Improving Genomic Evaluation Strategies In Dairy Cattle Through SNP Pre-Selection

P. Croiseau[*¶], C. Colombani[†], A. Legarra[†], F. Guillaume[‡], S. Fritz[§], A. Baur[§], R. Dassonneville[*‡], C. Patry[*§], C. Robert-Granié[†] and V. Ducrocq[*]

Introduction

The availability of the 54K SNP array has considerably changed the landscape of selection in dairy cattle. We are now able to use genomic prediction instead of pedigree-based genetic evaluations in selection programs. To perform genomic selection, various methods have been proposed with more or less success depending on the trait, the breed, the training population size…

In this paper, we investigate three different genomic evaluation methods: (i) the Genomic BLUP where the effect of each SNP is estimated by BLUP (Best Linear Unbiased Prediction) assuming known variances. (ii) the Elastic Net algorithm (EN), a variable reduction method. (iii) the Partial Least Squares algorithm which projects dependent variables onto a new space of lower dimension. EN and PLS were developed especially to solve the p>>n problem (very large number of dependent variables compared to the number of individuals).

These three methods were compared to pedigree-based BLUP. In a second step, we show how one can successfully improve the accuracy of the equation of prediction, whatever the method used, using pre-selection of the SNP based on results from QTL detection by linkage disequilibrium-linkage analysis..

Material and methods

Data: The data consisted in 678 Montbéliarde bulls and 1756 Holstein bulls genotyped with the Illumina Bovine SNP50 BeadChip®. With a minor allele frequency of 5%, 38959 and 41101 SNP were retained for the Montbéliarde and Holstein breed respectively. Mendelian segregation was checked. Large familial information about the animal was used to infer missing genotypes and phases with a very low error rate. All these steps were performed using the DualPHASE software (Druet and Georges 2009).

The data was divided into a learning population to estimate prediction equations and a validation population where prediction equations were used to predict the phenotypes (average future performance of daughters) and compared to the observed values. The learning population consisted in 451 Montbéliarde bulls and 1216 Holstein bulls while the validation population was composed by 227 Montbéliarde bulls and 540 Holstein bulls.

[*] INRA, UMR1313 - Génétique Animale et Biologie Intégrative, 78352 Jouy en Josas, France
† INRA, UR 631, Station d'Amélioration Génétique des Animaux, F-31320 Castanet-Tolosan, France
[‡] Institut de l'Elevage, 149 rue de Bercy, 75595 Paris, France
[§] UNCEIA, 149 rue de Bercy, 75595 Paris, France
[¶] This work has been financed by ANR project AMASGEN

Five traits were studied using DYD (Daughter Yield Deviation) (Mrode and Swanson 2004): Milk yield, Fat yield, Fat percent, Protein yield and Protein percent. DYD were weighted by their variance using their Effective Daughter Contribution (EDC).

Genomic evaluation methods: Genomic BLUP (GBLUP) fits allelic effects as random, with known variance. An equivalent model, integrating a genomic relationship matrix was used (VanRaden 2008).

$$G = ZZ'/2\sum p_i\left(1-p_i\right).$$

The same variance components as in the regular genetic evaluation are used.

The second method used is the Elastic Net approach (EN) (Zou and Hastie 2005). As for other penalized regression approaches, EN reduces the number of variables by removing, from the complete set of SNP, the ones which have an effect to weak to be significant. The underlying idea is that markers with small effect generate noise for the estimation of the markers of interest. Consequently, after removing them, we get a better estimation of the remaining set of SNP. EN corresponds to a combination of Ridge Regression (RR, equation 1) and LASSO procedures (equation 2). It has been shown that, in presence of correlated explanatory variables (i.e. linkage disequilibrium), RR retains all the predictors and distributes the estimated effect among them while LASSO retains only one predictor and removes the others (Zou and Hastie, 2003 and 2005). In this context, the use of the EN algorithm (equation 3) provides a more flexible tool. In the EN algorithm, an extra parameter α, which takes values in [0,1], is used to weight RR and LASSO penalties. With α=0, a complete LASSO model is defined whereas with α=1, we are in a complete RR context.

$$\hat{\beta}_{RR} = \arg\min\left\{\sum_{i=1}^{n}\left(Y_i - X_i\beta\right)^2 + \lambda\sum_j \beta_j^2\right\} \quad (1)$$

$$\hat{\beta}_{LASSO} = \arg\min\left\{\sum_{i=1}^{n}\left(Y_i - X_i\beta\right)^2 + \lambda\sum_j \left|\beta_j\right|\right\} \quad (2)$$

$$\hat{\beta}_{EN} = \arg\min\left\{\sum_{i=1}^{n}\left(Y_i - X_i\beta\right)^2 + \lambda\left(\alpha\sum_j \beta_j^2 + \left(1-\alpha\right)\sum_j \left|\beta_j\right|\right)\right\} \quad (3)$$

β is the vector of SNP effect, Y is the phenotype, X is the vector of genotypes.

Recently, Friedman *et al.* (2008) proposed a fast algorithm which use the cyclical coordinate descent algorithm, computed along a regularization path. This algorithm was implemented in an R package named "glmnet" (http://cran.r-project.org/web/packages/glmnet/index.html).

The last approach used was the Partial Least Squares regression (PLS) (Wold 1966; Hastie *et al.* 2001). PLS creates orthogonal score vectors (also called latent vectors) by maximizing the covariance between the response variable and the latent variables. The underlying assumption of PLS is that the observed data are generated by a system which is driven by a small number of latent (not directly measured) variables. The main idea is to perform successive regressions by projections onto latent structures to uncover hidden underlying biological effects. The number of desired dimensions must be provided. This algorithm was developed in an R package named "integrOmics" (http://cran.r-project.org/) (Lê Cao *et al.* 2009). However, the relationship between animals (with EN and PLS) is not accounted for.

SNP Pre-selection: The impact on the prediction equations of a pre-selection of the SNP used in the genomic evaluations was assessed. Pre-selection was based on results of QTL detection. This QTL detection combines a Linkage Disequilibrium analysis and Linkage

Analysis (LDLA) (Druet *et al.* 2008). It is a haplotype based method using a window of 6 SNP. From this LDLA analysis, Likelihood Ratio Test statistics for each point in the genome were obtained. Two criteria were retained to qualify a LRT peak. To be a LRT peak, the haplotype must have the highest LRT value within a window of plus or minus 25, 50, 100 or 200 SNP (window size) and a LRT higher than either 3, 5, 7 or 9 (LRT threshold). For instance, for milk yield, depending on value for the combination of parameters, we observe a number of LRT peaks included between 19 and 211 in Montbéliarde and between 27 and 239 in Holstein. Then, in a final step, the 50 SNP around the central SNP of each LRT peak were included in the GBLUP, EN or PLS analyses. Several values of numbers of SNP to include in the evaluation method were tested and 50 SNP gave the best results (not shown).

Calculation of the quality of the prediction equation:

Prediction equation were applied to the animals of the validation population to get estimated DYD (DYD_{est}). Then, the weighted correlation between DYD_{est} and observed DYD obtained after progeny test (DYD_{obs}) was computed, with weights equals to the EDC.

Results and discussion

First, pedigree-based BLUP, GBLUP, EN and PLS were applied using the whole set of SNP (table 1). For EN and PLS, results shown correspond to the "best" combination of parameters found by a grid search (for example a window size of 50 and a LRT threshold of 5 for EN).

It can be seen that genomic methods improves the correlation between DYD_{est} and DYD_{obs} compared to a pedigree-based BLUP approach based on a pedigree. Among the different genomic approaches, the best results were obtained with GBLUP and EN for the Montbéliarde breed and with PLS and EN for the Holstein breed. In table 1, the gain in accuracy varies across methods from 8.9% for Milk yield to 19.6% for Protein percent in Montbéliarde and from 1.1% for Milk yield to 35.1% for Fat percent in Holstein.

Table 1: Correlation between DYD_{est} and DYD_{obs} in Montbéliarde and Holstein breed obtained using pedigree-based BLUP, GBLUP, PLS and EN.

		Milk	Protein	Fat	Protein %	Fat %
Montbéliarde	pedigree-based BLUP	0.273	0.276	0.355	0.214	0.372
	GBLUP	0.350	**0.410**	**0.460**	**0.410**	0.340
	PLS	0.313	0.376	0.450	0.386	0.363
	Elastic-Net	**0.362**	0.404	0.374	0.364	**0.508**
Holstein	pedigree-based BLUP	0.423	0.330	0.317	0.449	0.390
	GBLUP	0.420	0.310	0.340	0.440	0.590
	PLS	**0.434**	**0.344**	0.376	0.501	0.594
	Elastic-Net	0.400	0.280	**0.436**	**0.591**	**0.741**

Table 2: Correlation between DYD_{est} and DYD_{obs} in Montbéliarde and Holstein breed obtained using pedigree-based BLUP, GBLUP, PLS and EN after a QTL-detection based SNP pre-selection.

		Milk	Protein	Fat	Protein %	Fat %
Montbéliarde	pedigree-based BLUP	0.273	0.276	0.355	0.214	0.372
	GBLUP	**0.490**	0.460	0.460	**0.480**	0.510
	PLS	0.458	0.473	0.492	0.485	0.521
	Elastic-Net	0.488	**0.549**	**0.495**	0.449	**0.566**
Holstein	pedigree-based BLUP	0.423	0.330	0.317	0.449	0.390
	GBLUP	**0.490**	**0.390**	0.490	0.590	0.660
	PLS	0.398	0.324	0.459	0.607	0.693
	Elastic-Net	0.480	0.387	**0.507**	**0.641**	**0.719**

The same analysis was run after SNP pre-selection based on QTL detection (table 2). Over all methods, a strong increase of the correlation whatever the trait studied was observed in almost all cases. Gains between strategies (ie between table 1 and 2) ranged from 3.5% for Fat yield to 13.9% for Protein yield on Montbéliarde and from -2.2% for Fat percent to 7.1% for Fat content on Holstein. Fat percent with EN on Holstein was the only trait which gave a lower correlation after SNP pre-selection. The best results were obtained either with GBLUP or EN. Compared to pedigree-based BLUP, gains in accuracies ranged between 14% for Fat yield and 27.3% for Protein yield in Montbéliarde and between 6% for Protein yield and 32.9% for Fat percent in Holstein.
The important point in this analysis is that, almost all the time, a significant gain of correlation was observed when a SNP pre-selection based on a QTL detection analysis was implemented, compared to the situation when the complete set of SNP was used in the genomic model.

Conclusion

The use of the complete set of SNP to predict a relevant prediction equation leads to an over parameterization of the model even when variable reduction methods are used. To alleviate this problem, we propose a pre-selection of the SNP based on a QTL detection procedure. With such a two-step approach, results were improved whatever the trait and whatever the breed. This strategy also offers a great advantage from a computing point of view. The limited number of SNP to deal with allows considering any population size and any number of SNP.
The choice of the method for an optimal estimation of SNP effect is still questionable, first because various methods give promising results, and second, because the list of approaches used in this study is not exhaustive. However, it must be stressed that different methods using a different number of SNP in the validation population could lead to approximately the same correlation between DYD_{est} and DYD_{obs}. If the number of QTL for each trait is unknown, it is probably more realistic to include hundreds rather than thousands or ten of thousands of SNP.

References

Druet, T., Fritz, S., Boussaha, M., Ben-Jemaa, S. *et al*. (2008) *Genetics* 178:2227-2235.

Druet, T., and Georges, M. (2009)

Friedman, J., Hastie, T. and Tibshirani, R. (2008) *Standford Univ*ersity.

Hastie, T., Tibshirani, R. and Friedman, J. (2001) *Springer-Verlag*. New York.

Lê Cao, K., González, I. and Déjean, S. (2009) *Bioinformatics.* 25:2855-2856.

Mrode, R. A. and Swanson, G. j. T. (2004) *Livestock Production Science*. 86:253-260.

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

Wold, H. (1966) *Academic Press*. New York.

Zou, H., and Hastie, T. (2003) *Department of Statistics*. Stanford University.

Zou, H., and Hastie, T. (2005) *Roy. Stat. Soc.* Ser. B 67:301-320.