Comparison of Genomic Selection Approaches in Brown Swiss within Intergenomics

***P. Croiseau*[1,] *F. Guillaume*[2] *and S. Fritz*[3]**

[1] INRA, UMR 1313, GABI, 78350 Jouy-en-Josas, France
[2] Institut de l'Elevage, 149 rue de Bercy, 75012 Paris, France
[3] UNCEIA, 149 rue de Bercy, 75012 Paris, France

Abstract

The European Brown Swiss federation, in collaboration with Interbull, funded and managed a project named Intergenomics. The goal of this project is to perform genomic evaluations of sires based on a joint analysis of all the genotypes collected around Europe. To date, six countries are involved in Intergenomics and according to the country, between 3 and 15 traits are available. In this study, we propose to compare a panel of 4 genomic selection approaches to the pedigree-based BLUP (Best Linear Unbiased Predictor). Among these 4 methodologies, performances of the genomic BLUP (GBLUP) were compared to 2 bayesian approaches (Bayesian LASSO and Bayes Cπ) and a variable selection approach (Elastic Net or EN). Except the GBLUP, the other genomic selection approaches deal with the p>>n problem (number of Single Nucleotide Polymorphism or SNP (p) is much higher than the number of bulls (n)).

We compare the correlations between observed and predicted deregressed proofs for the different traits, the different country scales and the different methods. Compared to the pedigree-based BLUP, genomic selection approaches allow a gain in correlation between 6.5 and 20.9%. Bayesian LASSO, Bayes Cπ and EN give the best results with a gain of correlation around 3% compared to a GBLUP. The slope of regression is also lowest with these three methods than with the pedigree-based BLUP and the GBLUP. Consequently, over the different country scale, the mean number of traits which validate the interbull test (slope of regression between 0.8 and 1.2) is lowest for the pedigree-based BLUP (6.4 traits in average) than for the Bayesian LASSO, Bayes Cπ and EN (between 7.8 and 8 traits in average).

Key words: genomic selection, dairy cattle, Intergenomics

Introduction

The European Brown Swiss federation, in collaboration with Interbull, funded and managed a project named Intergenomics. The goal of the project is to perform genomic evaluations of sires based on a joint analysis of all the genotypes collected around Europe. To date, six countries are involved in Intergenomics and between 3 and 15 traits are available depending on the country. In this study, we propose to compare a panel of 4 genomic selection approaches. Among these 4 methodologies, performances of the genomic BLUP (GBLUP) were compared to 2 bayesian approaches (Bayesian LASSO and Bayes Cπ) and a variable selection approach (Elastic Net or EN). All these genomic approaches were confronted to the pedigree-based BLUP (Best Linear Unbiased Predictor). Among the different results, the correlation between observed and estimated performances (deregressed proofs or DP) and the slope of regression were investigated for each trait in each country scale.

Materials and Methods

The data set consisted of 7041 progeny tested Brown Swiss bulls genotyped with the Illumina Bovine SNP50 BeadChip®. Since, the contribution of each country to the reference population is very different, we expect differences in t accuracies in function of the country scale. For a given trait, countries with a high contribution to the reference population are supposed to convert CD of abroad countries with a higher accuracy.

All the countries don't have the same number of traits evaluated so, to be able to compare results between countries, only common traits were considered. The Slovenian scale was removed because only the 3 production traits were available (Table 1).

Table 1. Number of traits evaluate in each country.

	Swiss	German	French	Italian	Slovenian	US
Nb of traits	15	15	13	12	3	15
production	3	3	3	3	3	3
type	9	9	8	7	0	9
functional	3	3	2	2	0	3

Among the 10 traits considered, there are 3 production traits (Fat, Milk and Protein), 6 type traits (foot angle, front teat length, rump angle, rear leg side view, rear udder height and stature) and 1 functional trait (direct longevity).

After a control quality based on minor allele frequency (1%), call rate (10%) and Hardy Weinberg Equilibrium test (10^{-4}), 42862 SNPs were retained. Mendelian segregation was checked. On the complete set of available animals, only bulls with genotypes and index who belong to a family of at least four animals with genotypes and index were retains. From the 7041 genotyped animals, this selection led to 4437 animals retained. For these animals, index and reliability were used to produce DP and equivalent daughter contribution (EDC). To infer missing genotypes and phases, DAGPHASE software, which is based on Beagle software, was used (Druet and Georges, 2008; Browning and Browning, 2009). A retrospective cross-validation scheme was chosen where animals of the training population are born before 2002 and they all have DP calculated in 2007. This training population contains between 445 and 3430 animals according to the trait. The animals of the validation population are born between 2002 and 2007. So, they have DP calculated in 2011. Prediction equations are produced using the training population and GEBV_{2007} are estimated for the validation population based on these prediction equations (Figure 1). Then, the weighted Pearson correlation is calculated between GEBV_{2007} and DP_{2011} where the weight is the EDC (Peers, 1996).

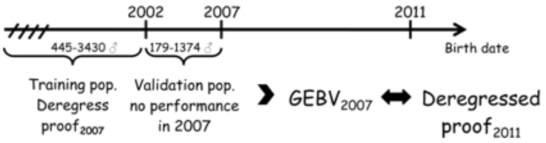
2002
2007
2011
445-3430 ♂
179-1374 ♂
Birth date
Training pop. Deregress proof2007
Validation pop. no performance in 2007
GEBV2007
Deregressed proof2011

Figure 1. Retrospective cross-validation scheme used to compare estimated and observed deregressed proofs.

Statistical Analysis

We use GS3 software (http://snp.toulouse.inra.fr/~alegarra/) to compare 3 genomic selection methodologies: the GBLUP (VanRaden, 2008), which uses the genomic relationship matrix, G (Habier *et al.*, 2007; VanRaden, 2008), instead of the pedigree-based relationship matrix, the Bayesian LASSO (Legarra *et al.*, 2011) and the BayesCπ (Kizilkaya *et al.*, 2010). These three methods allow including a polygenic component. Proportion between 10 and 90% were tested (by step of 10%) and in the results, the proportion which provides optimal correlations were shown.

In addition to these methods, a penalized regression approach was also tested, the Elastic Net (Zou and Hastie, 2003 and 2005). For the Elastic Net, the SNP pre-selection described in Croiseau *et al.*, 2011 was tested. No polygenic component could be included in the model with this approach.

Among the genomic selection methods tested, Bayesian LASSO, BayesCπ and EN were developed to solve the p>>n problem (the very large number of SNP compared to the number of animals). Nevertheless, Bayesian LASSO and BayesCπ led to an estimated effect for each SNP what is different of EN algorithm where only a part of the SNP panel will have an estimated effect and the other SNP effect will be set to 0.

Finally, all these methods were compared to a traditional pedigree-based BLUP. For the pedigree-based BLUP, genetic parameters were estimated using an Average Information-Restricted Expectation Maximization Likelihood (AI-REML) approach (Jensen *et al.,* 1996).

Results & Discussion

Table 2 presents the weighted correlation between $GEBV_{2007}$ and DP_{2011} for the 10 traits and for each country scale.

Table 2. weighted correlation between $GEBV_{2007}$ and DP_{2011} for the 10 traits and for each country scale.

traits	countries	BLUP	GBLUP	Bayesian Lasso	Bayes Cπ	EN
fat	**Swiss**	0.416	0.579	0.605	0.607	**0.612**
	French	0.388	0.544	0.584	0.585	**0.592**
	US	0.398	0.542	0.591	0.592	**0.600**
	Italian	0.423	0.582	0.610	0.618	**0.616**
	German	0.427	0.584	0.615	**0.624**	0.587
milk	**Swiss**	0.334	0.518	0.561	0.564	**0.582**
	French	0.309	0.476	0.530	0.531	**0.550**
	US	0.287	0.450	0.525	0.526	**0.551**
	Italian	0.343	0.526	0.562	0.562	**0.572**
	German	0.371	0.546	0.579	**0.580**	0.435
protein	**Swiss**	0.448	0.575	0.601	0.602	**0.603**
	French	0.385	0.510	0.544	0.545	**0.553**
	US	0.382	0.506	0.554	0.555	**0.565**
	Italian	0.436	0.569	0.593	0.593	**0.592**
	German	0.454	0.590	0.603	0.604	**0.626**
foot angle	**Swiss**	0.369	0.429	0.462	**0.559**	0.473
	French	0.320	0.391	0.418	0.419	**0.420**
	US	0.366	0.444	0.456	0.458	**0.466**
	Italian	0.370	0.418	0.454	**0.455**	0.454
	German	0.380	0.424	0.462	0.462	**0.647**
Front Teat Length	**Swiss**	0.486	0.616	**0.659**	0.657	0.656
	French	0.465	0.619	0.652	**0.659**	0.653
	US	0.466	0.617	0.652	**0.657**	0.653
	Italian	0.464	0.616	0.653	**0.656**	0.655
	German	0.467	0.617	0.651	**0.657**	0.556
Rump Angle	**Swiss**	0.418	0.520	0.551	**0.555**	0.547
	French	0.438	0.525	0.557	**0.558**	0.550
	US	0.432	0.524	0.552	**0.554**	0.547
	Italian	0.418	0.516	0.549	**0.552**	0.544
	German	0.441	0.540	0.561	0.563	**0.587**
Rear Leg Side View	**Swiss**	0.473	0.555	**0.574**	0.573	0.569
	French	0.419	0.502	**0.529**	0.528	0.521
	US	0.401	0.485	**0.516**	0.516	0.510
	Italian	0.408	0.496	**0.520**	0.519	0.513
	German	0.416	0.494	0.522	0.521	**0.528**
Rear Udder Height	**Swiss**	0.436	0.506	0.528	**0.529**	0.508
	French	0.345	0.421	**0.445**	0.444	0.433
	US	0.427	0.508	**0.527**	**0.527**	0.516
	Italian	0.405	0.480	**0.497**	**0.497**	0.491
	German	0.368	0.464	0.475	0.474	**0.551**
stature	**Swiss**	0.407	0.522	0.575	0.579	**0.601**
	French	0.362	0.501	0.576	**0.588**	0.587
	US	0.413	0.554	0.605	**0.618**	0.617
	Italian	0.406	0.553	0.604	**0.618**	0.613
	German	0.364	0.525	0.579	**0.593**	0.392
direct longevity	**Swiss**	0.282	0.288	0.293	0.293	**0.331**
	French	0.272	0.297	0.296	0.296	**0.316**
	US	0.266	0.314	0.321	0.323	**0.345**
	Italian	0.312	0.346	0.340	0.341	**0.352**
	German	0.354	0.391	0.363	0.363	**0.466**

We first observe that genomic selection improves correlations drastically whatever the trait and whatever the country scale. Then, depending on the trait, the optimal correlation is shared by Bayesian LASSO, BayesCπ and EN but for a same trait, the optimal correlation is very often given by the same method whatever the country scale, even if, most of the time, no strong difference appears between these 3 approaches.

Surprisingly, for the German scale, BayesCπ and EN results are very contrasted. For instance, for milk yield, EN is the best method for all the country scale except for the German scale where the correlation is 12 points lower what is not the case for BayesCπ. However, for rear udder height trait, the opposite situation appears. BayesCπ give the optimal correlation whatever the country scale except for the German scale where the EN improves the correlation of 7 points.

For a better overview of the genomic selection efficiency, the mean correlation per country scale and per method where respectively presented in table 3 and 4. The correlation was supposed to be dependant on the contribution of each country to the reference population. For a given trait, countries with a high contribution to the reference population are supposed to convert CD of abroad countries with a higher accuracy. Table 3 provides an idea of how contribution to the reference population impact correlations (other factors like the quality of the phenotype also impact this correlation). Consequently, the lowest correlation is obtained using the French scale, the country with the lowest contribution (about 200 animals) and the highest correlation is obtained using the Swiss population, the country with the highest contribution (about 2000 animals). However, the correlation variation is not very high.

Table 3. mean correlation over the 10 traits using BayesCπ method in function of the country scale.

country scale	mean correlations
Swiss	0.552
French	0.515
US	0.532
Italian	0.541
German	0.544

To compare the efficiency of the different genomic selection approaches, the mean correlation over the 10 traits (all countries scale confounded) and the absolute deviation to 1 of slope of regression were presented on table 4. Concerning the correlation, as expected, the pedigree-based BLUP gives the lowest one and all genomic selection methods improve this correlation of at least 10 points. The GBLUP comes in second position and is clearly less efficient than the 3 others methods (Bayesian LASSO, BayesCπ and EN) which improve correlation of 3 points. Concerning the slope of regression, there is clearly 2 groups of methods, from one hand, the pedigree-based BLUP and the GBLUP with an absolute deviation to 1 of 0.18 in average (over the 10 traits), and, from the other hand, the Bayesian LASSO, BayesCπ with a deviation of 0.11 and EN which is a bit less efficient with a deviation of 0.13.

Table 4. mean correlation and absolute deviation to 1 of slope of regression over the 10 traits (all country scale confounded) in function of the genomic selection approach.

	mean correlations	Absolute deviation to 1 of slope of regression
pedigree-based BLUP	0.393	0.184
GBLUP	0.502	0.182
Bayesian LASSO	0.533	0.110
BayesCπ	0.537	0.109
EN	0.536	0.133

Obviously, these results have some consequences on the validation of the Interbull test. For the routine genomic evaluations, only traits where the slope of regression is between

0.8 and 1.2 are validated (Mantysaari *et al.*, 2010). Table 5 shows the number of traits which validate the Interbull test for each method. If no method allows validating the Interbull test for all traits, Bayesian LASSO, BayesCπ and EN clearly validate more traits. However, for GBLUP, Bayesian LASSO and BayesCπ, the polygenic component part retained is the one which maximize the correlation independently of the slope of regression and for the EN, a large panel of combination of parameters were tested and the retained one is based on the same criteria. So, for the traits which do not validate the Interbull test, it is possible to retain the best correlation among the solutions which validate the Interbull test.

Table 5. Number of traits which validate the Interbull test for each method and each country scale.

country scale	pedigree-based BLUP	GBLUP	Bayesian LASSO	Bayes Cπ	EN
Swiss	5	7	7	7	7
French	6	4	9	9	8
US	6	7	8	8	6
Italian	7	5	8	8	9
German	6	9	8	8	9
mean	6	6.4	8	8	7.8

We would like to know if the optimal polygenic component part is dependant of the trait. Table 6 show the mean polygenic component part retained over the 5 country scale for each trait. For GBLUP, the optimal polygenic component part is 10% for the big majority of the traits. Only the direct longevity deviates noticeably from this value. For Bayesian LASSO but particularly for BayesCπ, the optimal polygenic component part varies a lot according to the trait and, is higher than for GBLUP. For instance, with BayesCπ, Read Udder Height needs a polygenic component part of 44% and 52% for Bayesian LASSO. Others traits like direct longevity or foot angle required a polygenic component part higher than 30%.

Table 6. Mean of the polygenic component part retained over the 5 country scale for each trait and for each method.

	traits	GBLUP	Blasso	BayesCpi
production	fat	10%	10%	14%
	mil	10%	10%	12%
	pro	10%	18%	22%
type	fan	10%	10%	42%
	ftl	10%	16%	30%
	ran	10%	10%	22%
	rls	14%	24%	20%
	ruh	12%	52%	44%
	sta	10%	10%	18%
functional	dlo	22%	30%	36%

Conclusions

This study proposes a comparison of the genomic selection approaches in Brown Swiss for 10 traits. Among the countries which share data, the contribution to the reference population vary between around 200 and 2000 animals what can have an impact on the accuracy to convert CD of abroad animals and consequently, on the efficiency of the genomic selection approach. In this context, a comparison of the results for each country scale is proposed.

Compared to a pedigree-based BLUP, genomic selection allows a gain in correlation between GEBV and DP which range between 6.5 and 20.9 points. Among the set of genomic selection approaches tested, Bayesian LASSO, BayesCπ and EN give the best results with a gain in correlation near 3 points compared to a GBLUP. Moreover, with this set of three methods, the slope of regression is closer to 1 than for the pedigree-based BLUP and GBLUP and a higher number of traits validate the Interbull test.

As expected, correlations obtained in the French scale (where the contribution to the reference population is the lowest) are, in average over the 10 traits, 4 points lower than for the Swiss scale (where the contribution is the higher).

Acknowledgements

This work was perform within the 'UMT 3G'.

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.

Croiseau, P., Legarra, A., Guillaume, F., Fritz, S., Baur, A., Colombani, C., Robert-Granié, C., Boichard, D. & Ducrocq, V. 2011. Fine tuning genomic evaluations in dairy cattle through SNP pre-selection with the Elastic-Net algorithm. *Genetics research 93:6,* 409-417.

Druet, T. & Georges, M. 2009. A Hidden Markov Model Combining Linkage and Linkage Disequilibrium Information for Haplotype Reconstruction and QTL Fine Mapping. *Genetics 184,* 189-198.

Habier, D., Fernando, R.L. & Dekkers, J.C. 2007. The impact of genetic relationship information on genome-assisted breeding values. *Genetics 177,* 2389-2397.

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.

Legarra, A., Robert-Granié, C., Croiseau, P., Guillaume, F. & Fritz, S. 2011. Improved Lasso for genomic selection. *Genet Res. 93,* 77-87.

Mantysaari, E., Zengting, L. & Van Raden, P. 2010. Interbull Validation Test for Genomic Evaluations. *Interbull Bulletin 41,* 17-21.

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.

Zou, H. & Hastie, T. 2003. Regression shrinkage and selection via the Elastic Net, with application to microarrays. *Technical report, Standford University.*

Zou, H. & Hastie, T. 2005. "Regularization and variable selection via the Elastic Net." *J. R. Statist. Soc. B. 67,* 301-320.