322. Finetuning hyper-parameters increases the prediction accuracy in single-step genetic evaluation

M. Neshat[1,2,4*], Md.M. Momin[1,2,3,4], B. Truong[1], J.H.J. van der Werf [5], S. Lee[6] and S.H. Lee[1,2,4*]

[1]Australian Centre for Precision Health, Univ. of South Australia Cancer Research Institute, University of South Australia, Adelaide, SA, 5000, Australia; [2]UniSA Allied Health and Human Performance, Univ. of South Australia, Adelaide, SA, 5000, Australia; [3]Department of Genetics and Animal Breeding, Faculty of Veterinary Medicine, Chattogram Veterinary and Animal Sciences University (CVASU), Khulshi, Chattogram, 4225, Bangladesh; [4]South Australian Health and Medical Research Institute (SAHMRI), University of South Australia, Adelaide, SA, 5000, Australia; [5]School of Environm. and Rural Science, Univ. of New England, Armidale, NSW, Australia; [6]Div. of Animal Breeding and Genetics, National Institute of Animal Science (NIAS), South Korea; mehdi.neshat@unisa.edu.au; hong.lee@unisa.edu.au

Abstract

Single step genomic best linear unbiased prediction (HBLUP) has been widely used in livestock breeding. The HBLUP method (e.g. BLUPf90) requires hyper-parameters to combine genomic and pedigree relationships and these should be adequately initialised to maximise the accuracy of genomic prediction. In this study, we assess the performance of HBLUP, using various values of hyper-parameters in simulated genomic data. We show that the tuning parameter (tuning GRM relative to the pedigree-based numerator relationship matrix) considerably increases prediction accuracy, confirming previous studies. The scale factor, α, which scales the allele effect size by its frequency, also affects accuracy and the optimal scale factor can vary for each trait. In conclusion, fine-tuning the hyper-parameters of HBLUP is necessary to maximize prediction accuracy and the scale factor should be considered.

Introduction

Single step genomic best linear unbiased prediction uses a H-matrix that is a harmonised matrix of a pedigree-based numerator relationship matrix (NRM) and a genomic relationship matrix (GRM). The method is known as HBLUP. The H-matrix allows us to use the information of non-genotyped individuals in genomic prediction, using a data augmentation technique (see Legarra *et al.* 2009; Misztal *et al.* 2009). HBLUP has been widely used in the genetic evaluation of livestock (McMillan *et al.* 2017).

In HBLUP, there are several (hyper) parameters that can determine the performance of HBLUP.

Firstly, blending is important because it ensures GRM being a positive definite matrix (VanRaden 2008) thereby avoiding numerical problems in HBLUP (Legarra *et al.* 2009). Secondly, tuning is important because it adjusts the scale of GRM relative to that of NRMbefore inverting. Given that GRM is based on samples in the last few generations (genotyped individuals) whereas pedigree has been recorded from the founders, tuning can correct for this scale difference (Legarra *et al.* 2009; Miztal *et al.* 2009). Thirdly, parameters required to construct GRM may be important. A pairwise genomic relationship is the product of scaled genotypic coefficients of two random individuals (VanRaden 2008; Yang *et al.* 2010). Speed *et al.* (2012) generalised these forms, introducing a scale factor (α) that can determine the genetic architecture of a complex trait (aka heritability model).

In this study, we investigate the hyper-parameters required to estimate GRM to see how they affect the accuracy of HBLUP. First, various values of tuning and blending hyper-parameters were applied and compared to assess the performance of HBLUP, using simulated data with various scenarios of the historical population. In the analyses, we used the direct average information algorithm (Lee *et al.* 2006; Yang *et al.*

R.F. Veerkamp and Y. de Haas (eds)

DOI: 10.3920/978-90-8686-940-4_322, © M. Neshat *et al.* 2022

2011) that is robust to non-positive definite GRM so that we can assess all sorts of blending values including blending = 1. This is followed by exploring various values for the α of GRM that can determine the genetic architecture of a complex trait (heritability model).

Materials & methods

Simulated data. QMSim software (Sargolzaei *et al.* 2009) was used to simulate the historical population for 100 generations. In each generation, 50 males and 500 females were randomly selected and mated, generating 1000 offspring. Nine thousand biallelic markers in total were simulated, which were equally distributed across 30 chromosomes. Phenotypes were simulated for a complex trait, varying the scale factor (α=0 or -0.5) (Gowane *et al.* 2019). For the HBLUP analyses, the last five generations were recorded for phenotype and pedigree (n=5,000) and the last two generations were genotyped (n=2,000). To assess HBLUP accuracy, 1000 individuals were randomly selected from the last two generations as the target dataset, and the remaining 4,000 samples (1000 genotyped and 3,000 ungenotyped) were used as the discovery dataset.

Bending and tuning of GRM in HBLUP. The blending process can adjust the GRM to be a positive definite matrix (Legarra *et al.* 2009, Misztal *et al.* 2009). The adjusted GRM, referred to as $G_{blended}$, can be expressed as:

$$G_{blended} = \theta G + (1-\theta)A_{22} \quad \forall\ 0 \le \theta \le 1 \tag{1}$$

where θ is a positive coefficient to achieve a balance between GRM and the part of NRM (A_{22}) that is corresponding to the numerator relationships between the genotyped individuals.

Subsequently, the tuning process can be applied to adjust the scale of the GRM relative to that of the NRM. Following Legarra *et al.* (2014), the adjusted GRM is tuned as

$$G_{tuned} = \beta G_{blended} + \omega \mathbf{J} \tag{2}$$

where J is a matrix with the same size of GRM, and all elements are equal to one, and ω and β are computed as:

$$\omega = \frac{(I A_{22}\, I - I \hat{G} I)}{n^2} \qquad \beta = \frac{\frac{\left[\sum_{i=1}^{n} A_{22_{i,i}} - I' A_{22} I\right]}{n}}{\frac{\left[\sum_{i=1}^{n} G_{i,i} - I' G I\right]}{n}} \tag{3}$$

where I is an array with the size of $n \times 1$ and all values equal to one.

Genomic relationship matrix (GRM) and scale factor (α). It is assumed that the relationship between genetic variance and allele frequency can change depending on the evolutionary forces such as selections, mutation, migrations, and genetic drift. In the following equation, we can see that the variance of the ith genetic variant (v_i) is a function of the allele substation effect (β_i) and the allele frequency (p_i), which can be written as (Momin *et al.* 2021)

$$Var(v_i) = 2p_i(1-p_i)^{1+2\alpha} \times {\beta_i}^2 \tag{4}$$

where α is the scale factor that can determine the relationship between genetic variance and allele frequency, i.e. the genetic architecture of a complex trait. In the infinitesimal model (Falconer and Mackay 1996), α is assumed to be zero for all traits. An alpha value of -0.5, assumes that the genetic variance of the causal variant has a uniform distribution across tits minor allele frequency spectrum. However, it has been

reported that the value of α can vary, depending on trait and population (Momin *et al.* 2021; Speed *et al.* 2017; Speed *et al.* 2012). The generalised form of the GRM with the hyper-parameter of alpha (Speed *et al.* 2012) can be written as

$$G_{ij} = \frac{1}{d}\sum_{l=1}^{L}\left[(x_{il} - 2p_l)(x_{jl} - 2p_l)\right][2p_l(1-p_l)]^{2\alpha} \quad (5)$$

where G_{ij} is the genomic relationship between the i^{th} and j^{th} individuals, the number of SNPs is L and d is the expected diagonals and computed as:

$$d = L \cdot \mathbb{E}[(x_{il} - 2p_l)^2[2p_l(1-p_l)]^{2\alpha}] \quad (6)$$

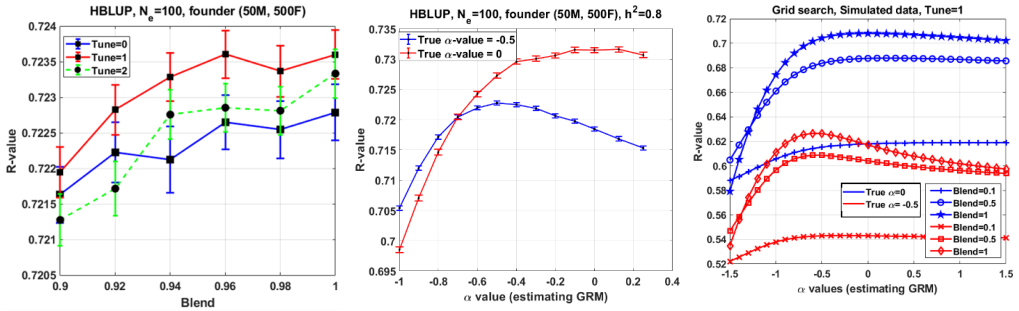
Statistical models. In this paper, we used genome-based restricted maximum likelihood (GREML) and HBLUP, based on linear mixed models, to predict individual breeding values (Gao *et al.* 2012).

Results

Figure 1a shows that the tuning process significantly improves the prediction accuracy (Pearson corelation coefficient (R-value) between true and estimated breeding value), confirming previous studies. Blending with θ<1 is not really improving the performance of HBLUP. The tuning parameter (tune=1, Eq. 3) based on Legarra *et al.* (2009) performed better than tune=2 (Eq. 3 except for β=1). Figure 1b shows that the choice of the α value (scale factor) is important and an optimal α can improve the prediction accuracy. As expected, the highest prediction accuracy is achieved when using the true α value. As shown in Figure 1c, the best configuration of the hyper-parameters could be obtained, using a grid search that considered blend, tune and alpha (scale factor) simultaneously.

Discussion

A blending strategy was not effective in improving the accuracy of HBLUP for various historical population scenarios (result not shown) although one study reported that blending could increase the prediction accuracy in dairy cattle (Gao *et al.* 2012). The second observation was that the tuning methods on the H-matrix could significantly increase the prediction accuracy. We compared three tuning methods



(a) **(b)** **(c)**

Figure 1. HBLUP accuracy and hyper-parameters. (a) The HBLUP accuracy (R-value) improves when using tune=1 (Equation 3) or tune=2 (Equation 3 except β=0 Vitezica *et al.* 2011). However, blending (θ<1) would not increase the accuracy for this simulated dataset. The error bars are 95% CI over 3,000 replications. (b) The HBLUP accuracy increases when using the true α values used in the simulation. The error bars are 95% CI over 3,000 replications. (c) The HBLUP accuracy for a single simulation replicate using a grid search method where the prediction accuracy was measured from 5-fold cross validation.

including tune=1 (based on Legarra's method (2009)), tune=2 (based on Vitezica *et al.* 2011), and tune=0 (i.e. without tuning). It is well known that alpha can vary across populations and traits (Momin *et al.* 2022). We show that the effect of α on the HBLUP accuracy was considerable. Our findings confirm that fine-tuning the hyper-parameters of HBLUP is necessary where the scale factor, a novel hyper-parameter in the context of HBLUP, should be considered. It is recommended that a grid search or similar optimisation algorithm should be used to find the best configuration for these hyper-parameters including blending, tuning and alpha.

Acknowledgements

This study is supported by Cooperative Research Program for Agriculture Science and Technology Development (PJ0160992022) from the Rural Development Administration, Republic of Korea.

References

Gao H., Christensen O.F., Madsen P., Nielsen U.S., Zhang Y., Lund M.S., and Su G., (2012). *Genet Sel Evol*, *44*(1): 1-8. https://doi.org/10.1186/1297-9686-44-8.

Gowane G.R., Lee S.H., Clark S., Moghaddar N., Al-Mamun H.A., and van der Werf, J.H. (2019). J Anim Breed Genet, 136(5): 390-407. https://doi.org/10.1111/jbg.12420

Lee S.H., and Van Der Werf J.H. (2006). *Genet. Sel. Evol*, *38*(1): 1-19. https://doi.org/10.1186/1297-9686-38-1-25

Legarra A., Aguilar I. and Misztal I. (2009). *J Dairy Sci*. 92(9): 4656-63. https://doi.org/10.3168/jds.2009-2061

McMillan A.J., and Swan A.A. (2017). In *Proc Assoc Advmt Anim Breed Genet*. Fitzroy, Australia

Misztal I., Legarra A., and Aguilar I. (2009). *J Dairy Sci*, 92(9): 4648-4655. https://doi.org/10.3168/jds.2009-2064

Momin M.M., Shin J., Lee, S., Truong B., Benyamin B., and Lee S.H. (2021). *bioRxiv*. https://doi.org/10.1101/2021.09.16.460619

Sargolzaei M. and Schenkel F.S. (2009) *Bioinformatics*, 25(5): 680–681. https://doi.org/10.1093/bioinformatics/btp045

Speed D., Cai N., Johnson M.R., Nejentsev S. and Balding D.J., (2017). *Nature Genetics*, 49(7): 986-992. https://doi.org/10.1038/ng.3865

Speed D., Hemani G., Johnson M.R. and Balding D.J., (2012) *Am J Hum Genet*. 91(6): 1011-1021. https://doi.org/10.1016/j.ajhg.2012.10.010

VanRaden P. M. (2008). *J Dairy Sci*, 91(11): 4414-4423. https://doi.org/10.3168/jds.2007-0980

Vitezica Z.G., Aguilar I., Misztal I., and Legarra, A. (2011). *Genet Res*, *93*(5): 357-366. https://doi.org/10.1017/S001667231100022X

Yang J., Benyamin B., McEvoy B.P., Gordon S., Henders A.K., *et al.*, 2010. *Nature Genetics*, *42*(7): 565-569. https://doi.org/10.1038/ng.608