Comparison of methods for genomic prediction of breeding values using haplotype versus single-SNP genotypes in Holstein cattle

Z. Karimi1, L.F. Brito1, M. Sargolzaei1,2 & F. S. Schenkel1

1 Centre for Genetic Improvement of Livestock, Department of Animal Biosciences, University of Guelph, N1G 2W1, Guelph, Ontario, Canada
schenkel@uoguelph.ca (Corresponding author)
2 The Semex Alliance, N1H 6J2, Guelph, Ontario, Canada

Summary

A promising approach for genomic evaluation is the use of haplotype blocks instead of individual single nucleotide polymorphisms (SNPs) to construct the genomic relationship matrix (G). This study aimed to compare genomic prediction reliabilities using different methods to construct haplotype blocks, such as using linkage disequilibrium (LD; measured as r²), fixed length or a combination of both to construct G instead of using individual SNPs. A total of 20,206 Holstein bulls genotyped with 50K SNP panel with available de-regressed estimated breeding values for milk, protein and fat yields, and fat and protein percentages were used. Generally, reliabilities from haplotype-based methods were similar to or slightly higher than from individual SNPs, except for fat percentage. Moreover, constructing haplotype blocks based on LD using a r² threshold from 0.2 to 0.5 yielded similar prediction reliabilities compared to constructing haplotype blocks using a fixed length of 5 adjacent SNPs. A stricter or more lenient r² decreased prediction reliability.

Keywords: genomic selection, gblup, haplotype blocks

Introduction

The development of novel methodologies for genomic prediction of breeding values has driven substantial improvements in the accuracy of genomic breeding values, and therefore, placed genomic selection as the major landmark in modern livestock breeding programs. A promising approach for genomic evaluation is the use of haplotype blocks to construct the genomic relationship matrix instead of using individual single nucleotide polymorphisms (SNPs) (Hayes et al., 2007; Meuwissen et al., 2014; Cuyabano et al., 2014). A haplotype block can be defined as the combination of adjacent genomic markers. Different methods for defining haplotype blocks have been proposed, such as based on levels of linkage disequilibrium (LD) among SNPs, combining a fixed number of adjacent SNPs into a block, and defining haplotype blocks using runs of homozygosities (Luan et al., 2014; Hayes et al., 2007; Ødegård and Meuwissen, 2015). LD-based haplotype blocks are hypothesised to be more useful for genomic predictions compared to haplotype blocks based on the fixed number of adjacent SNPs, as they group markers that are most likely inherited together and minimize the number of haplotype alleles, as discussed by Cuyabano et al. (2014). Therefore, this study aimed to investigate whether or not LD-based haplotype blocks could yield more accurate genomic breeding values compared to using either fixed-length haplotype blocks or individual SNP genotypes in Holstein cattle.
Materials and methods

Genotypic, pedigree and phenotypic data

Official estimated breeding values (EBVs) and 50K panel genotypes for 20,206 Holstein bulls born from 1960 to 2011 were provided by the Canadian Dairy Network (CDN, Guelph, Ontario). The genotyping quality control removed SNPs that: deviate from Hardy-Weinberg equilibrium (HWE, \( p\)-value \( \leq 10^{-8} \)), with minor allele frequency (MAF) lower than 0.01, located in non-autosomal regions, and with call rate lower than 0.98. A total of 44,318 SNPs remained for further analyses. The traits analyzed included milk (MILK), fat (FAT) and protein (PROT) yields and fat (FAT\%) and protein (PROT\%) percentages.

Statistical analyses

Haplotypes were inferred using FImpute software (Sargolzaei et al., 2014). Afterwards, haplotype blocks were constructed based on three different approaches: 1) based on pairwise LD information; 2) based on a fixed number of adjacent SNPs; and 3) combination of LD-based haplotype blocks and individual SNPs not assigned to haplotype blocks or with fixed length haplotype blocks of 5 adjacent SNPs not assigned to haplotype blocks. Pairwise \( r^2 \) was calculated using SNP1101 software (Sargolzaei, 2014) and used as a measure LD. For LD-based approach, SNPs were grouped in the same haplotype block if the LD between every two adjacent SNPs was greater than or equal to a certain \( r^2 \) threshold (i.e., 0.1, 0.2, 0.3, 0.4, 0.5, 0.6 and 0.7). For haplotypes based on a fixed number of SNPs, 5 adjacent SNPs were included in each haplotype block based on a preliminary investigation that show this number of SNPs to be generally adequate (data not shown). In order to compare the approaches, genomic predictions were carried out using either single-SNP based relationship matrix (GS) or haplotype based relationship matrices (GH) constructed according to VanRaden (2008). GH was calculated after converting haplotypes to pseudo-SNPs (Meuwissen et al., 2014). The dataset was divided as training and validation sets. Bulls (n = 605) born between 2010 and 2011 with official proofs in 2016 were included in the validation, while the training set included bulls born from 1960 to 2007 (n = 19,601). The reliability of genomic predictions was assessed as the squared Pearson correlation between genomic estimated breeding values (GEBVs) and de-regressed EBVs.

Results and discussion

The total number of non-blocked SNPs increased from 19,919 to 40,509 by increasing pairwise \( r^2 \) threshold from 0.1 to 0.7 for constructing LD-based haplotype blocks. In a similar study, Cuyabano et al. (2014) reported a substantial increase in the total number of non-blocked SNPs, when increasing \( D^* \), as threshold for defining haplotype blocks, using high density genotype data. The total number of LD-based haplotype blocks decreased drastically from 9,040 to 2,986 as the pairwise \( r^2 \) threshold became stricter by increasing it from 0.1 to 0.7. In addition, the total number of haplotype alleles resulting from LD-based haplotype blocks decreased from 39,661 to 14,343 when pairwise \( r^2 \) threshold was increased from 0.1 to 0.7. Regarding to the combined LD-based and fixed-length haplotype blocks, the total number of haplotype blocks decreased from 16,336 to 11,433 by increasing pairwise \( r^2 \) threshold from
0.1 to 0.7. However, total number of haplotype alleles resulted from the combined LD-based and fixed-length haplotype blocks increased from 86,611 to 98,980.

Comparison of alternative LD-based haplotype block approaches

Table 1 presents the observed reliability from different methods for constructing haplotype blocks and using them to build GH. They included LD-based haplotypes only, LD-based haplotypes + individual SNPs, and LD-based haplotypes + fixed-length 5-SNP haplotypes. Table 1 shows results for the LD threshold of $r^2=0.2$ only. The method that combined LD-based haplotype blocks and fixed-length of haplotype blocks of 5 adjacent SNPs resulted in the highest reliabilities compared to the other two methods. Therefore, keeping the non-blocked SNPs, while converting them to the fixed–length haplotype blocks seems a better strategy compared to removing or keeping them as individual SNPs in the dataset.

Table 1. Prediction reliabilities based on LD-based haplotype blocks only, LD-based haplotype blocks with non-blocked SNPs, or using combined LD-based and fixed-length haplotype blocks for milk yield (MILK), fat yield (FAT), protein yield (PROT), fat percentage (FAT %), and protein percentage (PROT%) using 50K genotypes data in Holstein cattle.

<table>
<thead>
<tr>
<th>Trait</th>
<th>LD-based haplotypes ($r^2 = 0.2$)</th>
<th>LD-based haplotypes ($r^2 = 0.2$) + SNP</th>
<th>LD-based haplotypes ($r^2 = 0.2$) + fixed-length haplotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>MILK</td>
<td>0.587</td>
<td>0.596</td>
<td>0.597</td>
</tr>
<tr>
<td>FAT</td>
<td>0.555</td>
<td>0.551</td>
<td>0.564</td>
</tr>
<tr>
<td>PROT</td>
<td>0.513</td>
<td>0.530</td>
<td>0.532</td>
</tr>
<tr>
<td>FAT%</td>
<td>0.741</td>
<td>0.731</td>
<td>0.742</td>
</tr>
<tr>
<td>PROT%</td>
<td>0.770</td>
<td>0.778</td>
<td>0.780</td>
</tr>
</tbody>
</table>

Comparison of the use of individual SNPs, fixed-length haplotype blocks, and LD-based haplotype blocks

Table 2 shows the observed prediction reliabilities using different LD threshold criteria for constructing haplotype blocks in conjunction with fixed-length 5-SNP haplotypes, as well as the observed prediction reliabilities using only fixed length 5-SNP haplotypes or only individual SNP genotypes for the 5 traits considered. Except for FAT%, fixed length SNP haplotypes resulted in slightly higher prediction reliabilities compared to the individual SNP-based method, except for FAT%. LD-based haplotypes with $r^2$ threshold from 0.2 to 0.5 showed all similar results, which were also similar to the observed reliability of the fixed-length haplotype method. In general, as the pairwise $r^2$ threshold became stricter or more lenient, the prediction reliabilities for all traits dropped. A similar study using simulated data revealed that for some traits, using LD-based haplotypes might lead to higher levels of accuracy compared to using fixed-length haplotype blocks (Calus et al., 2008).

Conclusions

In general, reliabilities of haplotype-block based approaches were similar to or slightly higher than that from individual SNPs, except for FAT%. In addition, constructing haplotype blocks based on LD information under threshold of $r^2$ from 0.2 to 0.5 resulted generally in similar
prediction reliabilities compared to constructing haplotype blocks using a fixed length of 5 adjacent SNPs. A lower or higher \( r^2 \) threshold decreased prediction reliabilities. Further investigation will consider the ranking of the top bulls using the different approaches to construct the genomic relationship matrix, additional traits, and the use of high density genotypes.

### Table 2. Prediction reliabilities based on single SNPs (GS) and haplotype blocks with alternate criteria to define the haplotype blocks (FL5: Fixed-length haplotype of 5 SNPs; LD+FLH: LD-based haplotypes considering various \( r^2 \) thresholds + fixed-length 5-SNP haplotypes) for milk yield (MILK), fat yield (FAT), protein yield (PROT), fat percentage (FAT %), and protein percentage (PROT%) using 50k genotype data in Holstein cattle.

<table>
<thead>
<tr>
<th>Trait</th>
<th>GS</th>
<th>FL5 (( r^2 = 0.1 ))</th>
<th>LD+FLH (( r^2 = 0.2 ))</th>
<th>LD+FLH (( r^2 = 0.3 ))</th>
<th>LD+FLH (( r^2 = 0.4 ))</th>
<th>LD+FLH (( r^2 = 0.5 ))</th>
<th>LD+FLH (( r^2 = 0.6 ))</th>
<th>LD+FLH (( r^2 = 0.7 ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>MILK</td>
<td>0.594</td>
<td>0.605</td>
<td>0.593</td>
<td>0.597</td>
<td>0.594</td>
<td>0.596</td>
<td>0.598</td>
<td>0.549</td>
</tr>
<tr>
<td>FAT</td>
<td>0.557</td>
<td>0.564</td>
<td>0.549</td>
<td>0.564</td>
<td>0.553</td>
<td>0.551</td>
<td>0.563</td>
<td>0.549</td>
</tr>
<tr>
<td>PROT</td>
<td>0.523</td>
<td>0.541</td>
<td>0.526</td>
<td>0.532</td>
<td>0.525</td>
<td>0.527</td>
<td>0.533</td>
<td>0.530</td>
</tr>
<tr>
<td>FAT%</td>
<td>0.753</td>
<td>0.737</td>
<td>0.718</td>
<td>0.742</td>
<td>0.739</td>
<td>0.742</td>
<td>0.745</td>
<td>0.714</td>
</tr>
<tr>
<td>PROT%</td>
<td>0.757</td>
<td>0.783</td>
<td>0.774</td>
<td>0.780</td>
<td>0.782</td>
<td>0.784</td>
<td>0.788</td>
<td>0.775</td>
</tr>
</tbody>
</table>

### Acknowledgments

This research was supported in main part by Agriculture and Agri-Food Canada and by additional contributions from Dairy Farmers of Canada, the Canadian Dairy Network and the Canadian Dairy Commission under the Agri-Science Clusters Initiative.

### List of references