Proc. Assoc. Advmt. Anim. Breed. Genet. ***17****: 227-230*

GENOME BASED GENETIC EVALUATION AND GENOME WIDE SELECTION USING SUPERVISED DIMENSION REDUCTION BASED ON PARTIAL LEAST SQUARES

G. Moser[1], R.E. Crump[1,2], B. Tier[1,2], J. Sölkner[1,3], K.R. Zenger[1,3], M.S. Khatkar[1,3], J.A.L. Cavanagh[1,3] and H.W. Raadsma[1,3]

[1]CRC for Innovative Dairy Products, Level 1, 84 William Street, Melbourne, VIC 3000
[2]Animal Genetics and Breeding Unit, University of New England, NSW 2351
[3]ReproGen, Faculty of Veterinary Science, The University of Sydney, Camden, NSW

SUMMARY

The method of partial least squares was applied to the prediction of genetic merit using whole genome scan data consisting of 10715 SNP. The method is particularly suited to data sets that have many more markers than observations and in which markers are collinear due to high linkage disequilibrium. A SNP ranking method was applied to select a subset of markers which have equal predictive power compared to using all SNP simultaneously.

INTRODUCTION

The advent of high density SNP typing platforms in cattle has opened the possibility to perform genome-wide selection (GWS) without the need of QTL and pedigree analysis in predicting molecular breeding values (MBV). A challenging problem connected with whole-genome data is that it contains typically many more variables (single nucleotide polymorphisms, SNP) than observations (estimated breeding values, EBV). It is not uncommon to collect genotype information on several thousand SNP markers for only a few hundred individuals. Since most traditional multivariate techniques are not applicable with such high-dimensional genomic data special techniques such as variable selection or dimension reduction are required.

A powerful approach for analyzing high-dimensional SNP data is supervised dimension reduction based on partial least squares (PLS). As a supervised approach, it uses the response variable of interest in the dimension reduction step which often makes it more efficient in prediction problems than the unsupervised principal component analysis. In PLS, dimension reduction and regression are performed simultaneously.

MATERIAL AND METHODS

The data comprised 10715 SNP typed in 1546 dairy bulls born between 1955 and 2001. High-reliability EBV for several traits of the progeny tested bulls were supplied by the Australian Dairy Herd Improvement Scheme.

Partial Least Squares. PLS (Garthwaite 1994) is a multivariate regression method that relates the data matrix (X, descriptors) to a response that can be either singular (y) or multiple (Y). The PLS method seeks to uncover a small number of latent variables from a much larger set of correlated descriptors. PLS can also been seen as an ordinary regression method for finding the matrix of regression coefficients B where both SNP (X) and EBV (Y) are decomposed into latent variables,

$$X = t_1 p'_1 + t_2 p'_2 + ... + t_A p'_A + E = TP' + E$$
$$Y = t_1 c'_1 + t_2 c'_2 + ... + t_A c'_A + F = TC' + F$$

where A is the number of latent variables, t is a score vector for X, p is the loading vector for X, c is the loading vector for Y, E is the residual matrix for X and F is the residual matrix for Y. In order to obtain latent components, PLS maximises the covariance between EBV and a linear combination of the SNP markers $t = Xw$, where t is the score vector, X is the data matrix and w is the weight vector (Garthwaite 1994).

Validation of models. Internal validation of data using cross-validation was performed to determine a model's predictive capacity and to determine the optimal model complexity (i.e. number of latent components). The Mean Squared Error of prediction (MSEP) was used as the objective function in model complexity selection. The learning data set L was randomly divided into K segments L_k, $k = 1 .. K$, of roughly equal size. The k-fold cross-validation estimate is

$$MSEP_{cv.K} = \frac{1}{n_L} \sum_{k=1}^{K} \sum_{i \in L_k} \left(f_k(x_i) - y_i \right)^2$$

where f_k is the predictor trained on L / L_k, i.e., all observations not in L_k. To test if the model over fits the data a permutation method was applied that randomly assigned EBV values to animals.

SNP selection. A selection approach based on the latent components of the PLS model uses the weight vectors w. The influence of SNP k for the a-th PLS component is defined as a function of w^2_{ak}. VIP (variable importance in projection) is the accumulated sum over all PLS dimensions of the variable influence:

$$VIP_{Ak} = \sqrt{\sum_{a=1}^{A} \left(w_{ak}^2 * SSY_a \right)}$$

where SSY_a is the sums of squares explained by PLS dimension a.

RESULTS AND DISCUSSION

A summary of the results obtained by PLS for 7 traits is shown in Table 1. A small number of latent components (4-8) accounted for a large proportion of the EBV variance (77% - 94%). Less than 10% of the variance of the marker matrix is explained by the model. This indicates that a large proportion of the SNP information is redundant due to high linkage disequilibrium over longer genomic distances. For chromosome 6 of the same data, Khatkar *et al.* (2006) found that significant linkage disequilibrium extends to 13 megabases.

The critical issue in developing a good model is generalisation. Models that are too complex may fit the noise, not just the signal, leading to overfitting. Such a model may well describe the relationship between SNP and EBV of the sires used to develop the model, but may subsequently fail to predict new data. For example, a PLS model was fitted to data of randomly permuted EBV values of the trait fertility. The derived model with 4 latent components explained 62% of the EBV variance compared to 77% in the original data. However, the cross-validated R^2 was less than .01 compared to .67 in the original data. It appears that the predictive power of the derived models is high and that these provide valid predictions of molecular breeding values in new bulls.

Table 1. Fit of PLS model for various traits using 1546 bulls and 10715 SNP markers. The optimal model complexity was derived by 10-fold cross validation

Trait	Number of latent components	Proportion of variance accounted for	
		EBV	SNP
APR[A]	6	91.64	7.06
ASI[B]	6	90.95	7.13
Protein kg	7	94.07	7.60
Milk Yield	7	91.70	7.69
Fat %	8	92.05	8.66
Overall Type	4	78.67	5.59
Cow Fertility	4	77.36	5.73

[A]Australian Profit Ranking [B]Australian Selection Index

Using reduced SNP sets provides faster and more cost-effective genotyping and allows the application of statistical methods which can not handle data with more predictors than response variables. In PLS SNP can be ranked on their VIP value, which reflect the importance of the marker in the model both with respect to their correlation to the EBV and with respect to the marker data. A simple forward selection strategy to minimise the cross-validated prediction error can then be applied to develop sets of non-redundant SNP that are useful in predicting breeding values.

Figure 1 shows an example of the SNP selection process for the trait Overall Type. To form an external validation set 200 bulls were randomly selected and excluded from the PLS analysis. The weight values of the markers are normalized in the unit interval. The weight distribution of the PLS model with n = 10715 SNP (Figure 1a) is characterised by only a few SNP with high VIP values, which are selectively chosen in the selection steps. A subset of 121 SNP is obtained with a much less concentrated distribution of weight values (Figure 1b). The predictive performance of the reduced SNP set was very similar to the model utilizing all SNP. Applied to predicting MBV in the independent test set of 200 bulls, the complete SNP data model gave a correlation between EBV and MBV of 0.72 compared to 0.69 for the 121 SNP set.

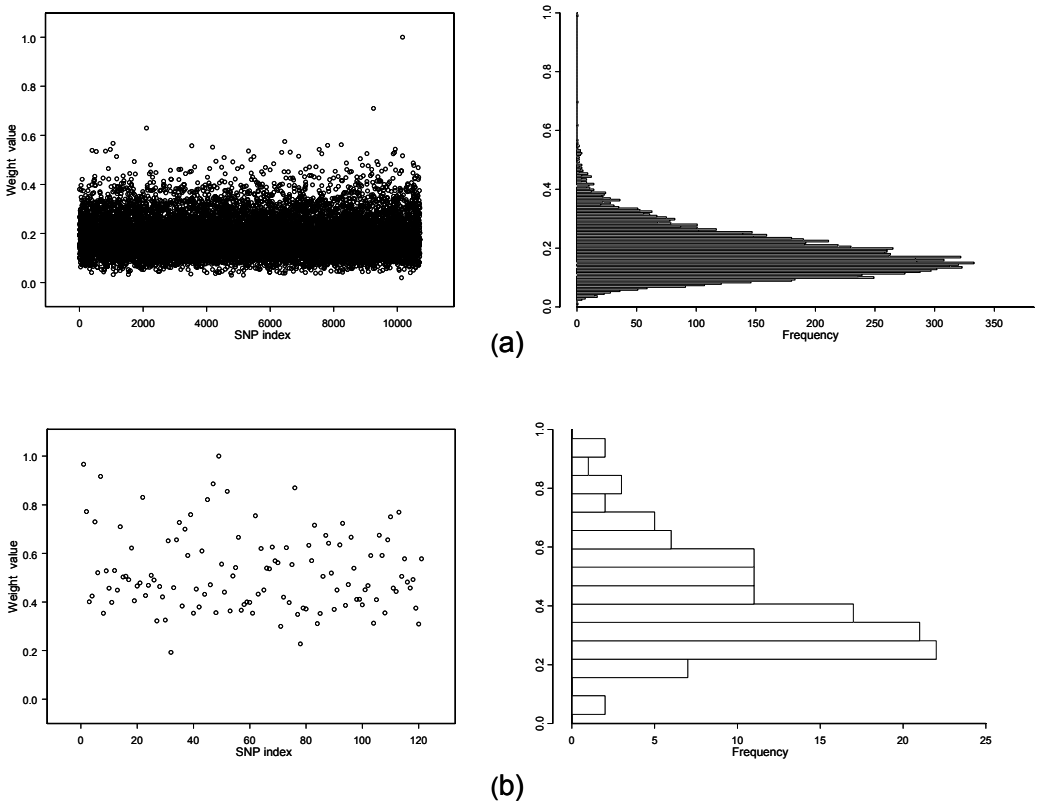

Figure 1. SNP weight distribution using the complete data of 10715 SNP (a) and a reduced set of 121 SNP trained by forward selection of SNP according to their VIP value (b).

Acknowledgments
This work is part of the Co-operative Research Centre for Innovative Dairy Products project 1.4a. The Australian Dairy Herd Improvement Scheme provided the EBV data.

References

Garthwaite, P. H (1994) *JASA* **89**:122
Khatkar M.S., Collins, A., Cavanagh, J.A.L., Hawken, R.J., Hobbs, M., Zenger, K.R., Barris, W., McClintock, A.E., Thomson, P.C., Nicholas F.W., and Raadsma, H.W. (2006) *Genetics* **174:**79