Genomic Prediction Using Linkage Disequilibrium and Co-segregation

A.S. Leaflet R2818

Xiaochen Sun, Research Assistant; Rohan Fernando, Professor; Dorian Garrick, Professor; Jack Dekkers, Professor, Department of Animal Science, Iowa State University

Summary and Implications

A linear mixed model fitting both genome-wide co-segregation (CS) and linkage disequilibrium (LD) was developed to improve accuracy of genetic prediction for pedigreed populations of unrelated families that have half sibs represented in both training and validation. Cosegregation was modeled as the effects of genome-wide 1-centimorgan haplotypes that one individual inherits from pedigree founders through identity-by-descent, while LD was modeled as allele substitution effects of all marker genotypes. Prediction accuracy of the LD-CS method was compared to the accuracy of three LD methods – GBLUP, BayesA and BayesB, using simulated datasets of varying numbers of paternal half sib families. Results show that the LD-CS method tended to have higher accuracy than any of the LD methods. With an increase in the number of families, the accuracy of the LD-CS method persisted, while the accuracy of the LD methods dropped. The results indicate that by fitting CS explicitly, the LD-CS method has higher and more consistent prediction accuracy than LD methods.

Introduction

The accuracy of breeding values estimated based on genome-wide high-density SNP genotypes is expected to be higher than estimates obtained from pedigree-based methods, under the assumption that QTL underlying a quantitative trait are in linkage disequilibrium (LD) with SNP markers. Substantially higher accuracies with genomic prediction have been frequently observed in simulation studies, but prediction accuracies from real data of several livestock species are reported to be only slightly higher and sometimes worse than accuracies from pedigree-based methods, especially for prediction of families or breeds not well represented in the training data. One possible reason is that small effective population size of most current livestock populations' results in long-range LD within families, which is not consistent in phase or strength across families. A model based on co-segregation (CS) information can address this problem because CS models the effects of the haplotypes that an individual inherits from pedigree founders through identity-by-descent and does not explicitly rely on LD. This study developed a linear mixed model that fits both LD and CS information (LD-CS) to improve accuracy of genomic prediction for pedigreed populations that comprise unrelated families. The LD-CS method is compared to three commonly used LD methods – GBLUP, BayesA and BayesB – on simulated datasets with a half sib design and different numbers of families.

Materials and Methods

The LD-CS method models the genomic estimated breeding value (GEBV) of an individual as separate LD-GEBV and CS-GEBV. The LD-GEBV is calculated by summing up allele substitution effects from all SNP genotypes; and the CS-GEBV is calculated by summing up effects of all 1-centimorgan haplotypes that the individual inherits from pedigree founders through identity-by-descent. Bayesian inference based on Markov chain Monte Carlo was used to estimate allele substitution effects and founder haplotype effects.

Datasets of paternal half sib families from different numbers of sires were simulated to compare LD-CS with LD methods. Each sire was mated to 20 independent dams with one offspring per mating. Within each sire family, 10 half sibs were used for training to predict the EBV for the other 10 half sibs. All individuals had phenotypes for a quantitative trait with heritability 0.5 and genotypes for 2,000 SNP that evenly covered 2 chromosomes, each of 100 cM in length.

Results and Discussion

Prediction accuracies for four datasets with 1, 2, 10, or 100 half sib families are in Table 1. The LD-CS method had significantly higher accuracy than all LD methods, except for the dataset with one half sib family. With an increase in the number of families, the accuracies of LD-CS method persisted, while that of the LD methods dropped. Accuracies of LD methods that fitted more SNP in the model (GBLUP and BayesA) tended to have higher accuracies than those fitting fewer SNP (BayesB), because marker genotypes can capture some CS effects.

In conclusion, by fitting CS explicitly, the LD-CS method can result in higher and more consistent prediction accuracy across families than LD methods. The LD-CS method is therefore more suitable for real populations with complex family structures.

Acknowledgments

We gratefully acknowledge the support from USDA AFRI Competitive Grants No. 2010-65205-20341 from the National Institute of Food and Agriculture, and Grant No. R01GM099992-01A1 from National Institutes of Health.

Table 1. Prediction accuracy of LD-CS method and of LD methods (namely GBLUP, BayesA and BayesB) for varying numbers of simulated half sib families. Results are the average of 32 replicates for each number of sires.

Number of Sires	LD-CS	GBLUP	BayesA	BayesB
1	0.345	0.337	0.350	0.347
2	0.535	0.524	0.523	0.526
10	0.503	0.464	0.464	0.463
100	0.546	0.480	0.433	0.327