Proc. Assoc. Advmt. Breed. Genet. 21: 149-152

IMPROVING THE ACCURACY OF ACROSS BREED GENOMIC PREDICTIONS

M.E. Goddard[1,3], I.M. MacLeod[1,2,3], S. Bolormaa[3], B.J. Hayes[3,4], and K.E. Kemper[1]

[1] Faculty of Veterinary & Agricultural Science, University of Melbourne, Victoria, Australia
[2] Dairy Futures Cooperative Research Centre, La Trobe University, Bundoora VIC 3083, Australia
[3] AgriBio, Dept. Economic Dev., Jobs, Transport & Resources, Victoria, Australia
[4] Biosciences Research Centre, La Trobe University, Victoria, Australia

SUMMARY

Genomic predictions derived from one breed but applied in another breed typically have low accuracy due to SNP x breed interactions (due to either QTL x breed interactions or differences in LD between breeds) and due to differences between breeds in QTL allele frequency. In this paper we discuss the importance of these two factors and the implications for livestock breeding.

INTRODUCTION

Genomic selection (Meuwissen *et al.* 2001) has been very successful in predicting the breeding value of animals from DNA marker data. It works best in Holstein cattle for milk production traits where there is a large amount of high quality data on which to train the prediction and the animals where the prediction is to be used (target animals) are closely related to the training population (often their sons). As the test animals become less closely related to the training population the accuracy of prediction declines (Habier et al 2010) and when they are of different breeds, the accuracy is typically low (eg. Kemper et al 2015a). Unfortunately there are many traits where we do not have a large training dataset within every breed. For instance, the expense of measuring feed conversion efficiency limits the size of training datasets. Therefore we would like to use a multi-breed training dataset to maximize the number of training animals and to predict breeding value in animals of a breed included in the training data or even a breed not included in the training data. In these situations the low accuracy of across breed prediction is a severe disadvantage. Alternately, if a method was available in which across breed prediction was of high accuracy, it seems likely that within breed prediction would also be more robust and not affected by the degree of relationship between training and test animals. In this paper we consider reasons for the low accuracy of across breed prediction and what might be done to increase the accuracy.

ACROSS BREED ACCURACY OF GENOMIC PREDICTIONS

Information from another breed can be used in prediction in two ways. Firstly, Brondum *et al.* (2012) and Khansefid *et al.* (2014) showed that the accuracy of prediction in breed B could be increased by including in the statistical model SNPs that were associated with the trait in breed A but by estimating the effect of the SNP entirely within breed B. This implies that some of the same QTL segregate in both breeds. Secondly, the accuracy can be improved slightly by estimating the effect of each SNP across all target breeds (Bolormaa *et al.* 2013a, Hoze *et al.* 2014, Makgahlele *et al.* 2013) and some accuracy is obtained even in a breed not included in the training population (Kemper *et al.* 2015a). For instance, Kemper *et al.* (2015a) found the accuracy in Australian Red cattle for milk production traits averaged 0.3 based on a training population of Holstein and Jerseys. They also reported that the regression of phenotype on genomic EBV (bias) was 0.6 on average indicating that the EBVs exaggerated the predicted differences between animals.

REASONS FOR LOW ACCURACY

There are two reasons for low across breed prediction accuracies – SNP x breed interactions and differences in QTL allele frequency between breeds. Khansefid *et al.* (2014) analysed residual

feed intake using genomic relationship matrices and found the SNP variance and the SNP x breed variance to be about equal. SNP x breed interactions could be due to QTL x breed interactions or differences between breeds in linkage disequilibrium (LD) between QTL and SNPs.

The extent of QTL x breed interactions, due to non-additive gene effects, is largely unknown. It is similar to sire by breed interaction which is seldom above 0.2 of the genetic variance, so it seems unlikely that QTL x breed would explain 0.5 of the genetic variance. One type of non-additive variance is dominance which Bolormaa *et al.* (2015) estimates to explain 5% of phenotypic variance across a number of traits in beef cattle.

Differences in LD between breeds depend on the distance which separates the QTL and SNP. GBLUP uses LD over long distances within a breed but this LD breaks down between breeds (deRoos *et al.* 2008). Therefore this will explain some of the SNP x breed interaction found by Khansefid *et al.* (2014). Differences in LD occur even at short distances in the case of a QTL mutation which has occurred since the breeds diverged. In this case, an ancestral haplotype may exist in one breed with the ancestral QTL allele and in the other breed with the mutant QTL allele. Thus even with sequence data and methods such as BayesR (Erbe *et al.* 2012), SNP x breed interaction will occur for recent QTL mutations unless the QTL mutation itself is used.

Differences in QTL allele frequency will occur due to selection and drift. QTL with minor allele frequency (MAF) near to 0.5 contribute more to genetic variance and have their effect estimated more accurately than QTL with low MAF. This increases the accuracy of prediction within a breed. However, if the MAF is low in the training population but high in the target population, the QTL is important to genetic variance in the target population but its effect will be estimated poorly. This will reduce the accuracy of the genomic prediction in the target population. The extreme case of this phenomenon is when the QTL segregates in the target population but not in the training population. In that case, no estimate of its effect can be made and the variance it explains will be totally missed. This can happen if the QTL mutation is recent and has occurred in one breed since the breeds diverged. It can also occur if the QTL is old but has become fixed in the training population but not the target population. This extreme case, where the QTL segregates only in one breed, places an upper limit on the potential accuracy attainable using across breed genomic prediction. Kemper *et al.* (2015b) recently estimated that about ½ the QTL discovered for milk production traits in Holstein also segregated in Jersey cattle. Even among random SNPs in sequence data, 20% do not segregate in both Holsteins and Jerseys (Kemper et al 2015b).

The factors that cause low accuracy of across breed predictions can also cause bias. For instance, if a QTL mutation has occurred in the training population since the breeds diverged, the genomic predictions will predict that the QTL contributes to variance in the target population when it does not segregate. Kemper *et al.* (2015b) documented several examples where Holstein-only QTL were predicted for Jersey cattle when the QTL did not segregate in that breed.

AGE AND BREED DISTRIBUTION OF QTL

From the discussion of factors causing low accuracy it emerges that two closely related parameters are important - the age of QTL mutations and the range of breeds in which they segregate. Since the mutation causing a QTL (a QTN) is seldom known, the age and distribution of QTL is not well understood. For neutral mutations we can estimate their average age from the ratio of the heterozygosity per site (in cattle this is typically about 0.001) to the heterozygosity introduced each generation by mutation (2 x mutation rate e.g. 2×10^{-8}) which gives an average age of 50,000 generations or well before domestication of cattle and sheep. This average disguises a large range in age from new mutations in the last generation to very old ones.

QTL are unlikely to be neutral and so selection will modify this average age. One estimate of average age is the genetic variance (e.g. $0.5V_E$) divided by the variance added by mutation each generation (e.g. $0.001V_E$) or about 500 generations. However, this average includes detrimental

mutations of large effect which are soon eliminated from the population, so the average age of those that are segregating is likely to be much greater than 500 generations and to vary greatly about this average. The age of QTL can also be estimated from the length of a common haplotype which surrounds the mutant allele. For myostatin mutations causing double muscling O'Rourke *et al*. (2010) estimated their age to be <100 generations. Consistent with this, each mutation segregates in one or a few related breeds.

Kemper *et al*. (2015b) examined the length of haplotype surrounding QTL for milk traits in Holstein. They found examples of QTL that appeared recent (800 generations) and occurred in Holsteins but not Jersey and others that appeared old (12000 generations) and occurred in both breeds. By comparison de Roos *et al*. (2008) estimated the age of the Holstein-Jersey divergence at 400 generations. There were also QTL that appeared old but did not segregate in Jerseys. This is expected to occur because breeds of *Bos taurus* cattle have suffered some inbreeding since they diverged and consequently lost a proportion of polymorphisms including QTL. Of 11 QTL in Holstein, 6 also segregated in Jerseys.

Saatchi *et al.* (2014) found 4 QTL for weight that segregated in several breeds of US beef cattle and we have found QTL in the same position in Australian beef cattle. On the other hand, Bolormaa *et al.* (2013b) concluded the QTL seldom segregate in both *B. taurus* and *B. indicus*, which diverged perhaps 100,000 generations ago.

Thus we conclude that while many QTL segregate in multiple *B. taurus* breeds, some QTL segregate only in some breeds, either because they are recent mutations or because they are old but fixed in some breeds.

IMPLICATIONS FOR LIVESTOCK BREEDING

We conclude that the poor accuracy of genomic prediction when the training dataset comes from one breed and the predictions are applied to another breed is due to a combination of QTL x breed interactions, differences in LD between breeds and differences in QTL allele frequency between breeds (especially when a QTL segregates only in some breeds). What strategies can be used to overcome this problem?

One strategy is to do all prediction within breed. This is simple to implement because a low density SNP panel (e.g. 50K) is satisfactory and simple statistical methods, such as BLUP, can be used. However, this strategy cannot be implemented in all cases and even where it can be used it has disadvantages – some 10% of the genetic variance is not explained by a 50K SNP panel and the predictions may not be robust when applied to target animals not closely related to the training population (MacLeod *et al.* 2014a).

The alternative strategy is to use a multi-breed training population. This requires dense SNPs, ideally sequence data, and a statistical method which can find and utilise the causal mutations or markers in near complete LD with them. Figure 1 shows results for a mixed breed reference (Holstein and Jersey) where accuracy was evaluated in 3 validation sets that differed in their relatedness to the reference. We compared accuracy for different SNP densities: including a combined set of high density markers and imputed sequence in and near coding regions (SEQ). Instead of real phenotypes, we simulated 4000 QTL into the real genotypes (h^2=0.6) so that the true breeding value was known. We found accuracy increased as the density of SNP increased to SEQ and this was most apparent in the target population that was least related to the training population (Australian Red). However, there was still a drop in accuracy even with the QTL included in the SEQ data. This must be partly a result of the BayesR analysis spreading the effect of a single QTL across several SNP in strong LD with the QTL in the training population. Figure 1 also demonstrates that the BayesRC method (Macleod *et al.* 2014b), which is similar to BayesR but includes a broadly defined biological prior, also increases the accuracy of across breed genomic predictions.

As causal mutations or markers in near perfect LD with causal mutations are discovered we will be better able to assess the importance of non-additive genetic effects causing QTL x breed interactions and fit them in the model if necessary.

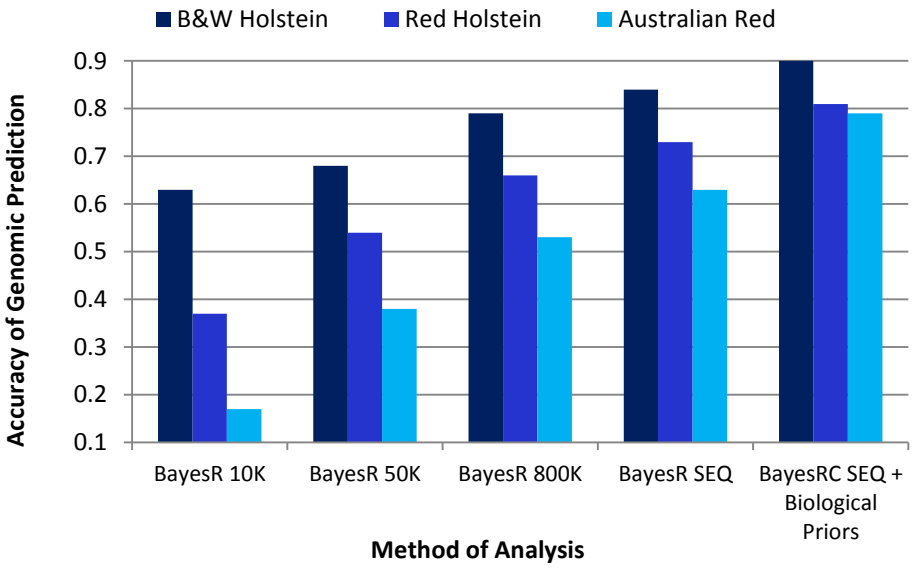

Figure 1. Accuracy of genomic prediction for simulated QTL using different densities of genotypes [(10K, 50K, 800K SNP or including sequence variants (SEQ)] with BayesR or BayesRC. Validation populations were either closely related to the reference (Black & White [B&W] Holstein), somewhat related (Red Holstein) or a different breed (Australian Red)

REFERENCES

Brondum, R.F., Su, G., Lund, M.S., Bowman, P.J. *et al.* (2012) *BMC Genomics* **13**:543.
Bolormaa, S., Pryce, J.E., Kemper, K.E., Savin, K. *et al.* (2013a) *J Anim Sci* **91:**3088.
Bolormaa, S., Pryce, J.E., Kemper, K.E., Hayes, B.J. *et al.*(2013b) *Gen Sel Evol* **45:**43.
Bolormaa, S. Pryce, J.E., Zhang, Y., Reverter, A. *et al.* (2015) *Gen Sel Evol* **47:**26.
de Roos, A.P.W., Hayes, B.J., Spelman, R.J. and Goddard, M.E. (2008) *Genetics* **179:**1503.
Erbe, M., Hayes, B.J., Matukumalli, L.K., Goswami, S. *et al.* (2012) *J Dairy Sci* **95:**4114.
Habier, D., Tetens, J., Seefried, F. R., Lichtner, P., & Thaller, G. (2010). *Gen Sel Evol*. **42**:5.
Hozé C, Fritz S, Phocas F, Boichard D., *et al.* (2014) *J Dairy Sci* **97**:3918.
Kemper, K.E. Reich, C.M., Bowman, P.J., vander Jagt, C.J. *et al.* (2015a) *Gen Sel Evol* **47:**29.
Kemper, K.E. Hayes, B.J., Daetwyler, H.D. and Goddard, M.E. (2015b) *J Anim Breed Genet* **132:**121.
Khansefid, M., Pryce, J.E., Bolormaa, S., Miller, S.P. *et al.* (2014) *J Anim Sci* **92:**3270.
MacLeod, I.M., Hayes, B.J., Goddard, M.E. (2014a). *Genetics* **198:**1671.
MacLeod, I.M., Hayes, B.J., Vander Jagt, C.J., Kemper, K.E. *et al*. (2014b). *Proc. 10th World Congress Genetics Applied to Livestock Production*
Makgahlela ML, Mäntysaari EA, Strandén I, *et al.* (2013) *J Anim Breed Genet* **130:**10.
Meuwissen, T.H.E., Hayes, B.J. and Goddard, M.E. (2001) *Genetics* **157:**1819.
O'Rourke, B.A., Greenwood, P.L., Arthur, P.F. and Goddard, M.E. (2013). *Anim. Genet* **44:**86.
Saatchi, M., Schnabel, R., Taylor, J. and Garrick, D. (2014) *BMC Genomics* **15:**442.