GENOMIC PREDICTION USING SEQUENCE DATA IN A MULTIBREED CONTEXT

M.S. Lund[1], I. van den Berg[1,2,3] and D. Boichard[2]

[1] Center for Quantitative Genetics and genomics, Department of Molecular Biology and Genetics, Aarhus University, Tjele, Denmark
[2] INRA, UMR1313 GABI, 78350, Jouy-en-Josas, France
[3] AgroParisTech, Paris, France

SUMMARY

Sequence data can potentially increase the reliability of multi breed genomic prediction by containing causative variants or markers in in high linkage disequilibrium (LD) with those. Sequence data does, however, also contain a large number of variants in low LD with the causative mutations, limiting the potential increases in prediction reliability when the full sequence would be used directly for genomic prediction. The objective of this study was to use sequence variants to increase the reliability of multi breed prediction in dairy cattle. First, a simulation study based on real sequence data was carried out to investigate how sequence variants can improve the reliability of across breed prediction. The simulation study used the regression of genomic relationships at causative mutations on genomic relationships at prediction markers to measure the loss in prediction reliability as a consequence of using markers in imperfect LD. It was concluded that it is important to use only variants very close to the causative mutations. In the second part a number of two component Bayesian SNP BLUP models were used, where the first component mainly model variation within the breeds, while the second component model covariance across the breeds. Here, sequence variants selected from a multi breed GWAS for production traits were used as prediction markers in the second component. Different models and selection strategies were compared. Large increases in reliability, up to 0.10, were observed for multi breed prediction using QTL variants compared to within breed prediction using only 50K markers. Our results show that using a selective number of sequence variants can result in large increases in reliability, but careful selection of the variants is essential

INTRODUCTION

The reliability of genomic prediction is highly dependent on the size of the reference population. While for some breeds, for example Holstein, there are large national and international reference populations available, reliabilities in other breeds can be limited due to the smaller size of the reference populations. Smaller breeds can potentially benefit from the large reference populations available for some breeds by multi breed prediction. In practice, however, multi breed prediction only results in substantial increases in reliability compared to within breed prediction when closely related breeds are combined (Lund *et al.,* 2014). One reason for this could be that linkage disequilibrium (LD) is only conserved over short distances across breeds, and therefore, the density of marker chips is insufficient to allow across breed prediction (de Roos *et al.*, 2008). While the markers on the high density (HD) chip are dense enough for across breed prediction, the HD chip did not result in substantial increases compared to the 50K chip. The reason is likely that increasing the density to HD or full sequence does not only add variants closer to the causative mutations, but also variants in low LD with the causative mutations. Unless only variants in complete LD with the causative mutations are used, a loss in prediction reliability occurs (de los Campos *et al.*, 2013). By including QTL variants selected from the sequence, Brøndum *et al.* (2015) found increases in reliability up to 5% for within breed prediction of production traits in dairy cattle. Because LD is conserved over shorter distances across breeds than within breed, such an approach can potentially be more beneficial for across breed and multi breed prediction than for

within breed prediction. Methods for multi breed predictions must be able to 1) capture the genetic variance within breed without introducing noise from private variants or SNP associations only present in one (potentially dominating) breed and 2) capture covariance across breeds by markers in very close LD with causative variants segregating in multiple breeds.

The objective of this study was to use sequence variants to increase the reliability of multi breed prediction in dairy cattle. First, a simulation study based on real sequence data was carried out to investigate how sequence variants can improve the reliability of across breed prediction. Subsequently, sequence variants associated with QTL detected for milk, fat and protein yield were used for multi breed prediction in three dairy cattle breeds.

MATERIALS AND METHODS

For the simulation study, realised sequences on chromosome 1 of 122 Holstein, 27 Jersey, 28 Montbéliarde, 23 Normande and 45 Danish red bulls were used. Causative mutations were randomly sampled from 1,475,541 bi-allelic SNP and indels on chromosome 1, or from all variants with a minor allele frequency (MAF) below 0.10. The number of causative mutations was 10, 50, 100 or 250. Different sets of prediction markers were compared, with all variants from the 50K or HD chip, only the 50K or HD variants closest to each causative mutations, or sequence variants in two 1 Kb intervals on either side of each causative mutation. In the latter scenarios, the distance between intervals and causative mutations varied from 1 base to 1 Mb, and the intervals contained either all variants or only the variants with a MAF of at least 0.10.

For each scenario, two genomic relationships matrices were constructed for each breed and each pairwise combination of breeds, using either the causal loci, or the prediction markers. Genomic relationship matrices were scaled using the allele frequencies computed using the genotypes of all individuals in the genomic relationship matrix. Subsequently, the loss in R^2 was computed following de los Campos et al. (2013):

$$\overline{R}^2_{n+1,y} \leq R^2_{n+1,y}\left[1-(1-b_{n+1,y})^2\right],$$

where, for individual *n+1*, the difference between the prediction ($\overline{R}^2_{n+1,y}$) using markers in imperfect LD with the causative mutations and the prediction ($R^2_{n+1,y}$) if prediction markers were in perfect LD with causative mutations is quantified by the reliability factor (RF) $1-(1-b_{n+1})^2$. The *b* in the RF is the regression coefficient of the genomic relationships at prediction markers on the genomic relationship markers at the causative mutations. First, *b* was computed for each individual. Subsequently, RF was computed within replicate, using the *b* averaged across individuals. Finally, RF was averaged across replicates.

The second part of the study used imputed sequences and deregressed proofs (DRP) for milk, fat and protein yield from 5,852 French Holstein, 5,411 Danish Holstein, 1,203 Danish Jersey and 937 Danish Red bulls. First, bulls genotyped with the 50K chip were imputed to HD. For the French data, this step was performed using Beagle 3.0.0, while for the Danish breeds, IMPUTE2 was used. Subsequent imputation to whole-genome sequence was for all breeds done using IMPUTE2. The reference used for imputation to sequences of the Danish bulls consisted of the bulls in run 4 of the 1000 bull genome project, while for the imputation of the French bulls, a combined French-Danish reference set was used. The latter consisted of 122 Holstein, 27 Jersey, 28 Montbéliarde, 23 Normande and 45 Danish Red bulls.

A number of Bayesian SNP BLUP models were run. As prediction markers, either the 50K markers, or 50K markers in one component and sequence variants selected from a multi breed GWAS (van den Berg *et al.*, these proceedings) were used in a second component. Different selection strategies were compared. Selecting either all variants with a p-value in the multi breed

GWAS below 10^{-10}, 10^{-14} or 10^{-20}, or selecting maximum 1, 10 or 25 variants per intervals of 1, 2 or 10 Mb. Genomic breeding values were estimated using a Bayesian SNP BLUP model. Both single trait models, using a breed effect to account for differences between breeds, and multi trait models, fitting the same trait in different breeds as different correlated traits were used. The single trait models contained a within breed 50K component (ST-WB50K), a multi breed 50K component (ST-MB50K), or multi breed 50K and QTL components (ST-MB50K-MBQTL). The multi trait models contained a multi breed 50K component (MT-MB50K), multi breed 50K and QTL components (MT-MB50K-MBQTL) or a within breed 50K and a multi breed QTL component (MT-WB50K-MBQTL). For all models, marker effects and variance components were estimated using Bayz software, with a MCMC chain of 50,000 iterations, discarding the first 10,000 as burn-in. Reliabilities were estimated as the squared correlation between DRP and GEBV, divided by the mean reliability of DRP in the test population.

RESULTS AND DISCUSSION

RF decreased rapidly when the distance between prediction markers and causative mutations increased. This decrease was larger for across breed prediction than within breed prediction. Figure 1 shows the RF as a function of the distance between causative mutations and prediction markers for across breed prediction. Sequence variants on an interval on a similar distance to the causative mutations as the closest 50K or HD marker resulted in a larger RF when all 50K or HD variants were used, while RF was largest when only the closest 50K or HD markers were used. This shows that, in order to benefit from full sequence data, it is important to use only variants in high LD with the causative mutations, rather than using all sequence variants.

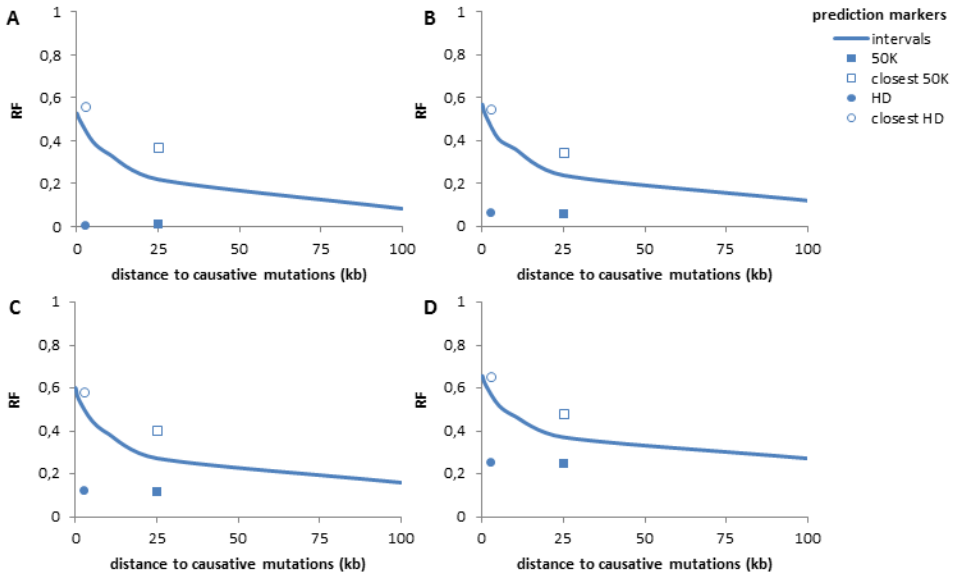

Figure 1. Average across breed reliability factor (RF) computed from intervals or from SNP from the 50K or HD chips, for different numbers (*c*) of causative mutations. A: *c*=10, B: *c*=50, C: *c*=100, D: *c*=250.

Using sequence variants selected from a multi breed GWAS resulted in substantial increases in reliability for all breeds and traits. The reliability was, however, highly sensitive to the set of

prediction markers used. Maximum increases compared to within breed prediction using 50K markers ranged from 0.042 for fat yield in Jersey to 0.105 for milk yield in Jersey. For all breeds and traits, the highest reliabilities were obtained when the number of variants per QTL interval was limited. Selecting many variants per QTL risks the selection of variants that are not in LD across breed, and, thereby, lowers the reliability. While for both Danish and French Holstein, best results were obtained with single trait models, the multi trait models generally resulted in higher reliabilities in Jersey and Danish Red. Because most of the individuals in the data were Holstein, Holstein had a much larger effect on the estimated marker effects than Jersey and Danish Red. Although some QTL are shared across breeds, this is not the case for all QTL, and markers associated with QTL segregating in Holstein but not in the other breeds could introduce noise. A multi trait Bayesian variable selection that would allow different sets of prediction markers to influence different breeds, could potentially lead to larger increases in reliability than those observed here.

Table 1. Scenarios with largest prediction R^2 for each breed and trait. Δ is the difference with within breed prediction using 50K markers. HOLDK = Danish Holstein, HOLFR = French Holstein, JER = Jersey, RDC = Danish Red.

Breed	Trait	50K	Best scenario	Δ
HOLDK	Milk	0.440	ST-MB50K-MBQTL10-25/1	0.087
	Fat	0.475	ST-MB50K-MBQTL20-25/1	0.103
	Protein	0.388	ST-MB50K-MBQTL10-1/1	0.055
HOLFR	Milk	0.327	ST-MB50K-MBQTL14-25/1	0.079
	Fat	0.367	ST-MB50K-MBQTL20-25/1	0.097
	Protein	0.372	ST-MB50K-MBQTL14-25/10	0.065
JER	Milk	0.299	MT-WB50K-MBQTL20-1/10	0.105
	Fat	0.161	ST-MB50K-MBQTL10-10/10	0.042
	Protein	0.219	MT-WB50K-MBQTL20-1/10	0.049
RDC	Milk	0.136	MT-MB50K-MBQTL20-10/1	0.073
	Fat	0.114	MT-MB50K-MBQTL20-25/10	0.075
	Protein	0.093	MT-WB50K-MBQTL14-10/10	0.059

Our results, both from simulation and real data, show that using a selective number of sequence variants can result in large increases in reliability for multi breed prediction in dairy cattle, but careful selection of the variants is essential.

ACKNOWLEDGEMENTS

IB benefited from an Erasmus-Mundus fellowship and a grant by Apisgene, within the framework of the European Graduate School in Animal Breeding and Genetics. This research was supported by center for Genomic Selection in Animals and Plants (GenSAP) funded by The Danish Council for Strategic Research).

REFERENCES

Brøndum R.F., Su G., Janss L., Sahana G., Guldbrandtsen B., Boichard D. and Lund M.S. (2015) *J. Dairy Sci.* **98**:4107.

de los Campos G., Vazquez A.I., Fernando R., Klimentidis Y.C. and Sorensen D. (2013) *PLoS Genet.* **9**:e1003608.

de Roos A.P.W., Hayes B.J., Spelman R.J. and Goddard M.E. (2008) *Genetics* **179**:1503.

Lund M.S., Su G., Janss L., Guldbrandtsen B. and Brøndum R.F. (2014) *Livest. Sci*. **166**:101.