

The Polish Genomic Breeding Value Estimation Project

J. Szyda^{a,b}, A. Żarnecki^c and S. Kamiński^d

^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 *National Research Institute for Animal Production, Balice, Poland*

^d *Department of Animal Genetics, University of Warmia and Mazury, Olsztyn, 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 and, consequently, of genomic breeding values (DGVs). Out of 54 000 total SNPs, 46 267 SNPs were selected for the DGV estimation model, based on their minor allele frequency (≥ 0.01) and call rate ($\geq 90\%$). The average minor allele frequency among selected SNPs amounts 0.26 and the average call rate is 99.75%. The DGV prediction is based on a mixed model containing deregressed national proofs as a dependent variable and a random effect of uncorrelated SNPs. The residuals are weighted by the reciprocal of effective daughter contributions. In the Polish DGV project 30 traits are considered, comprising production, conformation, udder health, fertility, as well as functional longevity. An overall correlation between DGV and conventional proofs for the bulls from the training data set is very high and amounts to 0.97 for milk yield.

1. Introduction

The Polish direct/genomic breeding value (DGV) estimation project has been initiated in 2008 in the frame of the MASinBULL consortium. There are four institutions involved in the project. University of Warmia and Mazury is responsible for collecting and maintenance of DNA samples, management of SNP genotyping, the molecular analysis and organizational coordination of the project. The Wrocław University of Life Sciences is responsible for design and maintenance of the DGV data base, data editing, its statistical analysis involving estimation of DGVs and their reliability. The Wrocław University of Life Sciences also does research related to the estimation of DGVs. The National Research Institute for Animal Production provides conventional breeding values (EBVs), their accuracy and pedigree data. The project is financially supported by Bydgoszcz Animal Breeding and Insemination Center.

2. Materials and Methods

2.1. Animals

The training data set, used for the estimation of additive effects of SNPs consists of 1 227 Polish Holstein-Friesian bulls, born between 1987 and 2003. Figure 1 shows the frequency distribution of bulls across birth years. The animals were selected for genotyping based on two major criteria: first – the accuracy of their EBVs should be sufficiently high, second – EBVs should be representative for the population of all bulls active in Poland. The reliability of EBV is a function of effective daughter contribution (EDC) and figure 2 shows the distribution of genotyped bulls across different EDC classes, based on milk yield. The 84.1% of bulls from the training data set is represented by between 40 and 230 effective daughters, there is also a relatively large group of 141 (11.5%) of heavily used bulls with EDC over 360. Furthermore, based on data for milk yield evaluation, Figure 3

shows that the EBVs of bulls from the training data set very well represent the overall EBVs. One major discrepancy comprises bulls with

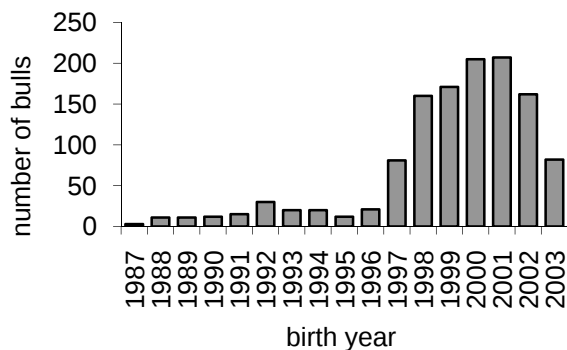


Fig. 1 Distribution of genotyped bulls across birth years.

extremely low EBVs, which are underrepresented in the training data set.

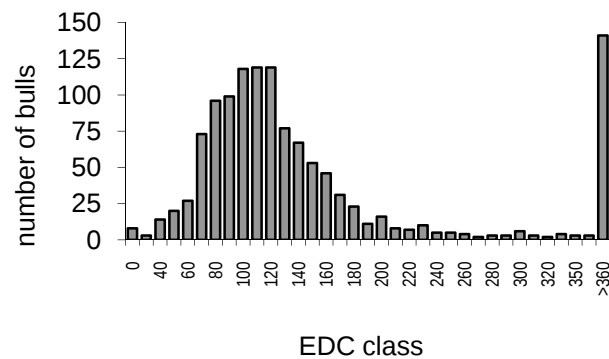


Fig. 2 Distribution of genotyped bulls across effective daughter contribution (EDC) classes for milk yield.

2.2. Traits

Altogether 30 traits are considered for the estimation of DGV. They comprise 3 production traits: milk, fat and protein yields; a

set of 21 traits describing conformation; udder health, which is expressed by the somatic cell score; 4 fertility related traits; and functional longevity.

2.3. Genotypes

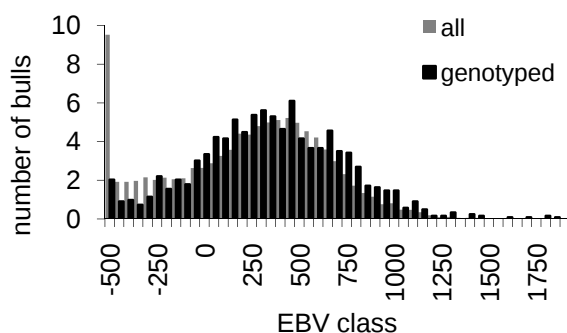


Fig. 3 Distribution of genotyped bulls across estimated breeding value (EBV) classes for milk yield.

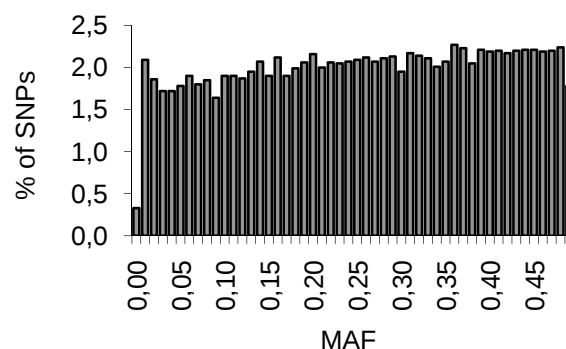


Fig. 4 Distribution of selected SNPs across minor allele frequency (MAF) classes.

Genotypes originate from the Illumina BovineSNP50 Genotyping BeadChip, which consists of 54 000 SNPs. In the training data set the SNP selection was based on their polymorphisms expressed by the minor allele frequency (MAF) and their technical quality expressed by call rate. 46 267 SNPs passed the

selection criteria set by $MAF \geq 0.01$ and at least 90% call rate. Average call rate obtained for our data was high and amounted to 99.66% and 99.75% for all and for selected SNPs, respectively. The average MAF with 0.23 for all SNPs and 0.26 for selected SNPs, was also relatively high. The distribution of SNPs

among MAF classes, is approximately uniform, showing that the half of selected SNPs have a high MAF over 0.26 (Figure 4). Still, 50.75% of all SNPs and 42.67% of selected SNPs have only two genotypes available.

2.4. DGV estimation

The following mixed model was used for the estimation of SNP effects: $\mathbf{y} = \boldsymbol{\mu} + \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, where $\mathbf{y}$ $1 \times N_g$ represents vector of deregressed EBVs, $\boldsymbol{\mu}$ is a general mean, $\mathbf{X}$ is a $N_g \times N_{\text{SNP}}$ design matrix for SNP genotypes, which is parameterised using -1, 0, or 1 for a

homozygous, a heterozygous, and an alternative homozygous SNP genotype respectively, $\boldsymbol{\beta}$ is a random $1 \times N_{\text{SNP}}$ vector of additive SNP effects assuming $\sim N\left(0, \mathbf{I} \frac{\hat{\sigma}_a^2}{N_{\text{SNP}}}\right)$, and $\mathbf{e}$ is a $1 \times N_g$ vector of residuals with $\sim N(0, \mathbf{D}\sigma_e^2)$. $\hat{\sigma}_a^2$ is an assumed additive polygenic variance and the diagonal matrix $\mathbf{D}$ contains the reciprocal of EDC on the diagonal. The model dimensions are defined by the number of genotyped animals $N_g=1\ 227$ and the number of SNPs $N_{\text{SNP}}=46\ 267$.

DGV is defined as the sum of additive effects of all SNPs estimated from the above model.

3. Results and Discussion

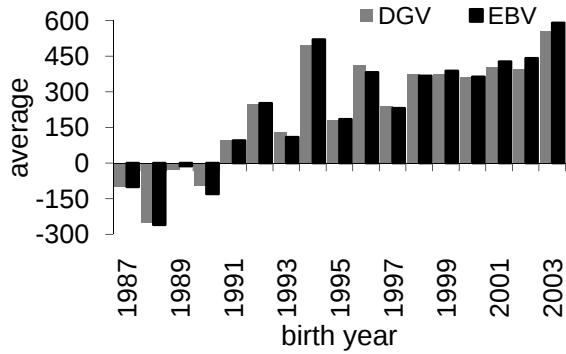


Fig. 5 Distribution of an average direct (DGV) and classically estimated (EBV) breeding value across bulls' birth years.

The average EBV for milk yield among genotyped bulls amounted to 371.04 ± 385.12 kg milk and was somewhat higher than the average DGV which was equal 357.95 ± 358.86 kg. There was some variation in the average estimates both of DGV and EBV across bulls' birth years, which reflects the genetic trend observed in the data, but it is notable that both estimates of additive genetic merit remained very similar (Figure 5). The low DGV and EBV for birth years 1987-1991 is related to the fact that only a small number of bulls born in each year within this period was genotyped (between 3 bulls in 1987 and 15 bulls in 1991),

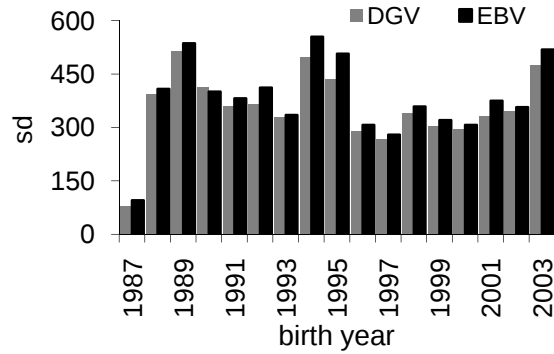


Fig. 6 Distribution of a standard deviation of direct (DGV) and classically estimated (EBV) breeding value across bulls' birth years.

which may not be representative for the whole population. Figure 6 shows that the standard deviations of DGV and EBV across bulls' birth years do not show a clear trend. EBVs (from 95.54 kg to 554.72 kg) always exhibited slightly higher variation than DGVs (from 80.23 kg to 513.18 kg), but generally remained on the similar level.

Overall correlation between EBV and DGV were very high and amounted to 0.97! It is noteworthy, that although the correlation varied across bulls' birth years, it always remained higher than 0.90 (Figure 7). The

lowest correlation of 0.93 was estimated for 1992 while the highest correlation estimates equal 0.998 were obtained for 11 bulls born in 1989 and for 15 bulls born in 1991. Some, albeit lower, variation in correlations was also observed across EDC classes as shown in Figure 8. Since the most accurate DGV and EBV estimates were obtained for the highest number of effective daughters - the last class

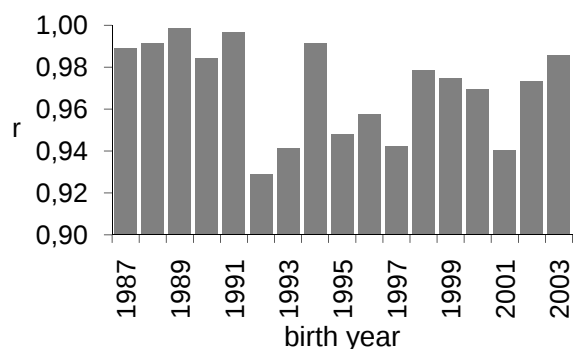


Fig. 7 Distribution of correlation coefficients between direct (DGV) and classically estimated (EBV) breeding values across bulls' birth years.

also exhibited the highest correlation of 0.997. As expected, higher EDC classes showed higher correlations in general, except class of 400-499 effective daughters and 600-699 effective daughters. The two aforementioned classes contain relatively few bulls - 24 and 16, respectively. This may account for the lower correlations.

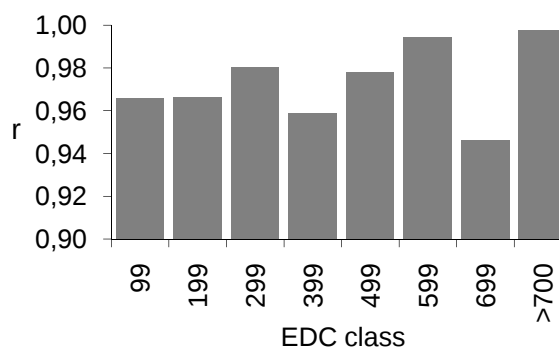


Fig. 8 Distribution of correlation coefficients between direct (DGV) and classically estimated (EBV) breeding values across effective daughter contribution (EDC) classes.

4. Conclusions

The high correlations between EVBs and DGVs for milk yield obtained for the training data set are promising for the future application of genotypic information in the early prediction of additive genetic merit of animals. However, further research is needed in order to find out whether the high EBV, DGV correlations will also be the case for traits with low heritability.

Acknowledgement

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