Prediction Of Genomic Breeding Values Under Various Strategies For SNP Selection And Various Statistical Models

K. Żukowski [*], S. Kamiński, A. Żarnecki, J. Szyda[†*]

Introduction

Many strategies were used to reduce the dimensions of the genomic breeding value prediction model either through selection of the most informative single nucleotide polymorphisms (SNPs) (Khatkar et al., 2008; Habier et al., 2009; Żukowski et al., 2009) and/or through the choice of the appropriate statistical model (Meuwissen et al., 2001; vanRaden, 2008; Moser et al., 2009; Żukowski et al., 2009). A proper consideration of those two factors is especially important in the context of designing a smaller and cheaper SNP microarray, which could then be used in a much greater extent for the genotyping and then selection within dairy populations.

The main aim of this study was to use the real data from the Polish Holstein-Friesian population in order to compare various approaches to SNP selection from the Illumina BovineSNP50 BeadChip and various statistical models for the prediction of genomic breeding values, based on approximately 3 000 SNPs.

Material and methods

Animals and Phenotypes. The analysed data comprises 1 216 Polish Holstein-Friesian bulls whose phenotypic records were approximated by estimated breeding values for milk, protein and fat yields and were obtained from the national release from February 2009.

Genotypic Data. Each of the 1 216 bulls is genotyped at 54 001 SNPs using the Illumina BovineSNP50 BeadChip (Illumina, 2008). SNP preselection was based on minor allele frequency (MAF<0.01) and genotype call rate (>90%), moreover 3.2% of SNPs (1 746) were excluded because they were not mapped to any chromosome. The total number of SNPs used for further analyses comprised 44 709 SNPs.

Data Set Selection. In the total, seven SNP subsets of a similar size, but constructed using different criteria, were generated (Table 1).

[*] Department of Genetics and Animal Breeding, Wrocław University of Environmental and Life Sciences, Wrocław, Poland.
[†] Institute of Natural Sciences, Wrocław University of Life Sciences, Wrocław, Poland

Table 1: Subsets of SNPs used in the analysis.

Subset	#SNPs	Selection criteria
S1	3 000	randomly selected 100 SNPs from each chromosome
S2	2 513	SNPs uniformly distributed within chromosomes, the number of SNPs for each chromosome depends on the number of identified QTLs (Hu et al., 2007)
S3	2 981	SNPs uniformly distributed across a genome
S4	2 994	SNPs uniformly distributed across base pairs
S5	3 000	SNPs with the highest estimates for milk yield* and $r^2 < 0.8$
S6	3 000	SNPs with the highest estimates for stature* and and $r^2 < 0.8$
S7	3 000	SNPs with the highest estimates for type* and $r^2 < 0.8$

*as estimated by Szyda et al. (2009)

Statistical Analysis. The following models were used to estimate SNP effects from different subsets of data:

Model with fixed SNP effects and a random polygenic effect:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{Xb} + \mathbf{Z}_{\alpha}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$$

where: **y** is a vector of deregressed estimated breeding values for milk yield, protein yield, or fat yield; **μ** is the overall mean; **b** a is vector of fixed additive SNP effects; **X** is the design matrix for fixed SNP effects with the elements given by -1, 0 and 1 for an SNP genotype 11, 12 (21) and 22 respectively; **α** is a vector of random additive polygenic effects of genotyped animals assuming $\boldsymbol{\alpha} \sim N\left(0, \mathbf{A}\hat{\sigma}_{\alpha}^2\right)$ with **A** representing the polygenic relationship matrix and $\hat{\sigma}_{\alpha}^2$ is the estimate of an additive polygenic variance originating from the whole active population of dairy cattle in Poland, which amounts to 0.33 for milk yield and 0.29 for protein and fat yields; $\mathbf{Z}_{\boldsymbol{\alpha}}$ is a design matrix for **α**; **ε** is a vector of random errors assuming $\varepsilon \sim N\left(0, \mathbf{D}\sigma_{\varepsilon}^2\right)$ with **D** being a diagonal matrix with reciprocal of effective daughter contributions (EDC) for i-th bull and σ^2_{ε} denoting error variance. This model was applied to SNP data from each chromosome separately, so that altogether 29 evaluations were needed to obtain the estimates of all SNPs.

Model with random uncorrelated SNP effects:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{Z}_{a}\mathbf{a} + \boldsymbol{\varepsilon}$$

where: **a** is a vector of random additive SNP effects assuming $\mathbf{a} \sim N\left(0, \mathbf{I} \cdot \frac{\hat{\sigma}_{\alpha}^2}{\#SNP}\right)$, with **I** being an identity matrix and #SNP represents the number of SNPs in the model; $\mathbf{Z}_{\boldsymbol{a}}$ is a design matrix for SNPs effects with the elements of -1, 0 and 1 for an SNP genotype 11, 12(21) and 22 respectively; all the other model parameters are as defined above.

Model with a random correlated SNP effects:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{Z}_{a}\mathbf{a} + \boldsymbol{\varepsilon}$$

This model differs from the above model in only one aspect - it assumes a correlation between SNPs through: $\mathbf{a} \sim N\left(0, \mathbf{LD} \cdot \frac{\hat{\sigma}_{\alpha}^{2}}{\#SNP}\right)$ with **LD** being a matrix of r^2 coefficients between pairs of linked SNPs estimated using PLINK (Purcell et al., 2008) with values less than 0.2 truncated to zero.
DGV is defined as the sum of additive effects of SNPs estimated from the above models:

$$\hat{\mathbf{g}} = \mathbf{X}\hat{\mathbf{b}} + \mathbf{Z}\hat{\mathbf{a}}$$

Results and discussion

Correlation between estimated breeding values and direct genomic values for milk yield, protein yield and fat yield are shown in Figure 1, Figure 2 and Figure 3 respectively. The results show generally high correlations between estimated breeding values and direct genomic values and are similar for all considered traits. The highest correlation values were obtained for subset with the highest estimates for milk yield and r^2<0.8 for a model with random uncorrelated SNP effects, what coincides with results of previous simulation studies (Meuwissen et al., 2001; Moser et al., 2009). Differences between models with random uncorrelated and correlated SNP effects were very small with advantage of the former model. In all cases correlations were above 0.9.

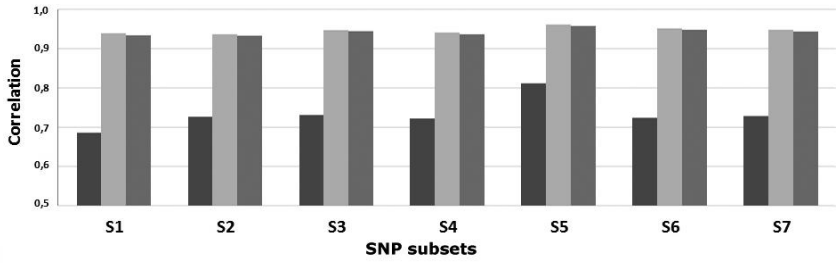

Figure 1: Correlation between EBV and DGV for milk yield for different data subsets.
For each data set: left column - model with fixed SNP effects and a random polygenic effect, middle column - model with a random uncorrelated SNP effects and right column - model with a random correlated SNP effects.

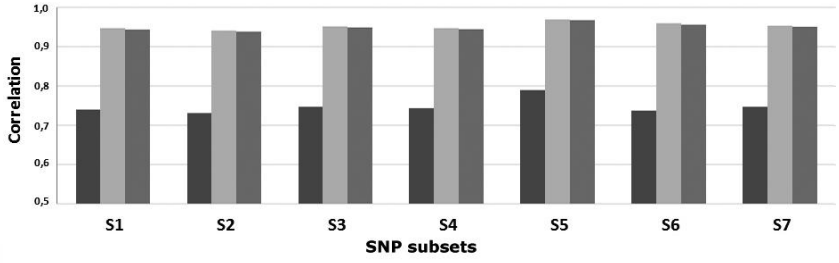

Figure 2: Correlation between EBV and DGV for protein yield for different data subsets. For each data set: left column - model with fixed SNP effects and a random polygenic effect, middle

column - model with a random uncorrelated SNP effects and right column - model with a random correlated SNP effects.

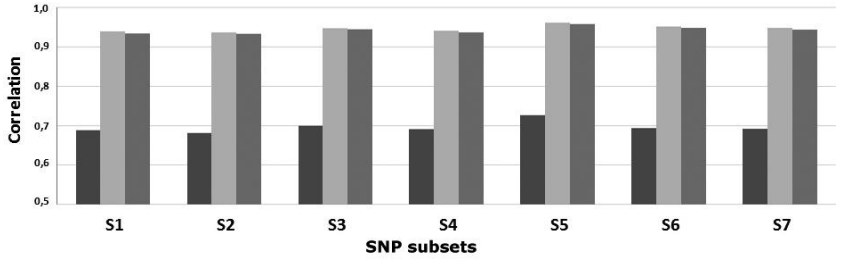

Figure 3: Correlation between EBV and DGV for fat yield for different data subsets. For each data set: left column - model with fixed SNP effects and a random polygenic effect, middle column - model with a random uncorrelated SNP effects and right column - model with a random correlated SNP effects.

Conclusion

Generally, for prediction purposes models with random SNP effects are superior to the model with fixed SNP effects. The bad performance of a model with fixed SNPs effects was related to overestimated SNPs effects. The incorporation of information about the covariance structure between SNPs did not have any noticeable influence on the correlations. Comparing different data sets, the subset with the highest estimates for milk yield and $r^2<0.8$ showed the highest correlations and thus would be recommended for a small SNP chip as long as selection for production level is of main interest. Another promising approach would be to use information on identified QTL. 3 000 SNPs is statistically and economically supported number to be considered for a small chip which could be widely used for genomic selection on a level of a whole population.

Acknowledgements

The project is realized by MASinBULL and financially supported by Bydgoszcz Animal Breeding and Insemination Center.

References

Habier, D., Fernando, R.L., and Dekkers, J.C.M., (2009) *Genetics*, 182(1): 343–353.

Hu Z., Fritz E.R., Reecy J.M. (2007) *Nucleic Acids Research*, 2007, 35: D604–D609.

Illumina, 2008.

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

Moser, G., Tier B, Crump RE, *et al.* (2009) *Genet Sel Evol.* 31;41(1):56.

Khatkar, M.S., Nicholas F.W., Collins A.R., *et al.* (2008) *BMC Genomics.* 9:187.

Purcell, S., B. Neale, K. Todd-Brown, *et al.* (2007) *Am. J. Hum. Genet.*, 81:559-575.

Szyda J., Żarnecki A., and Kamiński K., (2009) *Interbull Bulletin*. 39:43-46.

VanRaden P.M. (2008) *Journal of Dairy Science*. 91(11):4414-23.

Żukowski K., Suchocki T., Gontarek A., Szyda J. (2009) *BMC Proc.* 3(Suppl 1): S13.