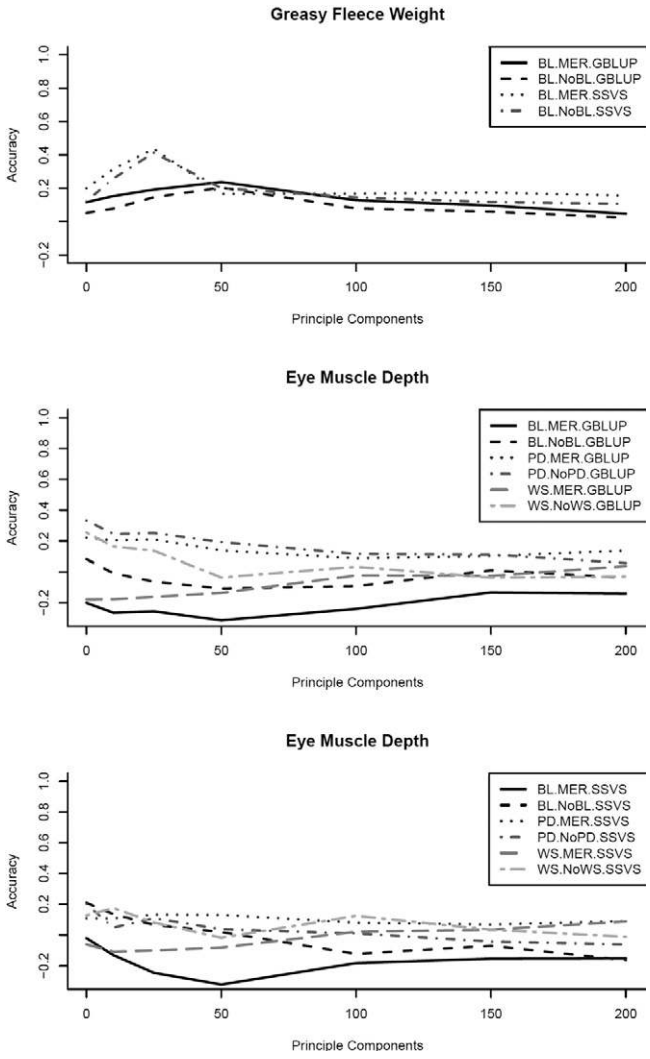

Figure 7. Accuracy when predicting across breeds in greasy fleece weight (top) and eye muscle depth (middle, bottom) with a Merino-only reference population (MER) or reference populations which exclude the breed to be predicted (NoXX) with methods genomic best linear unbiased prediction (GBLUP) and Bayesian stochastic search variable selection BayesSSVS (SSVS) , where XX is either BL for Border Leicester, PD for Polled Dorset, or WS for White Suffolk.

SNP densities, individuals which are more related to the reference population will be predicted better than more distantly related individuals.

LITERATURE CITED

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

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

Daetwyler, H. D., F. S. Schenkel, M. Sargolzaei, and J. A. B. Robinson. 2008. A genome scan to detect quantitative trait loci for economically important traits in Holstein cattle using two methods and a dense single nucleotide polymorphism map. J. Dairy Sci. 91:3225−3236.

De Roos, A. P. W., B. J. Hayes, and M. E. Goddard. 2009. Reliability of genomic breeding values across multiple populations. Genetics 183:1545−1553.

Freedman, M. L., D. Reich, K. L. Penney, G. J. McDonald, A. A. Mignault, N. Patterson, S. B. Gabriel, E. J. Topol, J. W. Smoller, C. N. Pato, M. T. Pato, T. L. Petryshen, L. N. Kolonel, E. S. Lander, P. Sklar, B. Henderson, J. N. Hirschhorn, and D. Altshuler. 2004. Assessing the impact of population stratification on genetic association studies. Nat. Genet. 36:388−393.

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

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

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

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

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, K. L. Verbyla, and M. E. Goddard. 2009. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet. Sel. Evol. 41:51−59.

Kemper, K. E., D. L. Emery, S. C. Bishop, H. Oddy, B. J. Hayes, S. Dominik, J. M. Henshall, and M. E. Goddard. 2010. The distribution of SNP marker effects for faecal worm egg count in sheep, and the feasibility of using these markers to predict genetic merit for resistance to worm infections. Genet. Res. 93:203−219.

Liu, Z., F. Seefried, F. Reinhardt, S. Rensing, G. Thaller, and R. Reents. 2011. Impacts of both reference population size and inclusion of a residual polygenic effect on the accuracy of genomic prediction. Genet. Select. Evol. 43:19.

MacLeod, I. M., B. J. Hayes, K. W. Savin, A. J. Chamberlain, H. C. McPartlan, and M. E. Goddard. 2010. Power of a genome scan to detect and locate quantitative trait loci in cattle using dense single nucleotide polymorphisms. J. Anim. Breed. Genet. 127:133−142.

Marchini, J., L. R. Cardon, M. S. Phillips, and P. Donnelly. 2004. The effects of human population structure on large genetic association studies. Nat. Genet. 36:512−517.

McVean, G. 2009. A genealogical interpretation of principal components analysis. PLoS Genet. 5:e1000686.

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

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

Mortimer, S. I., J. H. J. van der Werf, R. H. Jacob, D. W. Pethick, K. L. Pearce, R. D. Warner, G. H. Geesink, J. E. Hocking Edwards, G. E. Gardner, E. N. Ponnampalam, S. M. Kitessa, A. J. Ball, and D. L. Hopkins. 2010. Preliminary estimates of genetic parameters for carcass and meat quality traits in Australian sheep. Anim. Prod. Sci. 50:1135−1144.

Oddy, V. H., T. J. Longhurst, F. W. Nicholas, J. F. Maddox, and M. B. McDonagh. 2005. Sheep Genomics—Core Technologies and Resources. Assoc. of the Advancement of Animal Breeding and Genetics, Christ Church, New Zealand.

Patterson, N., A. L. Price, and D. Reich. 2006. Population structure and eigenanalysis. PLoS Genet. 2:e190.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38:904−909.

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

R Core Developent Team. 2010. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Safari, E., N. M. Fogarty, and A. R. Gilmour. 2005. A review of genetic parameter estimates for wool, growth, meat and reproduction traits in sheep. Livst. Prod. Sci. 92:271−289.

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

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

Verbyla, K. L., B. J. Hayes, P. J. Bowman, and M. E. Goddard. 2009. Technical note: Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle. Genet. Res. 91:307−311.

Won, S., J. B. Wilk, R. A. Mathias, C. J. O'Donnell, E. K. Silverman, K. Barnes, G. T. O'Connor, S. T. Weiss, and C. Lange. 2009. On the analysis of genome-wide association studies in family-based designs: A universal, robust analysis approach and an application to four genome-wide association studies. PLoS Genet. 5:e1000741.

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

Yang, J., T. A. Manolio, L. R. Pasquale, E. Boerwinkle, N. Caporaso, J. M. Cunningham, M. de Andrade, B. Feenstra, E. Feingold, M. G. Hayes, W. G. Hill, M. T. Landi, A. Alonso, G. Lettre, P. Lin, H. Ling, W. Lowe, R. A. Mathias, M. Melbye, E. Pugh, M. Cornelis, B. S. Weir, M. E. Goddard, and P. M. Visscher. 2011. Genome partitioning of genetic variation for complex traits using common SNPs. Nat. Genet. 43:519−525.