SNP Selection For Low-density Assays For Genomic-Enabled Predictions Using Parent Averages

A.I. Vazquez[*†], G.J.M. Rosa[†], K.A. Weigel[†], D. Gianola[†‡], D.B. Allison[*]

Introduction

Progeny tests can provide highly accurate predictions of transmitting ability (**PTA**) for selection candidates. Nevertheless, they have the down-side of being expensive and time consuming. The recent availability of high-density genotyping (**HD**) of single nucleotide polymorphisms (**SNP**) has made possible research on incorporating whole genome data for selection for many livestock species. Those HD genotypes show the sharing of chromosome sections between individuals and are being applied for genomic selection (Meuwissen *et al.* 2001) attaining good reliability (VanRaden *et al.* 2009). For example, the correlation between PTA and genome-enabled predictions (**GPTA**) for net merit (**NM**) was 0.62 using a 50K SNP assay (Weigel *et al.* 2009). Unfortunately, massive genotyping with HD platforms is still not feasible in many situations due to the market cost per unit (e.g., US$ 225 for the BovineSNP50 Beadchip by Illumina). A low-density assay (**LD**) with a subset of informative SNPs is of great interest to genotype a larger number of animals for a lower price. The PTA and GPTA have shown an increasing correlation as the number of SNPs used increases, with a slowdown in the increase above 1,500 SNPs for NM (Weigel *et al.* 2009), milk, protein, fat, productive life, somatic cell score and daughter pregnancy rate (Vazquez *et al.* 2009). The former study showed that a GPTA using 1,500 SNPs had a correlation with PTA 10% smaller than the correlation achieved using HD platform. The selection of the best SNPs for NM provides an efficient set for multi-trait selection objectives. Evenly spaced SNPs across the chromosome had performance below that of target SNPs, but with the advantage of representing the whole genome (Habier *et al.* 2009). We also evaluated the gain in accuracy of the GPTA in HD and LD assay settings when considering parent averages for the trait being predicted. This study (a) evaluated different strategies for selecting subsets of SNPs; (b) quantified the predictive ability of SNP panels of different sizes; (c) evaluated the performance of strategies with predictions based on parent averages; and (d) compared those strategies with the GPTA that also uses parent averages of the sires.

Material and methods

Data. This study used USDA data of 3,715 sires genotyped with the Illumina Bovine SNP50 Bead Chip and their NM PTAs from the Animal Improvement Programs Laboratory at USDA-ARS Beltsville Agricultural Research Center. Models were trained with the sires

[*] Biostatistics Department, Section on Statistical Genetics, University of Alabama-Birmingham, USA.
[†] Dairy Science Department, University of Wisconsin-Madison, USA.
[‡] Animal Sciences Department, University of Wisconsin-Madison, USA.

born before 1999 (n=3,305) using their 2003 PTAs. The predictive ability of the models was assessed using the remaining bulls, born on or after 1999 (n=893), and their 2008 PTAs. Each locus was coded with "zero" or "two" if the SNP was either of the homozygous and "one" for the heterozygous locus. Loci with minor allele frequencies below 5% or with 10% or more of missing values were removed. Missing SNPs were imputed based on the most probable genotype indicated by conditional probabilities on the neighboring SNPs. The final data set contained information on 32,518 SNPs for the HD platform.

Statistical analyses. Standardized NM PTAs, $\mathbf{NM} = \left(NM_1,...,NM_n\right)'$ were regressed on SNP genotypes, $\mathbf{x}_i = \left(x_{i1},...,x_{ip}\right)'$, and on standardized parent-average NM PTAs, $\bar{p}_i$, $i \in \left(1,...,n\right)$, using a Bayesian LASSO model (Park and Casella 2008). The likelihood function is:

$$p\left(\mathbf{NM} \mid \beta_0, \boldsymbol{\beta}_1, \beta_2, \sigma_\varepsilon^2\right) = \prod_{i=1}^{n}\left(NM_i \mid \beta_0 + \mathbf{x}_i'\boldsymbol{\beta}_1 + \bar{\mathrm{p}}_i\beta_2, \sigma_\varepsilon^2\right), \qquad \text{Eq.(1)}$$

where β_0 is an effect common to all subjects; $\boldsymbol{\beta}_1 = \left(\beta_{1,1},...,\beta_{1,p}\right)'$ is a vector of marker effects, β_2 is the regression of NM on $\bar{p}_i$, and σ_ε^2 is the variance of model residuals. The prior distribution of model unknowns was as follows:

$$p\left(\beta_0, \boldsymbol{\beta}_1, \beta_2, \sigma_\varepsilon^2, \boldsymbol{\tau}^2, \lambda\right) \propto \left[\prod_{j=1}^{p} N\left(\beta_{1,j} \middle| 0, \sigma_\varepsilon^2\tau_j^2\right)\right] \chi^{-2}\left(\sigma_\varepsilon^2 \middle| S_\varepsilon, df_\varepsilon\right)$$
$$\times \left[\prod_{j=1}^{p} Exp\left(\tau_j^2 \middle| \lambda\right)\right] p\left(\lambda \middle| \alpha_1, \alpha_2, K\right), \qquad \text{Eq. (2)}$$

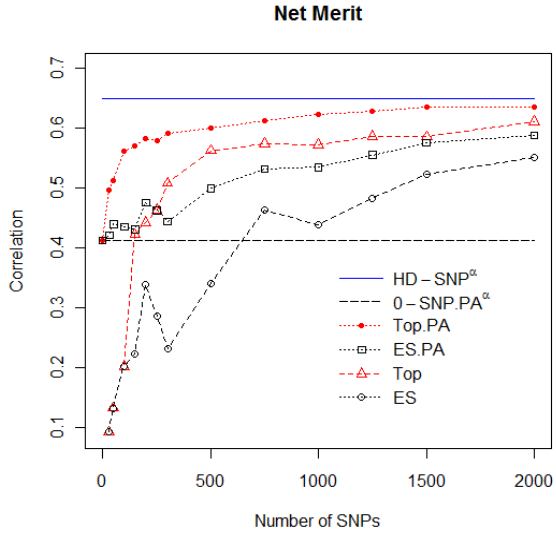
where $N\left(\beta_{1,j} \middle| 0, \sigma_\varepsilon^2\tau_j^2\right)$ is a normal density assigned to the j^{th} marker effect, $j \in \left(1,..., p\right)$, $\chi^{-2}\left(\sigma_\varepsilon^2 \middle| df_\varepsilon, S_\varepsilon\right)$ is a scaled-inverted Chi-square density with degree of freedom df_ε and prior scale S_ε, assigned to the residual variance; $Exp\left(\tau_j^2 \middle| \lambda\right)$ is an exponential prior for the scale parameters τ_j^2, and $p\left(\lambda \middle| \alpha_1, \alpha_2, K\right) \propto Beta\left(\lambda K^{-1} \middle| \alpha_1, \alpha_2\right)$ (see, de los Campos *et al.* 2009). Samples from the posterior distribution were obtained using the Gibbs sampler made available in de los Campos *et al.* (2009). In our application $df_\varepsilon = 1$, $S_\varepsilon = 0.5$, $\alpha_1 = \alpha_2 = 1.4$ and $K = 500$. Inferences were based on 55,000 samples obtained after discarding 10,000 as burn-in.

First, we fitted models regressing the PTAs on all markers (32,518), with (**HD-SNP.PA**) and without (**HD-SNP**) parent averages, together with a model with only the parent averages (**0-SNP.PA**) in the training set to the 2003 PTAs. Then, two strategies were used to select informative SNPs, including: (1) top SNPs selected based on the absolute value of their estimated effects on the HD-SNP model, and (2) SNPs evenly spaced (**ES**) along the whole genome. For each selection strategy, different panel sizes were evaluated including 30, 50, 100, 150, 200, 250, 300, 500, 750, 1000, 1250, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000, 7500, 10000, 15000, 20000, 25000, 30000 and the full set 32,518 SNPs. Later, we

fitted models with fewer SNPs on the subsets including parent average information (**Top.PA** and **ES.PA** for strategies 1 and 2, respectively) and excluding parent averages (**Top** and **ES** for strategies 1 and 2, respectively). Finally, we assessed predictive ability in the testing set via the correlation between GPTAs and the 2008 PTAs for all the models. The predictive abilities of the models were evaluated *versus* the predictive ability of the HD-SNP.PA, the HD-SNP and the 0SNP.PA models.

Results and Discussion

The correlations between the GPTAs and PTAs were 0.412, 0.649 and 0.650 for 0-SNP.PA, HD-SNP and HD-SNP.PA, respectively (Figure 1). The figure includes lower and upper bounds for goal correlations. Genomic predictions were highly correlated with PTAs, being 58% superior to the correlation with the predictions from parent averages; predictions from 0-SNP.PA were the 100 base. Similarly, VanRaden *et al.* (2009) reported a predictive correlation of 0.33 for parent average and 0.53 for GPTAs, both for NM in Holsteins. The average breeding value of the parents accounts for one half of the variance of the genetic values of the offspring; the other half is accounted for the Mendelian segregation. These results seem to indicate that the molecular marker data follows Mendelian segregation, at least partially. The use of parent average combined with the HD panel (HD-SNP.PA) did not improve GPTAs (0.650 and 0.649 for HD-SNP.PA and HD-SNP, respectively).



$^{\alpha}$ HD-SNP and 0-SNP.PA have a fixed number of SNPs.

Figure 1: Correlations between predictive transmitting ability and genomic predictions obtained with: the high density assay (HD-SNP); the parent average (0-SNP.PA); the top SNPs for HD-SNP with (Top.PA); without (Top) parent averages; evenly spaced SNPs with (ES.PA); and without (ES) parent averages, by set sizes of SNPs.

On the other hand, in regressions with the LD-SNP sets, there was a great advantage of combining parent averages with the SNPs, especially for small sets (Figure 1). The

correlation between PTA and GPTA from the regression with 300 SNPs improved from 0.51 for the Top SNPs model to 0.59 for the Top.PA model, and from 0.23 for the 300 SNPs ES regression to 0.44 for the 300 SNPs ES.PA. The strategy that achieved the largest correlations with less SNPs was Top.PA, which had a gap with the model HD-SNP of 9% when using just 300 SNPs as predictors, and a gap smaller than 5% in the 1,000 SNPs subset. The second best strategy was the Top set with a gap of 22% with the HD-SNPs when the set had 300 SNPs, and a gap smaller than 5% with a set of 2,500 SNPs. The ES.PA had a gap with the HD-SNPs model of 32% with a set of 300 SNPs, and that gap was about 5% with 5,000 SNPs. Finally, the ES set that did not consider the parent average with 300 SNPs had a gap with the HD-SNPs model of 64%, and that gap was about 5% with 7,500 SNPs.

In small sets for LD platforms, parent averages had higher impact in improving the correlations. As the set size increased, the differences became smaller and the use of parent average was less important. When the sizes of the sets were larger than 200 or 300 SNPs, selecting the SNPs as the Top ones for the trait had more impact than using the parent average (Figure 1). For large sets, there was no difference between the strategies used; e.g., all models with 10,000 SNPs had a gap with the HD-SNPs model of approximately 1%. However, a LD platform without parent averages and without the most important SNPs for the trait could perform even below the parent average prediction, e.g. 300 SNPs ES yield a correlation of 0.23, almost half of a simple prediction with parent averages (0.42).

Conclusions

Lower density platforms with target SNPs can provide reasonable predictive ability of the genetic merit of animals (0.51 *vs.* 0.23 for Top SNPs and ES when parent averages are not considered). The ES sets should include a larger number of SNPs (about 3K to 5K), or consider imputation techniques to improve predictive ability. Additionally, if the parent average information is available, a platform as small as 300 SNPs combined with those averages could attain predictions of about only 9% below a HD platform (0.59). The improvement of using GPTA instead of parent average for prediction was 58% superior for this trait; this impact may be even larger in lower heritability traits. These results applied to NM in this sample of the Holstein population and may not extend to other samples or traits with different heritabilities.

References

de los Campos, G., Naya, H., Gianola, D., *et al.* (2009). *Genetics,* 182:375-385.

Habier, D., Fernando, R.L. and Dekkers, J.C.M. (2009). *Genetics,* 182:343–353.

Meuwissen, T.H.E., Hayes, B.J., and Goddard, M.E. (2001). *Genetics,* 157:1819-1829.

Park, T. and Casella, G. (2008). *J. Am. Stat. Assoc.,* 103:681-686.

VanRaden, P.M., Van Tassell, C.P., Wiggans, G.R., *et al.* (2009). *J. Dairy Sci.,* 92:16-24.

Vazquez, A.I., Rosa, G.J.M., Weigel, K. A., *et al.* (2009). *J. Dairy Sci.* 92:125.

Weigel, K.A., de los Campos, G., González-Recio, O., *et al.* (2009). *J. Dairy Sci.* 92:5248-5257.