Accuracy of genomic prediction within and across populations for nematode resistance and body weight traits in sheep

Citation for published version:

Digital Object Identifier (DOI):
10.1017/S1751731114000081

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Animal

Publisher Rights Statement:
This is a PDF file of an unedited manuscript that has been accepted for publication. The publisher version is available at: http://dx.doi.org/10.1017/S1751731114000081

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Accuracy of genomic prediction within and across populations for nematode resistance and body weight traits in sheep

V. Riggio 1,a, M. Abdel-Aziz 2,a, O. Matika 1, C.R. Moreno 3, A. Carta 4, and S.C. Bishop 1

1 The Roslin Institute and R(D)SVS, University of Edinburgh, Easter Bush, Midlothian EH25 9RG, Scotland, UK
2 Department of Animal and Fish Production, College of Agriculture and Food Sciences, King Faisal University, Al-Ahsa, 31982, Saudi Arabia
3 INRA, UR631, Station d’Amélioration Génétique des Animaux, BP 27, F-31326, Castanet-Tolosan, France
4 Settore Genetica e Biotecnologie, AGRIS Sardegna, Olmedo, Sassari 07040, Italy

a Equal contributors

Corresponding author: Valentina Riggio. Email: valentina.riggio@roslin.ed.ac.uk

Short title: Genomic predictions for sheep nematodes and weight

Abstract
Genomic prediction utilizes SNP chip data to predict animal genetic merit. It has the advantage of potentially capturing the effects of the majority of loci that contribute to genetic variation in a trait, even when the effects of the individual loci are very small. To implement genomic prediction, marker effects are estimated with a training set
including individuals with marker genotypes and trait phenotypes; subsequently
genomic estimated breeding values (GEBV) for any genotyped individual in the
population can be calculated using the estimated marker effects. In this study we
aimed to: i) evaluate the potential of genomic prediction to predict GEBV for
nematode resistance traits and body weight in sheep, within and across populations;
ii) evaluate the accuracy of these predictions through within-population cross-
validation; and iii) explore the impact of population structure on the accuracy of
prediction. Four datasets comprising 752 lambs from a Scottish Blackface population,
2,371 from a Sarda x Lacaune backcross population, 1,000 from a Martinik Black-
Belly x Romane backcross population, and 64 from a British Texel population were
used in this study. Traits available for the analysis were faecal egg count for
Nematodirus and Strongyles and body weight at different ages or as average effect,
depending on the population. Moreover, immunoglobulin A was also available for the
Scottish Blackface population. Results show that GEBV had moderate to good
within-population predictive accuracy, whereas across-population predictions had
accuracies close to zero. This can be explained by our finding that in most cases the
accuracy estimates were mostly due to additive genetic relatedness between
animals, rather than linkage disequilibrium (LD) between SNP and QTL. Our results,
therefore, suggest that genomic prediction for nematode resistance and body weight
may be of value in closely related animals, but that with the current SNP chip
genomic predictions are unlikely to work across breeds.

**Keywords:** genomic prediction, population structure, nematode resistance, body
weight, sheep
Implications

Genomic prediction utilizes SNP chip data to predict animal genetic merit. Using data from several populations, our results suggest that genomic prediction may be of value for nematode resistance and body weight in closely related animals, but with current technologies it is unlikely to work across populations. Genetic relatedness between animals and population structure affect these estimates and need to be taken into consideration before considering implementation.

Introduction

Traditional genetic improvement has relied on the use of phenotypes together with the knowledge of the pedigree of each animal to estimate its breeding value. This has led to genetic gains in most farmed species; especially with ‘easy-to-measure’ production traits. However, the efficiency decreases when traits are difficult to measure, have a low heritability, or cannot be quickly, inexpensively and correctly measured. An example is nematode resistance, assessed using indicator traits such as faecal egg count (FEC), which is critically important for the sheep industry.

To overcome this issue, there has long been an interest in using simply inherited genetic markers to increase the rate of genetic gain (Dekkers and Hospital, 2002). However, for many quantitative traits, such as production and health traits, a large number of loci appear to affect the trait, with each of them individually explaining only a limited proportion of the total genetic variance (Hayes and Goddard, 2001, Sanna et al., 2008, Kemper et al., 2011). Genomic selection (GS) has the advantage of potentially capturing the effects of the majority of loci that contribute to genetic variation, even when the effects of the individual loci are very small (Hayes et al., 2009a). With GS, first marker effects are estimated with a training set (TS) which
includes individuals with marker genotypes and trait phenotypes; genomic estimated
breeding values (GEBV) of any genotyped individual in the population can then be
calculated using the estimated marker effects (Habier et al., 2007). The resulting
GEBV, therefore, exploit associations between markers and QTL through linkage
disequilibrium (LD) and linkage, along with the capture of pedigree relationships
between animals (Habier et al., 2007).

Accessing sufficient animals to both train and validate GEBV remains challenging in
practice, and cross-validation with individuals from the same population is often used
to assess the accuracy of the GEBV (Habier et al., 2007). However, validation
studies can be also performed using separate phenotyped and genotyped
populations (Hayes et al., 2009a, Luan et al., 2009, Su et al., 2010), with an accuracy
which depends on the genetic relationship of the validation set to the TS (Habier et
al., 2007, Habier et al., 2010). This is possible because markers used in the
statistical models to estimate marker effects also capture additive genetic
relationships between individuals (Cockerham, 1969, Ritland, 1996), therefore, even
if markers are not in LD with QTL, the accuracy of GEBV will still be non-zero.
However, animals more closely related to those included in the TS are expected to
obtain more reliable predictions (Habier et al., 2007, Legarra et al., 2008, Sonesson
and Meuwissen, 2009).

At present, the accuracy of GEBV has been evaluated in experiments involving
several livestock species, such as dairy (Harris et al., 2008, Hayes et al., 2009b) and
beef (Saatchi et al., 2011) cattle populations, chicken (González-Recio et al., 2009),
and sheep (Daetwyler et al., 2010b, Daetwyler et al., 2012a, Daetwyler et al., 2012b,
Duchemin et al., 2012). Apart from the study of Kemper et al. (2011), the use of high
density genomic information to select for nematode resistance in sheep has received
less attention. Therefore, the aims of this study were to: i) evaluate the potential of GS to predict GEBV for nematode resistance traits, as well as body weight, both within and across populations; ii) evaluate the accuracy of these predictions through within-population cross-validation; and iii) explore the impact of population structure within population, by decomposing the accuracy of genomic prediction into component parts.

Material and methods

Four datasets comprising 752 lambs from a Scottish Blackface (SBF) population, 2,371 ewes from a Sarda x Lacaune (SAR) backcross population, 1,000 lambs from a Martinik Black-Belly x Romane (MBR) backcross population, and 64 lambs from a British Texel (BT) population were used in this study. As shown in the principal components plot of the SNP chip markers reported in Supplementary Figure S1, the four populations are genetically distant. Genomic predictions were conducted firstly within population, using the SBF data. This was because of the availability of both pedigree and SNP marker data, along with several traits, allowing us to potentially explore a variety of trait architectures as well as contributions of LD and linkage to genomic predictions. Secondly, an evaluation of across-population prediction was conducted using all four populations, albeit with limited phenotypes common across datasets.

Phenotype data

SBF data: The SBF lambs were bred over a period of three years (2001-2003), with traits measured including lamb weights (16 and 24 weeks, and average animal effect from a repeatability model excluding pedigree) and faecal egg counts (FEC) for Nematodirus and Strongyles collected at 16, 20 and 24 weeks of age, and their
average animal effects as well as plasma IgA (on 737 out of the 752 lambs). The population comprised F2 and double backcross lambs from two originally different lines, bred from 10 sires (half-sib family size = 11-146). More details on the data structure and the phenotypes are given in Riggio et al. (2013). Fecal samples were collected from the rectum of each lamb at the time of weighing and used for FEC assays, using the modified McMaster technique as described by Gordon and Whitlock (1939) and Bairden (1991). The activity of plasma IgA against a somatic extract of third-stage larvae from Teladorsagia was measured by indirect ELISA, as described by Strain et al. (2002), using blood samples collected at 24 weeks of age. The relative IgA activity was calculated according to the formula suggested by Sinski et al. (1995). The average animal effects were estimated by fitting a repeatability model to trait values across the different time points, and then standardized to a mean of 0 and a standard deviation of 1. FEC and IgA measurements were all right-skewed. Therefore, prior to analysis, FEC measurements were log-transformed by $\ln(\text{FEC}+x)$, where $x$ is a constant used to avoid the zero values, whereas IgA measurements were cube-root transformed.

Other populations: Phenotypes available on BT lambs were for FEC at 20 weeks for Strongyles and Nematodirus, and body weight at 24 weeks. A detailed description of the data was given in Matika et al. (2011). The phenotype available for the two remaining populations (SAR and MBR) was the “average animal effect” for Strongyles FEC. A detail description of the animals in the MBR population was given in Sallé et al. (2012), and for the SAR population in Sechi et al. (2009).

Genotype data

All animals from the four populations were genotyped using the 50k SNP chip. The SNP genotypes data were subjected to quality control (QC) measures, specific for
each population (see Supplementary Material S1). After QC, 42,841 SNPs were available for the SBF and BT populations, 44,859 for the SAR, and 42,469 for the MBR. Out of these SNPs, 38,991 were in common among the four populations and therefore used for further analyses.

Assessment of GEBV predictive value

SBF data: For the analysis within population, validation sets were obtained by masking the phenotype (i.e., setting the phenotype as “unknown”) for a defined number of individuals from the TS. The individuals whose phenotype was masked were selected in two different ways. The first way was through random selection: five non-overlapping cross-validation sets were created by randomly selecting 150 (152 for the fifth subset) lambs at a time, masking each phenotype only once. The second way was to select individuals belonging to specific families, to test the extent to which results differed depending on how related families were to the remaining families forming the TS.

Data were first analysed without fitting any polygenic or genomic effect, to correct for fixed effects. The following model was fitted:

\[ y_{ijlmn} = \mu + S_i + K_j + L_l + G_m + A_n + \beta DB + e_{ijlmn} \]

where, \( y_{ijlmn} \) is the phenotype of the \( n^{th} \) individual, \( S_i \) is the effect of the sex (male and female), \( K_j \) is the effect of the year of birth (2001 to 2003), \( L_l \) is the effect of the litter size (single or multiple), \( G_m \) is the effect of management group (two levels, corresponding to those born in the first 2 weeks of the lambing season and those born subsequently), \( A_n \) is the effect of age of dam (1 to 4 years), \( DB \) is a covariate effect of day of birth and \( \beta \) its regression coefficient, and \( e_{ijlmn} \) is the residual error.
The resulting adjusted phenotypes or residuals \((y^*)\) were then analysed using the ASReml package (Gilmour et al., 2009), fitting the model:

\[
y^* = \mu + Zg + e,
\]

where \(y^*\) is a vector of the adjusted phenotypic records, \(Z\) is a design matrix, \(g\) is a vector of random additive genomic effects distributed as \(N(0, \sigma_g^2 G)\), \(\sigma_g^2\) is the additive genetic variance, \(G\) is the genomic relationship matrix, and \(e\) is the vector of residuals. The \(G\) matrix was constructed using the method of VanRaden (2008). The genetic variance/covariance matrix and GEBV (i.e., \(\hat{g}\) ) of the SBF lambs in the TS were estimated by utilizing both phenotype and genotype information. The predicted genomic breeding values (PGEBV), i.e. GEBV calculated without phenotypic information on the individual, were estimated fitting the model described above but masking the phenotypes of each subset in turn. Thus, in addition to its GEBV, after analysing each randomisation, every individual had a PGEBV obtained from marker data alone from random masking of phenotypes, with a similarly obtained PGEBV following masking of families.

**Across populations:** Two combined datasets were used for across population predictions, with SBF, SAR and MBR making the first set (4,123 individuals) and SBF and BT making the other (816 lambs). In the former data, two populations were used as TS to predict the third one (i.e., SAR and MBR to predict SBF; SBF and SAR to predict MBR; and SBF and MBR to predict SAR). Moreover, to test for the impact of cross-family links on GEBV, two analyses were conducted in which a few half-sib family members were allocated to the TS and used as a connection with the rest of the half-sib family members in the validation set. In these analyses, either one or 10
lambs from each half-sib family from the SBF data were randomly chosen to be in the TS.

Accuracy and predictive values of PGEBV

Genomic prediction accuracies were calculated for each validation set (both within and across populations). Firstly, the Pearson correlations of PGEBV with the adjusted phenotypes ($r_{\hat{y}y}$) were calculated and the accuracy ($r_{\hat{y}y}$) for each validation set was estimated by dividing $r_{\hat{y}y}$ by the the square root of the heritability of each trait for that specific validation set:

$$\text{Accuracy} = \frac{r_{\hat{y}y}}{\sqrt{h^2_y}}$$  (Legarra et al., 2008).

The accuracy for each trait was then obtained by averaging the estimates across validation groups.

The sampling properties of the prediction accuracies were explored by repeating the overall within-SBF cross-validation analysis, described above, 10 times and calculating the accuracy separately for each replicate. For each replicate, a new randomisation was performed so that the individuals comprising each of the groups were different. The standard error of the accuracy was then estimated as the empirical standard deviation of the 10 accuracy values. This exercise was performed for the average animal effect for Strongyles FEC, as an example trait.

Two further sets of analyses were performed using SBF data, alone. Firstly, we calculated the correlation between GEBV and PGEBV. This case represents a situation where progeny’s performance is predicted from markers before the availability of phenotypes. Secondly, the cross validation prediction accuracy analysis
was also performed using pedigree-based EBVs, rather than genomic EBVs. This addresses the question of how, in this population, the accuracy of genomic predictions compares to the accuracy of pedigree-based predictions.

**Exploring contribution of population structure in the Scottish Blackface data**

To explore the contribution of population structure to the accuracies of the genomic predictions, several analyses were performed. Firstly, to determine the effectiveness of the $G$ matrix in capturing additive genetic effects relative to the $A$ matrix, we analysed the SBF data fitting both the $G$ matrix and the pedigree-based numerator relationship matrix $A$ using the following model:

$$y^* = \mu + Zv + Zg + e,$$

where the effects are as defined above, with $v$ being an additional vector of additive polygenic effects normally distributed as $N(0, A\sigma_a^2)$, with $A$ being the numerator relationship matrix.

Secondly, the contribution of population and genome structure to genomic prediction accuracies of the SBF population was assessed by fitting chromosome-specific $G$ matrices. Following the methodology of Daetwyler et al. (2012a), 26 chromosome specific $G$ matrices were calculated, using only the SNPs on each chromosome. Each chromosome was then fitted instead of the overall $G$ matrix. To measure the proportion of the total genetic variance explained by each chromosome, we also carried out an analysis fitting each chromosome and the $G$ matrix consisting of all SNPs minus those in that specific chromosome (which corresponds to fitting all chromosomes simultaneously). The following model was then fitted:

$$y^* = \mu + Zg_{\text{chr}} + Zg_{\text{rest}} + e,$$
where $g_{ch}$ and $g_{rest}$ are the vectors of additive genomic effects unique to the chromosome under investigation and to all remaining chromosomes, respectively.

The terms $g_{ch}$, $g_{rest}$ and $e$ were assumed to be normally distributed: $N(0, G_{ch} \sigma_{ch}^2)$ and $N(0, G_{rest} \sigma_{rest}^2)$, respectively. Here, $G_{ch}$ is the genomic matrix for one chromosome and $G_{rest}$ is the genomic matrix estimated from the rest of the genome excluding the unique fitted chromosome markers.

Insight into the components contributing to the accuracy can be gained by regressing the difference in phenotypic variance explained by individually vs. simultaneously fitted chromosomal $G$ matrices on chromosome length (Yang et al., 2011, Daetwyler et al., 2012a). This was given by this equation:

$$\sigma_{c(sup)}^2 - \sigma_{c}^2 = b_0 + b_1 L_c + e$$

where $\sigma_{c(sup)}^2$ is variance explained by each chromosome analysed individually and $\sigma_{c}^2$ the variance when the chromosome are analysed jointly, with $b_0$ being the intercept which represents the component due to relatedness amongst animals rather than tagged QTL, and $b_1$ the slope that relates genetic variance to chromosome length ($L_c$), i.e. tagged QTL. We calculated the proportion of the genetic variance explained by the population structure (i.e. additive genetic relatedness as opposed to QTL tagged by the SNP chip) by dividing $b_{0d}$ (intercept of the difference) with the intercept from regressing the variance explained by individually fitted chromosomes on chromosome length ($b_{0i}$).

**Results**

*Accuracy and predictive values of PGEBV*
SBF data: Correlations between PGEBV and adjusted phenotypes, with corresponding accuracies for each trait, for the cross-validation groups in the SBF population are reported in Table 1, together with the accuracies estimated using pedigree-based EBV. Correlations varied between groups, ranging from marginally negative (-0.027 in group 1 for *Nematodirus* FEC at 16 weeks) to positive and moderate (0.382 in group 5 for IgA). Moderate accuracies ($r_{gg}$) were observed, generally between 0.42 and 0.68, with the exception of the accuracy for *Nematodirus* FEC at 16 weeks (0.10), this being the trait with the lowest heritability. Accuracies using pedigree-based EBV ranged from 0.27 to 0.52, and were slightly lower than the genomic EBV accuracies for 9 of the 12 traits. The empirical standard error of the accuracy for *Strongyles* FEC average animal effect, estimated as the standard deviation of the accuracies across the 10 replicated cross validation, was 0.04.

Correlations between GEBV and PGEBV (Table 2), representing the relationship between genomic EBVs predicted with and without individual data were all strong and positive. The average value across all traits was 0.76.

Lower correlation estimates between phenotype and PGEBV were obtained when all members in one sire family were predicted from the remaining sire families in the SBF data (Table 3). However, differences were observed in relationship connectivity between families. For example, nematode resistance indicator trait results (i.e., both IgA and FEC) showed that the families which were more closely related to the remaining families in the TS were those with more accurate PGEBV. In particular, the half-sib family sired by ram 22 (i.e., Fam22), which is the most highly related to the remaining TS families (data not shown) showed the highest correlations. However, different results were found for body weight, suggesting that not only relatedness is
important but other factors (such as trait heritability or markers in LD with mutations affecting the trait) may play a part.

**Across populations:** The correlations between PGEBV and adjusted phenotype for the *Strongyles* average animal effect were -0.054, -0.030 and 0.005 for SBF vs. (MBR plus SAR), MBR vs. (SBF plus SAR) and SAR vs. (SBF plus MBR) datasets, respectively. The correlations between PGEBV and adjusted phenotypes for the BT data vs. SBF were -0.012, -0.010 and 0.067 for *Strongyles* and *Nematodirus* FEC at 20 weeks and for body weight at 24 weeks, respectively. In both analyses, the predictions for genetically distant groups were usually close to zero. However, when one or 10 lambs from each sire family from the SBF data were randomly chosen and included in the TS, the correlations between PGEBV and y* were slightly higher, and always positive with 0.129 and 0.070 for SBF vs. (MBR plus SAR plus 10SBF) and SBF vs. (MBR plus SAR plus 100SBF), respectively.

**Exploring contribution of population and genome structure**

The results of the analysis in the SBF data, fitting either the A or G matrix alone, or both together, are reported in Supplementary Table S1. For some traits the heritability estimates were either completely explained by the G matrix (i.e., IgA and *Nematodirus* FEC at 20 weeks) or the A matrix (*Strongyles* FEC at 20 weeks and *Nematodirus* FEC at 16 weeks) when the analysis was done fitting both G and A matrices. However, for the other FEC traits (both *Strongyles* and *Nematodirus*) there was a contribution from both matrices. In general there was little discernible pattern in these results. Moreover, the relative partitioning of genetic variation between the A and G matrices may be expected to vary as the number and size of families varies, thus it is difficult to draw general conclusions from these results.
For the SBF population, heritability estimates were also obtained either fitting only one chromosome or when simultaneously fitting one chromosome plus the whole $G$ matrix (results not shown). Although similar trends were observed, the proportions of genetic variation accounted for when fitting only one chromosome were always overestimated. However, in both cases it is possible to identify the chromosomes that explain most of the genetic variation of the traits.

We tested the hypothesis that fitting all $G_{ch}$ (i.e., chromosome-wide genomic matrices) simultaneously would result in each chromosome explaining a fraction of the total genetic variance proportional to its length, consistent with the polygenic assumptions underlying GBLUP. Whilst there was a weak tendency for this to be the case for most traits (as an example, Figure 1), the majority of the captured genetic variation appeared to be independent of chromosome length. This can be seen in Table 4 which reports intercept, slope, and $R^2$ for the three regressions (i.e., by fitting each chromosome individually, by fitting all chromosomes simultaneously, and the difference between the two) as well as the proportion of genetic variance explained by relatedness for all traits considered. These proportions (ranging from 0.39 to 0.98, with an average of 0.77) suggest that in most cases our accuracy estimates are mostly due to additive genetic relatedness, rather than LD between SNP and QTL. The $A$-matrix-derived heritabilities were compared to accuracies and proportion of genetic variance explained by relatedness ($b_{00}/b_{0i}$) for all nematode resistance indicator traits (results not shown). Amongst the *Strongyles* FEC and IgA results there was little discernible relationship between these variables. The *Nematodirus* traits were more variable, however they tended to have lower heritabilities and relatively large genetic effects (i.e. QTL) had previously been observed on some of
the smaller chromosomes (see Discussion) suggesting that the polygenic inheritance assumption was inappropriate for the *Nematodirus* traits.

**Discussion**

One of the objectives of the current study was to understand the dynamics of applying genomic selection to hard-to-measure traits using field data. We assumed two scenarios, with the first scenario having young animals selected from markers before their phenotypes can be measured and secondly, where we break the assumption that the animals of the TS and the validation sets are from the same population i.e., we explore situations where the animals vary from being closely related to unrelated. Therefore, we explored the possibility of using genomic predictions within and across populations; whilst prediction accuracies within a population were good, with a small empirical standard error, our results highlighted the difficulties of prediction using genetically distant individuals.

We also reported prediction accuracies estimated by using both the $G$ and the $A$ relationship matrix. The accuracies estimated with the $G$ matrix were usually higher than those with the $A$ matrix, suggesting an advantage in using genomic information for predictions, even when pedigree knowledge is available. The one case where the accuracies estimated with the $A$ matrix was substantially better, viz. *Nematodirus* FEC at 16 weeks, was for a trait for which heritability estimate was mostly explained by the $A$ matrix (Supplementary Table S1).

Although several studies on GEBV accuracy/reliability estimated from real data have been reported in the literature for cattle with GEBV reliabilities ranging from 18 to 78% (Harris *et al.*, 2008, Hayes *et al.*, 2009b, VanRaden *et al.*, 2009), fewer are reported for sheep. Our GEBV accuracies are similar to others obtained using a
medium-density markers chip of 15 to 79% for wool traits in Merino sheep (Daetwyler et al., 2010b), and 7 to 31% for carcass and meat quality traits in multi-breed sheep data (Daetwyler et al., 2012b). In a study on the Lacaune dairy sheep breed using different genomic methods, Duchemin et al. (2012) reported accuracies varying from 0.4 to 0.6, according to the traits (i.e. milk yield, fat content, and somatic cell scores), with minor differences among genomic approaches. These authors also showed that the inclusion of molecular information, as compared with traditional schemes, increased accuracies of EBV of young males at birth from 18 up to 25%, according to the trait (Duchemin et al., 2012). However, it has to be considered that the accuracy of the GEBV depends on the size of the population and on the heritability of the trait. For low heritability traits, a very large number of records will be required in the TS to subsequently achieve high accuracies of GEBV in unphenotyped animals. If we consider our SBF population, where the effective population size (Ne) is ~500 (Kijas et al., 2012), then according to the formula suggested by Daetwyler et al. (2010a) to achieve an accuracy of 0.6, we would need ~ 30,000 individuals for a trait with very low heritability (e.g., Nematodirus FEC at 16 weeks), and ~ 5,000 for a trait with moderate heritability (e.g., IgA).

The current study explored the contributions of LD and relatedness to the accuracies of genomic predictions. The heritability estimates obtained either fitting only one chromosome or when simultaneously fitting one chromosome plus the whole $G$ matrix showed that nematode resistance in sheep is a complex trait with contributions from many regions in the genome affecting these traits. However, with the exception of Nematodirus FEC at 16 weeks (Supplementary Figure S2; Riggio et al., 2013), the results favour a polygenic mode of inheritance, which is largely captured by additive relationships between animals. This is illustrated by the results
when a chromosome at a time was fitted, that overestimated the proportion of genetic variance explained as opposed to when one chromosome and the G matrix were simultaneously fitted. As highlighted by Daetwyler et al. (2012a), if the only contribution of the SNP to the accuracy of genomic prediction was through LD with QTL, and assuming a polygenic model, then a G matrix constructed from only the SNP on one chromosome should capture genetic variation in proportion to its length, assuming that there is no population stratification. However, this was not the case in our study. It was therefore clear that a large proportion of the accuracy of genomic prediction in the SBF population, at the current SNP density, is due to population structure, i.e. relatedness between animals. In other words, only a small proportion of the accuracy was due to LD between SNP and QTL.

This proposition was tested formally using the regression approach suggested by Yang et al. (2011). The intercept \( b_{0d} \) of the difference between the variance for each chromosome when analysed individually or simultaneously was highly significant for all traits \( P<0.0001 \), with the exception of body weight at 24 weeks \( P=0.09 \). On the other hand, the slope \( b_{1d} \) of the difference was significant only for some of the traits. These values show the importance of the relatedness in our SBF population, suggesting that most of our accuracy is probably captured by additive relatedness. The ratio \( b_{0d}/b_{0i} \) is a measure of the proportion of genetic variance explained by such relatedness (Yang et al., 2011), and with the exception of NFEC16, this measure was high \( 0.59-0.98 \) and therefore accounted for most of the variation in our SBF GEBV predictions. Of interest is the observation that accuracy and the component due to relatedness were largely independent of the A-matrix-derived heritability estimates (results not shown).
The impact of relatedness has been previously studied, and differences in accuracies have been ascribed to the number of relatives in the TS and the degree of additive-genetic relationships with training individuals (Habier et al., 2010). Legarra et al. (2008) analysed accuracies of GEBV for individuals either related or unrelated to the TS in a mouse population, concluding that markers were able to recover family information to some extent. Our choice of predicting all members of a single sire family from the remaining sire families in the SBF data was designed to reduce the upward biases of accuracies resulting from within-family prediction when half-sib families are randomly split between TS and validation sets. In this case we showed that the closer the individuals in the validation set are to the TS, the higher the accuracy. This is probably due in part to the fact that genomic predictions across closely related individuals capture linkage effects, whereas those across distantly related animals require LD between SNP and QTL. However, it should be noted that although we used distinct sire families with the SBF data, these families were in most part, also closely related.

We also estimated the accuracy achieved when predicting breeding values across populations. These across-population accuracies were very low, sometimes even negative. These low estimates may be explained by extension from our previous results. Firstly, much of the accuracy in the SBF dataset was due to additive genetic relationships between animals, as captured by the marker IBS relationships. This will not be possible in distant populations. Secondly, the component of accuracy due to LD between SNP and QTL is also likely to be low in distant breeds, as the linkage phase between SNP and QTL will differ randomly in different breeds. The more distant the relationship between individuals, the shorter the genomic distance over which phase will be consistent. This outcome is reinforced by the finding that the
accuracy achieved for across-population prediction was somewhat higher when a small number of animals from the population to be predicted were included in the TS.

It has been suggested that the use of a different method (i.e., BayesSSVS; Verbyla et al., 2009) could increase across-breed prediction, as it assigns SNP to either a distribution with very small variance (i.e. near 0) or one with a larger variance in the prediction model, unlike GBLUP which assumes that all SNP effects are sampled from distributions with the same variance (Daetwyler et al., 2012a). However, this suggestion pre-supposes that the same gene variants are segregating in different populations, and that the SNP density is sufficient for there to be consistent LD between marker and QTL in (some of) the different populations. It has been suggested that the number of SNP needed to predict unrelated individuals is equal to $10NeL$, where L is the length of the genome in Morgans (Meuwissen, 2009). In the SBF population, with $Ne$ of ~500 (Kijas et al., 2012) and L of approximately 27 Morgans, predictions for unrelated individuals would require at least 135,000 SNP.

This marker density may be achievable with the forthcoming high density sheep SNP chip.

In summary, we have applied genomic prediction techniques to nematode resistance and body weight data and found GEBV which, at first sight, appeared to have moderate to good within-population predictive accuracy, despite a relatively limited training set. However, much of the accuracy achieved appears to be a result of the markers capturing additive genetic relationships between animals in the population. This is reinforced by the observations that (i) the accuracy tends to drop when predictions are across more distantly related animals in the same population, (ii) across-population predictions have accuracies close to zero and (iii) some across-population accuracy can be recovered by including a small number of animals from
the target population in the training set. These results suggest that genomic prediction for nematode resistance and body weight may be of value in closely related animals, but with the current SNP chip genomic predictions are unlikely to work across breeds.

Acknowledgements

These results are obtained through the EC-funded FP7 Project 3SR-245140. French SNP data were funded by the SHEEPSNPQTL ANR project. Funding from the Regional Government of Sardinia contributed to the collection of Sardinian SNP and phenotype data. We also wish to acknowledge funding contributions from EADGENE_S, the BBSRC Institute Strategic Programme Grant at The Roslin Institute and the Scottish Government's Strategic Partnership for Animal Science Excellence (SPASE) initiative.

References


Table 1 Correlations between predicted genomic estimated breeding values and adjusted phenotypes and accuracies* for the random cross-validation groups both using the genomic relationship matrix and the pedigree-based relationship matrix in the Scottish Blackface population

<table>
<thead>
<tr>
<th>Trait</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
<th>Group 4</th>
<th>Group 5</th>
<th>Genomic-based accuracy</th>
<th>Pedigree-based accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>IgA</td>
<td>0.151</td>
<td>0.174</td>
<td>0.314</td>
<td>0.359</td>
<td>0.382</td>
<td>0.532</td>
<td>0.513</td>
</tr>
<tr>
<td>SFEC16</td>
<td>0.192</td>
<td>0.074</td>
<td>0.089</td>
<td>0.245</td>
<td>0.174</td>
<td>0.487</td>
<td>0.516</td>
</tr>
<tr>
<td>SFEC20</td>
<td>0.141</td>
<td>0.099</td>
<td>0.216</td>
<td>0.150</td>
<td>0.091</td>
<td>0.432</td>
<td>0.401</td>
</tr>
<tr>
<td>SFEC24</td>
<td>0.138</td>
<td>0.068</td>
<td>0.186</td>
<td>0.172</td>
<td>0.110</td>
<td>0.442</td>
<td>0.476</td>
</tr>
<tr>
<td>NFEC16</td>
<td>-0.027</td>
<td>0.059</td>
<td>0.071</td>
<td>0.034</td>
<td>-0.006</td>
<td>0.099</td>
<td>0.342</td>
</tr>
<tr>
<td>NFEC20</td>
<td>0.210</td>
<td>0.292</td>
<td>0.193</td>
<td>0.324</td>
<td>0.220</td>
<td>0.598</td>
<td>0.488</td>
</tr>
<tr>
<td>NFEC24</td>
<td>0.212</td>
<td>0.182</td>
<td>0.155</td>
<td>0.178</td>
<td>0.130</td>
<td>0.503</td>
<td>0.408</td>
</tr>
<tr>
<td>W16W</td>
<td>0.206</td>
<td>0.127</td>
<td>0.231</td>
<td>0.232</td>
<td>0.234</td>
<td>0.516</td>
<td>0.336</td>
</tr>
<tr>
<td>W24W</td>
<td>0.169</td>
<td>0.073</td>
<td>0.165</td>
<td>0.109</td>
<td>0.046</td>
<td>0.417</td>
<td>0.292</td>
</tr>
<tr>
<td>SFEC_av</td>
<td>0.319</td>
<td>0.179</td>
<td>0.254</td>
<td>0.303</td>
<td>0.175</td>
<td>0.540</td>
<td>0.442</td>
</tr>
<tr>
<td>NFEC_av</td>
<td>0.208</td>
<td>0.317</td>
<td>0.192</td>
<td>0.282</td>
<td>0.234</td>
<td>0.481</td>
<td>0.357</td>
</tr>
<tr>
<td>WW_av</td>
<td>0.149</td>
<td>0.147</td>
<td>0.195</td>
<td>0.136</td>
<td>0.057</td>
<td>0.684</td>
<td>0.270</td>
</tr>
</tbody>
</table>

*accuracy here is the average of the accuracies across validation sets, estimated as the correlation for each validation set divided by the square root of its heritability.
Table 2  Correlations between genomic estimated breeding values and predicted estimated genomic breeding values for the random cross-validation groups in the Scottish Blackface population

<table>
<thead>
<tr>
<th></th>
<th>Group1</th>
<th>Group2</th>
<th>Group3</th>
<th>Group4</th>
<th>Group5</th>
<th>average</th>
</tr>
</thead>
<tbody>
<tr>
<td>IgA</td>
<td>0.674</td>
<td>0.731</td>
<td>0.784</td>
<td>0.699</td>
<td>0.773</td>
<td>0.732</td>
</tr>
<tr>
<td>SFEC16</td>
<td>0.737</td>
<td>0.606</td>
<td>0.699</td>
<td>0.729</td>
<td>0.764</td>
<td>0.707</td>
</tr>
<tr>
<td>SFEC20</td>
<td>0.841</td>
<td>0.764</td>
<td>0.850</td>
<td>0.788</td>
<td>0.846</td>
<td>0.818</td>
</tr>
<tr>
<td>SFEC24</td>
<td>0.825</td>
<td>0.804</td>
<td>0.815</td>
<td>0.826</td>
<td>0.794</td>
<td>0.813</td>
</tr>
<tr>
<td>NFEC16</td>
<td>0.774</td>
<td>0.750</td>
<td>0.700</td>
<td>0.690</td>
<td>0.710</td>
<td>0.725</td>
</tr>
<tr>
<td>NFEC20</td>
<td>0.709</td>
<td>0.863</td>
<td>0.823</td>
<td>0.867</td>
<td>0.767</td>
<td>0.806</td>
</tr>
<tr>
<td>NFEC24</td>
<td>0.842</td>
<td>0.783</td>
<td>0.816</td>
<td>0.880</td>
<td>0.847</td>
<td>0.834</td>
</tr>
<tr>
<td>W16W</td>
<td>0.627</td>
<td>0.676</td>
<td>0.719</td>
<td>0.794</td>
<td>0.713</td>
<td>0.706</td>
</tr>
<tr>
<td>W24W</td>
<td>0.666</td>
<td>0.667</td>
<td>0.743</td>
<td>0.799</td>
<td>0.632</td>
<td>0.702</td>
</tr>
<tr>
<td>SFEC_av</td>
<td>0.811</td>
<td>0.697</td>
<td>0.777</td>
<td>0.769</td>
<td>0.795</td>
<td>0.770</td>
</tr>
<tr>
<td>NFEC_av</td>
<td>0.764</td>
<td>0.765</td>
<td>0.765</td>
<td>0.798</td>
<td>0.735</td>
<td>0.765</td>
</tr>
<tr>
<td>WW_av</td>
<td>0.661</td>
<td>0.779</td>
<td>0.828</td>
<td>0.830</td>
<td>0.750</td>
<td>0.770</td>
</tr>
</tbody>
</table>

Table 3 Correlations between predicted genomic estimated breeding values and adjusted phenotypes for families in the Scottish Blackface population

<table>
<thead>
<tr>
<th></th>
<th>Fam022</th>
<th>Fam058</th>
<th>Fam085</th>
<th>Fam161</th>
</tr>
</thead>
<tbody>
<tr>
<td>IgA</td>
<td>0.324</td>
<td>0.087</td>
<td>0.174</td>
<td>0.119</td>
</tr>
<tr>
<td>SFEC16</td>
<td>0.198</td>
<td>0.023</td>
<td>0.179</td>
<td>0.055</td>
</tr>
<tr>
<td>NFEC16</td>
<td>0.108</td>
<td>-0.055</td>
<td>0.036</td>
<td>0.018</td>
</tr>
<tr>
<td>W16W</td>
<td>-0.072</td>
<td>0.162</td>
<td>0.291</td>
<td>0.124</td>
</tr>
</tbody>
</table>

IgA: Immunoglobulin-A; SFEC16, NFEC16, and W16W: Strongyles and Nematodirus faecal egg count and body weight at 16 weeks
Table 4 Intercept, slope (i.e., proportion of phenotypic variance/Mb), and $R^2$ for the three regressions (i.e., by fitting each chromosome individually, by fitting all chromosomes simultaneously, and the difference between the two) as well as the proportion of genetic variance explained by relatedness ($b_{o2}/b_{o1}$) for all traits considered

<table>
<thead>
<tr>
<th>Trait</th>
<th>R²</th>
<th>Intercept</th>
<th>Slope</th>
<th>R²</th>
<th>Intercept</th>
<th>Slope</th>
<th>R²</th>
<th>Intercept</th>
<th>Slope</th>
<th>Difference R²</th>
<th>Intercept</th>
<th>Slope</th>
<th>b_{o2}/b_{o1}</th>
</tr>
</thead>
<tbody>
<tr>
<td>IgA</td>
<td>0.26</td>
<td>0.058***</td>
<td>0.00025**</td>
<td>0.06</td>
<td>0.001</td>
<td>0.00010</td>
<td>0.34</td>
<td>0.056***</td>
<td>0.00015***</td>
<td>0.98</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SFEC16</td>
<td>0.10</td>
<td>0.029**</td>
<td>0.00014</td>
<td>0.08</td>
<td>0.005</td>
<td>0.00011</td>
<td>0.02</td>
<td>0.024***</td>
<td>0.00003</td>
<td>0.84</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SFEC20</td>
<td>0.10</td>
<td>0.041***</td>
<td>0.00009</td>
<td>0.00</td>
<td>0.012*</td>
<td>-0.00002</td>
<td>0.25</td>
<td>0.029***</td>
<td>0.00010**</td>
<td>0.71</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SFEC24</td>
<td>0.06</td>
<td>0.039***</td>
<td>0.00006</td>
<td>0.02</td>
<td>0.008</td>
<td>0.00004</td>
<td>0.03</td>
<td>0.031***</td>
<td>0.00003</td>
<td>0.80</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NFEC16</td>
<td>0.00</td>
<td>0.025**</td>
<td>-0.00002</td>
<td>0.00</td>
<td>0.015</td>
<td>-0.00002</td>
<td>0.00</td>
<td>0.010***</td>
<td>0.00000</td>
<td>0.39</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NFEC20</td>
<td>0.44</td>
<td>0.063***</td>
<td>0.00020**</td>
<td>0.04</td>
<td>0.005</td>
<td>0.00005</td>
<td>0.56</td>
<td>0.058***</td>
<td>0.00015***</td>
<td>0.92</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NFEC24</td>
<td>0.06</td>
<td>0.047***</td>
<td>0.00008</td>
<td>0.01</td>
<td>0.016*</td>
<td>-0.00003</td>
<td>0.28</td>
<td>0.032***</td>
<td>0.00011**</td>
<td>0.67</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W16W</td>
<td>0.28</td>
<td>0.037***</td>
<td>0.00022**</td>
<td>0.00</td>
<td>0.009</td>
<td>-0.00001</td>
<td>0.46</td>
<td>0.028***</td>
<td>0.00024***</td>
<td>0.76</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W24W</td>
<td>0.41</td>
<td>0.022***</td>
<td>0.00018***</td>
<td>0.00</td>
<td>0.009</td>
<td>-0.00001</td>
<td>0.28</td>
<td>0.013</td>
<td>0.00020**</td>
<td>0.59</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SFECav</td>
<td>0.07</td>
<td>0.068***</td>
<td>0.00012</td>
<td>0.00</td>
<td>0.013</td>
<td>0.00001</td>
<td>0.17</td>
<td>0.056***</td>
<td>0.00011*</td>
<td>0.82</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NFECav</td>
<td>0.07</td>
<td>0.079***</td>
<td>0.00015</td>
<td>0.02</td>
<td>0.011</td>
<td>0.00007</td>
<td>0.11</td>
<td>0.068***</td>
<td>0.00008</td>
<td>0.86</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WWav</td>
<td>0.11</td>
<td>0.017**</td>
<td>0.00010</td>
<td>0.10</td>
<td>0.003</td>
<td>0.00008</td>
<td>0.01</td>
<td>0.015***</td>
<td>0.00002</td>
<td>0.85</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
IgA: Immunoglobulin-A; SFEC16, SFEC20, and SFEC24: faecal egg count at 16, 20 and 24 weeks for Strongyles; NFEC16, NFEC20, NFEC24: faecal egg count at 16, 20 and 24 weeks for Nematodirus; W16W and W24W: body weight at 16 and 24 weeks; SFEC_{av}, NFEC_{av}, WW_{av}: average animal effect for Strongyles and Nematodirus faecal egg count and for body weight

*P < 0.05; **P < 0.01; ***P < 0.001
Figure 1 Proportion of phenotypic variance explained per chromosome for Immunoglobulin-A (scattered points) and fitted regression (line). Chromosome fitted individually (top regression) or simultaneously (bottom regression). Middle regression results from plotting the difference between top and bottom regression.