Proceedings, 10th World Congress of Genetics Applied to Livestock Production

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

X. Sun[1], **R. L.Fernando**[1], **D. J. Garrick**[1], **J. C. M. Dekkers**[1]

[1]Iowa State University, Ames, IA, United States

ABSTRACT: Genomic prediction estimates QTL effects by exploiting LD. High LD can only occur when SNPs and QTL have similar minor allele frequencies (MAF). Marker panels tend to use SNPs with high MAF and will have limited ability to predict rare QTL. In practice, increasing SNP density has not improved prediction accuracy. This might be explained if a trait had many rare QTL. In such cases, linear models fitting haplotypes could have an advantage because haplotypes could be in complete LD with QTL alleles. SNP genotypes were simulated with 200 SNPs per cM. Genomic breeding values were predicted using either SNP genotypes or non-overlapping haplotypes. When QTL had low MAF, prediction accuracy from haplotype models were significantly higher than for SNP models. Results suggest that haplotype models can be an efficient alternative to SNP models especially when traits are controlled by many rare QTL.
Keywords: prediction accuracy; haplotype; rare variant

Introduction

Implementation of genomic evaluation into breeding programs has been successful because genomic prediction of breeding values is more accurate than pedigree-based parent average for many economically valuable traits. With the rapid progress in genotyping and next-generation sequencing technologies, high-density SNP genotypes have been collected for increasing numbers of animals through chip genotyping, genotyping-by-sequencing or imputation. Accuracy of genomic prediction is expected to increase with increasing SNP density due to the assumption that SNPs in high linkage disequilibrium (LD) with quantitative trait loci (QTL) or even the QTL themselves could be included in the panel, and hence can explain most of the additive genetic variance. However, results from both simulation and field data analyses show limited advantage in prediction accuracy of using 770K or sequencing SNPs over 50K SNPs (VanRaden et al. 2011, Erbe et al. 2012).

Given that most traits in breeding objectives have comprised survival, growth, or reproduction of the individual, they have undergone long natural and intense artificial selection, and QTL affecting such traits are likely to have low minor allele frequencies (MAF). SNPs that are included on SNP chips are usually chosen from sequencing and prototype genotyping of reference samples and have generally been chosen to have 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 if the additional SNPs have high MAF. Moreover, increasing SNP density exacerbates statistical and computational difficulties for linear models when fitting increasingly large numbers of SNPs.

Both the problems of incomplete LD and expensive computation could be addressed by fitting haplotypes constructed from phased SNP genotypes. First, although rare QTL cannot be in high LD with common SNPs, they can be in high LD with haplotypes (Goddard and Hayes, 2007). Second, with increasing SNP density, the number of observable unique haplotypes eventually asymptotes due to finite population size and becomes less than the number of SNPs, at which point haplotype models will have lower dimension than SNP models.

Previous studies on haplotype models for genomic prediction were based on haplotypes constructed from low density SNP genotypes, in which the LD between haplotype and QTL was incomplete (Calus et al. 2007, Villumsen et al. 2008, Hickey et al. 2012). Although these studies reported advantages in prediction accuracy of haplotype over SNP models for specific haplotype sizes and with modeling of similarity among haplotype alleles, the potential advantage of haplotype models in prediction accuracy and computational efficiency when SNP density approaches sequence data, where there is almost complete LD between haplotype and QTL alleles regardless of the MAF of QTL, has not been studied.

Thus, the objectives of this study were to investigate the effect of MAF of QTL on prediction accuracy and to test the hypothesis that prediction accuracy can be improved with less computational burden by fitting haplotypes.

Materials and Methods

Simulated datasets. The initial generations comprised a population with effective size 500 that was randomly mated for 500 generations to reach mutation-drift equilibrium, before being reduced to effective size 100 and randomly mated for another 100 generations to generate LD spanning longer genomic distances. The population was then expanded to 2,000 individuals in the following 20 generations to represent the base population. A random sample of 1,500 individuals from this population was sampled, of which 1,000 individuals were used for training and the remaining 500 for validation.

The genome comprised two chromosomes, each with length 100cM. Initially, 80,000 SNPs were evenly positioned on each chromosome and a sufficient number of QTL candidate loci were randomly positioned within every 1 cM chromosomal segment. All SNPs and QTL were bi-allelic with initial allele frequencies 0.5. QTL effects were randomly sampled from a Gamma distribution with scale 0.4 and shape parameter 1.66, and had equal chance to be

positive or negative. Mutation rate was 2.5×10^{-6} per locus per meiosis.

In the base population, 20,000 SNPs per chromosome and 1 QTL in each 1cM segment were randomly sampled according to different assumptions on MAF of SNPs and QTL. Two scenarios were simulated for the MAF of QTL: 1) all QTL had MAF > 0.06 (common QTL), and 2) all QTL had MAF between 0.01 and 0.06 (rare QTL). For both common and rare QTL scenarios, datasets were generated where all 40,000 SNPs had MAF > 0.06 (common SNP). Specifically for the rare QTL scenario, an additional dataset with all 40,000 SNPs having MAF > 0.01 was generated.

In the base population of size 2,000, the effects of the selected QTL were scaled to achieve a total genetic variance of 4.29. True breeding values (TBV) were calculated by summing up all QTL effects for a given individual. Normal random variables with mean zero and variance 10.0 to represent residual effects were added to TBV to generate phenotypic values for a trait with heritability 0.3. Twenty random replicates were simulated for each combination of scenarios of MAF of QTL and MAF of SNPs.

Statistical analyses. Genomic estimated breeding values (GEBV) for validation individuals were predicted using linear mixed models fitting SNP genotypes or haplotypes. Models BayesA and BayesB (Meuwissen et al., 2001) were used to estimate SNP allele substitution or haplotype effects.

In the analyses with models fitting haplotypes, the linkage phase of the 40,000 SNPs was assumed known without error for both training and validation individuals. This assumption is justified because high phasing accuracy could be achieved under simulated SNP density, (e.g. Browning and Browning, 2007). The haploid genome was divided into non-overlapping segments of 1.0cM or 0.2cM. Unique SNP haplotypes for each segment that had a frequency > 0.01 in the combined training and validation population with size 1,500 were defined as common haplotypes. Either all unique or only common SNP haplotypes were fitted in the model for genomic prediction. Those haplotypes only present in validation population had zero estimated effects, and they didn't contribute to prediction of GEBV.

Formulation of models BayesA and BayesB based on haplotypes (termed "BayesA_H" and "BayesB_H", respectively) was similar to Meuwissen et al. (2001), except every unique haplotype allele was considered to have a random effect with an independent *t* distribution as prior. Value of π in BayesB_H was defined as the proportion of unique SNP haplotypes that were not in LD with any QTL alleles, which was set to 0.97 and 0.95 when segment sizes were 1.0 and 0.2cM, respectively.

Point estimates for SNP allele substitution effects and haplotype effects were their posterior means estimated from Markov chain Monte Carlo samples with chain length 11,000 and the first 1,000 discarded as burn-in. Prediction accuracy of GEBV was represented by the Pearson correlation coefficient between GEBV and TBV in validation individuals.

Results and Discussion

Haplotype frequencies and concordance between SNP haplotypes and QTL alleles. Frequencies of unique haplotype alleles were calculated for one dataset with MAF of QTL and SNP > 0.06. With SNP haplotype size 1.0cM, the total number of unique haplotype alleles across all 1.0cM segments was 10,559, of which 1,628 were common haplotypes (Table 1). When haplotype size was 0.2cM, the numbers of all and common haplotype alleles were 11,069 and 3,722, respectively. Under mutation and random drift, only 15% and one third of haplotype alleles were common when haplotype size was 1.0 and 0.2cM, respectively (Table 1). The dimension of the haplotype model was one quarter of the dimension of the SNP model, and models fitting only common haplotype alleles had less than one tenth dimension of the SNP model, resulting in a potential 10-fold greater computational efficiency for haplotype models. Table 1 shows results from one simulated dataset where SNP density was 20 per cM, 10 times less dense than the aforementioned scenario, which was similar to Villumsen et al. (2008) and Hickey et al. (2012). When SNP density increased 10 fold, the number of unique haplotype alleles increased less than two fold and the number of common haplotype alleles stayed the same, which suggested that the dimension of haplotype model would not increase much with increased SNP density.

Table 1. Average number of unique (No. Allele) and common (No. Common) haplotype alleles, and the proportion of discordant (Discordant %) haplotype alleles across all genome segments in the scenario of common QTL and common SNPs.

Segment length	No. Alleles	No. Common	Discordant %
200 SNPs per cM			
1.0cM	52.8	8.1	0
0.2cM	11.1	3.7	0
20 SNPs per cM			
1.0cM	36.9	7.4	1.7%
0.2cM	5.8	3.1	2.7%

Linkage disequilibrium exploited by the haplotype model was investigated by the concordance between haplotype alleles and QTL alleles. Discordant haplotypes were defined as those that carried both the major and minor QTL allele within the haplotype region, which meant that the LD between haplotype and the QTL allele was incomplete. The proportion of discordant haplotypes among all unique haplotypes within the population is given in Table 1. When SNP density was 200 per cM, there were no discordant haplotypes, suggesting complete LD between haplotype and QTL alleles, while a small proportion of discordant haplotypes existed when SNP density was 20 per cM.

Prediction accuracy from SNP and haplotype models with different MAF of QTL. When SNP MAF > 0.06, prediction accuracies of SNP models were much higher for traits that were controlled by common QTL than for traits controlled by rare QTL (see first two columns of Table 2). This suggests that prediction accuracies from the

same SNP panel can vary between traits, depending on the MAF of QTL for the trait, and that traits for which the QTL have similar MAF as SNPs on the panel are expected to have relatively high accuracy. Including SNPs with MAF < 0.06 into the model could increase prediction accuracy of SNP models for traits controlled by rare QTL (third columns of Table 2). These results are in agreement with those of Druet et al. (2014), who found that prediction accuracy could be increased up to 30% using sequencing data when the trait was controlled by many rare QTL, because many more rare SNPs can be captured by sequencing data than SNP chips.

Table 2. Mean prediction accuracies[1] across 20 replicates

MAF QTL	> 0.06	0.01~0.06	0.01~0.06
MAF SNP	> 0.06	> 0.06	> 0.01
BayesA	0.778	0.491	0.647
BayesB	0.829	0.613	0.788
$BayesA_H$, 1.0cM[2], a[3]	0.729	0.652	0.665
$BayesA_H$, 1.0cM, c[3]	0.728	0.646	0.659
$BayesA_H$, 0.2cM[2], a	0.776	0.657	0.685
$BayesA_H$, 0.2cM, c	0.767	0.643	0.674
$BayesB_H$, 1.0cM, a	0.721	0.774	0.792
$BayesB_H$, 1.0cM, c	0.736	0.756	0.774
$BayesB_H$, 0.2cM, a	0.811	0.769	0.798
$BayesB_H$, 0.2cM, c	0.806	0.743	0.772

1 Standard errors of mean were less than 0.025.

2 1.0cM, haplotype models with segment size 1.0cM; 0.2cM, haplotype models with segment size 0.2cM.

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

Prediction accuracies from haplotype models generally followed a similar trend as accuracies from SNP models, but were less affected by the MAF of QTL. This could be explained by the fact that haplotypes tended to be in higher or complete LD with QTL than single SNPs, regardless of the MAF of QTL. Haplotype models had no advantage over SNP models when QTL were common variants, but had significant advantage when QTL were rare variants (Table 2). Results suggest that for those traits where the prediction accuracy hardly improves by increasing chip SNP density, haplotype models may give higher prediction accuracy due to capture of QTL alleles by complete LD.

Models that fitted 0.2cM haplotypes generally had higher prediction accuracy than models that fitted 1.0cM haplotypes. There are two possible reasons for this. First, smaller size genome segments had fewer unique haplotype alleles and hence a smaller number of effects to be estimated within one segment, resulting in more accurate estimates of unique haplotype effects because more data is available to estimate their effects. Second, compared with large size segments, recombinations happened less often within small size segments and hence the proportion of discordant haplotypes tended to be smaller. On the other hand, the size of segments needed be large enough to allow enough segregating alleles to be in high or complete LD with QTL alleles. One critical question for haplotype models is the optimal segment size to achieve complete LD while to keep the overall number of haplotypes small. Villumsen et al. (2009) reported that fitting 10-SNP haplotypes of length 1.0cM gave highest prediction accuracy with a simulated marker density of 10 SNPs per cM. The optimal segment size for haplotype models largely depended on SNP density, level of LD and effective population size, and hence needs to be determined for specific datasets.

In most scenarios, prediction accuracy only decreased marginally when rare haplotypes were excluded from the model. Since few data were available to estimate effects of rare haplotypes, the estimated effects would be shrunk to zero and, thus, excluding rare haplotypes from the model had only minimal effect on prediction accuracy. The advantage of excluding rare haplotypes is the significant improvement in computational efficiency since a large proportion of haplotypes is rare, thus could result in an up to 10-fold reduction in the dimensionality of the model.

Conclusion

Under SNP density similar to genotyping by a 770K SNP chip or sequencing, haplotype models were shown to have significantly higher prediction accuracy than SNP models for traits controlled by rare QTL, with much less computation effort required. Thus, haplotype models can be efficient alternatives to SNP models when SNP density is high because they result in prediction accuracies that are less sensitive to the MAF of the underlying QTL and are computationally more efficient.

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.

Literature Cited

Browning, S. R., and Browning, B. L., (2007). Am.J. Hum. Genet. 81:1084-1097.

Calus, M. P. L., Meuwissen T. H. E., de Roos A. P. W. et al. (2007). Genetics 178:533-561.

Druet, T., Macleod, I. M., and Hayes, B. J. (2014). Heredity 112:39-47.

Erbe, M., Hayes, B. J., Matukumalli, L. K. et al. (2012). J. Dairy Sci. 95:4114-4129.

Goddard, M. E., and Hayes, B. J. (2007). J. Anim. Breed. Genet. 124:323-330.

Hickey, J. M., Kinghorn, B. P., Tier B. et al. (2012). J. Anim. Breed. Genet. 130:259-269.

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Genetics 157:1819-1829.

VanRaden, P. M., O'Connell, J. R., Wiggans, G. R. et al. (2011). Genet. Sel. Evol. 43:10.

Villumsen, T. M., Janss, L., and Lund,M. S. (2008). J. Anim. Breed. Genet. 126:3-13.