Across Breed QTL Detection and Genomic Prediction in French and Danish Dairy Cattle Breeds

I. van den Berg [*] [†] [‡]**, B. Guldbrandtsen**[*]**, C. Hozé**[†‡§]**, R. F. Brøndum**[*]**, D. Boichard**[†‡] **and M. S. Lund**[*]
[*]Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, Tjele, Denmark, [†]INRA, UMR1313 Génétique Animale et Biologie Intégrative, Jouy-en-Josas, France, [‡]AgroParisTech, UMR1313 Génétique Animale et Biologie Intégrative, Paris, France, [§]UNCEIA, Paris, France

ABSTRACT: Our objective was to investigate the potential benefits of using sequence data to improve across breed genomic prediction, using data from five French and Danish dairy cattle breeds. First, QTL for protein yield were detected using high density genotypes. Part of the QTL detected within breed was shared across breed. Second, sequence data was used to quantify the loss in prediction reliabilities that results from using genomic markers rather than the causal variants. 50, 100 or 250 causative mutations were simulated and different sets of prediction markers were used to predict genomic relationships at causative mutations. Prediction of genomic relationships at causative mutations was most accurate when predicted by a selective number of markers within 1 Kb of the causative mutations. Whole-genome sequence data can help to get closer to the causative mutations and therefore improve genomic prediction across breed.

Keywords:
dairy cattle
across breed
genomic prediction
QTL

Introduction

Genomic selection has been rapidly integrated in dairy cattle breeding programmes over the past years. The accuracy of genomic predictions depends on several factors, including the size of the reference population and the amount of linkage disequilibrium (LD) between genomic markers and quantitative trait loci (QTL) (Goddard et al., 2009). Sharing of reference populations between breeds could benefit small breeds. The amount of long-range LD is, however, much lower across breed than within breed (de Roos et al., 2008), reducing the accuracy of genomic selection across breeds. Consequently, across breed predictions has resulted in improved predictions for closely and moderately related breeds (Brøndum et al., 2011) while little or no improvements are reported for distantly related breeds (Erbe et al., 2012). Increasing the marker density by the inclusion of sequence data could help to increase the accuracy of genomic predictions. The objective of this study was to investigate the potential benefits of using sequence data to improve genomic prediction across distantly related breeds. The study included two parts. A prerequisite for accurate across breeds predictions is that breeds share a substantial amount of the genetic variance. Therefore the first part studied the proportion of QTL shared across breeds. Assuming that QTL are shared, association between markers and QTL must be conserved across breeds. The second part of the study attempted to assess how close markers need to be to the causal variants to efficiently use this common variance in predictions. In this second part, sequence data was used to quantify the loss in prediction reliabilities that results from using different sets of genomic markers rather than the true causal variants (de los Campos et al., 2014), when reference animals are from another breed.

Materials and Methods

Data. The dataset used for QTL mapping consisted of 5642 Nordic Holstein, 3130 French Holstein, 1238 Danish Jersey, 2236 Montbéliarde, 1970 Normande, and 1019 Danish Red bulls. Bulls were genotyped with the Bovine SNP50 BeadChip® or the Bovine HD BeadChip®. For bulls genotyped with the 50K chip, HD genotypes were obtained by imputation. Imputation of Nordic Holstein, Danish Jersey and Danish Red was done with IMPUTE2 (Howie et al., 2009) and imputation of French Holstein, Montbéliarde and Normande (Hozé et al., 2013) with Beagle 3.0.0 (Browning et al., 2006). For the second part of the study, the estimation of genomic relationships within and across breed, sequences of 122 Holstein, 27 Jersey, 28 Montbéliarde, 23 Normande and 45 Danish Red bulls were used. Variant calling was performed using GATK (DePristo et al., 2011).

QTL detection. The following single marker sire model was used for QTL detection:

$$y_{ij} = \mu + S_i + bg_{ij} + e_{ij}$$

where y_{ij} equals the deregressed proof of protein yield for individual j with sire i, S_i the effect of sire i, g_{ij} the genotype (0, 1 or 2 for respectively homozygous for allele 1, heterozygous and homozygous for allele 2) of individual i with sire j and e the random residual. A significance threshold of $p \leq 10^{-6}$ was used to detect QTL within each breed. A QTL was considered as shared across two breeds if, for a QTL detected in the first breed, there was a marker with a p-value $\leq 10^{-5}$ within a distance of 1 Mb in the second breed.

Simulated causative mutations and prediction markers. To quantify the loss in prediction reliability resulting from using genomic markers rather than the causal variants, 50, 100 or 250 causative mutations were randomly sampled from all single nucleotide polymorphisms (SNP) segregating in at least one of the five breeds on chromosome 1. Part of the non-causative

SNP was used as prediction markers according to 15 scenarios. In scenarios 50K and HD, the markers were those from the Bovine SNP50 BeadChip® (3229 markers) and Bovine HD BeadChip® (46,243 markers): In scenarios n50K and nHD, an equivalent number of markers was randomly sampled on the sequence. In the 100K scenario, 100,000 SNP were randomly selected. In another batch of scenarios, markers were selected at different distances from each causal mutation. The 50K and HD scenarios selected the SNP from the 50K or HD chips closest to each causative mutation. In the 1b, 1Kb, 5Kb, 10Kb, 25Kb, 100Kb, 500Mb and 1Mb scenarios, all SNP in two 1 Kb intervals on both sides of the causative mutations were selected, with a distance between causative mutation and interval of 1b, 1Kb, 5Kb, 10Kb, 25Kb, 100Kb, 500Mb or 1Mb. Each scenario was repeated 50 times.

Genomic relationships. For each scenario, two genomic relationship matrices (VanRaden, 2008) were constructed, the first using the causative mutations and the second using the prediction markers. This was done for all breeds separately, as well as for all breed combinations (combining two breeds at the time). Subsequently, the loss in prediction reliability resulting from using genomic markers rather than the causal variants was assessed by estimation of the regression of genomic relationships at prediction markers on genomic relationships at causative mutations (*b_{n+1}*) following de los Campos et al. (2013):

$$\overline{G}_{n+1,i} = b_{n+1} G_{n+1,i} + \xi_{n+i,i}$$

where { $\overline{G}_{n+1,1},\ldots,\overline{G}_{n+1,n}$ } and { $G_{n+1,1},\ldots,G_{n+1,n}$ } are the genomic relationships between individual *n+1* and all other *n* individuals at respectively prediction markers and causative mutations and $\xi_{n+i,i}$ equals the residual orthogonal to $G_{n+1,n}$. The minimum reduction of prediction R^2 equals $1-(1-b_{n+1})^2$.

Results and Discussion

Table 1 shows the number of QTL detected within each breed and the proportion of these QTL shared across breeds. The number of QTL detected varied strongly between breeds, from 449 in Danish Holstein to 28 in Danish Red. This can be explained with the difference in number of individuals used for each breed. For example, while there were 5642 Danish Holstein individuals in the dataset, there were only 1019 Danish Red individuals. As a consequence, the power in the QTL detection for Danish Red was much lower than for Danish Holstein. This difference in power complicates the comparison of the proportion of QTL across breeds. The results do, however, suggest that a part of the QTL detected in one breed is shared with other breeds.

The regression of genomic relationships at prediction markers on genomic relationships at causative mutations (*b*) increased when the number of causative mutations increased. For example, when simulating 50, 100 or 250 causative mutations and using the 50K markers as prediction markers, *b* for Holstein equalled 0.23, 0.39 and 0.63 respectively. Similar increases were found for other breeds and scenarios. While increasing the number of prediction markers (from 50K to HD or from n50K to nHD → 100000) did not affect *b*, *b* was slightly higher when the 50K or HD markers were used than when an equal number of randomly selected SNP was used. For example, when simulating 100 causative mutations, *b* for Normande equalled 0.54 for the 50K and HD scenarios, and 0.50 for the n50K and nHD scenarios. The 50K and HD markers were not randomly selected SNP, but selected to maximise MAF over a range of breeds (Matukumalli et al., 2009). Therefore, it is not surprising that using these markers as prediction markers instead of random markers results in higher values of *b*. Across breeds, differences in *b* for different sets of prediction markers were small and inconsistent. Figure 1 shows *b* for the scenarios with 100 causative mutations and different sets of prediction markers. Across breeds, the highest values of *b* were found between Montbéliarde and Danish Red.

When prediction markers close to the causative mutations were used, *b* was higher than when SNP from the chips or random SNPs were used. Increasing the distance between the prediction markers and causative mutations resulted in a decrease in *b* (figure 2). This decrease was larger across breeds than within breed. The largest decrease took place between 1Kb and 25Kb, while increasing from 100Kb to 1Mb did only result in small additional decrease of *b*. For example, in the scenario with 100 causative mutations, when predicting across Jersey and Normande, *b* was 0.36 when using prediction markers a distance of 1b, decreased to 0.14 at 25 Kb, further decreased to 0.08 at 100 Kb and 0.04 at 1 Mb. Using the closest 50K or HD marker resulted in an average distance between causative mutations and markers of 24.8Kb and 2.6 Kb respectively. In these scenarios, values found for *b* were higher than when 1 Kb intervals on a similar distance was selected. For example, when simulating 100 causative mutations, and predicting across Montbéliarde and Danish Red, *b* was 0.26, while when using 1 Kb intervals on 25 Kb of the causative mutations resulted in a *b* of 0.19.

Conclusion

Our results suggest that there is a large potential to increase the reliability of genomic predictions using sequence data. First, a proportion of QTL detected within breed was shared across breed. Second, prediction of the genomic relationships at causative mutations was more accurate when prediction markers close to the causative mutations were used than when a large number of random markers or markers on the Illumina chips were used. This difference was more pronounced for across breed prediction than for within breed prediction. Genomic prediction across breed will thus be more accurate when a selective number of markers within 1 Kb of the causative mutations is included in the model. Whole-genome sequence data can help to get closer to the causative mutations and therefore improve genomic prediction across breed.

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. French HD genotypes originated from the Gembal (ANR-10-GENM-0014) project. French sequence data originated from the Cartoseq (ANR-10-GENM-0018), Reproseq (Apisgene) and Cartofine (ANR-05-GENANIMAL-007) projects. The Danish data was supported by the project 'Genomic Selection – from function to efficient utilization in cattle breeding' (grant no. 3412-08-02253).

Literature Cited

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

Brøndum, R. F., Rius-Vilarrasa, E., Strandén, I. et al. (2011). *J Dairy Sci* 94:4700-4704.

de los Campos, G., Vazquez, A. I., Fernando, R. et al. (2013). *PLoS Genet* 9(7):e1003608.

DePristo, M., Banks, E., Poplin, R et al. (2011). *Nature Genetics* 43:491-498.

de Roos, A. P. W., Hayes, B. J., Spelman, R. J. et al. (2008). *Genetics* 179 :1503-1512.

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

Goddard, M. E., Hayes, B. J. (2009). *Nat Rev Genet* 10:381-391.

Howie, B. N., Donelly, P., Marchini, J. (2009). *PLoS Genet* 5:e1000529.

Hozé, C., Fouilloux, M-N., Venot, E. et al. (2013). *Genet Sel Evol* 45:33.

Matukumalli, L. K., Lawley, C. T., Schnabel, R. D. et al. (2009). *PLoS ONE* 4(4):e5350.

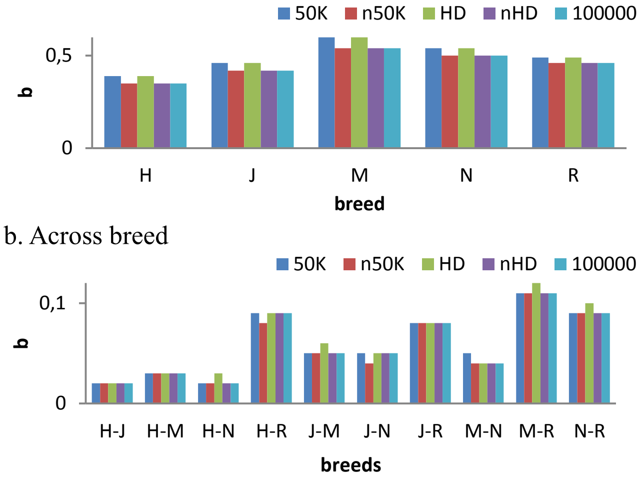
VanRaden, P. M. (2008). *J Dairy Sci* 91:4414-4423.

Table 1. Proportion of QTL detected for protein yield shared across breed. The diagonals represent the number of QTL with $p \leq 10^{-6}$ detected in the breed in the first column, and the off-diagonals the proportion of these QTL for which there is a SNP with $p \leq 10^{-5}$ within a distance of 1Mb found in the breed in the header.

breed	HD	HF	M	N	J	R
HD	449	0.36	0.26	0.08	0.15	0.10
HF	0.65	139	0.35	0.23	0.14	0.09
M	0.44	0.29	161	0.15	0.08	0.07
N	0.30	0.44	0.29	82	0.04	0.10
J	0.68	0.26	0.24	0.13	38	0.00
R	0.68	0.32	0.32	0.11	0.11	28

HD=Danish Holstein, HF=French Holstein, M=Montbéliarde, N=Normande, J=Jersey, R=Danish Red

Figure 1. Regression of genomic relationships at prediction markers on genomic relationships at 100 simulated causative mutations (b) on chromosome 1. Prediction markers are the markers from the 50K chip (50K), a constructed 50K chip (n50K), the HD chip (HD), a constructed HD chip (nHD) or 100000 randomly selected SNP (100000).

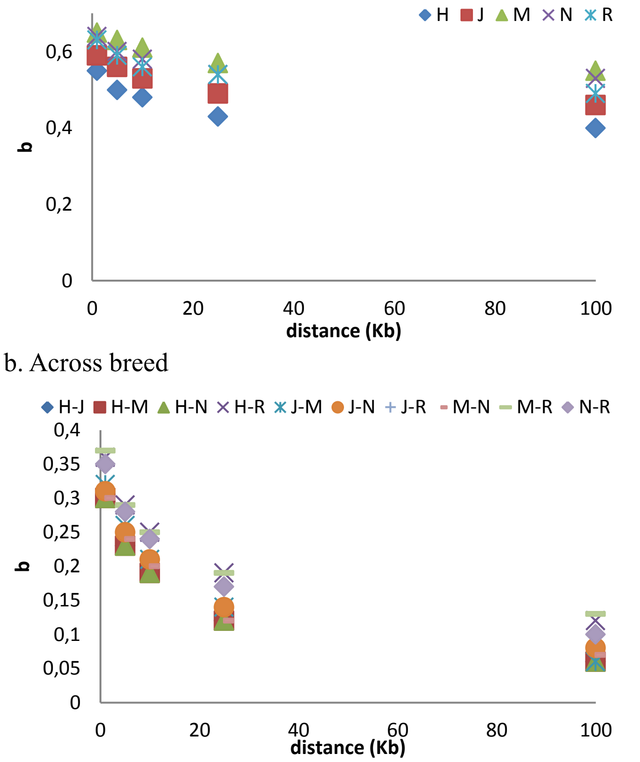


a. Within breed

b. Across breed

H=Holstein, J=Jersey, M=Montbéliarde, N=Normande, R=Danish Red

Figure 2. Regression of genomic relationships at prediction markers on genomic relationship at 100 simulated causative mutations (b) on chromosome 1. Prediction markers in 1Kb intervals on both sides of the causative mutations, with increasing distance between intervals and causative mutations were used.



a. Within breed

b. Across breed

H=Holstein, J=Jersey, M=Montbéliarde, N=Normande, R=Danish Red