Improved Accuracy of Genomic Prediction for Traits with Rare QTL by Fitting Haplotypes

A.S. Leaflet R3023

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

Summary and Implications

Genomic prediction estimates breeding values by exploiting linkage disequilibrium (LD) between quantitative trait loci (QTL) and single nucleotide polymorphisms (SNPs). High LD cannot occur when QTL and SNPs have different minor allele frequencies (MAF). Marker panels tend to use SNPs with high MAF and will have limited ability to predict rare QTL alleles. In practice, increasing SNP density has not improved prediction accuracy. A possible reason is that many traits are characterized by rare QTL. In that case, linear models fitting haplotypes could have advantage because haplotypes can be in complete LD with QTL alleles. SNP genotypes were simulated to resemble 600K chip for the bovine genome. Genomic breeding values were predicted using either SNP genotypes or non-overlapping haplotypes. When QTL had low MAF, the haplotype model had significantly higher accuracy than the SNP model. Results show that fitting haplotypes can improve the accuracy of genomic prediction for traits controlled by rare QTL.

Introduction

The accuracy of genomic prediction is expected to increase with SNP density due to the assumption that higher density panels will include SNPs with increased LD with QTL. Sequence based panels may even include the causal mutations in the panel. However, genomic prediction for field datasets showed limited improvement in accuracy when using 770K or sequence SNPs over 50K SNPs. These results suggest that the LD between QTL and SNPs may be low because many QTL have low MAF (rare QTL), while SNPs used for genotyping typically have moderate to high MAF. Since high or complete LD can only exist between two loci that have similar MAF, prediction accuracy for traits controlled by rare QTL is difficult to improve by increasing density of the SNP panel by adding SNPs with high MAF.

The problems of incomplete LD can be addressed by fitting haplotypes within short genomic windows that are constructed from phased SNP genotypes. With sufficient SNP density, QTL alleles can be in high or complete LD with haplotypes regardless of MAF of QTL. The objective of this study was to investigate the effect of MAF of QTL on prediction accuracy and to test the hypothesis that prediction accuracy can be improved by fitting haplotypes for traits controlled by rare QTL.

Materials and Methods

Datasets were simulated for an outbred population of 1,500 individuals. A random sample of 1,000 individuals without replacement was used as the training population to estimate SNP or haplotype effects, in order to predict genomic estimated breeding values (GEBV) of the other 500 individuals (validation). The simulated genome comprised two chromosomes each with length 100 cM. The simulated SNP density resembled 600K chips for the bovine genome. SNPs were selected with MAF either larger than 0.06 or larger than 0.01. One QTL was randomly placed in every 1.0 cM with MAF either larger than 0.06 (common QTL), or between 0.01 and 0.06 (rare QTL).

The haplotype model requires linkage phases of SNPs. These were assumed known in simulated datasets, but can be resolved accurately from high-density genotypes in field datasets, particularly with family data. The genome was partitioned into non-overlapping windows of 1.0 or 0.2 cM. Unique haplotypes in each window with frequency larger than 1% were defined as common haplotypes. GEBV were predicted using linear mixed models fitting either all or only common haplotype alleles. Bayesian method “BayesB” with $\pi = 0.99$ was used to estimate haplotype effects. Prediction accuracy was computed as correlation coefficient between GEBV and true breeding values in the validation population, and was compared to accuracy obtained from a SNP genotype based model (SNP model).

Results and Discussion

The accuracy of the SNP model was much higher for traits that were controlled by common QTL than for traits controlled by rare QTL when the MAF of SNPs was larger than 0.06 (Table 1). Including SNPs with MAF less than 0.06 increased the accuracy of the SNP model for traits controlled by rare QTL (Table 1). The haplotype model had no advantage over the SNP model when QTL were common, but had significant advantage when QTL were rare (Table 1). Fitting 0.2 cM haplotypes generally had higher accuracy than fitting 1.0 cM haplotypes. Decrease of accuracy was not significant when rare haplotypes were excluded from the model (Table 1).

In conclusion, at least for those traits controlled by rare QTL, fitting haplotypes may give higher accuracy due to near complete LD between QTL alleles and haplotypes.

Acknowledgments

This work was supported by the US Department of Agriculture, Agriculture and Food Research Initiative, National Institute of Food and Agriculture Competitive Grant No. 2012-67015-19420 and by National Institutes of Health Grant R01GM099992.

Table 1. Mean prediction accuracy and standard error (in parentheses) across 20 replicates of simulated datasets in each scenario of the minor allele frequencies (MAF) of QTL and SNPs

MAF of QTL	>0.06	0.01~0.06	0.01~0.06
MAF of SNPs	>0.06	>0.06	>0.01
SNP[1]	0.829 (0.007)	0.613 (0.025)	0.788 (0.013)
Haplotype[1], 1.0cM[2], a[3]	0.721 (0.009)	0.774 (0.011)	0.792 (0.010)
Haplotype, 1.0cM, c[3]	0.736 (0.008)	0.756 (0.012)	0.774 (0.010)
Haplotype, 0.2cM[2], a	0.811 (0.006)	0.769 (0.013)	0.798 (0.012)
Haplotype, 0.2cM, c	0.806 (0.006)	0.743 (0.013)	0.772 (0.012)

1 SNP, the SNP model; Haplotype, the haplotype model

2 1.0 cM, haplotype models fitting 1.0 cM haplotypes; 0.2cM, haplotype models fitting 0.2 cM haplotypes

3 a, haplotype models fitting all unique haplotypes; c, haplotype models fitting only common haplotypes