# A KERNEL METHOD FOR GENOME-WIDE SELECTION USING HAPLOTYPES

**G. Moser,[1,2], M.S. Khatkar[1,2] and H.W. Raadsma[1,2]**

[1]ReproGen, University of Sydney, Camden, NSW 2570
[2]Dairy Futures Cooperative Research Centre, Bundoora, 3083

## SUMMARY

A kernel method for genomic selection using haplotype information is described. A total of 2,144 Australian Holstein Friesian bulls were genotyped with the Illumina BovineSNP50 BeadChip. The accuracy of direct genomic values for 9 traits was computed as the correlation between the predicted direct genomic value (DGV) and the current genetic evaluations for a validation set of young bulls. A kernel method and random regression BLUP, both using SNP genotypes, were also analysed. The accuracy of DGV derived by the three methods was very similar. Using haplotype instead of SNP information did not increase the accuracies of DGV and the kernel method based on SNP gave the highest accuracy overall. Using only SNP genotypes to predict DGV has the advantage that determining the linkage phase of the haplotypes is not required.

## INTRODUCTION

In genomic selection (Meuwissen *et al.* 2001), selection decisions are based on direct genomic values (DGV) derived from high-density single nucleotide polymorphic (SNP) markers. Several studies have now shown that genomic selection is significantly more accurate than traditional selection of young animals based on pedigree information (Harris *et al.* 2008; Van Raden *et al.* 2008; Hayes *et al.* 2009; Moser *et al.* 2009).

A variety of methods have been suggested for the estimation of DGV and comparisons on real data have found very similar accuracies of prediction between methods (González-Recio *et al.* 2008, Moser *et al.* 2009). Methods to estimate DGV are usually implemented using information of individual SNP genotypes, however, haplotypes generally provide more information than individual SNP. So far, results of the accuracy of DGV calculated from haplotypes using real data have not been reported.

The objective of this study was to compare the accuracy of DGV calculated from hapolotypes to the accuracy of DGV calculated from SNP using field data on 9 traits in Australian Holstein Friesian cattle. The haplotype approach is based on a kernel method in which haplotype information is used to measure the genomic similarity between animals.

## MATERIAL AND METHODS

**Phenotype and genotype data.** A total of 2,144 bulls Australian Holstein-Friesian bulls with genotype and phenotype information were available for this study. Bulls were divided in a training data set of 1,847 bulls born between 1955 and 2004 and a validation set of 297 young bulls born between 2001 and 2004, which represented progeny test teams of Genetics Australia for 2007, 2008 and 2009. Of the 297 young bulls in the bull validation set, 240 (80.8%) were sired by bulls in the training set.

The phenotypes used were deregressed breeding values for protein percentage, fat percentage, Australian Selection Index (ASI), Australian Profit Ranking (APR) and survival, and daughter trait deviations for protein yield, fat yield, milk yield and overall type (Otype). The deregression procedure removed the contribution of relatives other than daughters to the breeding values.

Phenotypes together with reliability information were provided by the Australian Dairy Herd Improvement Scheme (ADHIS, http://www.adhis.com.au).

SNP genotypes were derived from the Illumina BovineSNP50 BeadChip (Illumina Inc., San Diego, USA). After quality control and omitting SNP located on the sex chromosomes, a total of 42,576 markers remained for the analysis.

**Analysis methods.** Prediction equations to generate DGV were estimated from the training set by random regression BLUP (R-BLUP) using SNP genotype information and by kernels methods based on either SNP or haplotype information.

In R-BLUP, regression coefficients are obtained from the solution of the weighted ridge regression equations:

$$\begin{bmatrix} \mathbf{1}'\mathbf{R}^{-1}\mathbf{1} & \mathbf{1}'\mathbf{R}^{-1}\mathbf{X} \\ \mathbf{X}'\mathbf{R}^{-1}\mathbf{1} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{X}+\mathbf{I}\boldsymbol{\lambda} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix},$$

where **X** is an $N_{\text{Anim}}\times N_{\text{SNP}}$ matrix of genotypes encoded as 0 (homozygote), 1 (heterozygote) or 2 (other homozygote), **I** is diagonal matrix with non-zero elements $\lambda_1, \lambda_2, \ldots, \lambda_{N_{\text{SNP}}}$. The penalty term $\lambda$, which is the same for all SNP, overcomes the problem of ill-conditioning when multicollinearity among columns in **X** causes **X'X** to be singular, or nearly so. **R** is a diagonal matrix with non-zero elements $r_{ii}$=(1/rel)-1, where rel is the reliability of the phenotype information. The optimal $\lambda$ was derived by cross-validation.

Kernel based methods use linear models to implement non-linear regression by mapping the input space to a higher dimensional feature space using kernel functions. One can then perform ridge regression in the feature space which gives the following square equations:

$$\begin{bmatrix} 0 & \mathbf{1}' \\ \mathbf{1} & \mathbf{K}+\mathbf{I}\boldsymbol{\lambda} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}'\mathbf{y} \\ \mathbf{y} \end{bmatrix},$$

where **K** is a $N_{\text{Anim}}\times N_{\text{Anim}}$ positive definite matrix with elements $K(\mathbf{x},\mathbf{x}')$, which is a kernel function that measures the 'genomic distance' between two animals. When using SNP information we applied a Gaussian kernel, i.e. $K(\mathbf{x},\mathbf{x}')=\exp(-\|\mathbf{x}-\mathbf{x}'\|^2/\sigma^2)$, where $\|\mathbf{x}-\mathbf{x}'\|$ is the euclidean distance and $\sigma^2$ scales the distance. For haplotypes we used the hamming distance as genomic distance measure, which is equal to the number of SNP loci at which alleles are different between a pair of haplotypes of a given length. Let $\mathbf{x} = (\mathbf{h}_1,\mathbf{h}_2)$ and $\mathbf{x}' = (\mathbf{h}_1',\mathbf{h}_2')$ be the haplotype pairs to be compared between two sires, the haplotype distance was then defined as, $H(\mathbf{x},\mathbf{x}') = \min\{d(\mathbf{h}_1,\mathbf{h}_1')+ d(\mathbf{h}_2,\mathbf{h}_2'), d(\mathbf{h}_1,\mathbf{h}_2')+ d(\mathbf{h}_2,\mathbf{h}_1')\}$. Based on $H(\mathbf{x},\mathbf{x}')$ we compute an exponential kernel as $\exp(-H(\mathbf{x},\mathbf{x}')/\sigma^2)$. The values of the tuning parameters of each method were obtained by cross-validation.

Haplotypes for each chromosome were phased using Beagle Version 3.2.1 (Browning and Browning 2007). The effect of haploytype length was investigated by constructing kernels using haplotypes including 100, 10 or 5 consecutive SNP.

All methods were implemented in FORTRAN90. The correlation coefficient between predicted DGV and the realized phenotypes was used to evaluate the accuracy of DGV prediction.

## RESULTS AND DISCUSSION

Correlations between predicted DGV and phenotypes of the young bulls in the validation set obtained by the various methods are shown in Table 1. Overall, accuracies of DGV were very similar between methods. However, the kernel method based on SNP genotypes slightly outperformed the haplotype kernel and R-BLUP. These findings are consistent with our previous analysis (Moser *et al.* 2009), where we also found very small differences between a range of different methods proposed for DGV prediction, but where support vector regression (SVR) based

**Table 1. Correlation between DGV and phenotypes in the validation animals derived by R-BLUP and kernel methods based on information from SNP or haplotypes of different length**

| | R-BLUP | Kernel method | | | |
|---|---|---|---|---|---|
| | | SNP | Number of SNP in haplotype | | |
| Trait | | | 100 | 10 | 5 |
| Protein | 0.497 | 0.523 | 0.483 | 0.497 | 0.503 |
| Fat | 0.522 | 0.545 | 0.467 | 0.509 | 0.504 |
| Milk | 0.528 | 0.547 | 0.510 | 0.531 | 0.529 |
| OType | 0.503 | 0.524 | 0.509 | 0.510 | 0.510 |
| Protein% | 0.621 | 0.635 | 0.617 | 0.617 | 0.618 |
| Fat% | 0.566 | 0.582 | 0.548 | 0.550 | 0.542 |
| ASI | 0.341 | 0.346 | 0.327 | 0.311 | 0.316 |
| APR | 0.490 | 0.507 | 0.470 | 0.475 | 0.494 |
| Survival | 0.076 | 0.169 | 0.147 | 0.141 | 0.135 |

on SNP genotypes gave the highest accuracies overall. The kernel method used here is very similar to SVR in Moser *et al*. 2009.

The largest difference between kernel methods and R-BLUP was observed for the trait survival, which has low heritability. The correlation between DGV and phenotypes was nearly twice as high for the kernel methods compared to R-BLUP. This indicates that more training animals may be required to obtain accurate prediction equations for traits with lower heritability when using R-BLUP

Kernel methods have the potential to capture high order non-linear interactions between genotypes. The kernel methods gave similar results compared to R-BLUP which assumes an additive model, suggesting that there was no additional predictive ability to be gained from these interactions. This could reflect the response variables in our data set, which are the averages of performance of bulls over a large number of daughters, so dominance effects are averaged out and only a proportion of the epistatic variance remains.

Using haplotype instead of SNP information did not improve accuracies of DGV. This could be partly due to the fact, that the construction of haplotypes inevitably contains errors. Different ways to derive haplotypes have been described and are implemented in a number of software programs. For phasing we used Beagle (Browning and Browning 2007) , as it provided the highest accuracy of correctly phased alleles in simulated data sets that were derived to model the true Australian Dairy population (Khatkar *et al.* unpublished). A characteristic of Beagle is that it only outputs the most likely phase. It would be straightforward to derive a haplotype based distance from several haplotype candidates and their corresponding frequencies.

We would expect that the performance of the haplotype kernel depends on the degree of LD of the data set as haplotypes can be inferred more precisely for high LD data sets. However, many animals in the training and test data share DNA segments from a small number of extensively used sires so that LD as measured by D' is high in our data (Khatkar *et al*. 2008). The fact that using haplotypes of longer length gave similar accuracies to shorter haplotypes indicates that the accuracy of haplotype construction was very high. High LD could also have contributed to the good performance of the SNP kernel, as SNP information may be sufficient to capture the information of distinct haplotypes. Furthermore, a relatively small number of SNP is sufficient to provide similar accuracy to that achieved with a high-density assay (Moser *et al*. 2010).

In contrast to R-BLUP, using haplotype instead of SNP information to build kernels does not increase the dimensionality of the system of equation, as the genomic information is contained in a matrix of dimension $N_{\text{Anim}} \times N_{\text{Anim}}$. This makes kernel methods particularly attractive for handling genomic information derived from very high-density SNP arrays or re-sequencing projects.

## CONCLUSIONS

We proposed kernel methods for genomic selection using high-density SNP assays. The kernel methods were at least as accurate as random regression BLUP. Overall, predictions using kernels based on SNP genotypes were slightly more accurate compared to kernels based on haplotype data. The advantage of using only SNP genotypes is that determining the linkage phase of the haplotypes is not required and the markers do not need to be mapped. The main disadvantage of using haplotypes in R-BLUP is that that the number of effects that needs to be estimated is considerably larger than that for the SNP model. As SNP density and training data size increase kernel methods will become more attractive for genomic selection, especially when using information of haplotypes.

In this work we only applied a single measure of genomic similarity based on haplotypes, however, kernel methods provide enormous flexibility to consider more biological aspects in the models. For example prediction accuracies might improve if the haplotype block structure is considered in the models.

## ACKNOWLEDGEMENT

The research was funded by the Dairy Futures Cooperative Research Centre. The data resource was generated by the CRC-IDP and the Department of Primary Industries, Victoria, with input from Drs, Phil Bowman, Amanda Chamberlain, Matthew Hobbs and Mrs Gina Attard. The semen samples were provided by Genetics Australia.

## REFERENCES

Browning S.R. and B L Browning B.L. (2007) *Am. J. Hum. Genet.* **81:**1084.

González-Recio O, Gianola D, Long N, Weigel K.A., Rosa G.J. and Avendaño S. (2008) *Genetics* **178:**2305

Harris B.L., Johnson D.L. and Spelman R.J. (2009). Interbull Meeting, Niagara Falls, USA.

Hayes B.J., Bowman P.J., Chamberlain A.J. and Goddard M.E. (2009). *J. Dairy Sci.*, **92:**433–443.

Khatkar M.S., Nicholas F.W., Collins A.R., Zenger K.R., Cavanagh J.A., Barris W., Schnabel R.D., Taylor J.F. and Raadsma H.W. (2008) *BMC Genomics* **9:**187.

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

Moser G., Tier B., Crump R.E., Khatkar K.S. and Raadsma H.W. (2009) *Genet. Sel. Evol.* **41:**56

Moser G., Khatkar K.S., Hayes B.J. and Raadsma H.W. (2010) *Genet. Sel. Evol*. **42:37**

Van Raden P.M., Van Tassel C.P., Wiggans G.R., Sonstegard T.S., Schnabel, R.D., Taylor J.F. and Schenkel F. (2008). *J. Dairy Sci.* **91:**4414