Incorporation of Correlation Between SNPs into the Genomic Evaluation Model

J. Szyda[a,b], S. Kamiński[c] , A. Żarnecki[d], K. Żukowski[a]
[a] *Department of Animal Genetics, Wrocław University of Life Sciences, Wrocław, Poland*
[b] *Institute of Natural Sciences, Wrocław University of Life Sciences, Wrocław, Poland*
[c] *Department of Animal Genetics, University of Warmia and Mazury, Olsztyn, Poland*
[d] *National Research Institute for Animal Production, Balice, Poland*

Abstract

The training data set consisting of 1 227 Polish Holstein Friesian bulls, born between 1987 and 2003, genotyped with the BovineSNP50 Genotyping BeadChip was used for the estimation of single nucleotide polymorphism (SNP) effects using two approaches - a conventional model in which SNPs are assumed as uncorrelated and a model where correlation between SNPs is expressed by pairwise linkage disequilibrium coefficients. Results show that the model without SNP correlation is superior for prediction of breeding values with the correlation between Direct Genomic Values and Estimated Breeding Values for milk yield for the training data set amounting to 0.98, while the model with SNP covariance only exhibiting the correlation of 0.59. On the other hand taking account for the covariance between SNPs allows for a much clearer differentiation between estimates of additive effects of particular SNPs and consequently for localising causal mutations.

1. Introduction

Recently many countries have incorporated the genomic information into their genetic evaluation systems (Hayes *et al.,* 2009, Loberg and Dürr 2009). Although a great variety of statistical models, estimation methods, Single Nucleotide Polymorphism (SNP) selection criteria, and dependent variable definitions were applied and compared, none of the approaches takes into account that the SNPs are correlated through physical linkage or selection. In the current study we tackle this methodological gap and explore the potential benefits and problems of using a more realistic genomic evaluation model with SNP covariance. The analysis is performed by comparing a models where all SNPs are summed as independent (no covariance) with a model, in which covariances between SNPs are considered.

2. Materials and Methods

2.1. Material

The data set, used for the estimation of additive effects of SNPs consists of 1 227 Polish Holstein-Friesian bulls, born between 1987 and 2003. Genotypes originate from the Illumina BovineSNP50 Genotyping BeadChip consisting of 54 001 SNPs of which 46 267 passed the selection criteria of Minor Allele Frequency ≥ 0.01 and at least 90% call rate imposed on our data set. This data was described in detail by Szyda *et al.* (2009). In the current analysis deregressed Estimated Breeding Values (EBV) for milk yield corresponding to the national evaluation release from April 2009 are used for the illustration of methodology and results.

2.2. Estimation and modelling of Linkage Disequilibrium

The pairwise Linkage Disequilibrium (LD) is expressed as a squared correlation coefficient (r^2) between allele counts observed at two SNPs and was calculated using the PLINK software (PLINK, Purcell *et al.,* 2007). This approach is computationally feasible for large data sets since it does not require haplotype reconstruction, but it provides only an approximation of the true LD. In our analysis genotypic data from all bulls were used with the underlying, simplified assumption that the individuals are unrelated.

2.3. Estimation of Direct Genomic Values

The following mixed model was used to estimate additive effects of the selected N_{snp}=46 267 SNPs for N_a=1 227 bulls with genotypes:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zg} + \mathbf{e},$$

where $\mathbf{y}$ [N_a] represents a vector of deregressed EBVs, $\mathbf{X}$ is a [N_axN_b] design matrix for fixed effects, $\mathbf{b}$ [N_b] is a vector of N_b fixed effects, which in the current model comprise only a general mean (N_b=1), $\mathbf{Z}$ is a [N_axN_{snp}] design matrix for SNP genotypes, which is parameterized as -1, 0, or 1 for a homozygous, a heterozygous, and an alternative homozygous SNP genotype respectively, $\mathbf{g}$ is a [N_{snp}] vector of random additive SNP effects, and $\mathbf{e}$ is a [N_a] vector of residuals with $\mathbf{e}\sim N(0, \mathbf{D}\hat{\sigma}_e^2)$ with $\mathbf{D}$ being a diagonal matrix containing the reciprocal of Effective Daughter Contributions on the diagonal. Two approaches towards modelling the covariance structure of $\mathbf{g}$ were considered: (i) without covariance - assuming $\mathbf{g}\sim N\left(0, \mathbf{I}\frac{\hat{\sigma}_a^2}{N_{snp}}\right)$, with $\mathbf{I}$ being an identity matrix and $\hat{\sigma}_a^2$ representing the additive genetic variance of milk yield and (ii) including covariance information - assuming $\mathbf{g}\sim N\left(0, \mathbf{G}_{LD}\frac{\hat{\sigma}_a^2}{N_{snp}}\right)$ with the off-diagonal elements of $\mathbf{G}$ consisting of pairwise LD coefficients between linked SNPs including only coefficients $r^2 \geq 0.80$ (smaller values were truncated to 0.00), the upper value of r^2 was truncated to 0.95. The corresponding mixed model equations (Henderson, 1984) have the following form:

$$\begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

with $\mathbf{R}$ represented by $\mathbf{D}\hat{\sigma}_e^2$ and $\mathbf{G}$ either by $\mathbf{I}\frac{\hat{\sigma}_a^2}{N_{snp}}$ or by $\mathbf{G}_{LD}\frac{\hat{\sigma}_a^2}{N_{snp}}$, depending on the SNP covariance model considered. The estimation of model parameters was based on the iteration on data technique through the Gauss-Seidel algorithm with residuals update (Legarra and Misztal 2008). Note, that for the model with SNP covariance the estimates of SNP effects from the no covariance models were used as starting values and only one iteration round was performed to obtained new estimates.

DGV is defined as the sum of additive effects of SNPs estimated from the above models: $\widehat{\mathbf{DGV}} = \mathbf{Z}\hat{\mathbf{g}}$.

3. Results and Discussion

3.1. Linkage Disequilibrium structure

The average genomewise r^2 among pairs of linked (i.e. located on the same chromosome) SNPs amounts to 0.098. As illustrated by Figure 1, the decay of LD with increasing intermarker distance is observed, still there are pairs of SNPs which exhibit high LD even if they are relatively apart one from another. The extend and rate of decay of LD observed in our data is very similar to those recently reported for Holstein-Friesian by The Bovine Hap Map Consortium (2009), showing that starting from an intermarker distance of covering up to 300 Kbp, pairwise LD remains at an approximately constant level of 0.10.

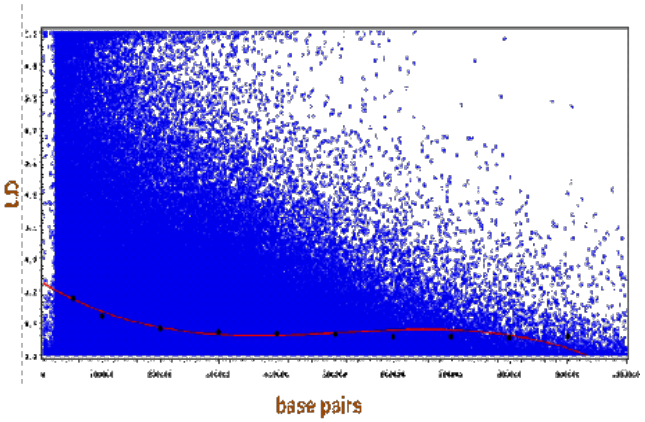

Fig. 1 Distribution of linkage disequilibrium measure (r^2) along the bovine genome. Black dots represent average r^2 for a given intermarker distance, red line is the cubic polynomial fitted to the averaged data.

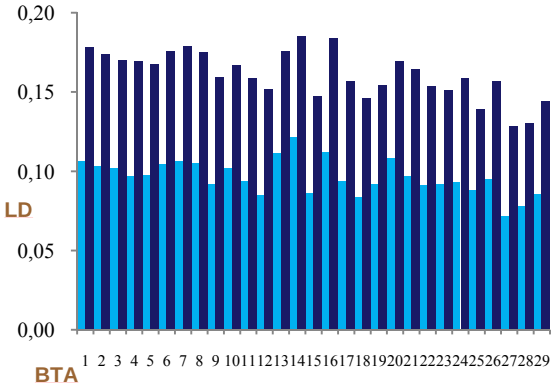

Fig. 2 Average r^2 (blue) and its standard deviation (navy blue) calculated separately for each chromosome.

While considering chromosomes separately some variation in average r^2 and its variability is observed, that, surprisingly, does not seem to be a clear function of a chromosome length (Figure 2). An interesting observation is that the highest average r^2 and the highest r^2 variability is attributed to BTA14 - the chromosome known to harbour DGAT1 - a gene of a very pronounced effect on production traits (Grisart *et al.,* 2002).

3.2. SNP effect estimates

The estimates of additive effects of all SNPs originating from both models are presented on Figure 3. Although the resolution of the graph does not allow for inferences on particular markers, it is evident that estimates from the model, which incorporates covariances between SNPs are more variable in a sense that there are more extreme (high and low) estimates, as compared to the values originating from the no covariance model. This feature of the covariance model is highly appreciated from the point of view of candidate gene mapping, since - as it is clearly seen on the graph, it prevents the true underlying genetic effect of an unknown gene from being equally distributed between SNPs of high LD.

Shortcomings and advantages of the LD based approach are visualised on Figure 4 using an example of the region of BTA14 in the neighbourhood of DGAT1. When LD between SNPs is very high there is not enough information for neither of the models to differentiate between their effects, but if no covariance is assumed in the model the genetic effect is distributed among multiple SNPs even if they are not in strong LD. Interestingly, the highest estimates of SNP effects based on our data set are not attributed directly to the DGAT1 region, but to SNPs located within other coding sequences such as: CYHR1 and NFKBIL2 (SNPs - ARS-BFGL-NGS-34135 and ARS-BFGL-NGS-94706 covering the region between 260 342 and 281 534 bp of BTA14); LOC506831, LOC524974, and MAPK15 (HAPMAP25384-BTC-001997 and HAPMAP24715-BTC-001973 covering 835 055-856 890 bp); as well as PTK2 (HAPMAP32970-BTC-064990 and HAPMAP24986-BTC-065021 covering 2 313 594-3 018 725 bp). Clearly, a larger data set with more recombinations is needed to break the high LD estimated between some marker pairs to enable better gene identification resolution.

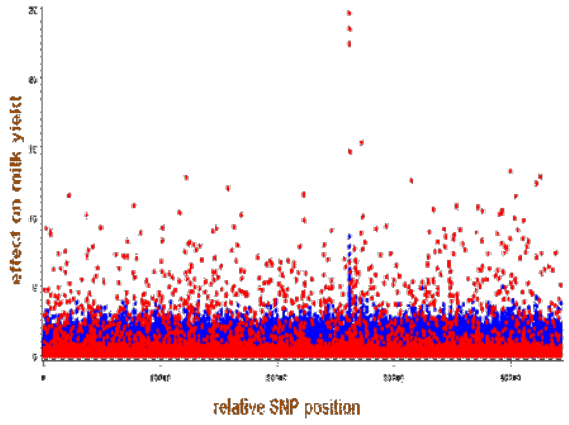

Fig. 3 Estimates of additive SNP effects on milk yield, based on the model without SNP covariance (blue dots) and the model with SNP covariance (red dots).

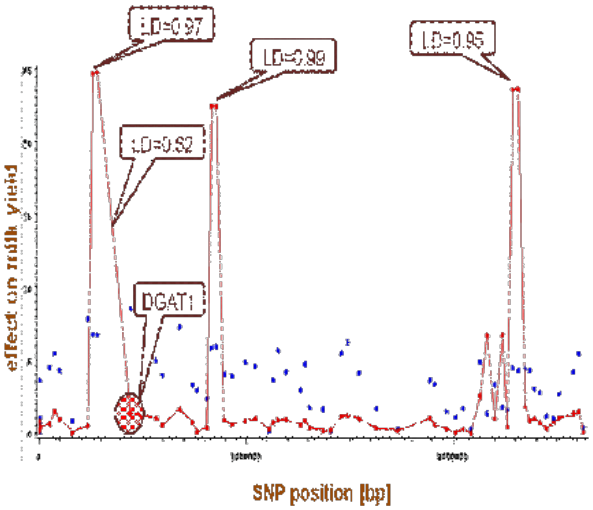

Fig. 4 Estimates of additive SNP effects on milk yield around the BTA14 region harbouring DGAT1. Model without SNP covariance (blue dots). Model with SNP covariance (red dots).

3.3. Correlation of DGV with EBV

The correlation between EBV and DGV among bulls from the training data set estimated using a no covariance model is very high and amounts to 0.98. However, when DGV is calculated based on the model with SNP covariance, the correlation drops down to 0.59, showing that for prediction of genomewide breeding value such a model is the inferior one.

4. Conclusions

Comparison of the performance of both applied approaches towards statistical modelling of multiple SNP effects shows that incorporating covariances between SNPs into the model provides better resolution for candidate gene selection and is thus a useful tool for identification of causal mutations and further on for designing a small SNP chip, which could be used in a large scale dairy cattle genotyping. On the other hand, neglecting covariance between SNPs and thus allowing for more variability of the particular estimates results in a much better prediction of the total additive genetic merit of an individual ant thus is a recommended approach for GBV calculation based on the currently widespread BovineSNP50 Genotyping BeadChip.

5. References

Grisart, B., Coppieters, W., Farnir, F., Karim, L., Ford C., *et al.* 2002. Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. *Genome Res. 12,* 222–231.

Hayes, P.J., Bowman, P.J., Chamberlain, A.J. & Goddard, M.E. 2009. Invited review: Genomic selection in dairy cattle: Progress and challenges. *J. Dairy Sci. 92,* 4008-4017.

Henderson, C.R. 1984. *Applications of Linear Models in Animal Breeding,* University of Guelph.

Legarra, A. & I, Misztal. 2008. Technical Note: Computing Strategies in Genome-Wide Selection. *J. Dairy Sci. 91,* 360-366.

Loberg, A. & Dürr, J. 2009. Interbull survey on the use of genomic information. *Interbull Bulletin 39*, 3-14.

PLINK (1.06) http://pngu.mgh.harvard.edu/purcell/plink/

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J. & Sham, P.C. 2007. PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet. 81,* 559-575.

Szyda, J., Żarnecki, A. & Kamiński, K. 2009. The Polish genomic breeding value estimation project. *Interbull Bulletin 39*, 43-46.

The Bovine Genome Sequencing and Analysis Consortium, *et al.* 2009. Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. *Science 324*, 528-532.

Acknowledgement

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