Performance of Genomic Prediction using Different Multi-Allelic Genomic Relationship Matrices for Genotyping-by-Sequencing (GBS) Data

Jie Kang[1,2], Phillip Wilcox[1], Michael Black[2], Stephen Byrne[3], Dan Milbourne[3], Marty Faville[4], Joanne Jacobs[4], Ken Dodds[4]
1. Department of Mathematics and Statistics, University of Otago, Dunedin, New Zealand
2. Department of Biochemistry, University of Otago, Dunedin, New Zealand
3. Teagasc, Crop Research Centre, Oak Park, Carlow, Ireland
4. AgResearch, Invermay Agricultural Centre, Mosgiel, New Zealand

Introduction

- Genomic prediction (GP) has been frequently implemented using GBS data since its introduction almost a decade ago.
- Genomic relationship matrices (GRM) can be derived from GBS data, and used with genomic best linear unbiased prediction (GBLUP) to perform GP.
- Short haplotypes ('ShortHaps') refer to multiple variants situated in small genomic segments such as those captured within GBS reads.
- ShortHaps are more advantageous than single nucleotide polymorphisms (SNPs) because they are more likely to be strongly associated with causal variants. Multi-allelic genomic relationship matrices (MA-GRMs) constructed using shortHaps are expected to be superior.

Aim

Investigate whether performance of genomic prediction can be further improved by using multi-allelic genomic relationship matrices (MA-GRMs).

Method

Three different approaches for developing MA-GRM were assessed, alongside pedigree (A) and SNP genotype (GRM) methods, using a simulated dataset based on GBS and phenotypic data from perennial ryegrass.

MA-GRM

1. Haplotype-dosage method (Meuwissen et al., 2014)

$$MAGRM_1 = XX'/(2p(1-p))$$

where X denotes the matrix of centred haplotype score (dosage), and p denotes the haplotype allele frequencies.

2. Haplotype similarity method (Ferdosi et al., 2016)

$$MAGRM_2 = K\Gamma K'/2$$

where K is the Kronecker product between an $n \times n$ identity matrix (n = number of individuals) and a 1×2 unit matrix; Γ is the genetic relationship matrix.

3. Linkage disequilibrium method (Mathew et al., 2018)

$$MAGRM_3 = XS^{-1}X'$$

where S is a block-diagonal matrix containing the (averaged) multiallelic variance-covariance matrix of haplotypes.

MA-GBLUP

$$y = X\beta + Zu + e$$

where y is a matrix of phenotypes, X and Z are incidence matrices, β and u are fixed and random effects with $Var(u) = G\sigma^2$ and e is the error term. In particular, G can be any relationship matrix, including those listed above.

References

Meuwissen et al. (2014). Genetics Selection Evolution, 46, 1-8.
Ferdosi et al. (2016). Genetics Selection Evolution, 48, 75.
Mathew et al. (2018). Heredity, 120, 356-368.
Hess et al. (2018). Proc 11th World Congr Genet Appl to Livest Prod, 365
Pavlidis et al. (2016). Theoretical and Applied Genetics, 131, 703-720.

Simulation

- ApeKI GBS data of 96 ryegrass samples (with a known pedigree, Fig. 1) were simulated using simGBS (Hess et al., 2018), based on the ryegrass reference genome described in Pavlidis et al. (2016).
- 448,450 out of 1,866,744 virtually digested GBS tags/fragments were retained after fragment size-selection at (65, 155) bp. 206,807 tags contained SNPs (3 SNPs per tag on average) with 676,590 SNPs in total.
- 1000 QTL from a quantitative trait with heritability of 0.1, 0.4, 0.7, 0.9 were simulated (1000 replicates of each scenario), with the QTL effects and residuals drawn from standard normal distributions.

Figure 1: Pedigree used in simulation

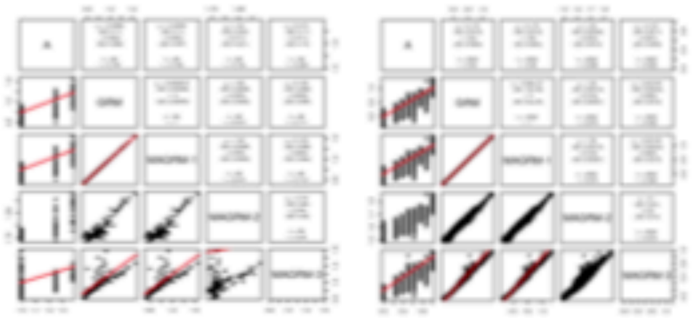

Figure 2: Pairwise comparison of diagonal (left) and off-diagonal (right) between five relationship matrices

Results & Discussion

- The pedigree (A), SNP genotype (GRM) and three haplotype-based relationship matrices (MA-GRM) returned consistent results (Fig. 2), where off-diagonals and diagonals of all relationship matrices are highly correlated (apart from the diagonals of MAGRM-3).
- Performance of GBLUP using different GRMs was almost identical (Fig. 3). However, the results may be constrained by the population size and make-up (including SNP density and linkage disequilibrium between SNPs).
- This study assumed GBS data has infinite depth - the impacts of low sequencing depth and other GBS-specific sources of variation need to be considered in the future.

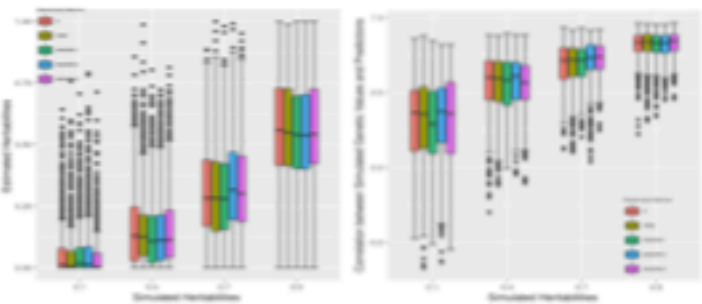

Figure 3: Performance of GBLUP using five relationship matrices 1. estimated heritabilities (left) 2. correlation between simulated genetic values and predictions (right)

Conclusion

1. There was a strong correlation between MA-GRMs and other relationship matrices.
2. The use of MA-GRMs did not improve performance of GP under current settings.

University of Otago · Teagasc · Walsh Scholarships · AgResearch · Pastoral Genomics