Testing different single nucleotide polymorphism selection strategies for prediction of genomic breeding values in dairy cattle based on low density panels

J. Szyda[1,2], K. Żukowski[1], S. Kamiński[3], A. Żarnecki[2]

[1]Department of Genetics, Wroclaw University of Environmental and Life Sciences, Wroclaw, Poland
[2]National Research Institute of Animal Production, Balice, Poland
[3]Institute of Animal Genetics, University of Warmia and Mazury, Olsztyn, Poland

ABSTRACT: In human and animal genetics dense single nucleotide polymorphism (SNP) panels are widely used to describe genetic variation. In particular genomic selection in dairy cattle has become a routinely applied tool for prediction of additive genetic values of animals, especially of young selection candidates. The aim of the study was to investigate how well an additive genetic value can be predicted using various sets of approximately 3000 SNPs selected out of the 54 001 SNPs in an Illumina BovineSNP50 BeadChip high density panel. Effects of SNPs from the nine subsets of the 54 001 panel were estimated using a model with a random uncorrelated SNPs effect based on a training data set of 1216 Polish Holstein-Friesian bulls whose phenotypic records were approximated by deregressed estimated breeding values for milk, protein, and fat yields. Predictive ability of the low density panels was assessed using a validation data set of 622 bulls. Correlations between direct and conventional breeding values routinely estimated for the Polish population were similar across traits and clearly across sets of SNPs. For the training data set correlations varied between 0.94 and 0.98, for the validation data set between 0.25 and 0.46. The corresponding correlations estimated using the 54 001 panel were: 0.98 for the three traits (training), 0.98 (milk and fat yields, validation), and 0.97 (protein yield, validation). The optimal subset consisted of SNPs selected based on their highest effects for milk yield obtained from the evaluation of all 54 001 SNPs. A low density SNP panel allows for reasonably good prediction of future breeding values. Even though correlations between direct and conventional breeding values were moderate, for young selection candidates a low density panel is a better predictor than a commonly used average of parental breeding values.

Keywords: 3K chip; genomic selection; prediction; single nucleotide polymorphism

Since Meuwissen et al. (2001) proposed the application of genetic values predicted from a large number of single nucleotide polymorphisms (SNPs) for selection, many studies related to development of methodology and practical application of genomic selection have been conducted (for recent reviews see Hayes et al., 2009; Calus, 2010; Liu et al., 2011). Challenges remain, however: (*i*) in view of rapidly growing sizes of training data sets, how to deal with the large dimensions of the statistical model used for estimation of SNP effects; and (*ii*) in view of widespread genotyping of all members of active populations and all incoming selection candidates, how to reduce genotyping costs. While dimension reduction can be realized by the choice of an appropriate statistical model, a primary way of cost reduction is the application of cheaper, low density SNP panels. A currently widespread procedure in genomic evaluations is to use sparse commercially available 3K or 6K SNP panels geno-

Supported by the Bydgoszcz Animal Breeding and Insemination Centre (Poland).

typed on very many selection candidates for the imputation to the standard 50K panel (Habier et al., 2009; Van Raden et al., 2009; Weigel et al., 2009; Dassonneville et al., 2012; Mulder et al., 2012). Recently Dassonneville (2012) and Mulder (2012) compared how various SNP selection procedures influence the quality of imputation. Our study is focused on a more basic question in checking how well different sets of sparse SNPs are able to capture the true additive genetic variability of production traits and on investigating predictive ability of low density panels without imputation. In particular, we use real data from the Polish Holstein-Friesian population in order to compare various strategies of selecting approximately 3000 SNPs out of the Illumina BovineSNP50 BeadChip panel, and prediction of genomic breeding values using model with random uncorrelated SNP effects.

MATERIAL AND METHODS

Animals

The training data set comprised 1216 Polish Holstein-Friesian bulls. Additionally, 622 bulls were used for validation. Bulls were born between 1987 and 2004. The majority of bulls in the training data set (87%) were born between 1997 and 2003, while in the validation dataset between 2001 and 2004 (54%). The age distribution of the genotyped animals is presented in Figure 1. For training bulls the daughter effective contributions (EDCs) ranged between 5 and 2756 effective daughters for production traits.

Dependent variables

Phenotypic records were represented by deregressed estimated breeding values (dEBV) for milk, protein, and fat yields, and were obtained from the national conventional routine genetic evaluation. EBVs were calculated using the random regression test day model following Strabel et al. (2005) and deregressed following the approach described by Jairath et al. (1998). For the training dataset EBVs ranged between –691.0 and 1795.0 kg (± 377.9 kg) for milk yield, between –40.4 and 56.2 kg (± 10.5 kg) for protein yield, and between –31.8 and 54.0 kg (± 12.9 kg) for fat yield. The distribution of EBVs very well represented the distribution of EBVs for the whole population of Holstein-Friesian bulls active in Poland (Figure 1).

Genotypes

DNA was extracted from semen collected and maintained in the DNA repository at the Institute of Animal Genetics, University of Warmia and Mazury. SNP genotypes were obtained using the Illumina BovineSNP50 BeadChip (revision 1) containing 54 001 SNPs.

The selection of SNPs for the statistical analysis was based on the genotypes observed among 1216 bulls from the training dataset. In the first step, SNPs which were not mapped to any chromosome, had a minor allele frequency below 0.01, or showed a call rate below 90%, were discarded. As a result 7734 SNPs, making 14.32% of the total 54 001 SNPs, were excluded from further consid-

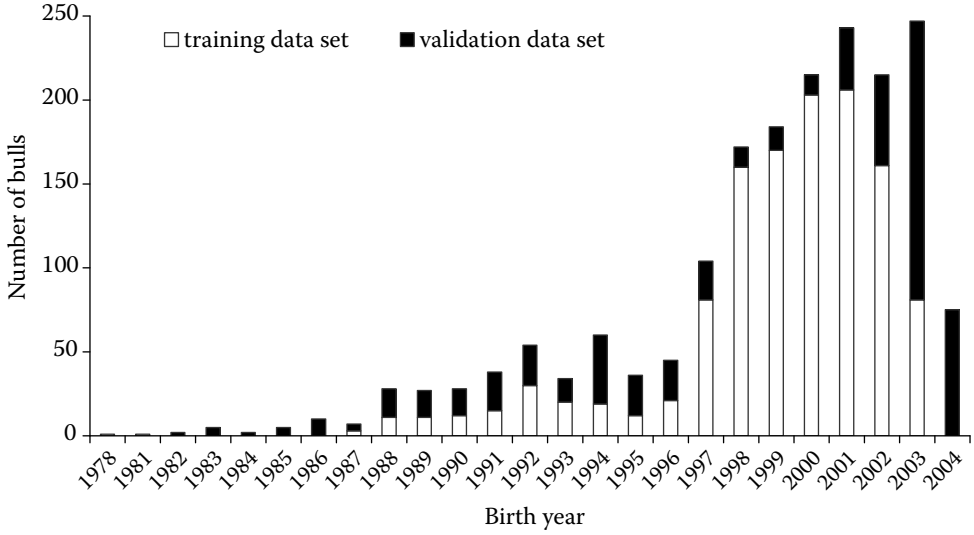

Figure 1. Distribution of genotyped bulls across birth years. Grey bars represent animals from the training data set, black bars represent animals from the validation data set

eration. The remainder of 46 267 SNPs was used for construction of low density panels.

Estimation of linkage disequilibrium

Pairwise linkage disequilibrium between linked SNPs was expressed by the r^2 statistics (r^2 was estimated using PLINK software package (Purcell et al., 2007)):

$$r^2 = \frac{D^2}{p_A(1 - p_A)\, p_B(1 - p_B)}$$

where:

D = deviation from Hardy-Weinberg equilibrium

p_A, p_B = frequencies of the more frequent allele at SNPs A and B

Construction of 3K SNP panels

Nine SNP subsets were generated using the following criteria:

Subset S1 (3000 SNPs)**.** 100 SNPs randomly selected from each chromosome.

Subset S2 (2513 SNPs)**.** The number of SNPs on each chromosome was set proportionally to the number of QTL listed by the Cattle Quantitative Trait Locus Database release from September 2009 (Hu et al., 2007). The SNPs were approximately evenly spaced on each chromosome. Note that here the highest number of SNPs (245) was selected for BTA14 while the lowest (4) for BTAX.

Subset S3 (2981 SNPs)**.** Uniformly distributed across the SNP ranking along the genome, which corresponds to selection of approximately every 15th SNP. As a consequence the highest number of SNPs (192) represented BTA01 and the lowest number (38) represented BTAX.

Subset S4 (2994 SNPs)**.** SNPs uniformly distributed along the nucleotide sequence. The highest number of SNPs (187) was located on BTA01, and the fewest SNPs (33) were selected for BTAX.

Subset S5 (2976 SNPs)**.** The number of SNPs on each chromosome was proportional to the number of segments of approximately equal length. Within each segment one SNP with the highest minor allele frequency was selected. The highest number of SNPs (183) represented BTA01 and the lowest number (49) represented BTA28.

Subset S6 (3000 SNPs)**.** SNPs selected based on their highest estimates for milk yield as obtained for the high density SNP set using a univariate, multiple SNP, mixed model by Szyda et al. (2009). An additional criterion was that the pairwise linkage disequilibrium between SNPs, expressed by r^2, is below 0.80. For **S6** the highest number of 171 SNPs was mapped to BTA01, and the fewest SNPs (41) were located on BTA28.

Subset S7 (3000 SNPs)**.** SNPs selection was as for **S6** except that estimated effects on stature instead of on milk yield were considered. The highest (172) and lowest (51) numbers of SNPs were selected for BTA1 and BTA25, respectively.

Subset S8 (3000 SNPs)**.** SNPs selection was as for **S6** except that estimated effects on type instead of on milk yield were considered. The highest (180) and lowest (53) numbers of SNPs were selected for BTA1 and BTA25, respectively.

Subset S9 (2886 SNPs)**.** SNPs corresponding to the Illumina GoldenGate Bovine3K Genotyping BeadChip except for SNPs on BTAY. For this subset the highest (175) number of SNPs was located on BTA01, and the lowest (48) on BTA25.

Table 1. Number of common SNPs (above diagonal) and percentage of common SNPs (below diagonal) between data subsets

	S1	S2	S3	S4	S5	S6	S7	S8	S9
S1		185	205	212	204	215	195	218	179
S2	7.36		134	162	172	176	174	161	169
S3	6.88	4.50		201	188	203	195	190	215
S4	7.08	5.41	6.71		211	209	195	189	210
S5	6.86	5.79	6.32	7.10		285	289	262	343
S6	7.17	5.87	6.77	6.97	9.59		478	572	256
S7	6.50	5.80	6.50	6.50	9.72	15.93		1820	234
S8	7.27	5.37	6.33	6.30	8.81	19.07	60.67		247
S9	6.20	5.86	7.45	7.28	11.54	8.87	8.11	8.56	

The number and percentage of SNPs common across subsets are shown in Table 1. For most subset combinations the percentage of common SNPs is below 9%. Similarity was the highest, over 60% (1820 SNPs), between **S7** and **S8**, and relatively high between **S6** and **S8** (19.07%, 572 SNPs) as well as between **S6** and **S7** (15.93%, 478 SNPs).

Estimation of SNP effects

The model with random uncorrelated SNP effect was used to estimate SNP effects from different subsets of data. The model includes:

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

where:

y = vector of dEBV for milk, protein, or fat yield (m: 1)

μ = overall mean

a = vector of random additive SNP effects assuming $a \sim N(0, \mathbf{I}(\hat{\sigma}_a^2/n))$, with $\mathbf{I}$ being the identity matrix and $\hat{\sigma}_a^2$ representing the additive genetic variance (n: 1)

$\mathbf{Z}$ = corresponding design matrix with elements of −1, 1, and 0 and for two homozygous and a heterozygous SNP genotype respectively

ε = vector of random errors assuming $\varepsilon \sim N(0, \mathbf{D}\sigma_\varepsilon^2)$ with $\mathbf{D}$ being a diagonal matrix with the reciprocal of effective daughter contributions corresponding to dEBV and σ_a^2 denoting error variance (m: 1)

n = number of SNPs

m = number of genotyped animals

The model assumes that all of the observed additive genetic variance is due to the random effect of SNPs. The additive genetic variance ($\hat{\sigma}_a^2$) was not estimated but was assumed as known, based on the estimates used in the Polish national genetic evaluation model for a corresponding trait. Estimation of the parameters underlying considered model was based on solving the mixed model equations:

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

where:

$\mathbf{X}^T$ = design matrix for fixed effects (vector of 1)

$\mathbf{Z}^T$ = design matrix for random SNP effects

$\mathbf{G}$ = covariance matrix for random SNP effects expressed by $\mathbf{I}(\hat{\sigma}_a^2/n)$

$\mathbf{R}$ = residual covariance matrix expressed by $\mathbf{D}(\hat{\sigma}_e^2)$

$\hat{\boldsymbol{b}}, \hat{\boldsymbol{g}}$ = vectors of fixed effects represented by a general mean and random effects represented by SNP effects, respectively

$\boldsymbol{y}$ = vector of dEBV

Estimation of direct genomic values

A direct genomic value of i^{th} bull (DGV_i) is defined as the sum of additive effects of SNPs:

$$\hat{DGV}_i = \sum_{j=1}^{n} \mathbf{Z}_{ij}\hat{a}_j$$

where:

$\mathbf{Z}_{ij}$ = element of the design matrix for SNP effects corresponding to animal i and SNP_j

a_j = estimate of the additive effect of SNP_J

RESULTS AND DISCUSSION

Correlations between EBV and DGV

For the training data set Pearson's correlations between deregressed estimated breeding values and direct genomic values for production traits are shown in Table 2. The correlation ranged from

Table 2. Pearson's correlation coefficients between estimated breeding values and direct genomic values for bulls from the training dataset

Trait	Subset									54K SNP chip[1]
	S1	S2	S3	S4	S5	S6	S7	S8	S9	
Milk yield	0.94	0.94	0.95	0.94	0.95	0.98	0.96	0.95	0.91	0.98
Protein yield	0.95	0.94	0.95	0.95	0.95	0.97	0.96	0.95	0.91	0.98
Fat yield	0.94	0.94	0.95	0.94	0.95	0.96	0.95	0.95	0.90	0.98

[1]results were estimated for Szyda et al. (2011) study, note that there were differences in validation data set structure between Szyda et al. (2011) and the present study

Table 3. Pearson's correlation coefficients between estimated breeding values and direct genomic values for bulls from the validation dataset

Trait	Subset									
	S1	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	54K SNP chip[1]
Milk yield	0.46	0.43	0.41	0.40	0.37	0.45	0.39	0.39	0.27	0.38
Protein yield	0.45	0.45	0.44	0.42	0.39	0.46	0.41	0.43	0.36	0.37
Fat yield	0.32	0.39	0.28	0.27	0.32	0.39	0.39	0.32	0.25	0.32

[1]results were estimated for Szyda et al. (2011) study, note that there were differences in validation data set structure between Szyda et al. (2011) and the present study

0.90 to 0.98. Comparing the subsets, the values were always the highest for **S6** (milk yield 0.98, protein yield 0.97, fat yield 0.96) and the lowest for **S9** (milk yield 0.68, protein yield 0.69, fat yield 0.64). The corresponding correlations based on the high density panel were 0.98 for the three production traits (Szyda et al., 2011).

The correlations for the validation data set are shown in Table 3. As expected, the correlations were weaker than for the training data set and ranged between 0.25 (**S9**, fat yield) and 0.46 (**S1**, milk yield), which was less than the corresponding correlations of 0.38 for milk yield, 0.37 for protein yield, and 0.32 for fat yield, estimated for the high density panel

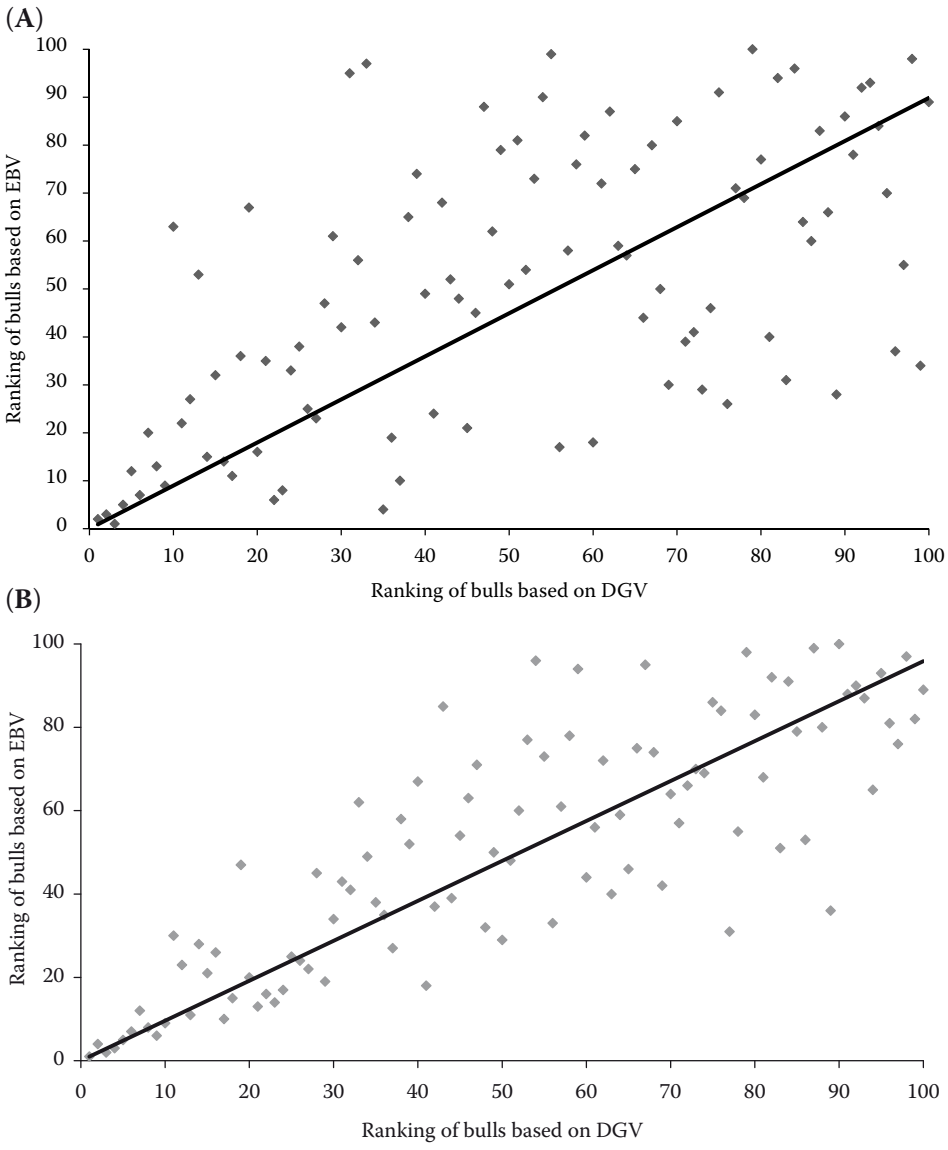

Figure 2. Rank correlations between top 100 bulls as ranked based on estimated breeding values (EBV) and direct genomic values (DGV) for milk yield, showing the worst **S9** (**A**) and the best **S6** (**B**) subsets

Table 4. Spearman's rank correlations coefficients between estimated breeding values and direct genomic values for top 100 bulls from the training dataset

Trait	Subset							
	S1	**S2**	**S3**	**S4**	**S6**	**S7**	**S8**	**S9**
Milk yield	0.68	0.73	0.75	0.67	0.83	0.77	0.75	0.59
Protein yield	0.66	0.62	0.69	0.70	0.80	0.75	0.74	0.62
Fat yield	0.64	0.68	0.73	0.63	0.75	0.69	0.70	0.65

Table 5. Regression coefficients of estimated breeding values on direct genomic values for the training dataset

Trait	Subset							
	S1	**S2**	**S3**	**S4**	**S6**	**S7**	**S8**	**S9**
Milk yield	1.10	1.08	1.09	1.09	0.95	1.06	1.06	1.12
Protein yield	1.09	1.08	1.09	1.08	0.99	1.06	1.05	1.11
Fat yield	1.07	1.07	1.08	1.09	0.99	1.05	1.05	1.10

by Szyda et al. (2011). Comparing the subsets, only **S9** resulted in markedly weaker correlations. The higher correlations observed are a consequence of different validation data set structures between Szyda et al. (2011) and the current study. Bulls in both validation data sets do not overlap – the former study used validation bulls which are younger than training bulls, the current study used validation bulls belonging to the same generations (1997–2003) as the training bulls (1987–2004). Although no large differences between traits were found, the correlations were somewhat weaker for fat yield.

Rank correlations between EBV and DGV for the training data set

Spearman's rank correlations among the top 100 bulls ranked based on EBV are shown in Table 4. Considering differences among subsets, the correlations were the strongest for **S6** and the weakest for **S9,** the same as observed for Pearson's correlations for milk yield. Figure 2 shows the rank comparisons for the worst subset **S9** and the best subset **S6.** The results indicated that for the best subset the two rankings overlapped, especially for the top 15 animals. For the worst scenario both rankings overlapped only for the top 10 animals.

Table 6. Regression coefficients of estimated breeding values (EBV) on direct genomic values and of EBV on parent average (PA) for the validation dataset

Trait	Subset								
	S1	**S2**	**S3**	**S4**	**S6**	**S7**	**S8**	**S9**	PA
Milk yield	0.64	0.56	0.52	0.56	0.57	0.50	0.50	0.42	0.53
Protein yield	0.62	0.63	0.62	0.59	0.58	0.56	0.57	0.52	0.40
Fat yield	0.42	0.50	0.37	0.34	0.47	0.47	0.40	0.36	0.19

Regression of EBV on DGV

Considering the training data set, for all combinations of traits and subsets the regression coefficients of EBV on DGV were close to the expected value of one (Table 5). The optimal results were obtained for subset **S6**. As expected, the regression coefficients were always lower for the validation bulls (Table 6). Trait with the lowest regression coefficients was fat yield (0.36–0.50) and the trait with the highest coefficients was protein yield (0.52–0.63). For the best subset **S2** regression coefficients were lower by 0.41 for milk yield, 0.07 for protein yield, and 0.12 for fat yield than if using the high density panel. Subset **S9** resulted in lower regression coefficients than the other subsets. Still, for most of the subset-trait combinations they were higher than a corresponding regression coefficient involving parent average (PA).

Comparison of SNP subsets

Based on the imposed data preselection criteria (minor allele frequency and call rate), the average SNP reduction over all chromosomes was 85.49%

and ranged from 77.57% for BTX to 89.51% for BTA20. This set was used to form the eight SNP subsets considered in this study. For **S1**, **S3**, and **S4** an arbitrary selection approach was used, where the selection of SNPs was related neither to Linkage Disequilibrium (LD) between them, nor to the information on genetic architecture underlying the traits under study. It can be hypothesized that such an arbitrary selection of SNPs may lead to loss of a substantial part of genetic information, and in consequence reduce the predictive ability of a model. An intermediate approach was represented by **S2**, for which SNPs at each chromosome were selected proportionally to the total (regardless of the trait) number of QTL reported in the QTLdb. The largest number of QTL (245) was reported for BTA14, and consequently many more SNPs were selected for this chromosome than in the remaining subsets. The arbitrary approach presented in **S3**, **S4**, and **S5** was used widely in imputation studies where good SNPs arrangement among imputed SNPs is the key to high imputation accuracy (Dassonneville et al., 2012; Mulder et al., 2012). **S6**, **S7**, and **S8** were not generated arbitrarily but were based on linkage disequilibrium and SNP effect estimates. Although r^2 was one of the selection criteria, the average r^2 in those subsets, which varied from 0.11 to 0.12, was higher than for the remaining subsets, indicating that the high estimates from previous studies picked SNPs in high LD. Moreover, the minor allele frequency for those three subsets was also higher; it varied from 0.31 to 0.32 instead of 0.26 to 0.27 as found for the remaining subsets, showing that preselection yielded more informative SNPs. Finally, **S9** representing the Illumina Bovine3K BeadChip contains SNPs selected based on even spacing on the cattle genome and their minor allele frequency.

Although Van Raden et al. (2009) indicated that careful preselection of SNPs is the key to higher prediction accuracy, in our study there were no marked differences in correlations between EBV and DGV across subsets. As it could have been expected, the correlations were the highest for **S6**, **S7**, and **S8**, especially for **S6**. A similar tendency has also been estimated by Vazquez et al. (2010). However, if a standard SNP set is expected to be used across the whole range of traits, the advantage of **S6**, **S7**, and **S8** would be close to zero, genetic correlations between production and many other traits subject to routine recording. A possible way of circumventing the problem is to select SNPs based on a composite selection index and to include parent average into the prediction model, as proposed by Vazquez et al. (2010). Moreover, we argue that in order to avoid an overrepresentation of SNPs remaining in high LD with genes having a strong effect on selection index (like *DGAT1*) an additional criterion for a low density panel should be pairwise LD between SNPs. Weigel et al. (2009) also estimated that subsets selected based on SNP effects were associated with better predictive ability than arbitrarily selected subsets, although they reported a much more pronounced difference than the one we found. Interestingly, despite having almost 500 fewer SNPs than the other subsets, **S2** did not show lower correlations, indicating that it was well able to capture underlying genetic variation of the analyzed traits. On the other hand, the commercially available **S9** usually showed lower correlations.

Comparison of SNP effect estimation models

Models with a random SNP effect have been investigated in many studies and always gave strong correlations between EBV and DGV as well as Bayesian models (Meuwissen et al., 2001; Habier et al., 2007; Moser et al., 2009; Van Raden et al., 2009). The only drawback of such models is the complexity of estimating model parameters, especially for large data sets, which are based on BLUP or on the Bayesian principle. The important advantage of the BLUP approach used in our study is that it requires a simple assumption of a constant variance component for each SNP and thus it is independent of the unknown number of QTL underlying an analyzed trait (Daetwyler et al., 2010). In the context of the selection of SNPs based on their effects estimated from a high density chip, the effects from a Bayesian approach cannot be directly interpolated to a low density chip (Habier et al., 2009). To circumvent this problem, either equal variances for all SNPs can be assumed, as was done in our study, or the effects of all SNPs can indirectly be incorporated into the low density panel through probabilities of descent of the missing marker genotypes (Habier et al., 2009).

Comparison of model predictive ability

As expected, the correlation coefficients estimated within the training data set were generally

very strong, since the same bulls that provide EBVs are used for estimating SNP effects. Yet the estimated rank correlation coefficients showed that the ranking of best bulls based on EBV did not reflect the ranking based on DGV, a feature very unfavourable from the breeding industry perspective. A lower correlation was especially evident for the 100 bulls with the top EBVs. However, when fewer top bulls (up to 15) are involved, the rankings based on EBV and DGV showed a very good agreement.

As already shown by many studies using both simulated (Meuwissen et al., 2001; Calus and Veerkamp, 2007; Habier et al., 2007; Muir, 2007) and real (Hayes et al., 2008; Moser et al., 2009; Habier et al., 2010) data sets, the correlations between EBV and DGV for validation bulls are much lower. For a real data set, a composite trait "lifetime net merit" and a smaller SNP panel, Weigel et al. (2009) obtained stronger correlations (0.25–0.57) than the ones estimated from our validation data (0.22–0.47). Habier et al. (2009) also estimated higher accuracies using a simulated data set. The accuracies obtained in their study appear to be independent of the number of simulated QTL, which is in agreement with our result based on real data, where similar correlations were obtained for milk yield and fat yield, a trait dominated by a gene with a large effect (*DGAT1*). This is caused by limited allocation of large effects to a single gene in the model with random SNP effects in comparison to Bayesian models and a limited size of the data set. Also correlations estimated by Vazquez et al. (2010), ranging between 0.50 and 0.65, are stronger than values obtained in our analysis. The lower correlations estimated in our study may have been due to differences in: (*i*) heritability – Habier et al. (2009) assumed very high heritability of 0.5 in contrast to 0.30 for milk yield and 0.29 for fat and protein yields estimated for the Polish Holstein Friesian population (http://www-interbull.slu.se/national_ges_info2/framesida-ges.htm), (*ii*) size of the training data set – 3305 bulls in Weigel et al. (2009) and Vazquez et al. (2010) versus 1216 bulls in our data set, (*iii*) structure of our validation data set with some of the validation bulls being older than bulls from the training data set.

An important difference from the training dataset is that the correlations obtained for production traits clearly varied across traits, with the lowest correlations obtained for fat yield (Table 1). This trend was previously demonstrated by Hayes et al. (2008) and Vazquez et al. (2010) for evenly spaced SNP panels. The loss in accuracy for fat yield could be explained by estimating the effect of *DGAT1*, or more precisely, of SNPs in strong LD with the gene. If the effects of the gene, which has a predominant impact on fat yield, are not precisely estimated in the training data set, or if recombination alters LD between SNPs and *DGAT1*, the accuracy estimated for the validation bulls decreases more rapidly than is observed for traits with a pure polygenic inheritance mode.

The regression coefficients of EBV on DGV for the validation set varied from 0.34 to 0.63 and were lower than the regression coefficients for the training dataset. Moser et al. (2009) gave similar results for young Holstein-Friesian bulls. The lower correlation coefficients for the validation dataset are perhaps also due to the age range of validated and training bulls. The main group of DGV predicted animals was born between 2003 and 2004. However, in contrast to other studies, in this work there were no situations where all older bulls were in the training data and all younger bulls were in the test data. A few bulls in the validation set were born earlier than the bulls from the training set, such as the group of bulls born before 1987. Thus we have to deal with backward DGV prediction of animals whose EBVs were estimated, on average, from more than 100 daughters. Moser et al. (2009) suggested that accuracies derived by cross-validation are likely to overestimate the actual accuracies of future predictions for some traits in young bulls. The solution to this problem might be to use combined SNP and pedigree information or to use **H** matrix incorporating full pedigree information (Legarra et al., 2009).

CONCLUSION

Using a real data set from the Polish population of Holstein Friesian dairy cattle, we showed that applying a low density panel consisting of approximately 3000 SNPs allows for reasonably good prediction of EBVs for production traits. At least for most of the tested combinations higher regression coefficients were obtained for regressing DGV on EBV than PA on EBV, which indicates that even a low density SNP panel is a better predictor of EBV for selection candidates than PA. In practice, an optimal solution would be to use a combination of both sources of information in the

form of Genomically Enhanced Breeding Values. In view of results of the present and previous studies (Szyda et al., 2011), in order to be able to describe the additive genetic variation of production traits using a low density panel, special attention should be paid to (*i*) the genotyping quality of SNPs, in order to maximize the informativeness of genotypes, and (*ii*) the selection of markers capturing the underlying major genes either through their high estimates on a composite selection index or through their high LD to known QTL.

Acknowledgement

The study is implemented within the frame of the MASinBULL project.

REFERENCES

Calus M.P.L. (2010): Genomic breeding value prediction: methods and procedures. Animal, 4, 157–164.

Calus M.P.L, Veerkamp R.F. (2007): Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. Journal of Animal Breeding and Genetics, 124, 362–368.

Daetwyler H.D., Pong-Wong R., Villanueva B., Woolliams J.A. (2010): The impact of genetic architecture on genome-wide evaluation methods. Genetics, 185, 1021–1031.

Dassonneville R., Fritz S., Ducrocq V., Boichard D. (2012): Short communication: Imputation performances of 3 low-density marker panels in beef and dairy cattle. Journal of Dairy Science, 95, 4136–4140.

Habier D., Fernando R.L., Dekkers J.C.M. (2007): The impact of genetic relationship information on genome-assisted breeding values. Genetics, 177, 2389–2397.

Habier D., Fernando R.L., Dekkers J.C.M. (2009): Genomic selection using low-density marker panels. Genetics, 182, 343–353.

Habier D., Tetens J., Seefried F.R., Lichtner P., Thaller G. (2010): The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genetics Selection Evolution, 19, 42–45.

Hayes B.J., Bowman P.J., Chamberlain A.C., Verbyla K., Goddard M.E. (2008): Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genetics Selection Evolution, 24, 41–51.

Hayes B.J., Bowman P.J., Chamberlain A.J., Goddard M.E. (2009): Genomic selection in dairy cattle: progress and challenges. Journal of Dairy Science, 92, 433–443.

Hu Z.L., Fritz E.R., Reecy J.M. (2007): AnimalQTLdb: a livestock QTL database tool set for positional QTL information mining and beyond. Nucleic Acids Research, 35, D604–D609.

Jairath L., Dekkers J.C.M., Schaeffer L.R., Liu Z., Burnside E.B., Kolstad B. (1998): Genetic evaluation for herd life in Canada. Journal of Dairy Science, 81, 550–562.

Legarra A., Aguilar I., Misztal I. (2009): A relationship matrix including full pedigree and genomic information. Journal of Dairy Science, 92, 4656–4663.

Liu Z., Seefried F.R., Reinhardt F., Rensing S., Thaller G., Reents R. (2011): Impacts of both reference population size and inclusion of a residual polygenic effect on the accuracy of genomic prediction. Genetics Selection Evolution, 43, 19.

Meuwissen T.H.E., Hayes B.J., Goddard M.E. (2001): Prediction of total genetic value using genome-wide dense marker maps. Genetics, 157, 1819–1829.

Moser G., Tier B., Crump R.E., Khatkar M.S., Raadsma H.W. (2009): A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genetics Selection Evolution, 41, 56.

Muir W.M. (2007): Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. Journal of Animal Breeding and Genetics, 124, 342–355.

Mulder H.A., Calus M.P.L., Druet T., Schrooten C. (2012): Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle. Journal of Dairy Science, 95, 876–889.

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. American Journal of Human Genetics, 81, 559–575.

Strabel T., Szyda J., Ptak E., Jamrozik J. (2005): Comparison of random regression test-day models for Polish Black and White Cattle. Journal of Dairy Science, 88, 3688–3699.

Szyda J., Żarnecki A., Kamiński S. (2009): The Polish genomic breeding value estimation project. Interbull Bulletin, 39, 47–50.

Szyda J., Żarnecki A., Suchocki T., Kamiński S. (2011): Fitting and validating the genomic evaluation model to Polish Holstein-Friesian cattle. Journal of Applied Genetics, 52, 363–366.

VanRaden P.M., VanTassell C.P., Wiggans G.R., Sonstegard T.S., Schnabel R.D., Taylor J.F., Schenkel F.S. (2009): Invited review: Reliability of genomic predictions for North American Holstein bulls. Journal of Dairy Science, 92, 16–24.

Vazquez A.I., Rosa G.J.M., Weigel K.A., de los Campos G., Gianola D., Allison D.B. (2010): Predictive ability of

subsets of single nucleotide polymorphisms with and without parent average in US Holsteins. Journal of Dairy Science, 93, 5942–5949.

Weigel K.A., de los Campos G., González-Recio O., Naya H., Wu X.L., Long N., Rosa G.J.M., Gianola D. (2009): Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. Journal of Dairy Science, 92, 5248–5257.

Received: 05–16–2012
Accepted after corrections: 11–19–2012

Corresponding Author

Joanna Szyda, Associate Professor, Wroclaw University of Environmental and Life Sciences, Department of Genetics, Kożuchowska 7, 51-631 Wroclaw, Poland
Tel.: +48 713 205 758, fax +48 713 205 758, e-mail: joanna.szyda@up.wroc.pl