THE ACCURACY OBTAINED FROM REFERENCE POPULATIONS FOR GENOMIC SELECTION

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

[1]School of Rural & Environmental Science, University of New England, Armidale, NSW, 2351 Australia
[2]Australian Centre for Precision Health, University of South Australia Cancer Research Institute, University of South Australia, Adelaide, SA, 5000 Australia

SUMMARY

For the design of breeding programs it is important to understand how trait measurement translates into selection accuracy. The introduction of genomic selection has created new challenges, in particular in relation to designing reference populations and valuing information sources for their contribution to genetic gain. The accuracy of genomic prediction depends on trait heritability, the number of phenotypes used (on genotyped animals) and the 'effective number of chromosome segments' that need to be estimated. The latter parameter is challenging to estimate but can in principle be derived from the variation in relationships between the reference set and the target animal. This paper attempts to validate that theory based on real data, with the aim to develop further insight into the value of a certain reference set for the genomic prediction of a certain target animal.

INTRODUCTION

Genomic selection has become an integral part of breeding programs. The information about genetic merit obtained from genomically tested animals depends on the accuracy of the genomic test itself, and that from various other sources of information such as performance data on an animal itself and (or) its relatives. There is good selection index theory about the value of various information sources, and the accuracy of estimated merit we can expect if we combine them in a prediction framework such as Best Linear Unbiased Prediction (BLUP). However, we are still struggling to get a good handle on the information that we can expect from a genomic test. A better understanding of the components that drive the accuracy of a genomic test is important, not only for the breeder who needs to decide whether to invest in it, but also for those setting up reference populations to facilitate a higher accuracy of genomic testing. Investment in reference population occurs through individual breeders or breeder groups, breed societies, and funding bodies. It is important to be able to value the contributions of different information sources, the possible advantages of further increasing the size of the reference population and the usefulness of a certain reference set for animals with varying degrees of relationship to that reference.

The purpose of this paper is to review the theory that has been proposed to predict the accuracy of genomic prediction and to validate this theory with some examples involving real data. This might lead to a way forward on how to decide about the size and structure of reference populations and how to value them in prediction of genetic merit in the context of breeding programs.

THEORY ON THE ACCURACY OF GENOMIC TESTING

The most frequently cited formula to predict the accuracy of genomic testing comes from Daetwyler *et al*. (2008), who proposed:

$$r_{\hat{g},g} = \sqrt{\frac{h^2}{h^2 + M_e/N}} \qquad [1]$$

where h^2 is the trait heritability, N is the number of individuals with an observed phenotype as well as genotype, and M_e is the 'effective number of chromosome segments'. The formula is remarkably simple. It is based on the accuracy of estimating a random effect, which is $N/(N+\lambda)$, where λ is the ratio of the residual variance (V_e) and the variance of the effect to be estimated. Under a polygenic model quantitative trait loci (QTL) are spread across the whole genome, each with a small effect. The variance of each independent chromosome segment is the V_A/M_e, where V_A is the additive genetic variance. When estimating one segment at a time then V_e is approximately equal to the phenotypic variance and $\lambda \cong M_e/h^2$, such that (1) is equivalent to $N/(N+\lambda)$. This will give a slight underestimation of accuracy if all segments are estimated jointly and $V_e < V_p$.

Further papers by Goddard (2009) and Goddard *et al*. (2011) have refined the theory, e.g. by accounting for lower density marker panels, where the LD between markers and QTL is insufficient such that the proportion of the genetic variance 'captured by markers' is $b = M/(M_e + M)$, where M is the number of genetic markers, and $r_{\hat{g},g} = \sqrt{bh^2/(h^2+M_e/N)}$. Note that with very many markers b approaches 1. For a given M_e and high values of b, there is limited dispute about predicting genomic accuracy. However, approximations for M_e vary widely, and various formulae have been presented all leading to quite different results (Table 1). In fact, variation between predictions of genomic accuracy almost entirely depend on the approximation of M_e.

Table 1. Predicted accuracy of genomic test ($r_{\hat{g},g}$), assuming 2500 observations (N), heritability h^2=0.30, Effective population size N_e = 250; average chromosome length L=1; number of chromosomes k=30, and number of markers M=50,000

	Reference and approximation for M_e				
	Daetwyler *et al.* 2008	Goddard 2009	Goddard *et al.* 2011	Meuwissen *et al.* 2013	Lee *et al.* 2017
Parameter[1]	$2N_eLk$	$2N_eLk/\ln(4N_eLk)$	$2N_eLk/\ln(N_eL)$	$2NeLk/\ln(2N_e)$	Eq(11)
Me	15000	1455	2717	2414	611
$b= M/(M_e + M)$	1.00	0.97	0.95	0.95	0.99
$\lambda = M_e/h^2$	50000	4991	9548	8434	2060
$\sqrt{(N/(N+\lambda))}$	0.22	0.58	0.46	0.48	0.74
$r_{\hat{g},g}$	0.22	0.57	0.44	0.47	0.74

[1] M_e = Effective number of chromosome segments; b= Proportion of genetic variance captured by markers; λ = variance ratio of residual and that of one chromosome segment; $\sqrt{(N/(N+\lambda))}$ is accuracy for b=1.

In the theory described so far the approximations of M_e assume the reference as a homogenous population where all individuals are more or less equally related to each other. However, genomic predictions are more accurate if the genomic relationship between the target animal and the reference population is higher (Habier *et al*. 2007; Clark *et al.* 2012). Van der Werf *et al.* (2015) noted that most reference populations are heterogeneous in their relationship towards the target animals they predict, i.e. some individuals in the reference are much more related to the target individual than others. They demonstrated in a simple model how a small group of more related individuals can contribute more information than a very large group of distantly related individuals. Heterogeneity also exists if the reference population consists of different breeds or crossbreds. Wientjes *et al*. (2015) have proposed deterministic prediction methods to accommodate information from different populations, where they also account for genetic correlations between populations being less than one.

The variation in relatedness is often hard to predict in advance in real world examples, and a pragmatic approach can be taken by looking at the variation in realised genomic relationships between the members of the reference population and the target individual to be predicted (Goddard *et al*. 2011). This 'empirical' M_e value derived from variation in genomic relationships implies that the M_e parameter is related to the data set used for genomic prediction rather than being a population parameter, e.g. related to a certain breed. Lee *et al.* (2017) showed via simulation of a full sib population structure that the variation in genomic relationship ($var(g_{ij})$) gives a reliable estimate of M_e as $M_e = 1/var(g_{ij})$. Using this M_e value in the Daetwyler formula gave satisfactory approximations of accuracy. However, calculating M_e from variation in relationships seemed to over predict the accuracy of a genomic test when simulating a typical nucleus breeding program with a nested full-sib/half sib design across multiple generations (Jack Dekkers, pers. comm). Van den Berg *et al*. (2019) also found over prediction when applying it to simulated and real data from mixed breeds of dairy cattle.

VALIDATING THEORY WITH EMPERICAL RESULTS

It is difficult to validate the genomic prediction theory in real data based on outcomes of industry genetic evaluations such as BREEDPLAN or LAMBPLAN because these are based on so-called single-step models where information via genomic relationships is combined with information through pedigree relationships. Moreover, these evaluations are based on multiple trait models where information from correlated traits is included in the estimated breeding value (EBV). To quantify the accuracy of the genomic test in a more designed way we compared the prediction of genomic breeding value accuracy for three different traits, with varying heritability, and using the same reference population and two different validation sets. We derived M_e from the variance in relationships (Lee *et al*. 2017) of the off-diagonal block of the genomic relationship matrix, i.e. between animals in the reference and animals in the validation set, and derived the predicted accuracy using [1]. The reference population consists of 5000 animals from multiple breeds from the CRC information Nucleus and MLA reference flocks. The validation population refers to 300 purebred merinos and 300 crossbred Border Leicester x Merino crosses. Predicted accuracies were compared with empirical accuracies derived from the correlation between predicted genomic breeding values and adjusted phenotypes of animals in the validation set, divided by h. Results are shown in Table 2.

The results show an obvious overestimation of the accuracy when using the variation in relationships to estimate the M_e value. A likely reason is that the reference population consists of multiple breeds, giving a much larger variation in relationship relative to using a purebred reference. Note that the accuracy is evaluated after correction for breed effects, i.e. it is a within breed accuracy. An accuracy 'across breeds' is much larger as from genotype data `it is relatively easy to predict differences between breeds, or genetic groups within breeds. A next step is therefore to correct the G-matrix for effects of population structure by taking out a number of principal components, i.e. using $G^* = G - \Sigma E_i E_i' d_{i,}$ where d_i is an eigenvalue of G and E_i is the associated eigenvector. Further testing can also occur using purebred reference populations, although such populations can still have an underlying group structure that needs to be taken into account. Van den Berg *et al*. (2019) also concluded that the variance in genomic relationships overestimated the accuracy, when they compared reference populations with various numbers of individuals from different breeds. They proposed an alternative method that seemed to be useful to predict accuracy from reference populations from combining breeds. However, there is also a need to evaluate the value of adding within breed cohorts to the reference, where these cohorts may vary in their relationship to the animals that are targeted in prediction.

Table 2. Realized genomic prediction accuracy and theoretical accuracy predicted from variation in relationships and effective number of chromosomes (M_e) for two validation sets and using a multi-breed reference population[1]

Test Set	Var(g_{ij})	Me=1/Var(g_{ij})	Predicted accuracy[2]	Realized accuracy[3]
BL x Merino	0.001989	502.7	0.86	0.21
Merino	0.001840	543.6	0.85	0.29

[1] Using a multi-breed reference set of N = 5000 animals, trait is post weaning weight; $h^2 = 0.28$

[2] Accuracy predicted using the Daetwyler formula [1] and the estimated value for M¬e¬.

[3] Realized accuracy is correlation between predicted genomic breeding value and observed phenotype (corrected for fixed effect), divided by the square root of heritability.

CONCLUSIONS

Further work is needed to validate the theory of deriving genomic prediction accuracy from the variation in genomic relationships, and to put a value on adding particular information sources to the reference population for genomic prediction. Although this approach requires a matrix with realised genomic relationships, it provides information about the contribution of various information sources, and this may be used to predict contributions of future cohorts. Moreover, this approach is flexible and can allow animals from multiple breeds or crossbreds.

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**: 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 Dekkers J.C.M. (2007) *Genetics* **177**: 2389.
Meuwissen T., B. Hayes and M. Goddard (2013) *Annu. Rev. Anim. Biosci*. **1**: 221.
Lee S. H., Clark S.A. and van der Werf J. H. J. (2017) *PLoS One* **12**: e0189775.
Van der Werf J.H.J, Clark S.A., Lee S.H. (2015) *Proc. Assoc. Advmt. Anim. Breed. Genet*. **21**:
Van den Berg I., Meuwissen T.H.E., MacLeod I.M. and Goddard M.E. (2019). *J. Dairy Sci*. **102**:1.
Wientjes Y.C.J., Bijma P., Veerkamp R.F. and Calus M.P.L. (2016) *Genetics* **202**: 799.