PREDICTING GENOMIC SELECTION ACCURACY FROM HETEROGENEOUS SOURCES

J.H.J. van der Werf[1,2], S.A. Clark[1,2] and S.H. Lee[1,2]

[1]School of Environmental and Rural Science, University of New England, Armidale, Australia, 2351
[2]Cooperative Research Centre for Sheep Industry Innovation, Armidale, 2351

SUMMARY

We predict genomic selection accuracy from a heterogeneous reference population that contains close relatives, herd- or flock mates and individuals from the wider population, using an established theory. The various sources of information were modeled as different and independent reference populations with different effective sizes. We show that information on close relatives can have a substantial effect on genomic prediction accuracy. We also show the increase of the genomic prediction accuracy to be less reliant on higher marker density or total reference population size when there are more closely related individuals to predict from. Conversely, the value of close relatives is smaller when the total reference population size is larger. Our modelling is useful to assess the value of a population reference versus a breeder's own reference, based on own animals genotyped.

INTRODUCTION

Genomic selection requires a reference population of individuals having information on both genotype and phenotype. The accuracy of genomic prediction depends on various parameters, including size of the reference sample, its genetic structure and the genetic architecture of the trait of interest. An important parameter is the effective size of the population. The effective population size is a predictor of the effective number of chromosome segments that are represented in the population. Theoretical predictions have usually considered a homogeneous population. However, in most practical applications, the reference population used for genomic predictions possibly consists of many subpopulations, e.g. breeds, lines or strains within a breed and part of the reference population maybe be directly related via pedigree to the animals to be predicted. Hence, reference populations consist of individuals that vary in relatedness to each other and to the target animals to predict. The distinction could be relevant for a breeder with genotyped individuals to assess the importance of own measurement versus that in the wider population.

Clark *et al.* (2012) showed that genomic predictions are more accurate if the genomic relationship between the target animal and the reference population is higher. Habier *et al* (2013) distinguished between three types of information in genomic prediction; linkage disequilibrium, additive-genetic relationships and co-segregation of QTL predicted from marker genotypes within a pedigree. They argued that it would be useful to understand how these sources contribute to the accuracy of genomic predictions, especially when designing reference populations for breeding programs. They show these contributions via simulated examples but did not provide simple predictions for them. Hayes *et al.* (2009) also considered the influence of relationships on genomic prediction. They followed the same approach as the general theory, i.e. by considering the number of independently segregating chromosome segments within families. They showed the accuracy of genomic prediction from varying sizes of full- and half- sib families, but did not consider the information from combined sources. We propose a simple approach to assess the importance of various sources of information used for genomic prediction in animal breeding.

MATERIALS AND METHODS

Predicting genomic selection accuracy. The accuracy of genomic breeding values (GBV) based on DNA marker genotypes can be predicted from theory (e.g. Daetwyler *et al*., 2008; Goddard, 2009; Goddard *et al*., 2011), assuming that prediction is based on a reference population of animals with phenotypes and genotypes for the same DNA markers, and these markers are linked to quantitative trait loci (QTL). Based on the infinitesimal model, the accuracy depends on *i)* the proportion of genetic variance at QTL captured by markers and *ii)* the accuracy of estimating marker effects. The proportion of genetic variance at QTL captured by markers (b) depends on LD between markers and QTL, which in turn depends on the number of markers (M) and the number of 'effective chromosome segments' (M_e); $b = M/(M_e + M)$. Prediction of M_e is not easy and various approximations have been presented by largely the same authors (Goddard, 2009; Hayes *et al.,* 2009, Goddard *et al*., 2011, Meuwissen *et al*., 2013). We will use $M_e = 2N_ELk/\ln(2N_E)$ (Meuwissen *et al*., 2013), where N_E = effective population size; L = average chromosome length; k = number of chromosomes. The accuracy of estimating marker effects depends on the captured genetic variance as a proportion of the total variance ($b.h^2$), the number of (unrelated) animals observed in the reference population (T), and M_e. The accuracy is the variance of the estimated (random) marker effects (q) as a proportion of the variation in true marker effects: $V(\hat{q})/V(q)$. This term is estimated as $\theta/(1+\theta)$, where $\theta = Th^2b/M_e$. Reliability of GBV is then $r^2 = b.V(\hat{q})/V(q)$ and the accuracy is the square root of this value.

Effective population size in a heterogeneous population. A critical parameter in the accuracy of genomic prediction is the effective population size (N_E). It is not easy to define 'population' in many practical cases and it is not possible to represent a reference population by a single value for N_E. We propose a very simple model relevant for breeding programs for beef cattle or sheep. For the prediction of an individual within a herd/flock we consider three sources of information based on animals that are measured and genotyped *1)* N_1 individuals from a certain breed but not closely related to the target animal, *2)* N_2 herd/flock mates of the target animal and *3)* N_3 close relatives of the target animal. We will refer to these sources of information as *breed*, *flock* and *relatives*, respectively. This is, of course, a simplified representation of heterogeneity, but a useful start to consider the contribution of each to overall prediction accuracy. We consider these three subsets as populations that differ in relatedness to the target animal as well as to each other, to be modeled as three different populations with different effective size, indicated as N_{E1}, N_{E2}, and N_{E3}, and a different number of chromosome segments, i.e. differing also in the size of the segments shared amongst each other and with the target animal. Each of these sources provides an estimation of breeding value and the reliability (r^2_i) of each GBV_i can be calculated as above. The three information sources are combined as $GBV = \Sigma GBV_i$ by using $cov(GBV_i, GBV_j) = r^2_i.\ r^2_j.V_A$, and $cov(GBV_i, a) = r^2_i.V_A$, where a is the true breeding value and V_A is the additive genetic variance. The accuracy of the GBV can then be calculated using standard selection index theory.

Study Design. For each of the three resources contributing to genomic prediction we varied values for N_{Ei} and N_i and marker density. We compared accuracy of GBV from just *breed* with predictions that included also information from *flock* and *relatives*. The total number in the reference population was kept equal between such comparisons. We evaluated the contribution of each information source as 'value of variate', defined as the relative loss in accuracy if that resource was removed. The trait heritability was 0.25.

RESULTS AND DISCUSSION

In a base scenario we assumed a population with a large diversity, N_{E1} = 1000, e.g. similar to the Merino population. Subsets of flock mates and relatives were represented by N_2 = 400 and N_3 = 50, with effective size N_{E2} = 50 and N_{E3} = 8. This scenario represents a lower value for the *breed* information source due to its large diversity, and a large number of individuals in the *flock* and *relatives* information sources. Results are shown in Figure 1, showing that the *flock* and *relatives* resources contribute substantially to the prediction accuracy, especially when the accuracy of the *breed*

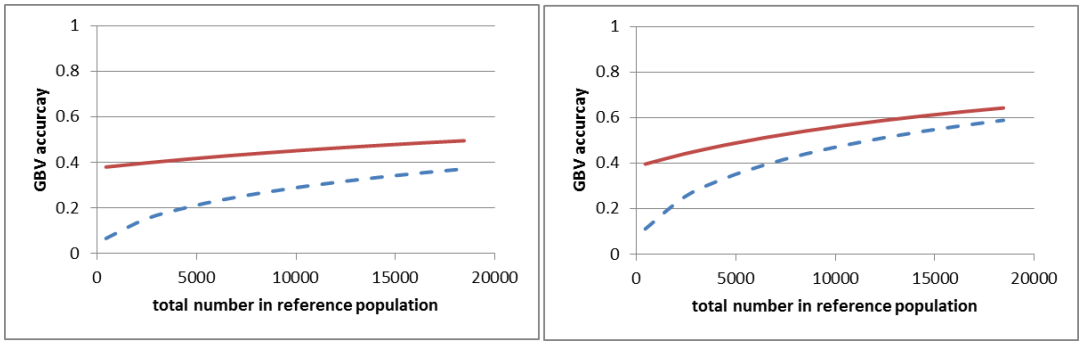

Fig. 1. Accuracy of GBV depending on total reference population size for low (Nmarkers=12k, left) and high (Nmarkers = 500k, right) marker density, comparing 'with' (continuous line) and 'without' (dashed line) information on *flock* and *relatives*.

resource is low. This is the case with low N_1 and with low marker density coupled with high population diversity (N_{E1}). The influence of *flock* and *relatives* decreases with large N_1 and also with higher marker density. Further comparisons are summarized in Table 1. The results show that for populations with lower N_{E1} the contribution of *flock* and *relatives* declines rapidly. If the contribution of *flock* and *relatives* is smaller due to less own data being available (lower values for N_2 and N_3) then their influence decreases accordingly, but it can still be substantial for small N_1 and high N_{E1}.

Overall the results illustrate that the GBV accuracy is likely higher than predicted based on the size in the reference population and the effective population size of the breed, due to information from relatives and more closely related individuals in the flock or herd. The effect will be larger when the information from the wider *breed* resource is of lower value, e.g. for smaller reference populations, or breeds with higher diversity. The effect of marker density is more notable if the breed diversity is high (high N_{E1}). The information from the own flock genotyping and recording can contribute substantially, and even if the numbers are relatively low (low N_2 and N_3) if the *breed* resource is small (e.g. N_1 = 2000). The assumption about N_{E2} and N_{E3} have some effect on the observed differences, e.g. when N_{E3} increases from 8 to 16 in the first case, the accuracy increase (diff) reduced from 95% to 87% and when N_{E2} increases from 50 to 100, the increase is further reduced to 64%.

The purpose of this study was to use a simple model to estimate of the importance of information on closer relatives in genomic prediction. This is relevant for breeders that have developed their own reference population. The value of this own reference can be substantial, unless a fairly large breed reference is available, and the value would be higher for more diverse breeds such a Merino.

Table 1 Value of the various information sources, accuracy of GBV with and without the *flock* and *relatives* information sources[2] and the relative accuracy difference (diff).

	Value of information source[1]					
N_1	*breed*	*flock*	*relatives*	GBV_acc_with	GBV_acc_wo	diff[3]
N_{E1}=1000, N_2=400, N_3=50						
2,000	16%	52%	21%	0.428	0.220	95%
5,000	31%	39%	15%	0.471	0.318	48%
10,000	45%	26%	10%	0.528	0.420	26%
N_{E1}=1000, N_2=100, N_3=10						
2,000	48%	36%	12%	0.279	0.205	36%
5,000	68%	19%	6%	0.357	0.309	15%
10,000	79%	11%	4%	0.445	0.414	7%
N_{E1}=200, N_2=400, N_3=50						
2,000	45%	26%	10%	0.528	0.448	18%
5,000	62%	12%	5%	0.640	0.599	7%
10,000	72%	5%	2%	0.739	0.718	3%

[1] Percent decrease in accuracy if this information source was removed. Note that these do typically not add up to 100%.

[2] N_{E2} = 50, N_{E3} = 8, Marker density = 50k.

[3] Difference between prediction accuracy with and without information from flock and relatives

CONCLUSIONS

This work shows a simple approach for modeling genomic prediction in a heterogeneous reference population by considering several subpopulations that differ in effective size. The model allows quantification of the importance of the own flock or herd information versus the wider breed information used for genomic prediction. We show that as a result of using some information from relatives, the increase of prediction accuracy with increasing the size of the wider reference population, or increasing marker density, maybe lower than expected. The validity of the approach needs to be tested with simulated as well as real data.

ACKNOWLEDGEMENTS

SHL is partly funded by the Australia Research Council (DE130100614)

REFERENCES

Clark, S.A., Hickey, J.M., Daetwyler, H.D. and van der Werf, J.H.J. (2012) *Genet. Sel. Evol*. **44**: 4.
Daetwyler, H.D., Villanueva, B. and Woolliams, J.A. (2008) *PLoS ONE* **3**(10):e3395
Goddard, M.E. (2009) *Genetica* **136**: 245.
Goddard, M.E., Hayes, B.J. and Meuwissen, T.H.E. (2011) *J. Anim. Breed. Genet.* **128**: 409.
Habier, D., Fernando, R.L. and Garrick, D.J. (2013) *Genetics* **194**: 597
Hayes, B.J., Visscher, P.M. and Goddard, M.E. (2009) *Genet. Res*. **91**:47.
Meuwissen, T.H.E., Hayes, B.J. and Goddard, M.E. (2013) **Annu. Rev. Anim. Biosci**. **1**: 221.