Comparison of Genomic Selection Approaches for Small Breeds

C. Hozé [1, 2], S. Fritz[2], D. Boichard[1], V. Ducrocq[1] and P. Croiseau[1]
[1] *INRA, UMR 1313, GABI, 78350 Jouy-en-Josas, France*
[2] *UNCEIA, 149 rue de Bercy, 75012 Paris, France*

Abstract

Within breed genomic selection based on medium SNP density (50K chip) is now routinely implemented in a number of large cattle breeds. However, building large enough reference populations remains a major challenge for many medium or small breeds. The high density BovineHD BeadChip® (HD) containing 777 609 single nucleotide polymorphism (SNP) developed in 2010 is characterized by short distance linkage disequilibrium expected to be maintained across breeds. Therefore, combined multi-breed reference populations can be envisioned. In France where genomic evaluations are only implemented in the three main dairy breeds, a HD-reference of 1869 animals from these 3 breeds was built. Then, 29 091 50K-genotypes from national genomic evaluation were imputed to high density to form a large HD-reference population. This population was used to develop a multi-breed genomic evaluation and compare genomic selection strategies for small breeds.

In this study, we chose to use a large breed (the Normande breed) to mimic a small breed in order to have a large enough validation population and better compare genomic selection approaches. Three training sets containing respectively 1597, 404 and 194 bulls and a unique validation dataset of 394 animals (the youngest Normande bulls) were formed. For each training set, three approaches were compared: pedigree-based BLUP, within-breed BayesCpi and multi-breed BayesCpi in which the reference population was formed by the Normande training dataset and 4989 Holstein and 1788 Montbéliarde bulls.

We computed the correlations between observed and predicted daughter yield deviations (DYD) for six traits and the different approaches. Compared to pedigree-based BLUP, genomic selection approaches provided an average gain in correlation ranging from 6.7 to 7.6% with the smallest reference population and up to 20% with the largest reference population. Multi-breed genomic selection gave the best results in all situations with an average gain in correlation of 3% compared to within-breed genomic selection. However, the increase in correlation is limited when the within-breed reference population is already large and the achieved accuracies are clearly higher. The slope of regression was closer to one when the number of individual in the reference population increased and was similar between multi-breed genomic evaluations and within-breed genomic evaluations. These results showed that multi-breed genomic selection can be an appealing strategy for small breeds.

Key words: genomic selection, genomic evaluation, multi-breed evaluation, high-density genotype

Introduction

In France, a very large number of animals from the main dairy breeds have been genotyped with the Bovine SNP50 BeadChip® (50K) and it is now possible to predict breeding values of animals at birth with high accuracy. In breeds with a limited number of progeny-tested bulls, assembling a large enough reference population is a real challenge. Under the assumption that linkage disequilibrium (LD) is conserved across breeds, reference populations from different breeds can be combined to increase the reference population size. As LD is not maintained across breeds with the classically used 50K chip, a denser chip, the BovineHD BeadChip® (HD), containing 777K SNPs was developed to detect conserved LD across breeds. Large HD-reference populations are now available to study multi-breed genomic selection. Here, within-breed and multi-breed genomic evaluations were compared to investigate their benefit for genomic selection of small breeds.

Materials and Methods

Due to the low number of individuals in the reference population, measuring accuracy of genomic selection in small breeds is difficult. Indeed, correlations based on a few dozen animals are difficult to interpret. For this reason, we chose to use a large dairy breed to mimic a small breed and develop a multi-breed genomic selection method. This strategy offers the opportunity to study several training population sizes while using a unique reasonably large validation set.

An initial HD dataset consisting of respectively 535, 527 and 773 progeny tested bulls from Normande (NO), Montbéliarde (MO) and Holstein (HO) breeds genotyped with the Illumina Bovine HD BeadChip® was used to impute HD genotypes from national 50K genotype databases. A quality control based on call rate (10%) and Hardy Weinberg equilibrium test (10^{-4}) was performed. In total, 706 791 SNPs were retained. Mendelian segregation was also checked. Close to thirty thousand 50K genotypes from national evaluations were available and feasibility of their imputation to HD was studied. Achieved accuracy with Beagle software (Browning and Browning, 2009) was higher than 99% in large breeds (Hozé *et al.*, 2013) and therefore HD-genotypes were imputed. After imputation, a HD-reference population of 29 091 bulls, mainly composed of Holstein bulls from the Eurogenomics consortium was available. On this complete set, only bulls with daughters in France and an equivalent daughter contribution (EDC) higher than five were retained leading to a reference population of 8768 animals (Table 1).

Table 1. Number of animals in the reference population for each breed.

	HD-genotyped	Reference population
Montbéliarde	527	1788
Holstein	773	4989
Normande	535	1991

Five production traits (milk yield, fat yield, protein yield, fat percentage and protein percentage) and one functional trait (somatic cell score) were considered. Phenotypes used were daughter yield deviation (DYD) derived from national genetic evaluations.

To compute efficiency of genomic selection for small breeds, a cross-validation scheme was implemented where the Normande breed was chosen to mimic a small breed. The validation population consisted of the 394 youngest 20% bulls from the Normande population and three training populations including 1597, 404 and 194 Normande bullsrespectively, were formed.

An Average Information Restricted Expectation Maximization Likelihood (AI-REML) approach (Jensen *et al.*, 1996) was used to estimate genetic parameters and compute estimated breeding values (EBV) with a traditional pedigree-based BLUP. Genomic evaluations were performed using BayesCpi (Kizilkaya *et al.*, 2010) implemented in GS3 software (*http://snp.toulouse.inra.fr/~alegarra*). The proportions of polygenic residual component and SNP with a non-zero variance were fixed to 30% and 1% respectively. Computational requirements for an HD-analysis were too high and convergence was too slow for a precise estimation of these parameters and the retained values were based on a preliminary study performed on 50K datasets (not shown). Then, the weighted Pearson correlation was computed between EBV and DYD, the weights being the EDC (Peers, 1996).

Each training dataset was also used for multi-breed genomic evaluation. In this case, the Normande training population was combined with the 1788 Montbéliarde and 4989 Holstein bulls. DYD were centered within breed and divided by genetic standard deviation of the breed (Table 2). A breed effect was also included in the evaluation model. Again, a weighted Pearson correlation was computed between EBV and DYD.

Table 2. Genetic parameters of the traits for each breed.

Trait	Breed	Phenotypic standard deviation	Genetic standard deviation
Milk (kg)	MO	1188	651
	NO	1093	599
	HO	1307	716
Fat yield (kg)	MO	46.89	25.68
	NO	45.98	25.19
	HO	50.46	27.64
Protein yield (kg)	MO	37.02	20.28
	NO	34.97	19.15
	HO	36.57	20.03
Fat content	MO	3.20	2.26
	NO	3.91	2.76
	HO	4.94	3.50
Protein content	MO	1.89	1.34
	NO	1.99	1.41
	HO	2.13	1.51
Somatic cell score	MO	2.11	0.91
	NO	2.07	0.90
	HO	2.27	0.98

Results & Discussion

Table 3 presents the weighted correlation between observed and predicted DYD for the 6 traits and the average slope of the regression of observed DYD on predicted DYD for a training dataset with 194 animals

Table 3. Weighted correlations between observed and predicted DYD, average correlation and slope for the 6 traits using BLUP, within-breed GS and between breed GS with 194 Normande bulls.

194 bulls in training population	BLUP	Within-breed BayesCpi	Multi-breed BayesCpi
Milk	0.13	0.17	0.23
FY	0.08	0.17	0.24
PY	0.13	0.17	0.24
Fat %	0.22	0.29	0.36
Protein %	0.24	0.35	0.35
SCS	0.28	0.32	0.29
Average correlation	0.18	0.25	0.29
Average slope	0.52	0.54	0.59

We first observe that genomic selection drastically improves correlations whatever the trait and slightly increases regression slopes. As expected, a higher gain in correlation is observed for fat yield and fat percentage which are traits influenced by large QTLs.

When we compare multi-breed vs within-breed GS, we observe an average increase of 3.8% in correlation. However, this increase is trait-dependent: the correlation is lower with multi-breed GS for SCS and there is no difference between the two approaches for protein percentage.

Whatever the approach, the average slope of regression clearly deviates from 1 and the average correlation is quite low. This may be the result of a low relationship between the training dataset and (part of) the validation dataset. Indeed, here more than 50% of animals in the validation population do not have their sire nor their grandsires in the training dataset.

To observe the impact of the reference population size of the breed on our results, we used the same approaches for a training dataset of 404 bulls. Results are presented in Table 4.

Table 4. Weighted correlations between observed and predicted DYD, average correlation and slope for the 6 traits using BLUP, within-breed GS and between breed GS with 404 Normande bulls.

404 bulls in training population	BLUP	Within-breed BayesCpi	Multi-breed BayesCpi
Milk	0.23	0.31	0.35
FY	0.30	0.39	0.40
PY	0.24	0.33	0.35
Fat %	0.31	0.40	0.48
Protein %	0.38	0.47	0.49
SCS	0.40	0.42	0.43
Average correlation	0.31	0.39	0.42
Average slope	0.66	0.73	0.71

As expected, average correlations increased when the number of individuals in the reference population increased. Higher correlations are partly explained by a better

estimated polygenic component due to a higher proportion (83%) of validation animals with their sire and their two grandsires in the training population. Again, GS approaches performed better than pedigree-BLUP but differences between multi-breed GS and within-breed GS were smaller. The increase in correlation was 2.3% in average and varied from 1 to 8% depending on the trait. Regression slopes were closer to 1 which suggests that low slopes were mainly due to the low number of animals in the training dataset.

To investigate if multi-breed GS could be interesting for a breed with an already large reference population, therefore, all animals of the Normande breed reference population (except those of the validation dataset) were considered. Results are presented in table 5.

Table 5. Weighted correlations between observed and predicted DYD, average correlation and slope for the 6 traits using BLUP, within-breed GS and between breed GS with 1597 Normande bulls.

1597 bulls in training population	BLUP	Within-breed BayesCpi	Multi-breed BayesCpi
Milk	0.32	0.48	0.50
FY	0.35	0.49	0.52
PY	0.30	0.49	0.51
Fat %	0.35	0.64	0.65
Protein %	0.40	0.63	0.64
SCS	0.43	0.56	0.57
Average correlation	0.36	0.56	0.57
Average slope	0.76	0.88	0.86

Again correlations were higher with an increased number of animals in the training population and GS approaches performed better than pedigree-based evaluation. The average increase in correlation with GS was around 20% when it was 7.5% with lower reference population size. Here, the benefit of multi-breed genomic selection was rather low. A gain was observed for all traits but the average gain in correlation was only 1.6%. The relatively large reference population size for the Normande breed allowed an accurate estimation of QTL effect and EBV prediction. We also observe average slopes of regression closer to 1.

Conclusions

This study proposes a comparison of genomic selection approaches for small breeds and six traits. It focuses on the benefit of multi-breed genomic selection for breeds with less than five hundreds animals in their reference population. We observed an increase in correlation between 5 and 9 % with genomic selection compared to pedigree BLUP and the lowest gain was observed for a training population of less than 200 animals. Compared to within-breed genomic selection, multi-breed genomic selection allows a gain in correlation between EBV and DYD which ranges between 1 and 8 percent. The highest increase was observed for traits influenced by large QTL (fat yield and fat percentage). It showed that conserved linkage disequilibrium across breeds is observed and that multi-breed genomic selection may benefit to breeds with small reference population size. However, gain in correlation with multi-breed GS is low when reference population is already large. Computing time required by the GS3 software with HD genotype was high. Methods and parameters used here were consequently optimized for a specific medium density dataset and would probably need to be adapted for complementary analyses. Further investigation on homogenization of phenotypes between breeds and multi-breed genomic selection methods is also required. Moving to haplotypic-based methods should improve multi-breed genomic selection through an increased level of linkage disequilibrium and a better estimation of QTL effect.

Acknowledgements

This work was perform within the “Unité Mixte de Technologie Gestion Génétique et Génomique des populations bovines” (UMT3G) of INRA GABI and is part of the GEMBAL project, funded by the Agence Nationale de la Recherche (ANR-10-GENM-0014), APISGENE, Races de France and

INRA "AIP Bioressources". The Eurogenomics consortium provided most of the Holstein HD genotypes. Most 50K genotypes originated from the Cartofine-ANR-05-GENANIMAL-007 project funded by ANR (French National Research Agency) and ApisGene, from the Eurogenomics consortium, and from genomic selection activity undertaken by the French cattle breeding companies, with LABOGENA as main genotyping lab.

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. *American journal of human genetics* 84.2, 210-223.

Hozé, C., Fouilloux, M.N.F., Venot, E., Guillaume, F., Dassonneville, R., Fritz, S., Ducrocq, V., Phocas, F., Boichard, D. & Croiseau, P. 2013. High density marker imputation efficiency in 16 French cattle breeds. *Genet. Sel. Evol.* in press

Jensen, J., Mantysaari, E.A., Madsen, P. & Thompson, R. 1996. Residual maximum likelihood estimation of (co)variance components in multivariate mixed linear models using average information. *J. Ind. Soc. Agric. Statistics 49*, 215-236.

Kizilkaya, K., Fernando, R.L. & Garrick, D.J. 2010. Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. *J Anim Sci. 88,* 544-551.

Peers, I. 1996. *Statistical Analysis for Education and Psychology Researchers.* Ed. The Falmer Press. Washington, DC.

VanRaden, P. 2008. Efficient methods to compute genomic predictions. *J Dairy Sci. 91*, 4414-4423.