Genomic Prediction using Haplotypes in Brown Swiss

M. Frischknecht[1,2*]***, F.R. Seefried***[1*]***, B. Bapst***[1]***, C. Flury***[2]***, H. Signer-Hasler***[2]***, D. Garrick***[3]***, C. Stricker***[4]***, Intergenomics Consortium***[5]***, R. Fries***[6]***, I. Russ***[7]***, J. Sölkner***[8]***, A. Bieber***[9]***, A. Bagnato***[10] ***and B. Gredler-Grandl***[1]

[1]*Qualitas AG, Chamerstrasse 56, 6300 Zug, Switzerland*
[2]*Bern University of Applied Sciences, School of Agricultural, Forest and Food Sciences HAFL, Länggasse 85, 3052 Zollikofen, Switzerland*
[3]*Iowa State University, Kildee 225, 50011 Ames, USA*
[4]*agn Genetics, Börtjistrasse 8b, 7260 Davos, Switzerland*
[5]*Interbull center, SLU - Box 7023, Uppsala S-75007, Sweden*
[6]*Technische Universität München, Liesel-Beckmann-Straße 1, 85354 Freising-Weihenstephan, Germany*
[7]*Tierzuchtforschung e.V., Senator-Gerauer-Str. 23, 85586 Poing, Germany*
[8]*University of Natural Resources and Life Sciences, Gregor-Mendel-Str 33, 1180 Wien, Austria*
[9]*Research Institute of Organic Agriculture (FiBL), Ackerstrasse 113, 5070 Frick, Switzerland*
[10]*University of Milan, Via Celoria 10, 20133 Milano, Italy*

* equal contribution

Abstract

In order to improve accuracy of genomic selection different approaches have been suggested. One possibility is to use haplotypes instead of SNPs. It is thought that by the usage of haplotypes the number of effects to estimate should be decreased and the accuracy should be increased because the haplotype should catch the causal variants better than from LD with SNPs. Different definitions of the length of haplotypes are possible. The haplotypes can either be determined by the number of SNPs in a haplotype, by the length in base pairs or by linkage disequilibrium (LD) measures. For this study we used four different definitions of haplotype lengths either based on physical length in bp or on LD measures. We used haplotypes with a length of 250kb or 1Mb, we defined the LD based groups in PLINK and either included or excluded SNPs that were not included in any LD block. We estimated genomic breeding values with each of these haplotype definitions and compared prediction accuracy to that achieved with 50K SNPs for four traits in Brown Swiss. The traits were protein yield, non-return rate 56 in heifers, somatic cell score and stature. Estimation of genomic breeding values was carried out applying a BayesC model. We found trait-specific differences in the ranking of the scenarios. However, differences in accuracies between scenarios within trait were relatively low and using haplotypes only marginally increased the accuracy of genomic breeding values. The number of variables to be fitted increased relative to the SNP model especially for scenarios where the haplotypes were defined by physical length.

Key words: Brown Swiss, haplotypes, genomic selection

Introduction

Genomic selection has been introduced into many cattle breeding programs over the last few years. The improvement of accuracy of genomic selection is still challenging. It has been found that using whole genome sequence data leads only to marginal increases in accuracy compared to 50K SNP chip data (e.g. Frischknecht *et al.*, 2016). An alternative strategy could be the use of haplotype alleles as covariates. It has been shown in HD data in Nordic Holstein population that the number of covariates can be decreased using LD-based haplotypes with different D’ thresholds (Cuyabano *et al.*, 2014). In that study it was determined that the optimal D’ for genomic prediction should be D’ ≥ 0.45. The authors found an increase in accuracy for predicting milk protein and fertility, but not for mastitis. A slight increase in accuracy and a decreased bias of genomic breeding values was found in the Danish Holstein population when predicting genomic breeding values with genealogy-based haplotypes for the same traits as in the study above (Edriss *et al.*, 2013). However, in that study the number of covariates increased. In a simulation study with densities similar to HD genotypes it has been found that the usage of haplotypes is especially beneficial if a trait is influenced by quantitative trait loci (QTL) with low minor allelic frequency (MAF) (Sun *et al.*, 2014).

In the study presented here, we evaluated the accuracy of genomic prediction in Brown Swiss cattle based on haplotypes derived from 50K SNP chip data.

Materials and Methods

Animals

We used genotypic and phenotypic data from the routine genomic evaluation of Brown Swiss in Switzerland of August 2015. Bulls in the reference population were born between 1959 and 2011 (Figure 1). We evaluated four traits: protein yield (PY), stature (STA), somatic cell score (SCS) and non-return rate 56 in heifers (NRH). The number of bulls in the reference population ranged from 2,018 to 5,294 depending on trait. The sets of bulls used per trait were selected according to a reliability threshold of the breeding values (0.5 for STA, 0.55 for NRH, 0.65 for SCS and PY). The validation set consisted of 250 to 600 of the youngest bulls, to perform forward prediction.

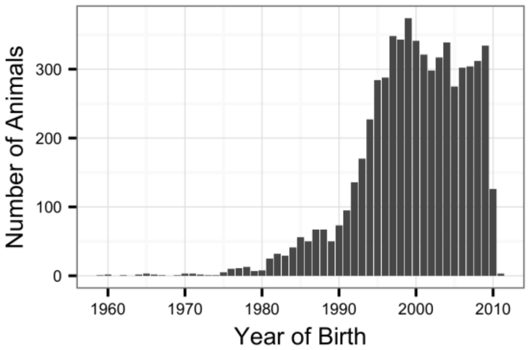

Figure 1. Age distribution of bulls used in the analysis of at least 1 trait.

Haplotype labelling

The 50K SNP chip data was formatted and filtered on the basis of quality measurements for routine genomic evaluation (40,636 SNPs). We phased the data with Beagle4 using default parameters (Browning and Browning 2009). We split the genotypes chromosome-wise and used the R-package GHap (Utsunomiya *et al.*, 2016) to label the haplotypes on each chromosome. Haplotype alleles with frequency less than 0.1% were excluded. We used four different definitions for the haplotype blocks – two having non-overlapping pre-defined length of 250kb (I) and 1Mb (II). The blocks were defined in GHap with the ghap.blockgen function. The other two approaches were constructed using PLINK1.9 (Chang *et al.*, 2015; Purcell *et al.*, 2007). The following parameter were used in order to include as many SNPs as possible in blocks (III):

```
--blocks no-pheno-req no-small-max-span
--blocks-max-kb 1000000
--blocks-recomb-highci 0.8
--blocks-strong-highci 0.8305
--blocks-strong-lowci 0.5005
```

The last haplotype block definition included the same haplotype blocks defined by PLINK1.9 as indicated above and additionally all SNPs that were not included in any block (IV). The construction of the haplotype blocks was done for each trait separately. We evaluated the prediction accuracy by comparison to caovariates based on the 50k SNP chip genotypes directly (0).

The haplotype labelling obtained from GHap corresponds to an additive SNP coding. This means that for each haplotype allele one column is written, where 0 means the animal does not carry this haplotype allele, 1 means this individual is heterozygous and carries one copy of this haplotype allele and 2 means this animal is homozygous for this haplotype. Thus one region in the genome covering one haplotype will take multiple columns according to the number of recoded haplotype alleles but one individual can only have a maximum of 2 alleles per haplotype block.

Genomic prediction

The above-mentioned coding was directly used in GenSel (Fernando and Garrick, 2009), for genomic prediction.

As input phenotypes we used deregressed breeding values (Garrick *et al.*, 2009). For each trait and method we first estimated π with a BayesCπ model and subsequently ran a BayesC analysis with the parameter obtained in the BayesCπ run. The following model was applied:

$$y=1^{'}m+Xb+e$$

$$\beta_i \begin{cases} = 0 \; with \; probability \; \pi \\ \sim N(0,\sigma^2_{SNP}) \; with \; probability \; (1-\pi) \end{cases}$$

where *y* is the vector of deregressed phenotypes, μ is the overall mean, *b* are the SNP or haplotype allele effects and *X* is a design matrix of covariate values for the SNPs of haplotype alleles, and *e* is the residual effect. The accuracy of genomic breeding values was calculated as the correlation between deregressed breeding values and estimated direct genomic breeding values for animals in the validation set. These correlations for the four methods using haplotypes and the control method using SNPs were compared.

Results & Discussion

Haplotype labelling

We investigated the number of haplotypes obtained from GHap and the frequency of the effect allele (Figure 2). For both parameters we found little difference between the four traits under investigation (results not shown). Therefore we pooled these two values for each scenario (0-IV) separately and only show the average values here. For the two methods based on physical length (I+II), we found that the number of covariates nearly doubled compared to 50K SNPs (0). Only the blocks obtained from PLINK1.9 (III) led to a decreased number of variables. However, a large number of SNPs (>25 000) were not included in any block. When adding these single SNPs to the haplotype covariates (IV), where we have again a slightly larger number of variables compared to (0).

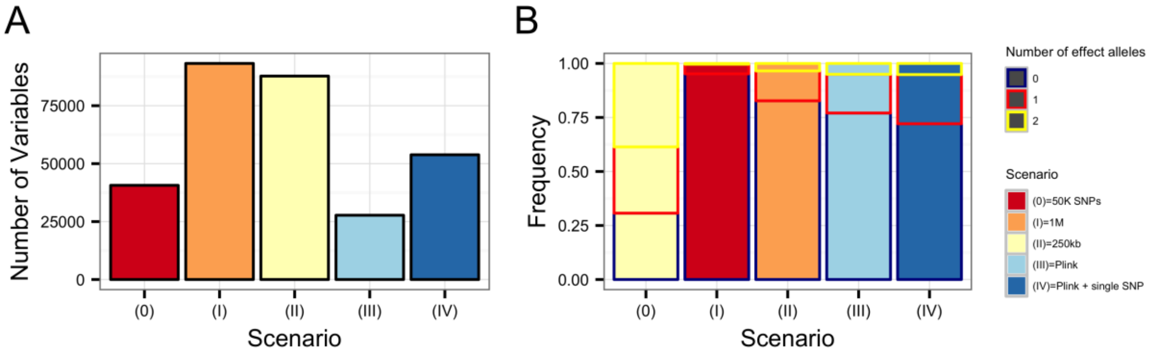

Figure 2. Analyses of the haplotype block properties. **A**: Number of variables for each haplotype block definition (mean across the four traits under investigation). 50k SNPs is for comparison the routinely evaluated 50k SNP chip data. **B**: Frequency of the effect allele.

The frequencies of the effect haplotypes are influenced by the number of haplotype alleles and the number of haplotype blocks along the genome. For the haplotype blocks with a length of 1M (I) we obtained the smallest proportion of effect alleles. This means, that we have a large number of haplotype alleles per haplotype block. We observed the same trend for the other three haplotype block definitions (II-IV), but less pronounced.

Genomic prediction using haplotypes

The accuracy of genomic prediction using 50K SNP chip data varied between traits (ranging between ~0.4 and ~0.65; Figure 3). We also found differences in accuracy in the different haplotype block definitions within trait. These differences are much lower than the differences between the traits. We observe that PY behaves differently than the other traits. For PY the genomic prediction using 50K SNP chip data (0) outperformed all haplotype prediction scenarios (I-IV). For the three other traits at least one of the haplotype predictions lead to higher accuracy than using the 50K SNP chip.

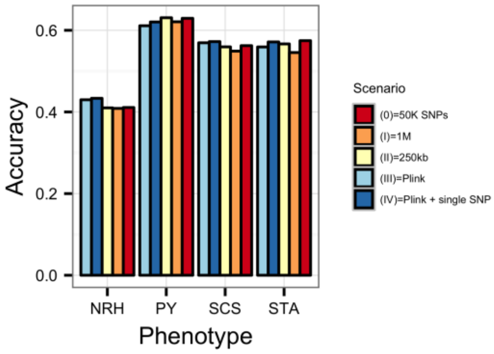

Figure 3. Accuracy of genomic prediction for each trait and haplotype block definition.

Generally, the 250kb haplotype block length (I) was slightly more accurate than the 1 Mb haplotype block (II). This is likely due to the larger number of variables for the 1M blocks (II). This increased number of covariates is associated with a large number of rare haplotype alleles for which the effects may not be estimated reliably. Accuracy was lowest for the haplotype blocks defined by PLINK1.9 (III) for all traits except PY. One reason for this could be, that so many SNPs are not included in any block, consequently they are excluded from the analysis and therefore a relatively large fraction of the genome is not considered at all for this analysis. This is supported by the fact that the addition of single SNPs to the haplotype blocks (IV) leads to higher accuracy.

Conclusions

For NRH and SCS the 250kb haplotype blocks (I) lead to highest accuracy, whereas for PY the 50K SNP chip scenario (0) was most accurate and for STA the PLINK1.9 blocks combined with single SNPs (IV) outperformed other methods. For a routine application it would be optimal to conduct the analysis for each trait with different haplotype block definitions. However, the increase in accuracy was only marginal.

Acknowledgements

This work has been performed within the frame of the Swiss Low Input Genetics (SLIG) project. We would like to thank the Swiss Commission for Technology and Innovation (CTI) for funding this project. We acknowledge the use of genotypes provided by the Intergenomics and the LowInputBreeds project and phenotypes provided by Braunvieh Schweiz.

References

Browning, B.L. & Browning, S.R. 2009. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Gen. 84*, 210-233.

Chang, C.C., Chow, C.C., Tellier, L.C.A.M., Vattikut, S., Purcell, S.M. & Lee, J.J. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience. 4*, 7.

Cuyabano, B.C.D., Su, G. & Lund, M.S. 2014. Genomic prediction of genetic merit using LD-based haplotypes in the Nordic Holstein population. *BMC Genomics. 15*, 1171.

Edriss, V., Fernando, R.L., Su, G., Lund, M.S. & Gulbrandtsen, B. 2013. The effect of using genealogy-based haplotypes for genomic prediction. *Genet Sel Evol. 45*, 5.

Fernando, R. & Garrick, D. 2009. GenSel – UserManual for a portfolio of genomic selection related analyses. Available: https://www.biomedcentral.com/content/supplementary/1471-2105-12-186-s1.pdf.

Frischknecht, M., Meuwissen, T.H.E., Bapst, B., Seefried, F.R., Flury, C., Garrick, D., Signer-Hasler, H., Stricker, C., Intergenomics Consortium, Bieber, A., Fries, R., Russ, I., Sölkner, J., Bagnato, A. & Gredler, B. 2016. Genomic prediction in cattle based on sequence data, in: *Book of Abstracts of the 67th Annual Meeting of the European Association for Animal Science.* Wageningen Academic Publishers, The Netherlands. pp. 104.

Garrick, D.J., Taylor, J.F. & Fernando R.L. 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genet Sel Evol. 41,* 55.

Sun, X., Fernando, R.L., Garrick, D.J. & Dekkers, J.C.M. 2014. Improved accuracy of genomic prediction for traits with rare QTL by fitting haplotypes. *Proceedings, 10th World Congress of genetics applied to livestock production.*

Utsunomiya, Y.T., Milanesi, M. & Utsunomiya, A.T.H. 2016. GHap: an R package for genome-wide haplotyping. *Bioinformatics. 32,* 18