Optimising a panel of single nucleotide polymorphisms for genomic prediction in New Zealand sheep

MA Lee[ab*], KG Dodds[c], S-AN Newman[c], AS Hess[c], D Campbell[b] and SM Clarke[c]

[a]Beef+Lamb NZ Genetics, PO Box 5501, Dunedin 9054, New Zealand; [b]Department of Mathematics and Statistics, University of Otago, PO Box 56, Dunedin, New Zealand; [c]AgResearch, Invermay Agricultural Centre, Private Bag 5004, Mosgiel, New Zealand

[*]Corresponding author: Email: michael.lee@otago.ac.nz

Abstract

A national genetic evaluation called the New Zealand Genetic Evaluation (NZGE) is undertaken weekly to support the genetic improvement of sheep via selective breeding in New Zealand (NZ). This is based on single step genomic best linear unbiased predictions (ssGBLUP), which allows animals that have genotypes and pedigree, or just pedigree, to be included in the same analysis (or evaluation). In theory, an optimal strategy is that the density and quality of SNPs used should be high enough to track all the haplotypes segregating in the population to be evaluated. Since genomic predictions have been used in NZ sheep, the content of the SNP panel used has decreased to a panel of 41K; despite that there are new SNP panels with more content available.. The effect of using this additional content and additional SNPs from an Ovine 600K chip was investigated using the trait fleece weight at 12-months. We found that a high density panel provided more accurate predictions. However, a SNP panel based on the 50K SNPs plus a relatively small subset of additional SNPs (e.g. from the 600K Chip), were as predictive as using the high density panel. We conclude that increased SNP density, beyond the 41K selected from the Ovine 50K SNP, chip will improve prediction accuracy. Our results suggest that the inclusion of SNPs from putative quantitative trait loci (QTL) may also improve prediction accuracy. However, further analyses will be helpful to distinguish the relative benefit to accuracy of increased SNP density across the genome and/or inclusion of SNPs associated with QTL.

Keywords: sheep, genetics, genomics, breeding, ssGBLUP, genomic prediction

Introduction

The process of genetic evaluation aims to take data (e.g. pedigree, phenotypes and fixed effects) to predicts breeding values. More-often the statistical method best linear unbiased prediction (BLUP) (Henderson et al. 1959; Henderson 1975) is used. In BLUP (Equation 1), the breeding values (**u**) are predicted from the information from phenotypes (**y**), fixed effects (**b**) and relationships between individuals, where **X** and **Z** are design matrices and **e** residual errors, assumed independent of the random effects, with variance proportion to I (identity matrix) and the variance of **u** is proportional to the animal relationship matrix.

Equation 1 BLUP

$$y = Xb + Zu + e$$

In genomic BLUP (GBLUP) the genotypes are used to create a genomic relationship matrix (GRM) in contrast to a pedigree relationship matrix (or numerator relationship matrix; NRM) for standard pedigree BLUP. In single step GBLUP (ssGBLUP) the NRM and GRM are combined to give matrix **H** that is used in the BLUP mixed model equations described by Henderson (Henderson 1975) to predict breeding values.

It is the inverse of the relationship matrices that are used in the mixed model equations (MME). For example, for ssGBLUP (Legarra et al. 2009), the inverse of **H** ($\mathbf{H}^{-1}$) is used in the MME (Equation 2) to obtain solutions for fixed effects ($\hat{b}$) and random effects or breeding values ($\hat{u}$), where λ is the ratio of the variance of residual additive effects over the variance of additive genetic effects. This implies that these parameters are known.

Equation 2 Mixed model equation described by Henderson

$$\begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z + H^{-1}\lambda \end{pmatrix} \begin{pmatrix} \hat{b} \\ \hat{u} \end{pmatrix} = \begin{pmatrix} X'y \\ Z'y \end{pmatrix}$$

Currently, NZGE for maternal breeds (*i.e.* Romney, Coopworth, Perendale and Composites of these breeds) uses single nucleotide polymorphisms (SNPs) based on the Illumina Ovine 50K SNP chip (Anonymous 2015). In practice, most animals are genotyped using a lower density chip and imputed up to 50K. The density of these lower density chips has increased with time from about 5,000 SNPs to 18,000 SNPs. Many of these additional SNPs are not on the 50K SNP chip, but are present on a higher density 600K chip.

Information from data such as whole-genome sequence (WGS) is increasingly being applied into genomic predictions, offering a potential for increased prediction accuracy by including causal mutations or single-nucleotide polymorphisms (SNPs) in strong linkage disequilibrium (LD) with causal mutations affecting the trait of interest. Research to include information from WGS into genomic predictions to increase prediction accuracy is a focus in many species. The hypothesis is that causal mutations or SNPs in strong linkage disequilibrium with causal mutations affecting the trait may increase prediction accuracy, particularly across breeds. It is impractical to use all markers available for routine NZGE; therefore a subset of those available is sought. The selection of an "ideal"

marker set is a challenging task. The longer term aim of this study was to understand the effect of using SNPs additional to those used routinely in NZGE and in particular those associated with quantitative trait loci (QTL). The longer term goal is to use this information to improve predictions and hence the profitability in farming. Here we report some initial results for the trait wool fleece weight at 12-months of age (FW12). Future work will investigate and contrast other traits with higher marker densities.

Materials and methods

Data

This study used historic data collected by NZ sheep farmers and researchers and prior information (e.g. models and parameters) provided by Beef+Lamb New Zealand Genetics (B+LNZG). The number of pedigreed animals in the evaluation was 10,751,607.

Genotype data from 6,706 animals that had both phenotype and/or progeny with phenotypes (FW12) were included in the evaluation. The number of animals available were low as there were much lower numbers of animals with 600K SNP panel data compared to 50K data. This genotype data included terminal sire breeds, which were sheep breed to produce lambs for meat (e.g. Sufolks, Texels and Sufteexs) and maternal sire breeds from which ewe replacements are taken from (e.g. sheep of breed Romney, Coopworth, Perendale or composites of these breeds). The SNP panels used and treatments are described in Table 1. The genotypes were from an Ovine 600K SNP chip (n=2,243) or imputed from lower density chips (n=4,463) using a reference panel from B+LNZG and using the software FImpute2 (Sargolzaei et al. 2014) or Beagle 5.0 (Browning et al. 2018). The mean accuracy of the imputed animals was 0.99 calculated as an allelic concordance. The accuracy of imputation for each animal was estimated by validation, where markers were removed from the animals genotyped and after imputation the removed markers compared to those imputed.

In a previous study using genotype data from 131,916 animals each with 13,209 SNPs we undertook a GWAS across 40 traits in NZ sheep, with sample sizes ranging from 1,316 to 105,248 animals - only 24 traits had significant QTL (data not shown). The goal of this GWAS was to use non-imputed SNPs to identify large QTL segregating in the population. We used the results from this study and data assembled from a number of more recent ovine SNP chips to investigate the effect of increased SNP content on prediction accuracy compared to a SNP chip based solely on the initial Ovine 50K chip.

The results for the SNP panels of 41K and 46K all used 5% of NRM and 95% GRM. The 41K panel is derived from the 50K Ovine SNP chip and the 46K panel is the 41K panel with the additional SNPs from the HD SNP Chip that are also on the more recent SNPs of density >15K used in NZ. The panel 41KKMM (Table 2) is the 46K panel with 5,292 SNPs of the most significant SNPs from the GWAS described above removed. These SNPs are putatively associated with QTL for FW12. The panel 45KMF is based on the 46K panel; except that the significant SNPs (Figure 1) and those flanking these SNPs within a genetic distance of 2,500,000 base pairs were removed. Consequently, 188, 210 and 185 SNPs were removed, respectively, on chromosomes 1, 3 and 8 (see Figure 1).

Computing and Software

Breeding values were predicted with BLUP and ssGBLUP by using preconditioned conjugate gradient in Mix99 (Strandén and Lidauer 1999). Other data handling used Linux shell commands or R (The_R_Development_Core_Team 2011) built with the Intel® Math Kernel Library.

Assessing Prediction accuracy of genomic prediction

The accuracy of evaluations was assessed by validation with 185 recent sires that had a mean (standard deviation) number of progeny with phenotypes of 36.26 (24). The year of birth of these sires ranged from 2014 to 2017. In the reduced dataset all of the phenotypes from the sires and their descendants were removed (*i.e.* un-recorded). The number of FW12 phenotypes in the full and reduced data set was, respectively, 2,022,154 and 1,946,506. The accuracy using the reduced data was compared as the correlation between the deregressed breeding values with the parent average removed (Garrick et al. 2009) calculated from a BLUP that included all of the data. A higher correlation for a given prediction with the reduced dataset implies a higher accuracy of prediction.

The percentage of NRM relative to the GRM used to calculate the H ranged from 5-75% for ssGBLUP evaluations with the highest density SNP panel (n=564,998) . Equation 3 and Equation 4 explain how this weighting is done, where the weighted GRM (G_w) is calculated from the GRM (G) and the NRM (A) depending on α (proportion of the NRM to be used).

Equation 3 Weighted G matrix

$$G_w = (1-\alpha)G + \alpha A$$

Equation 4 The H-inverse matrix used in the mixed model equations for ssGBLUP

$$H^{-1} = A^{-1} + \begin{pmatrix} 0 & 0 \\ 0 & G_w^{-1} - A_{22}^{-1} \end{pmatrix}$$

A description of the treatments investigated is given in Table 1. The model used for BLUP and ssGBLUP was identical except for the inverse of the covariance matrix (NRM versus **H**). The model and assumed genetic parameters used were proprietary to B+LNZG.

Results

A plot of the results from the genome wide association study (GWAS), described in materials and methods, based on 26,024 animals is given in Figure 1 for the trait FW12. There were three putative QTL regions that met a threshold of 0.05 divided by the product of the number of traits by the number of markers.

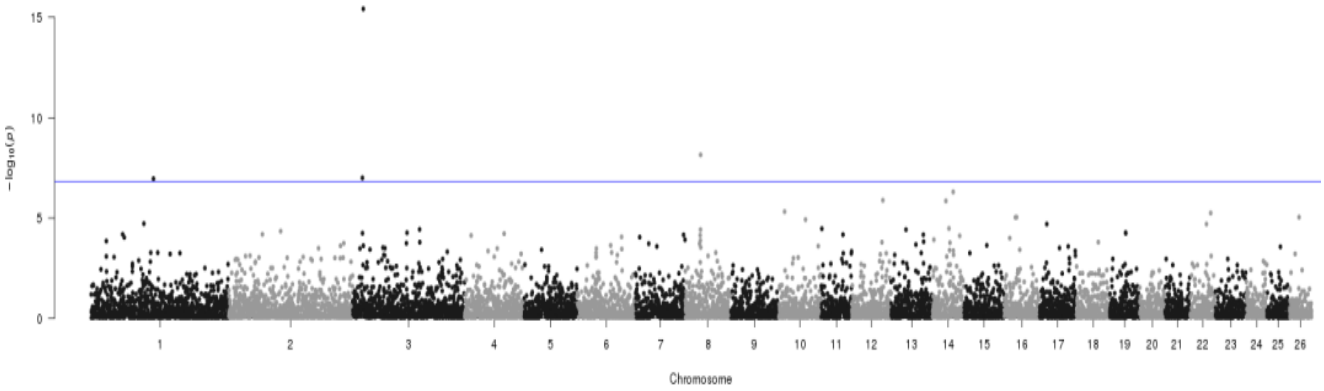
Table 1 Treatments and SNP panels used in this study. The treatments used for the different evaluations were: BLUP, pedigree only; ss0.75, ssGBLUP with a GRM calculated from 564,998 SNPs; ss0.5, ssGBLUP with a GRM calculated from 564,998 SNPs; ss0.25, ssGBLUP with a GRM calculated from 564,998 SNPs; ss0.05, ssGBLUP with a GRM calculated from 564,998 SNPs; ss0.05_41K, ssGBLUP with a GRM calculated from 40,881 SNPs; ss0.05_46K, ssGBLUP with a GRM calculated from 46,173 SNPs that consisted the same SNPs from the 41K panel and additional SNPs from the 564,998 panel; ss0.05_41KMM, 5,292 SNPs removed from the 46K panel that were the most significant SNPs from GWAS; ss0.05_45MF, the SNPs removed from the 46K panel that flanked three putative QTL detected by GWAS.

Treatment	SNP panel	Comment
BLUP	Non applicable	uses just pedigree
ss0.75	564998 SNPs from 600K SNP chip	H-matrix uses 75% A, 25% G
ss0.5		H-matrix uses 50% A, 50% G
ss0.25		H-matrix uses 25% A, 75% G
ss0.05		H-matrix uses 5% A, 95% G
ss0.05_41K	40,881 SNPs from 50K Chip	H-matrix uses 5% A, 95% G
ss0.05_46K	46,173 SNPs updated from HD and 50K	H-matrix uses 5% A, 95% G
ss0.05_41KMM	removed 5,292 GWAS SNPs from the 46,173 panel giving 40,881 SNPs	H-matrix uses 5% A, 95% G
ss0.05_45KMF	removed SNPs flanking 3 putative QTL	H-matrix uses 5% A, 95% G

Table 2 Validation results comparing different SNP panels and BLUP. Maternal Animals consisted of animals of breed Romney, Coopworth, Perendale and composites of these and Terminal Animals consisted of those of mainly Suffolk, Texel and composites of these. The Combined group included animals from both Maternal and Terminal breeds. See Table 1 for abbreviations.

Group	n	BLUP	ss0.75	ss0.5	ss0.25	ss0.05	ss0.05_41K	ss0.05_46K	ss0.05_41KMM	ss0.05_45KMF
Maternal Animals	96	0.23	0.27	0.28	0.3	0.31	0.29	0.31	0.29	0.3
Terminal Animals	89	0.56	0.57	0.56	0.55	0.5	0.5	0.5	0.49	0.5
Combined	185	0.56	0.57	0.58	0.57	0.55	0.54	0.54	0.53	0.54

Figure 1 Manhattan plot of GWAS results for the trait FW12. The horizontal line is a Bonferroni significance threshold of 0.05.



The validation results as correlations from comparing different treatments (see Table 1) are given in Table 2 for the full validation set and for the validation set split by breed type.

Discussion

Previous GWAS undertaken suggested for most current economically important traits there is not an abundance of large QTL, such as the myostatin mutation that affects muscling in sheep (Clop et al. 2006), segregating in the NZ sheep population (data not shown). This is relevant because an ongoing goal of research is to improve prediction accuracy by better using all data that is available. This might include WGS, prior knowledge on causal mutations etc.

Unsurprisingly, we found that a higher density panel gives more accurate prediction than a lower density (treatment ss0.05 was at least as good as any other method with 0.05 NRM weighting). The best correlation, 0.31, for maternal animals was seen using an α-value of 0.05 for the high density panel, whereas, for the lower density 41K chip based on the 50K Ovine SNP chip the correlation was 0.29. This was also shown in dairy cattle when the reliability of genomic breeding value was 0.5% and 1.0% higher using a 777K SNP chip compared to a 54K SNP chip respectively in Holstein and Red Dairy cattle (Su et al. 2012). The use of whole genome sequence for predictions, however, showed only a small increase in accuracy (1%) over a 60K panel in chickens (Heidaritabar et al. 2016). Practical and/or economic considerations would probably preclude the use

of such high marker density in routine genetic evaluation.

Prediction accuracy in diverse sheep populations increased when using variants selected from WGS data compared to standard 50K genotypes using GBLUP and Bayesian methods (Moghaddar et al. 2019). In the context of ssGBLUP the inclusion of additional SNPs, associated with QTL, from sequencing data could improve the accuracy for some, but not all, milk traits (Liu et al. 2020). They concluded that two more sophisticated ssGBLUP methods (e.g. weighted ssGBLUP) offered no significant advantage over standard ssGBLUP. In French dairy goats, weighted ssGBLUP was between 2 and 14% more accurate when a QTL was known to be segregating in the population, but otherwise less accurate or as accurate as, standard ssGBLUP (Teissier et al. 2019). In the context of NZGE, it would be difficult to implement weighted ssGBLUP, given the complexity of the population to be evaluated. Instead, the incorporation of markers associated with QTL may provide a better strategy for this population.

In relation to this we expected that the removal of the SNPs detected by GWAS would decrease prediction accuracy. In a comparison of ss0.05_41K with ss0.05_41KMM there was no difference for maternal animals and a modest (1%) decrease for terminal animals, when comparing the terminal animals with an α-value of 0.05. There was a 1% decrease in correlation between ss0.05_46K and ss0.05_45KMF for the maternal animals. This result is consistent with our expectation, but further analyses will need to be done to substantiate them.

In NZGE genotypes from maternal animals are evaluated separately of those from terminal animals. In this study, we combined the genotypes into one evaluation and analysed the results as correlations across and within these two groups. The results highlight that for ssGBLUP by combining animals from disparate breeds a trade-off needs to be made to optimally predict across all breeds. For example, the optimal percentage of the NRM to include in making **H** was 0.05 for maternal animals (see Table 2), whereas, for terminal animals this percentage (α in Equation 3) the correlation was worse than using BLUP (*i.e.* pedigree only). Using a higher percentage of NRM (less GRM) would be necessary to provide predictions for both the maternal and terminal animals that are more accurate than BLUP.

The inclusion of SNPs from GWAS was indicative of improving prediction accuracy and warrants further investigation (e.g. investigation across other traits). We were restricted by low sample size in this study as the number of progeny tested sires that had 600K genotypes was low. This meant that it was hard to interpret bias, where this is commonly analysed by comparing the slope from regressing the deregressed breeding value on the breeding value predicted, from the results so these results were omitted from this study. However, the results suggested that increased density might also decrease bias (data not shown).

We conclude that for the population and trait being evaluated the use of an increased SNP density, beyond the 41K selected from the Ovine 50K SNP chip, will improve prediction accuracy and that further investigation into including QTL effects into evaluations is warranted.

Acknowledgements

The authors gratefully acknowledge Beef + Lamb New Zealand Genetics, Levy paying farmers and Sheep breeders for contributing to this work. The research was also, in part, funded by Genomics Aotearoa.

References

Anonymous 2015. OvineSNP50 Genotyping BeadChip, In: Data Sheet: Agrigenomics.

Browning BL, Zhou Y, Browning SR 2018. A One-Penny Imputed Genome from Next-Generation Reference Panels. The American Journal of Human Genetics 103: 338-348.

Clop A, Marcq F, Takeda H, Pirottin D, Tordoir X, Bibe B, Bouix J, Caiment F, Elsen JM, Eychenne F, Larzul C, Laville E, Meish F, Milenkovic D, Tobin J, Charlier C, Georges M 2006. A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. Nature Genetics 38: 813-818.

Garrick DJ, Taylor JF, Fernando RL, 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genetics Selection Evolution 41: 55.

Heidaritabar M, Calus MP, Megens HJ, Vereijken A, Groenen MA, Bastiaansen JW 2016. Accuracy of genomic prediction using imputed whole-genome sequence data in white layers. Journal of Animal Breeding and Genetics 133: 167-179.

Henderson CR 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics 31: 423-447.

Henderson CR, Kempthorne O, Searle SR, von Krosigk CM, 1959. The estimation of environmental and genetic trends from records subject to culling. Biometrics 15: 192-218.

Legarra A, Aguilar I, Misztal I 2009. A relationship matrix including full pedigree and genomic information. Journal of Dairy Science 92: 4656-4663.

Liu A, Lund MS, Boichard D, Karaman E, Guldbrandtsen B, Fritz S, Aamand GP, Nielsen US, Sahana G, Wang Y, Su G 2020. Weighted single-step genomic best linear unbiased prediction integrating variants selected from sequencing data by association and bioinformatics analyses. Genetics Selection Evolution 52: 48.

Moghaddar N, Khansefid M, van der Werf JHJ, Bolormaa S, Duijvesteijn N, Clark SA, Swan AA, Daetwyler HD, MacLeod IM 2019. Genomic prediction based on selected variants from imputed whole-genome sequence data in Australian sheep populations. Genetics Selection Evolution 51: 72.

Sargolzaei M, Chesnais JP, Schenkel FS 2014. A new approach for efficient genotype imputation using information from relatives. BMC genomics 15: 478-478.

Strandén I, Lidauer M 1999. Solving large mixed linear models using preconditioned conjugate gradient iteration. Journal of Dairy Science 82: 2779-2787.

Su G, Brøndum RF, Ma P, Guldbrandtsen B, Aamand GP, Lund MS 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. Journal of Dairy Science 95: 4657-4665.

Teissier M, Larroque H, Robert-Granie C, 2019. Accuracy of genomic evaluation with weighted single-step genomic best linear unbiased prediction for milk production traits, udder type traits, and somatic cell scores in French dairy goats. Journal of Dairy Science 102: 3142-3154.

The R Development Core Team 2011. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.