Genomic prediction of health traits in humans: demonstrating the value of marker selection.

Genomic prediction of health traits in humans: demonstrating the value of marker selection.

M.L. Bermingham[1], R. Pong-Wong[2], A. Spiliopoulou[1], C. Hayward[1], I. Rudan[3], H. Campbell[3], A.F. Wright[1], J.F. Wilson[3], F. Agakov[4], P. Navarro[1] and C.S. Haley[1,2]

[1]MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, [2]The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, [3]Centre for Population Health Sciences, [4]Pharmatics Limited, UK.

ABSTRACT: In this study, we explored prediction of human height, high-density lipoproteins (HDL) and body mass index (BMI) using SNPs within a Croatian (N=2,186) and into a UK population (N=810) in Bayes-C (using Gibbs sampling) and G-BLUP frameworks. Correlation between predicted and observed trait values in 10-fold cross-validation was used to assess prediction accuracy. Using all available 263,357 SNPs, Bayes-C and G-BLUP had similar prediction accuracy across traits within the Croatian data, and for height and BMI when predicting into the UK population. However, Bayes-C outperformed G-BLUP in the prediction of less polygenic HDL into the UK population. Supervised feature selection allowed G-BLUP to achieve equivalent predictive performance to Bayes-C across all three traits with greatly reduced computational effort. Feature selection in the G-BLUP framework therefore provides a flexible and efficient alternative to computationally expensive Bayes-C for traits considered in this study.

Keywords:
Phenotype prediction
Feature section
Bayes-C
G-BLUP

Introduction

In animal breeding the emphasis lies on the prediction of genetic values to facilitate selection of individuals for breeding. In the field of human genetics, on the other hand the focus is on the prediction of phenotypes to optimize population or individual level health interventions (Kemper et al. (2012); de los Campos et al. (2013)). Therefore, one might ask what that can research from human data contribute to improving the efficiency of genomic selection in livestock species? The problem of predicting the genetic value is not dissimilar to predicting disease risk in humans, indeed many recent innovations from genomic prediction in cattle and other livestock species are now finding applications in the study of human genetics (de los Campos et al. (2010)). Animal breeders can likewise gain insights from the field of human genetics.

Complex diseases in man and animals are influenced by multiple genes and environmental factors (Hill et al. (2008)). Genome-wide association studies (GWAS) have identified thousands of SNPs associated with health-related traits, and thus provide a source of information about useful predictors for these traits (Donnelly (2008)). Making the best use of genotype data from these studies has been a major focus in recent years in the fields of genomic selection and phenotype prediction (Daetwyler et al. (2010); de los Campos et al. (2013); Wray et al. (2013)). In this study, we concentrate on how to increase the efficiency of genomic predictions for complex traits. One important issue is feature selection (i.e. selection of SNPs exhibiting non-redundant information) which could reduce model complexity and computational requirements.

The objective of this study was to investigate the effect of supervised feature selection on the performance of two widely used prediction methods: Bayes-C and genomic best linear unbiased prediction (G-BLUP).

Materials and Methods

In this study the complex traits height, high-density lipoproteins (HDL) and body mass index (BMI) were predicted using genomic markers within Croatian populations and into a UK population from Orkney

Data. Measurements of height, BMI and HDL cholesterol were obtained for all study participants. The genotypes in this study were generated using a dense Illumina SNP array. Following quality control, there were 263,357 autosomal SNPs and 2,996 phenotypic records available for inclusion in the analysis, 2186 from Croatia and 810 from the UK population sample from the Orkney Isles.

Feature selection. In an attempt to remove redundancy present in the in the genotype data, haplotype blocks were identified with PLINK software (version 1.07; Purcell et al. (2007)). Definition of haplotype blocks was based on the confidence interval method using the Haploview default settings (Gabriel et al., (2002)). All SNPs were partitioned into haplotype subsets. SNPs within a block were then simultaneously fitted as explanatory variables in a linear model where the dependent variable was the phenotype, and conditional trait-specific P-values extracted. The SNPs were ranked based on intra-block conditional association P-value, and the top 100, 500, 1,000, 5,000, 10,000, 50,000, 100,000, 150,000, 200,000, 250,000 markers were selected to generate the density-specific data sets.

Statistical analyses. All markers and the different subsets were used to predict genomic values for each trait in Bayes-C and G-BLUP frameworks. Tenfold cross-validation was used to evaluate model performance. Individuals in the Croatian data were randomly assigned (without replacement) to ten test datasets of roughly equal size (approximately 219 records). The models were trained on the remaining 90% of the Croatian data (approximately 1967 records). Prediction accuracy was calculated as the correlation of estimated genomic values on the observed phenotype of individuals in the test, and replication data. Bayes-C was implemented, under a Bayesian framework using Gibbs sampling in custom-made software (Nadaf et al. (2012)) and G-BLUP was implemented in ASReml (Gilmour et al. (2009)).

Results and Discussion

Table 1. Prediction accuracy estimates with 95% confidence intervals from G-BLUP and Bayes-C following 10-fold cross-validation using all 263,357 markers within the Croatian data, and into the Orkney replication data.

	Accuracy($_{\text{95\% confidence interval}}$)	
	G-BLUP	Bayes-C
	Croatian (test) data	
Height	$0.24_{(0.20\text{-}0.28)}$	$0.26_{(0.21\text{-}0.31)}$
HDL	$0.17_{(0.14\text{-}0.20)}$	$0.21_{(0.18\text{-}0.24)}$
BMI	$0.11_{(0.07\text{-}0.15)}$	$0.12_{(0.08\text{-}0.15)}$
	Orkney replication data	
Height	$0.07_{(0.06\text{-}0.08)}$	$0.05_{(0.04\text{-}0.06)}$
HDL	$0.02_{(0.01\text{-}0.02)}$	$0.14_{(0.10\text{-}0.18)}$
BMI	$0.08_{(0.07\text{-}0.09)}$	$0.06_{(0.05\text{-}0.07)}$

Bayes-C and G-BLUP had similar prediction accuracy across all traits within the Croatian data, and for the highly polygenic traits height and BMI when predicting into the Orkney data, when all 263,357 markers were used (Table 1). However, Bayes-C outperformed G-BLUP in the prediction of HDL (which is influenced by fewer quantitative trait loci than BMI and height) into the Orkney data. These findings are in accordance with reports indicating that Bayes-C provides higher prediction accuracies for less polygenic traits, whereas G-BLUP has been reported to be more accurate for highly polygenic traits (Coster et al. (2010); Daetwyler et al. (2010); de los Campos et al. (2013)).

The computational time required for a single training fold analysis in the G-BLUP method was 0.017 days, whereas the computational time requirement of Bayes-C using Gibbs sampling for the analogous analysis was over 3,000 times greater, at 64.11 days. We have used the Gelman-Rubin diagnostic to assess convergence of the sampler (Cowles and Carlin, 1996).

The Bayes-C framework did not provide substantial improvement in prediction accuracy following feature selection, over that observed in the G-BLUP framework However, feature selection allowed G-BLUP to achieve predictive performance that was not significantly less than that of Bayes-C with greatly reduced computational effort (Figure 1A-B). Thus with 10,000 selected SNPs when predicting HDL within Croatia, or with only 100 selected SNPs when predicting within Orkney, the accuracy of G-BLUP was similar to that of the Bayes-C method when the latter was at its best, i.e. when Bayes-C was using all SNPs. This is perhaps because the removal of redundancy at the marker level based on association signals with the phenotype of interest increases the average linkage disequilibrium between markers in the selected subset and quantitative trait. We show that feature selection guided by phenotype association information in the G-BLUP framework therefore provides a flexible and more efficient alternative to computationally time-consuming Bayes-C using the Gibbs sampler for all traits considered in this study.

Conclusion

Feature selection in the G-BLUP framework provides a flexible and more efficient alternative to computationally expensive Gibbs sampling in Bayes-C for all considered traits in this study. The small effective population size of livestock species means that the application of feature selection may provide even greater improvements in efficiency of genomic predictions for the industry, than observed in this study from human data.

Acknowledgements

This research was funded by grants from the biotechnology and biological sciences research council (UK) to C.H., Medical Research Council (UK) to A.W., from the Republic of Croatia Ministry of Science, Education and Sports to I.R. (108-1080315-0302), from the Chief Scientist Office of the Scottish Government to J.F.W. (CZB/4/276, CZB/4/438, CZB/4/710), the Royal Society to J.F.W. and the European Union framework program 6 EUROSPAN project to H.C. (contract no. LSHG-CT-2006-018947).

Literature Cited

Chen, F.-C. and Li, W.-H. (2001). A. J. Hum. Genet. 68: 444-456.

Coster, A., Bastiaansen, J.W.M., Calus, M.P.L., et al. (2010). Genet. Sel. Evol. 42:9.

Cowles, M.K., Carlin B.P. (1996). J of the Am. Stat. Assoc. 91:434.

Daetwyler, H.D., Pong-Wong, R., Villanueva, B. et al. (2010). Genetics 185, 1021-1031.

de los Campos, G, Gianola, D. and Allison, D.B. (2010). Nature Rev. Genet. 11: 880-886.

de los Campos, G., Hickey, J.M., Pong-Wong, R. et al. (2013). Genetics 193: 327-345.

Donnelly, P. (2008). Nature 456: 728-731.

Gabriel, S.B., Schaffner, S.F., Nguyen, H. et al. (2002). Science 296: 2225-2229.

Gilmour, A., Gogel, B., Cullis, B. et al. (2009). ASReml User Guide Release 3.0. VSN International Ltd. Hemel Hempstead, HP1 1ES, UK.

Hill, W.G. Goddard, M.E. and Visscher, P.M. (2008). PLoS Genet. 4: e1000008.

Kemper, K.E., and Goddard, M.E (2012). Understanding and predicting complex traits: knowledge from cattle. Hum. Mol. Gen. 21:45-51.

Leroy, G. Mary-Huard, T., Verrier, E. et al. (2013). Genet. Sel. Evol. 45: 1-10.

Long, N., Gianola, D., Rosa, G. et al. (2007). J.Anim. Breed. and Genet. 124:377-389.

Nadaf, J., Riggio, V., Yu, T.-P. et al. (2012). BMC Proc. p S6.

Purcell, S., Neale, B., Todd-Brown, K. et al. (2007). A. J. Hum. Genet. 81:559-575.

Renwick, JH (1971). *Annu. Rev. Genet.* 5: 81-120.

Tenesa, A. Navarro, P., Hayes, B.J. et al. (2007). Genome Res. 17:520-526.

Wray, N.R., Yang, J., Hayes, B.J. et al. (2013) Pitfalls of predicting complex traits from SNPs. Nature Rev. Genet. 14:507-515.

Zhang, Z., Ding, X., Lui, J., et al. (2011). J. Dairy Sci. 94:3642–365

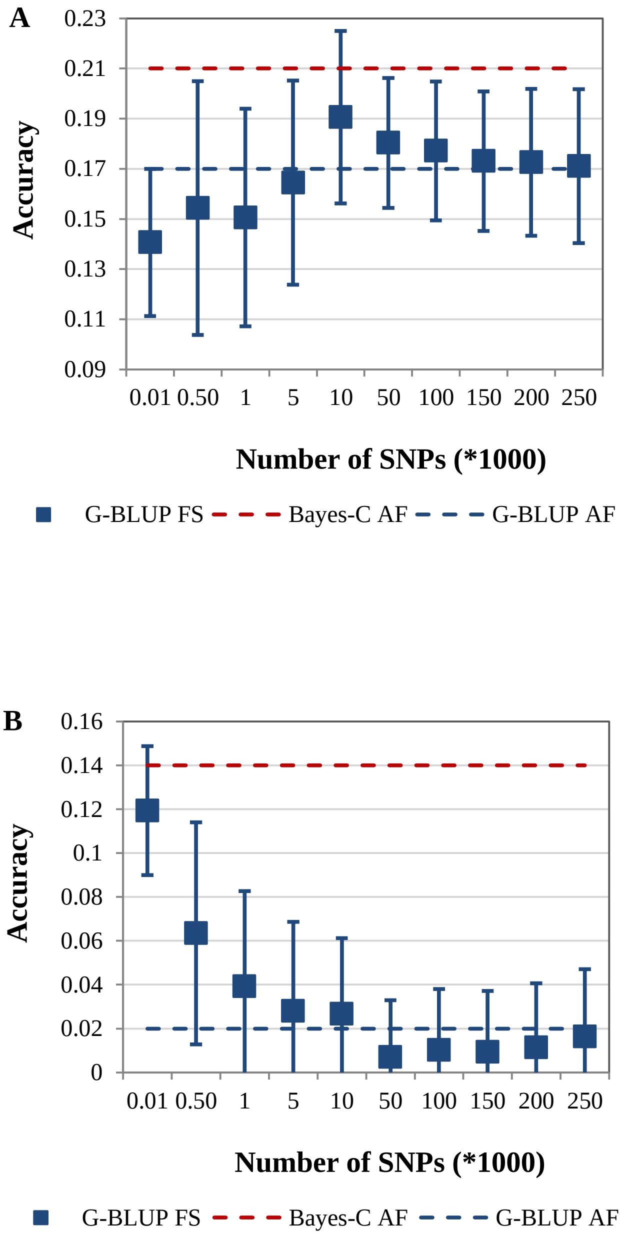

Figure 1.A-B.: Average prediction accuracy for high-density lipoprotein level (HDL) across the ten test folds in the Croatian (A) and replication ORCADES data (B) using the different marker densities selected using SNP selection in G-BLUP. The blue squares depict the mean (error bars show the 95% confidence interval) accuracy results across the ten folds from the different feature subset densities (FS). The broken red and blue lines depict the accuracy results from the full feature set of 263,357 markers (AF) in Bayes C and G-BLUP respectively.