Proceedings of the World Congress on Genetics Applied to Livestock Production, 11. 406

Stronger measures of genomic connectedness enhance prediction accuracies across management units

[1]Haipeng Yu, [1]Matthew L. Spangler, [1]Ronald M. Lewis, & [1,2]Gota Morota

[1]Department of Animal Science, University of Nebraska-Lincoln, Lincoln, Nebraska 68583
[2] morota@unl.edu (Corresponding Author)

Summary

Genetic connectedness assesses the extent to which estimated breeding values can be fairly compared across management units. Ranking of individuals across units based on best linear unbiased prediction (BLUP) is reliable when there is a sufficient level of connectedness due to a better disentangling of genetic signal from noise. Although a recent study showed that genomic relatedness strengthens the estimates of connectedness across management units compared to that of pedigree, the relationship between connectedness measures and prediction accuracies has been explored only to a limited extent. In this study, we examined whether increased measures of connectedness led to higher prediction accuracies evaluated by a cross-validation based on computer simulations. We found that the greater extent of connectedness enhanced accuracy of whole-genome prediction. The use of genomic information resulted in increased estimates of connectedness and improved prediction accuracies compared to those of pedigree-based models, especially when the numbers of markers and QTLs were large.

Keywords: cross-validation, genomic connectedness, genomic prediction, relatedness

Introduction

Genetic connectedness quantifies the extent of risk associated with the comparisons of estimated breeding values (EBV) across management units (Foulley et al., 1990). Best linear unbiased prediction (BLUP) of EBV can be fairly compared across units in the presence of a sufficient level of connectedness. On the other hand, an insufficient level of connectedness increases the risk of potential uncertainty in EBV comparisons when selecting individuals across units due to imperfect uncoupling of genetic signal from noise. Use of genomics can affect genetic evaluations by determining if EBV can be safely compared across management units and through developing prediction equations. In the former context, Yu et al. (2017) employed three measures of connectedness to examine the extent to which genomic information increases the estimates of connectedness. They found that the use of genomic relatedness improved genetic connectedness measures across management units compared to the use of pedigree relationships. However, it remains an open question as to whether increased connectedness observed by genomic relatedness also leads to increased prediction accuracy of genetic values across management units. The objectives of this study were to examine how choice of relationship matrices impact the estimates of connectedness under various simulated scenarios and to assess the relationship between connectedness level and

genome-enabled prediction accuracy. In addition, a guideline with respect to a sufficient level of connectedness is discussed.

Materials and Methods

Data simulation: Two replicates of genotypes and phenotypes were simulated by the QMSim software (Sargolzaei, 2009) with details summarized in Figure 1. Three phenotypes with heritability levels of 0.2, 0.5 and 0.8 were simulated with phenotypic variance of 1.0. Genotypic data mimicked the bovine genome were simulated for individuals (n = 2,530) in generations 3 to 10 coupled with 5,000 or 50,000 biallelic SNP markers evenly distributed across 29 pairs of autosomes with chromosome length of 2,333 cM. Additionally, 29 or 1,015 randomly distributed QTLs were simulated: the former is equivalent to one QTL per chromosome and the latter corresponds to 35 QTLs per chromosome. Only SNPs but not QTLs were used to infer measures of connectedness and to assess accuracy of prediction.

Management units simulation: The management units were simulated in two steps following Yu et al. (2017) (Figure 2). First, individuals were clustered into 10 distinctive groups by performing the k-medoid algorithm based on the $\mathbf{A}$ matrix and then clusters were assigned to management units. In total, six scenarios were considered in this study. In scenario 1, a completely disconnected design was simulated by assigning individuals within clusters 1 to 5 into management unit 1 (MU1) and clusters 6 to 10 into management unit 2 (MU2). Scenarios 2 to 6 varried from partially connected to completely connected designs, where the degree of genetic link was gradually increased by exchanging 10%, 20%, 30%, 40% and 50% of randomly sampled individuals between MU1 and MU2.

Genetic connectedness: The generalized coefficient of determination (CD) measures the precision of EBV (Laloë et al., 1996) and was also used in Yu et al. (2017). CD penalizes connectedness measurements if the genetic variability is too small within population. A summary CD of contrast between any management unit is defined as $\text{CD}(\mathbf{x}) = 1 - \lambda(\mathbf{x}'\mathbf{C}^{22}\mathbf{x})/(\mathbf{x}'\mathbf{K}\mathbf{x})$, where $\mathbf{K}$ is the relationship matrix and $\mathbf{x}$ is the vector of contrast (Laloë et al., 1996). For instance, a pair-wise comparison between i'th and j'th management units with $n_{i'}$ and $n_{j'}$ individuals, the contrast vector $\mathbf{x}$ will be set as $1/n_{i'}$, $-1/n_{j'}$ and 0 corresponding to individual belonging to i'th, j'th, and remaining units. A larger value suggests a stronger estimate of connectedness among management units.

Relationship matrix: Any kind of (semi) positive definite matrices can be used to define $\mathbf{K}$ (Morota, 2014). We used three types of $\mathbf{K}$ in this study constructed from different sources. The numerator relationship matrix ($\mathbf{K} = \mathbf{A}$) measures the expected additive genetic relationship coefficient between individuals on the basis of pedigree information. The construction of $\mathbf{A}$ matrix was based on tracing all individuals extending over 10 generations. In contrast, a genomic relationship matrix ($\mathbf{K} = \mathbf{G}$) measures the molecular similarity among individuals (VanRaden, 2008). One item that needs to be addressed when the $\mathbf{A}$ and $\mathbf{G}$ matrices are compared is that they are not on the same scale. The following $\mathbf{K} = \mathbf{G}^*$ matrix rescales $\mathbf{G}$ to the same base population as in $\mathbf{A}$ by adjusting the inbreeding coefficient level in $\mathbf{G}$ similar to that of $\mathbf{A}$, $\mathbf{G}^* = (1 - \bar{F})\mathbf{G} + 2\bar{F}\mathbf{J}$, where $\bar{F}$ and $\mathbf{J}$ refer to the average inbreeding coefficient of whole population in the $\mathbf{A}$ matrix and the n $\times$ n square matrix filled with 1, respectively (Powell et al., 2010).

Whole-genome prediction model: The relationship between connectedness and prediction accuracy was investigated with a standard BLUP model, $\mathbf{y} = 1\boldsymbol{\mu} + \mathbf{g} + \boldsymbol{\epsilon}$ where $\mathbf{y}$, $\boldsymbol{\mu}$, $\mathbf{g} \sim N(0, \mathbf{K}\sigma_g^2)$ and $\boldsymbol{\epsilon} \sim N(0, \mathbf{I}\sigma_\epsilon^2)$ refer to a vector of observed phenotypes, intercept,

random additive genetic effects, and residuals, respectively. The variance components σ_g^2 and σ_ϵ^2 represent variance of additive genetic effects and residual variance, respectively. The model was treated under a Bayesian framework according to Pérez (2014). The prediction accuracy was evaluated by two-fold cross-validation (CV), where two management units were treated as the training and testing sets. The variance components were inferred from the data and the predictive ability of the model was calculated as the Pearson correlation between predicted genetic values and observed phenotypes in the testing set. A Gibbs sampler was run for 10,000 iterations, where the first 2,000 samples were discarded as burn-in. A total of 8,000 samples coupled with a thinning rate of 5 were used to infer posterior means.

Results

The change of prediction accuracies with the increasing proportion of linked individuals quantified with CD of contrast is shown in Figure 3, where larger CD values suggest stronger connectedness. In general, the prediction accuracy improved when more individuals from the same clusters were assigned across units. Within each scenario, the estimates of CD increased up to scenario 3, but decreased from scenario 4 onward because CD penalized connectedness measures for reduced genetic variability. In Figure 3A with 29 QTLs and 5,000 markers, higher prediction accuracies were observed by using the **A** matrix than those using **G** for most scenarios. An exception was observed when **G** produced higher prediction accuracy than that of **A** in scenarios 5 and 6 for $h^2 = 0.5$ and in scenarios 4, 5,and 6 for $h^2 = 0.2$. Measures of **A**-based connectedness were similar or stronger than those of **G** for Scenarios 1, 2, 3, and 4. An analogous tendency was identified in Figure 3C with 1,015 QTLs and 5,000 markers, where slightly increased prediction accuracies and estimates of connectedness were observed with **A**. In contrast, with 29 QTLs and an increased number of markers (50,000), prediction accuracies were increased and the **G** yielded consistently larger measures of connectedness than those of the **A** (Figure 3B). Overall, **G** and $\mathbf{G}^*$ presented stronger estimates of connectedness and similar or higher prediction accuracies than those of **A**. Clearer differences were observed when increasing the number of QTLs to 1,015 (Figure 3D). Unlike other cases, the **G** matrix clearly yielded higher estimates of connectedness as compared to **A**. The performances of $\mathbf{G}^*$ were very similar to those of **G** in CD across all cases.

Discussion

We used contrasts of CD to investigate the relationship between connectedness and prediction accuracy. Prediction accuracy improved with increased capturing of connectedness between units. In general, prediction accuracy improved as more markers were used to infer a genomic relationship matrix and as more QTLs contributed to the genetic variation. These can be attributed to the fact that 1) the greater the number of markers, the better capturing of QTL relationships among individuals (Ober et al., 2012) and 2) genomic best linear unbiased prediction (GBLUP) performs better when the number of QTLs is large, because of its infinitesimal model assumption (Daetwyler et al., 2010). This result may change when an alternative whole-genome prediction model is used instead of GBLUP. For instance, a Bayes B type of model performs well when the number of QTL is small (Daetwyler et al., 2010). Measures of connectedness increased as more markers were used to characterize connectedness. Increased estimates of connectedness was also observed when there were more QTLs

although the number of SNPs remained constant (Figures 3A vs. 3C). This is likely because more accurate estimates of variance components were obtained when the number of QTLs is large or an increased number of QTLs will contribute to the possibility that a maker captures QTL effects. The extent of connectedness measured by CD and prediction accuracy from BLUP were non-linear as the proportion of individuals exchanged between the two units increased. This is because CD penalizes connectedness estimates when the amount of genetic variability across units was small. In general, **A**-based connectedness measures yielded higher connectedness and prediction accuracy than those of **G**- and $\mathbf{G}^*$-based connectedness when the numbers of QTLs and SNPs were small. However, the difference between them became more and more negligible when a h^2 was small. When the numbers of QTLs and SNPs were large (Figure 3D), the **G** and $\mathbf{G}^*$ matrices clearly outperformed that of **A** in prediction and also produced increased measures of connectedness.

In conclusion, increased connectedness measures and prediction accuracies were largely observed as more individuals from the same clusters were shared across management units. We found prediction accuracy improved with increased capturing of connectedness across units suggesting that increase in the accuracy of the EBV comparison is positively associated with increase in accuracy of CV-based prediction. The impact of genomics was more marked compared to pedigree when large numbers of markers and QTLs were used. While there is a need to establish increased levels of connectedness, simply increasing connectedness results in rapid decrease of relatedness variability which may not be desired in a breeding program. Use of CD allows us to find a connectedness level that gives a reasonable prediction accuracy while maintaining genetic diversity in a population.

List of References

Daetwyler, H. D., Pong-Wong, R., Villanueva, B., et al. (2010). The impact of genetic architecture on genome-wide evaluation methods. *Genetics*, 185(3):1021–1031.

Foulley, J., Bouix, J., Goffinet, B., et al. (1990). Connectedness in genetic evaluation. In *Advances in statistical methods for genetic improvement of livestock*, pages 277–308. Springer.

Laloë, D., Phocas, F., and Ménissier, F. (1996). Considerations on measures of precision and connectedness in mixed linear models of genetic evaluation. *Genet Sel Evol.*, 28:359.

Morota, G & Gianola, D. (2014). Kernel-based whole-genome prediction of complex traits: a review. *Frontiers in Genetics*, 5.

Ober, U., Ayroles, J. F., Stone, E. A., et al. (2012). Using whole-genome sequence data to predict quantitative trait phenotypes in drosophila melanogaster. *PLoS genetics*, 8(5):e1002685.

Pérez, Paulino & de los Campos, G. (2014). Genome-wide regression and prediction with the bglr statistical package. *Genetics*, 198(2):483–495.

Powell, J. E., Visscher, P. M., and Goddard, M. E. (2010). Reconciling the analysis of ibd and ibs in complex trait studies. *Nat Rev Genet.*, 11:800–805.

Sargolzaei, M & Schenkel, F. (2009). QMSim: a large-scale genome simulator for livestock. *Bioinformatics*, 25:680–681.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.*, 91:4414–4423.

Yu, H., Spangler, M., Lewis, R., et al. (2017). Genomic relatedness strengthens genetic

connectedness across management units. *G3: Genes, Genomes, Genetics*. Early online. doi: 10.1101/130138.

Figures

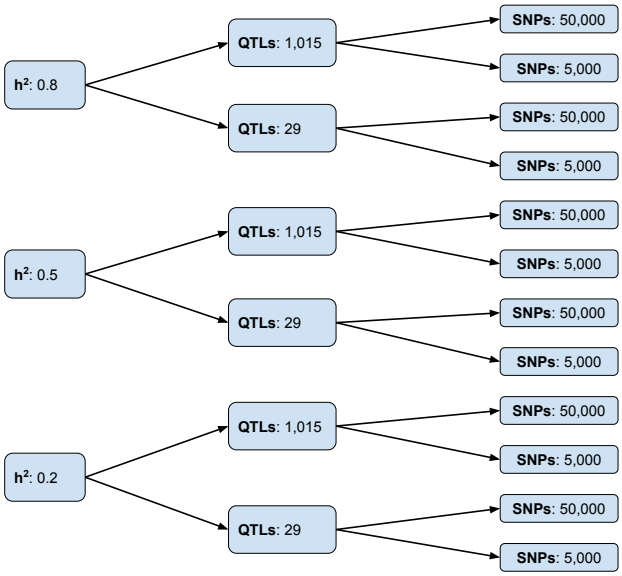

Figure 1: Genomic data simulation parameters. SNPs, QTLs and h^2 represent total single-nucleotide polymorphisms, quantitative trait loci, and trait heritability, respectively. Simulations were carried out across three different h^2 (0.8, 0.5 and 0.2), two different number of QTLs (1,015 and 29) and two different SNP densities (50,000 and 5,000).

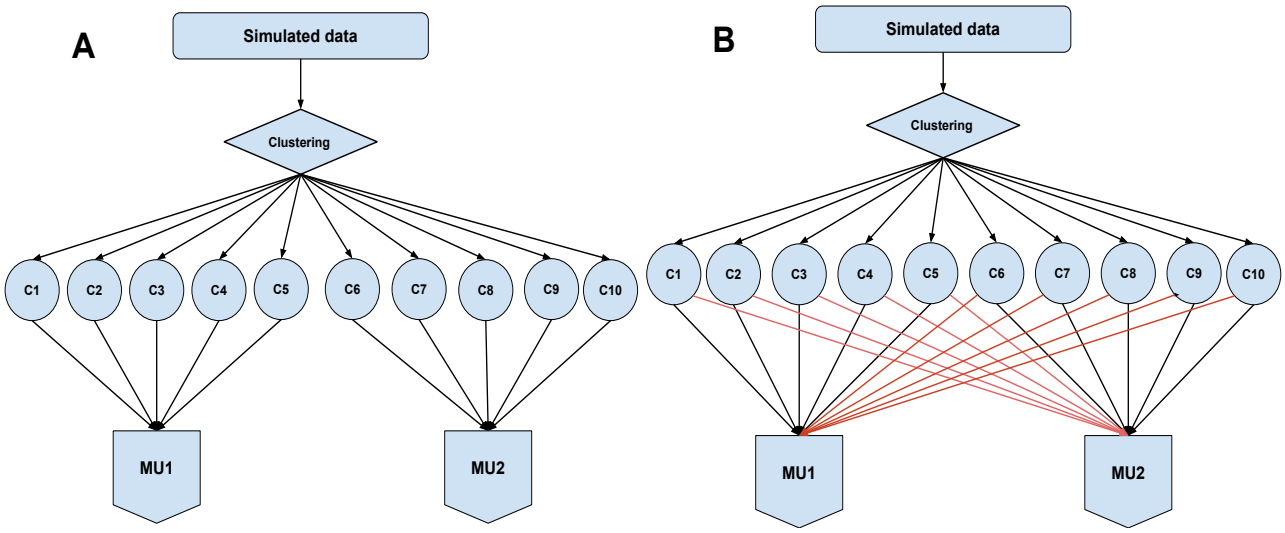

Figure 2: Management unit (MU) simulation scenarios. A: Scenario 1 (completely disconnected design). Individuals within clusters 1 to 5 were assigned to MU1 and clusters 6 to 10 were assigned to MU2. B: Scenarios 2 to 6 (partially connected to completely connected). The degree of connectedness was gradually increased by exchanging 10% (Scenario 2), 20% (Scenario 3), 30% (Scenario 4), 40% (Scenario 5) and 50% (Scenario 6) of randomly sampled individuals between MU1 and MU2. Scenario 6 corresponds to the completely connected design.

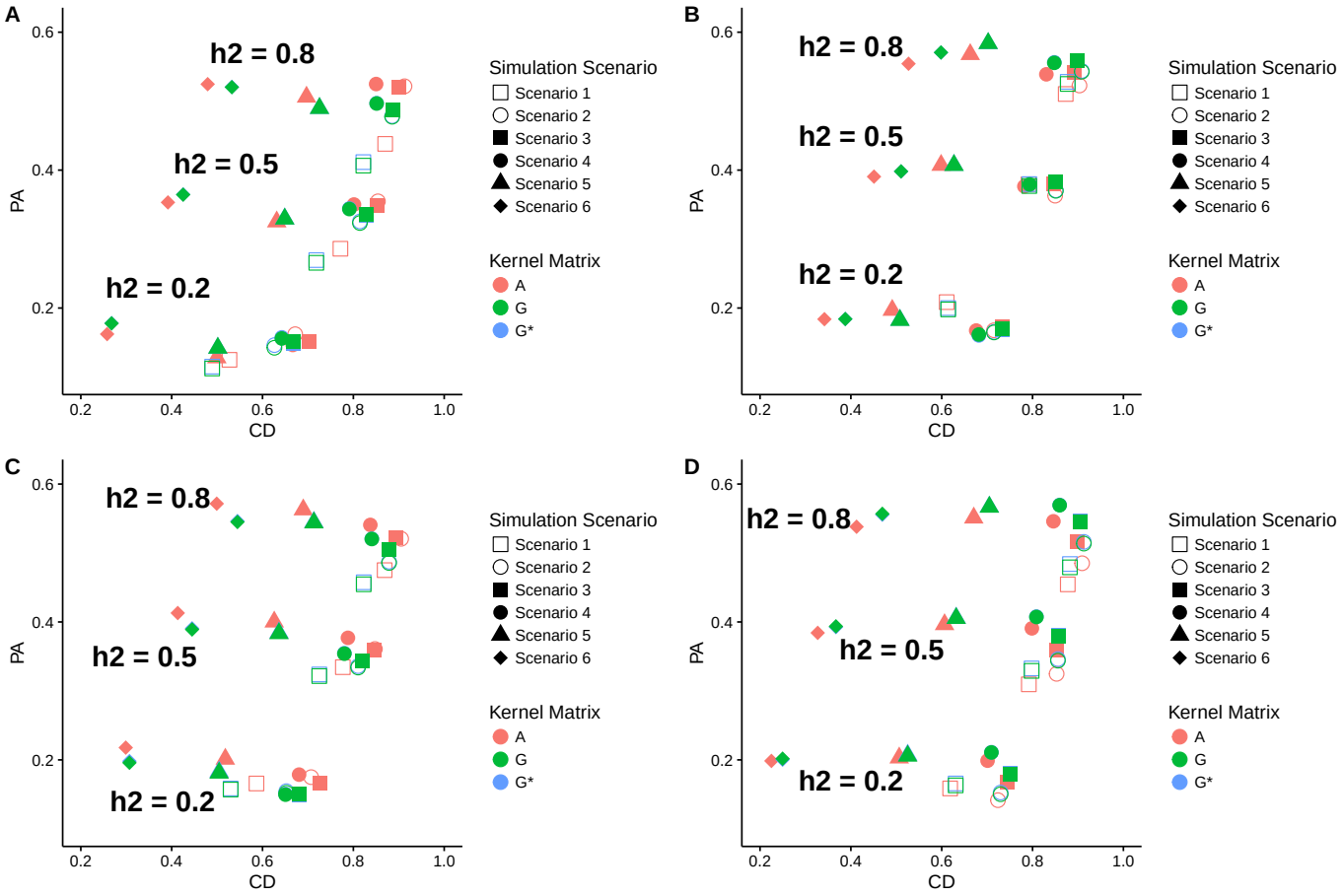

Figure 3: Relationship between connectedness and prediction accuracy. CD and PA denote coefficient of determination and prediction accuracy, respectively. PA was defined as the correlation between phenotypes and estimated breeding values $cor(\mathbf{y}, \hat{\mathbf{g}})$. Connectedness of pedigree-based $\mathbf{A}$, genome-based $\mathbf{G}$, and rescaled genome-based $\mathbf{G}^*$ within 6 management units simulation scenarios across 3 heritabilties were compared with their prediction accuracies in each graph. A: 29 QTLs and 5,000 markers. B: 29 QTLs and 50,000 markers. C: 1,015 QTLs and 5,000 markers. D: 1,015 QTLs and 50,000 markers.