Comparison of Accuracies of Genomic Prediction in French Limousin Cattle Population according to the Number of Markers and to Pedigree Relationship between Training and Validation Populations

M. Barbat [1], T. Tribout[2,3], R. Saintilan[1], E. Venot[2,3], M.N. Fouilloux[4], F. Phocas[2,3]
[1]UNCEIA, F-78350 Jouy-en-Josas, France, [2]INRA, UMR1313 GABI, F-78350 Jouy-en-Josas, France,
[3]AgroParisTech, UMR1313 GABI, F-75231 Paris, France, [4]Institut de l'Elevage - Idele, F-78350, France

ABSTRACT: Genomic breeding values were estimated in the French Limousin cattle breed for direct and maternal genetic effect traits routinely recorded at birth and weaning. A total of 1,646 bulls were genotyped with the Bovine SNP50 BeadChip® or the BovineHD BeadChip®. Their deregressed EBV were used in weighted analyses using a BayesC approach. Chip density and proportion of SNP with a non-zero effect did not affect the prediction accuracy. Accuracies of genomic EBV (GEBV) varied from 0.38 to 0.71 for direct effect traits and from 0.05 to 0.36 for maternal effect traits, when the validation animals had their sire in the training population. The GEBV of weakly related validation animals were on average 35% and 20% less accurate for direct and maternal effect traits, respectively. A prediction with only the 10,000 most important SNP resulted in similar GEBV accuracy than with all markers.
Keywords: beef cattle; genomic selection

Introduction

Genomic selection is a way to efficiently increase genetic gain by improving the accuracy of the breeding value estimates of young selection candidates that do not necessarily have their own performance record or progeny information. The accuracies of genomic estimated breeding values (GEBV) are keys to successful application of this technology in beef cattle populations. Many factors influence the accuracy of genomic selection, such as training population size and marker density (Erbe et al. (2012)), but also genetic structure of the population (Habier et al. (2011)).

The aim of this study is to investigate the role of such factors in the accuracy of genomic predictions for birth and weaning traits in Limousin breed, a beef cattle breed with a very large effective population size of about 2,000 (Bouquet et al. (2011)).

Materials and Methods

Genotype Data. A total of 1,646 registered French Limousin bulls were genotyped either with the Bovine SNP50 BeadChip® (50K) or with theBovineHD BeadChip® (777K). After quality control, 706,791 SNP of the 777K SNP chip were retained on 462 bulls and 37,634 SNP of the 50K SNP chip were retained on 1,184 bulls. Imputation of the 50K genotype data to 777K genotypes was performed using BEAGLE software (Browning and Browning (2009)). The SNP were mapped to the UMD 3.1 build of the bovine genome sequence assembled by the Center of Bioinformatics and computational Biology at the University of Maryland. Analyses were performed on 50K genotypes and on 777K genotypes (either true or imputed HD genotypes).

Phenotype Data. The 5 traits considered in French national genetic evaluation based on pre-weaning field data were studied: birth weight (BW), calving ease score (CE), weaning weight (WW), muscular development score (MD) and skeletal development score (SD). The response variables were deregressed estimate breeding values (DEBV) from traditional BLUP EBV according to Garrick et al. (2009) methodology. They were considered in weighted analysis to account for heterogeneous variances of the response variable due to a various amount of progeny records among genotyped animals. For BW, CE and WW, two different DEBV were considered for each phenotypic trait: direct DEBV for direct EBV and maternal DEBV for maternal EBV. Information about the number of records and reliability for each DEBV is presented in Table 1 for the full reference population and the two validation sets.

Table 1. Heritability (h^2), number (n) of animals in reference and validation populations and average reliability of DEBV (rel) for birth and weaning traits

Trait*	h^2	Full population		Close relatedness validation set		Weak relatedness validation set	
		N	rel	n	rel	n	rel
Direct							
BW	0.41	1,635	0.57	232	0.90	289	0.73
CE	0.04	1,622	0.23	232	0.57	289	0.37
WW	0.32	1,379	0.52	232	0.83	289	0.64
MD	0.22	1,305	0.44	232	0.76	289	0.55
SD	0.33	1,305	0.53	232	0.83	289	0.63
Maternal							
BW	0.04	683	0.33	232	0.33	166	0.43
CE	0.01	525	0.26	232	0.16	166	0.25
WW	0.06	601	0.34	232	0.29	166	0.42

* BW: Birth weight, CE: Calving ease, WW: Weaning weight, MD: Muscular development, SD: Skeletal development,

Population scenarios. Two different scenarios were tested depending on the degree of pedigree relationship between the training population and the

validation one. In the close relatedness (CR) scenario, the validation population was constituted of the animals which have their sire in the training set. In the weak relatedness (WR) scenario, the validation population was composed of animals which do not have their sire or progeny in the training set.

Statistical Models. Genomic prediction equations were derived using a BayesC approach (Habier et al. (2011)). The proportion π of SNP with a non-zero effect was fixed at 2 different values for each density of chip. A first π was tested to consider about 350 SNP fitted all together; the values were π=0.0005 for the HD genotypes (about 350 SNP) and π=0.01 for the 50K genotypes (about 380 SNP). A second π was tested to consider about 700 SNP fitted all together, corresponding to a scenario with a larger number of QTL. The values were π=0.001 for the HD genotypes (about 710 SNP) and π=0.02 for the 50K genotypes (about 750 SNP).

For each trait, the following model was fit for the response variable DEBV on training populations:

$$\mathbf{DEBV} = \mathbf{1}\mu + \mathbf{Ma} + \mathbf{e}$$

where **1** is a vector of 1, μ is the overall mean, ***M*** is an incidence matrix for marker genotypes. The genotypes are coded 0, 1 or 2 depending on the number of copies of a given marker allele the individual carried, **a** is a vector of marker effects, and **e** is a vector of residual effects.

Once the markers effects were estimated with BayesC method, the predicted genomic value (GEBV) of an individual was:

$$GEBVi = \sum_{j=1}^{J} M_{ij}\,\hat{a}_j$$

where $\hat{a}_j$ is the estimated effect of SNP *j*, $M_{ij}$is the genotype of individual *i* for SNP *j*, and *J* is the total number of markers.

The accuracy of genomic prediction was estimated for the validation population as the weighted correlation between DEBV and GEBV.

Analyses were performed using GS3 software (Legarra et al. (2013)). For each analysis, 50,000 iterations were run with a burn-in of 10,000 and a thin of 10.

Results and discussion

Accuracy of genomic predictions. Weighted correlations between DEBV and GEBV are presented in Table 2. The correlations for maternal effect traits were in general lower than for direct effect traits. It can be explained by the lower reliability and the lower number of records for maternal DEBV (Table 1). This was not the case for maternal versus direct effect GEBV for CE in the WR scenario, probably because no reliable results could be given for these traits with very low heritabilities and small number of animals in the reference population for maternal traits. Our results were in agreement with those of Saatchi et al. (2012) in the same breed but a bit lower because their validation criteria were genetic correlations whereas we used phenotypic correlations and because their reference populations were larger than ours (between 1,500 and 2,000 according to the traits).

For a given trait and a given population scenario, considering different markers densities and π values resulted in very limited changes in accuracies (<0.02) except for CE. For this trait, increasing the density of markers resulted in a strong increase (+0.10) or decrease (-0.04) in the GEBV accuracies for direct and maternal effects, respectively, in the WR population scenario. However the accuracy estimates for CE were very low and probably not reliable. To conclude, using HD instead of 50K genotypes did not allow any increase in accuracy of genomic predictions, despite the great genetic diversity of the Limousin population.

By contrast, pedigree relationship between the validation and training populations strongly influenced the prediction accuracy. When the animals of the validation population were closely related to the training set, the accuracies were in average higher (+ 0.20 for direct effect traits; +0.04 for maternal effect traits).

Because the value of π did not affect the accuracy of the prediction, the following results are only presented for the π corresponding to about 700 SNP with a non-zero effect.

The markers with large effect on genomic predictions. Less than 50 SNP per trait had an estimated effect greater than 1% of the phenotypic standard deviation, from 0 for all maternal effect traits and direct CE to 47 for BW. These results are presented for the 50K analyses on Figure 1. These numbers were even lower with the HD chip because the higher density of markers dilutes each SNP effect. Traits with higher heritability were those for which some markers captured a significant part of the DEBV variability. For a given trait, the number of SNP with large effects varied strongly across the CR and WR scenarios, although the proportion of common bulls in the two training sets was large. The markers with the largest effects were always among the 50 SNP which were the most often included in the genomic equation. Among these 50 most informative markers for each trait, only few SNP (between 2% for maternal WW and 32% for direct WW) were common to the two population scenarios. Further investigation is necessary to check the amount of close SNP with large effects that may have been co-selected between the 2 scenarios. However, one may wonder whether the genetic diversity of the Limousin breed is not too large to be able to fine map QTL explaining a large part of phenotypic variability for the entire Limousin population.

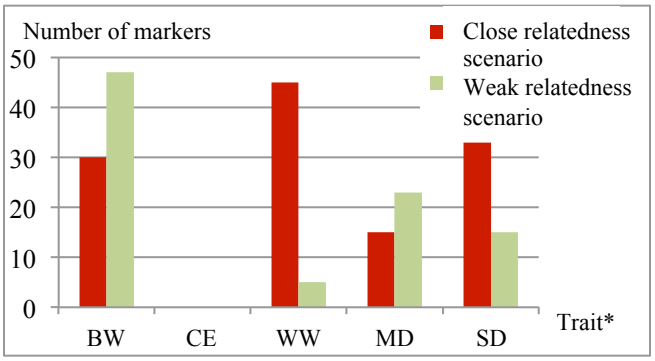

Figure 1. Number of markers with estimated effect larger than 1% of the DEBV standard deviation (50K genotypes)
* BW: Birth weight, CE: Calving ease, WW: Weaning weight, MD: Muscular development, SD: Skeletal development

The 10 000 most informative SNP. To assess the quality of GEBV based on a limited amount of SNP, GEBV of validation sets were predicted with only the 10,000 SNP per trait with the highest inclusion probability (most informative SNP) of the 50K chip. In that case, accuracies for direct GEBV (Table 3) and maternal GEBV (Table 4), were very closed to those obtained with all SNP (Table 2) whatever the population scenario. However, only 60 to 80% of direct genetic variances and 40 to 60% of maternal genetic variances were explained by these 10,000 most informative markers. At first glance, this suggests that using low density chip with about 10,000 markers could be sufficient to predict accurate GEBV for all traits. However, very few common SNP (7 for CR scenario and 15 for WR scenario) were shared by the 8 evaluated traits among the 10,000 most informative for each trait. This was mainly due to a very small number of common SNP shared by direct traits: 199 and 164 for the CR and WR scenarios respectively (Table 3). When considering only the small number of common SNP for direct traits, accuracy of genomic prediction dropped of 40% (for BW, WW and SD) to 70% (for CE and MD) for CR scenario, but was similar to those with the 10,000 most informative markers for WR scenario (except for CE and MD). The proportions of genetic variance explained by these common markers ranged from almost 0% for direct CE and MD to 5-10% for direct BW, WW and SD (Table 3). The proportion of genetic variance explained by the markers shared by the 3 maternal traits was very small (0.5 to 3%) despite the larger number of common markers (Table 4). However, the GEBV accuracies for maternal traits were very close when using only the common markers or the 10,000 most informative SNP.

Table 2. Accuracy of GEBV according to marker density, expected number of SNP with a non-zero effect (nSNP) and population scenario CR and WR

Scenario	CR				WR			
Chip density	50K		777K		50K		777K	
nSNP	380	750	350	710	380	750	350	710
Direct traits mean	**0.56**	**0.56**	**0.54**	**0.55**	**0.35**	**0.36**	**0.37**	**0.37**
BW	0.53	0.52	0.50	0.51	0.42	0.42	0.38	0.37
CE	0.39	0.39	0.38	0.38	-0.02	-0.02	0.09	0.08
WW	0.58	0.57	0.54	0.54	0.47	0.45	0.42	0.40
MD	0.67	0.69	0.69	0.71	0.33	0.38	0.38	0.40
SD	0.62	0.62	0.60	0.60	0.54	0.58	0.58	0.60
Maternal traits mean	**0.19**	**0.19**	**0.20**	**0.20**	**0.16**	**0.16**	**0.15**	**0.15**
BW	0.15	0.15	0.16	0.16	0.10	0.10	0.12	0.13
CE	0.35	0.35	0.36	0.35	0.17	0.17	0.14	0.13
WW	0.05	0.07	0.08	0.08	0.22	0.22	0.19	0.19

* BW: Birth weight, CE: Calving ease, WW: Weaning weight, MD: Muscular development, SD: Skeletal development

Table 3. Accuracies (acc) and proportion of genetic variance (pgv) explained by the 10,000 most informative markers for each trait and by the common markers among the 5 direct traits according to population scenario (50K genotypes)

Scenario	CR				WR			
Number of markers	199		10,000		164		10,000	
Parameter	acc	pgv	acc	pgv	acc	pgv	acc	pgv
Trait *								
BW	0.30	0.05	0.51	0.76	0.45	0.09	0.42	0.83
CE	0.12	0.003	0.39	0.62	0.31	0.008	-0.04	0.68
WW	0.34	0.11	0.57	0.73	0.40	0.08	0.45	0.79
MD	0.16	0.002	0.69	0.68	0.03	0.004	0.35	0.74
SD	0.41	0.09	0.62	0.76	0.50	0.06	0.56	0.80

* BW: Birth weight, CE: Calving ease, WW: Weaning weight, MD: Muscular development, SD: Skeletal development

Table 4. Accuracies (acc) and proportion of genetic variance (pgv) explained by the 10,000 most informative markers for each trait and by the common markers among the 3 maternal traits according to population scenario (50K genotypes)

Scenario	CR				WR			
Number of markers	1,273		10,000		1,828		10,000	
Parameter	acc	pgv	acc	pgv	acc	pgv	acc	pgv
Trait *								
BW	012	0.01	0.15	0.55	0.07	0.03	0.09	0.54
CE	0.32	0.005	0.35	0.45	0.05	0.01	0.17	0.45
WW	0.03	0.01	0.07	0.61	0.24	0.02	0.21	0.59

*BW: Birth weight, CE: Calving ease, WW: Weaning weight

Conclusion

The GEBV accuracies obtained were rather high despite the large genetic diversity of the French Limousin breed. However, because of the loss of accuracy with the WR scenario and the great differences within the large effects SNP between scenarios, one can wonder if the accuracy of the prediction will be as good as our results for new candidates having only distant relatives in the reference population.

To finish, our results show that within breed genomic predictions do not require the use of High Density genotypes. Besides, the conservation of good prediction accuracies for all traits with only the 10,000 most informative markers suggests that it could be possible to use a Low Density chip. This problematic has to be investigated because thanks to the lower price of this chip, its use in the field would be easier.

Literature cited

Bouquet, A., et al.. 2011. J. of Animal Sci. 89: 1719-1730

Browning, B. L., and Browning, S. R. (2009). The Am. J. Hum. Genet. 84: 210-223

Erbe, M., et al. (2012). J. Dairy Sci. 95: 4114-4129

Garrick, D. J., Taylor J. F., and Fernando R. L. (2009) Genet Sel Evol 41: 44-51

Habier, D., Fernando, R.F., Kizilkaya, K., Garrick, D.J. (2011). BMC Bioinformatics 12:186-197

Legarra, A., Ricard, A., Filangi, O. (2013). http://snp.toulouse.inra.fr/~alegarra.

Saatchi, M., Schnabel, R. D., Rolf, M. M., et al. (2012). Genet Sel Evol, 44: 38-47

Acknowledgements

GEMBAL project (ANR-10-GENM-0014) is funded by ANR, ApisGene, Races de France and INRA. The 50K genotypes originated from others ANR-ApisGene projects and from French cattle breeding companies' activity.