Animal Science Papers and Reports vol. 40 (2022) no. 2, 135-146
Institute of Genetics and Animal Biotechnology
of the Polish Academy of Sciences, Jastrzębiec, Poland

The impact of rare Single Nucleotide Polymorphism variants on the genomic evaluation of dairy cattle*

Tomasz Suchocki[1,2], Michalina Jakimowicz[1], Andrzej Żarnecki[2], Arkadiusz Dziech[3], Joanna Szyda[1,2]**

[1] Wrocław University of Environmental and Life Sciences, Department of Genetics, Biostatistics Group, Kożuchowska 7, 51-631 Wroclaw, Poland

[2] National Research Institute of Animal Production, Krakowska 1, 32-083 Balice, Poland

[3] Wrocław University of Environmental and Life Sciences, Department of Genetics, Kożuchowska 7, 51-631 Wroclaw, Poland

(Accepted April 12, 2022)

The experiment described in this study was designed to test the effect of rare variants on the genomic prediction in dairy cattle. Common polymorphisms are capable of explaining only a small proportion of the underlying genetic variation of complex phenotypes. Variants representing functional mutations with large effects on complex phenotypes are expected to be rare due to natural or artificial selection pressure. Therefore, it is important to check whether the use of rare variants could increase the accuracy of ranking of animals by providing a tool for more precise differentiation between bulls with a high additive genetic merit. The goal of our study was to verify whether including rare variants in a genomic selection model provides a more accurate description of the additive genetic background of traits under selection in dairy cattle. The number of animals used in the analyses varies and depending on the trait it ranged from 77,578 individuals for type trait to 100,650 individuals for somatic cell score. We used the linear mixed model to compare estimates of SNP effects for Holstein-Friesian cattle of the two data sets – a set containing only single nucleotide polymorphisms defined by minor allele frequency greater than 1%, which is routinely used in the Polish genomic evaluation system (46,216 SNPs), and a set containing SNPs selected based only on the call rate (54,378 SNPs). Based on the SNP estimates we also calculated Direct Genomic Values

*Calculations were carried out using resources provided by the Wroclaw Centre for Networking and Super-computing (http://wcss.pl), grant no. 509.
**Corresponding author: michalina.jakimowicz@upwr.edu.pl

(DGV) and Genomically Enhanced Breeding Values (GEBV) and compared them between both data sets. IAll the analyses were conducted for production, fertility, conformation and udder health traits. We also assessed the time required for the two most computationally demanding components of genomic selection, i.e. preparation of genotype data and estimation of SNP effects between those two data sets. The results of our study indicated that the analysis including rare variants resulted in changes in the individual ranking of the top 100 male and female candidates, whereas it had no effect on the outcome of the quality of EBV prediction as expressed by the Interbull validation test.

KEY WORDS: rare variants / genome-wide association study / validation test / SNP chip / genomic selection

Predicting phenotypes from genotype data is important for plant and animal breeding, as well as evolutionary biology. Genomic-based phenotype prediction has mostly been applied using data from single-nucleotide polymorphism (SNP) genotyping platforms. Usually, the set of markers included in the final analysis is edited based on minor allele frequency (MAF) and call rate. Such filtering leads to a result where additive effects of SNPs with rare genotypes are not considered in the analysis, so that the impact of such markers on estimated breeding values is neglected.

Rare genetic variants, i.e. polymorphisms with a lowminor allele frequency, typically below 1%, have been brought into focus in the context of genetic determination of complex traits [Bomba *et al.* 2017, Schaid *et al.* 2018]. This stems primarily from the so-called "missing heritability" indicated for most of the complex phenotypes measured in humans, indicating that the common polymorphisms are able to explain only a small proportion of the underlying genetic variation of such traits ranging between 1.5% and 50% [Manolio *et al.*, 2009]. Variants representing functional mutations with large effects on complex phenotypes are expected to be rare because of natural or artificial selection pressure against an unfavourable allele [Hayes and Daetwyler 2019]. The biological explanation is that since a mutation is functional, it is subjected to selection, which as a consequence affects population allele frequency more strongly than in the case of a neutral mutation [Frazer *et al.* 2009]. The effect of selection pressure is also strong on coding functional variants and would affect more fitness traits because of lower average heritability of those phenotypes.

Indeed, in human populations a number of studies has indicated associations between rare variants and complex traits [Sulem *et al.* 2011; Styrkarsdottir *et al.*, 2013]. Also, in yeasts (Saccharomyces cerevisiae) the importance of rare variants in phenotypes of quantitative traits was greater than might have been expected based on their occurrence (while only 27.8% variants were defined as rare, they constituted 51.7% of the median contribution for all traits). Moreover, quantitative trait loci (QTLs) commonly found in rare variants had larger substitution effects, while those with an abundance of common variants were less influential [Bloom *et al.* 2019]. Dairy cattle is a very good population facilitating verification of this hypothesis. It has undergone directional selection for production traits over many generations and has very good records of complex traits and familial relationships. Moreover, the recent success of genomic selection has provided extensive information on genotypes of

single nucleotide polymorphisms distributed over the whole genome, available for many individuals.

Therefore, the goal of our study was to verify whether including rare variants in a genomic selection model provides a more accurate description of the additive genetic background for traits under selection in dairy cattle. The analysis involved comparisons of two data sets – a set containing only SNPs defined by MAF greater than 1% and call rate over 99%, which is routinely used in the Polish genomic evaluation system, and a set containing SNPs selected based only on the call rate. For both data sets we compared (1) estimates of the effects of common SNPs; (2) changes in bull rankings based on genomically enhanced breeding values (GEBV); and (3) results of the Interbull validation test. The analysis also covered time required (CPU time) for the two most computationally demanding components of genomic selection: preparing genotype data and estimation of SNP effects between those two data sets.

Material and methods

Reference animals

The analyzed data set originated from the EuroGenomics Cooperative U.A. Holstein-Friesian dairy cattle population. For each bull born before 2010, pseudophenotypes were available in the form of deregressed breeding values (DRP) corresponding to the Interbull evaluation from April 2020. In the comparison we considered traits representing different functional groups, including one production, two fertility, two conformation, one udder health and one longevity trait. Specifically, the analyzed traits comprise protein yield (PRO), heifer conception rate (HCO), cow conception rate (CC1), stature (STA), type (TYP), somatic cell score (SCS) and functional longevity (DLO). The numbers of reference bulls for each of the considered traits are presented in Table 1. For all the traits except for TYP, the EuroGenomics reference population was used. For TYP, which is not evaluated internationally, we used the national reference population. Apart from different selection pressures (expressed by different weights in the total merit index) and different sizes of reference populations, traits were also selected to represent varying levels of heritability. HCO and CC1 are low-heritable traits, DLO, PRO, TYP and SCS are moderately heritable, while STA has high heritability. The heritability estimates corresponding to the Polish Holstein-Friesian population are presented in Table 1.

Most of the reference individuals (87%) were genotyped using the Illumina BovineSNP50 BeadChip Version 2. All individuals genotyped using other platforms were imputed to the above microarray using the Beagle software [Browning and Browning 2009]. Almost all the imputed animals were genotyped using the EuroG10K BeadChip v2-5. In the final analysis two data sets of SNP genotypes were used: (i) ORIG consisting of 46,216 SNPs representing the standard common SNP set used for the routine genomic evaluation in Poland, and (ii) RARE comprising 54,378 SNPs without preselection on MAF, including common and rare polymorphisms. The SNP

Table 1. Summary of analyzed sub-sets of individuals and trait characteristics

Trait	Number of bulls in the reference population born before 2010	Number of validation bulls born after 2010	Number of cows born after 2010	Heritability	Ratio of genetic variance for additive polygenic effect (%)
Protein yield	34,249	23,001	43,392	0.290	20
Heifer conception rate	31,509	23,553	43,392	0.027	40
Cow conception rate	33,534	23,448	43,392	0.028	40
Stature	33,299	23,566	43,392	0.540	30
Type	4,838	29,348	43,392	0.330	40
Longevity	21,795	25,998	43,392	0.173	40
Somatic cell score	34,168	23,090	43,392	0.320	20

selection criterion for ORIG comprised MAF of min. 0.01, while for the RARE data set SNPs were not preselected for MAF. For both data sets SNPs with unspecified genomic positions and with a call rate below 99% were removed.

Validation animals

For the trend validation of Genomically Enhanced Breeding Values (GEBVs) bulls born after 2010 were used, with pseudophenotypes expressed by DRPs from MACE (CC1, DLO, HCO, PRO, SCS, and STA) or DRP based on the national EBV (TYP). In addition, top 100 rankings of GEBVs estimated based on the ORIG and the RARE data sets for the validation bulls and cows born after 2010 were compared. All the bulls in the validation data set were originally genotyped using the Illumina BovineSNP50 BeadChip Version 2, while 93% of cows were genotyped using the EuroG10K BeadChip v2-5.

SNP effect estimation

The following mixed model [Szyda *et al.* 2011] was used to estimate the additive effects of SNPs:

$$y = \mu + Z_1 g + Z_2 a + \varepsilon \quad (1)$$

where y represents the vector of deregressed breeding values of the reference bulls (for all the traits except for TYP we used deregressed MACE EBVs calculated for the Polish scale and for TYP we used deregressed national EBVs), μ is the general mean, Z_1 is the design matrix for SNP genotypes, which is parameterized as -1, 0, or 1 for thea homozygous, heterozygous and an alternative homozygous SNP genotype, respectively, g is the vector of random additive SNP effects, Z_2 is the design matrix for a polygenic effect, a is the vector of random "residual" additive polygenic effects of bulls, which is important to reduce the inflation of genomic prediction with actual data and to account for the incomplete linkage disequilibrium between the SNPs and genes or causal mutations of the analyzed phenotypes [Liu *et al*. 2016]. ε is the vector of error terms with $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{D}\hat{\sigma}_e^2)$, where D is the diagonal matrix containing the reciprocal of bulls' effective daughter contributions (EDC; for all the traits except for TYP we used EDC from the MACE evaluation calculated on the Polish scale, for TYP

we used EDC from Poland) on the diagonal and $\hat{\sigma}_a^2$ representing the error variance. The covariance structure of g was assumed to be $\boldsymbol{g} \sim N\left(0, \boldsymbol{I}\frac{\hat{\sigma}_a^2}{N_{snp}}\right)$, with I being the identity matrix, $\hat{\sigma}_a^2$ representing the additive genetic variance of a given trait and NSNP being the number of SNPs used (here 46,216 for ORIG and 54,495 for the RARE data set) assigning the same small fraction of the polygenic variance to each of the NSNP polymorphisms. $\boldsymbol{a} \sim N(0, \boldsymbol{A}\hat{\sigma}_{a*}^2)$, where A is the numerator relationship matrix and $\hat{\sigma}_{a*}^2$ is the predetermined ratio of additive genetic variance for each of the traits, the same as assumed for the routine genomic evaluation in Poland. The variance ratio for each trait is presented in Table 1.

The estimation of parameters inof the above model was based on solving the mixed model equations [Henderson 1984]:

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{g}} \\ \hat{\boldsymbol{a}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X^T R^{-1} X} & \boldsymbol{X^T R^{-1} Z_1} & \boldsymbol{X^T R^{-1} Z_2} \\ \boldsymbol{Z_1^T R^{-1} X} & \boldsymbol{Z_1^T R^{-1} Z_1 + G_1^{-1}} & \boldsymbol{Z_1^T R^{-1} Z_2} \\ \boldsymbol{Z_2^T R^{-1} X} & \boldsymbol{Z_2^T R^{-1} Z_1} & \boldsymbol{Z_2^T R^{-1} Z_2 + G_2^{-1}} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{X^T R^{-1} y} \\ \boldsymbol{Z_1^T R^{-1} y} \\ \boldsymbol{Z_2^T R^{-1} y} \end{bmatrix},$$

where $\boldsymbol{R} = \boldsymbol{D}\hat{\sigma}_e^2$, $\boldsymbol{G_1} = \boldsymbol{I}\frac{\hat{\sigma}_a^2}{N_{snp}}$ and $\boldsymbol{G_2} = \boldsymbol{A}\hat{\sigma}_{a*}^2$. Consequently, the variance of y is then given by $\boldsymbol{Z_1 G_1 Z_1^T + Z_2 G_2 Z_2^T + R}$. The variance components of model (2) were not estimated and taken as known parameters from routine evaluation in Poland.

Model (1) is a component of the Polish routine genomic evaluation system of programs custom written using the SAS package 9.3 version and FORTRAN,. Rrun under the Suse Linux Bourn shell environment.

The effects of particular SNPs (g_i) were tested for significance, i.e. H_0: $g_i \neq 0$ vs. H_1: $g_i \neq 0$ using the Wald test:

$$W = \frac{\hat{g}_i}{SE(\hat{g}_i)}$$

where $\hat{g}_i$ is the estimate of SNP i and SE($\hat{g}_i$) is the standard error of effect $\hat{g}_i$. Because of the standard errors of individuals SNPs are not available whenand in calculating the Wald test for each SNP the same standard error was assumed. The null distribution of the W statistics is standard normal distribution. Because very often random effects are assumed to be normally distributed it is common practice to use W statistics to test significance of random SNP effects [Suchocki *et al.* 2020, Kosińska-Selbi *et al.* 2020].

Interbull validation test

In the validation we used two data sets: full and truncated. The full data set consisted of all available bulls with daughter information, whileand the truncated data set consisted only of bulls born before 2010. The genomic evaluation was validated by comparing the GEBVs of bulls born after 2010 to their DRPs obtained from the full data set [Mantysaari *et al.* 2010]. The bias of the genomic evaluation was estimated using the following weighted linear regression model:

$$y = \beta_0 + \beta_1 \cdot GEBV_r + e \quad (3)$$

where y represents the vector of deregressed breeding values for bulls, which have

effective daughter contributions higher than 20 in the full data and EDC equal to 0 in the truncated data, $GEBV_r$ is the vector of *GEBV* obtained for the truncated data set. The weights used in the covariance of the residual vector in Model (3) were expressed by:

$$w_i = \frac{EDC_i}{EDC_i + k}, \tag{4}$$

where EDC_i is the effective daughter contribution of bull i in the full data set and $k = \frac{4-h^2}{h^2}$. h^2 represents the heritability of the trait. The quality of DRP prediction based on GEBV defined by model (3) was compared to the quality of prediction based on the parental information (PI) expressed by the following model:

$$y = \beta_0 + \beta_1 \cdot PI_r + e \tag{5}$$

where PI_r is the pedigree index for the truncated data set. The validation test is passed by fulfilling the following conditions: (i) hypothesis H_0:β_1=$E(\beta_1)$, tested based on the *t* statistics, $t = \frac{|\hat{\beta}_1 - E(\beta_1)|}{SE(\beta_1)}$ is accepted aton the 5% significance level; note that β_1 and $SE(\beta_1)$ are estimated with model (3) and in the case of our data, where all the validation bulls are genotyped $E(\beta_1) = 1$, (ii) the coefficient of determination R^2 from model (3) is higher than R^2 from the model (5).

Comparison of different data sets

In view of the lack of independence of the data sets calculated based on ORIG and RARE for comparison we used the Spearman rank correlation coefficient.

Results and discussion

We observed differences in minimum MAF between the ORIG and the RARE data sets. For ORIG the minimum MAF was 0.011, while for RARE it was 0.002. We observed no differences between the mean, and maximum MAF between those data sets.

There are two most computationally burdened components of the genomic evaluation system: (i) estimation of the SNP effects, and (ii) preparation of pedigree and genomic data files. Using markers without preselection based on MAF in genomic selection increased the average computational time for the estimation of SNP effects by 14.9% on average, ranging from 6.8% for SCS to 28.0% for CC1. However, for the second most computationally burdened component of genomic selection, i.e. preparation of pedigree and genomic data files, there was no significant difference in computational time. The computational time of the analyses is presented in Figure 1. The estimators of SNP effects were consistent between evaluations based on common SNP present in the ORIG and RARE data sets. The Spearman rank correlation coefficients between SNP effects common to both data sets were at least 0.999 for all the considered traits. Additionally, no rare SNP effect was statistically significant based on the Wald test. Despite such a high correlation of SNP effects between the two data sets, we observed changes in the ranking of the top 100 candidate bulls

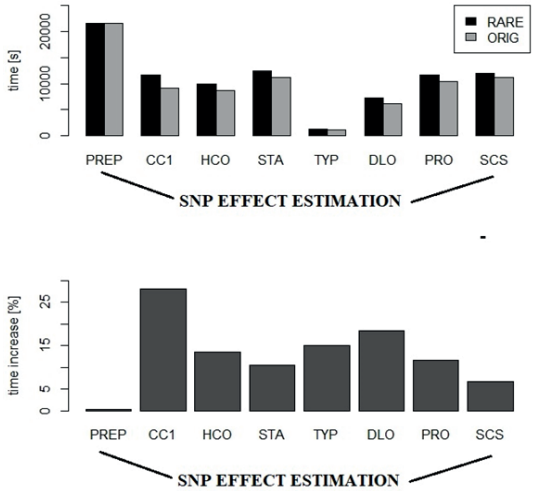

Fig. 1. Time in seconds [s] and time increase [%] for using the RARE data set in two main time-consuming elements of the genomic evaluation system. PREP = preparation of data sets for all traits, animals and SNPs. SNP ESTIMATION = SNP effect estimations.

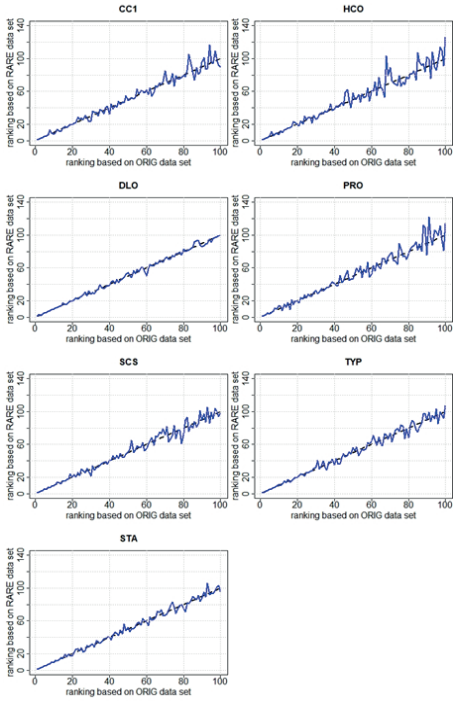

Fig. 2. Re-ranking of 100 top candidate bulls based on the RARE data set as compared to the ORIG data set.

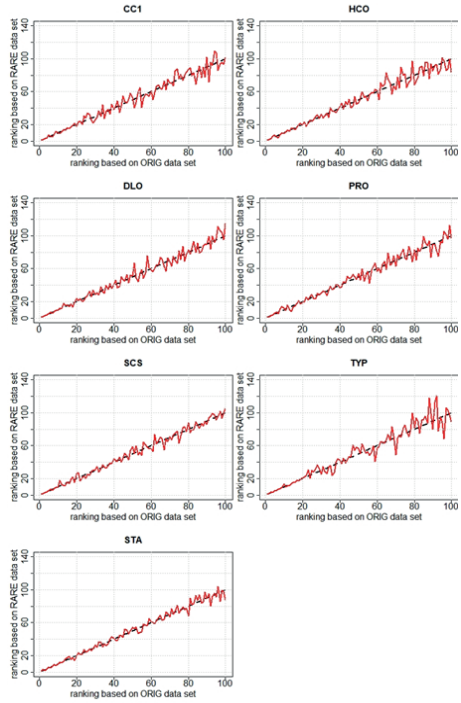

Fig. 3. Re-ranking of 100 top candidate cows based on the RARE data set as compared to the ORIG data set.

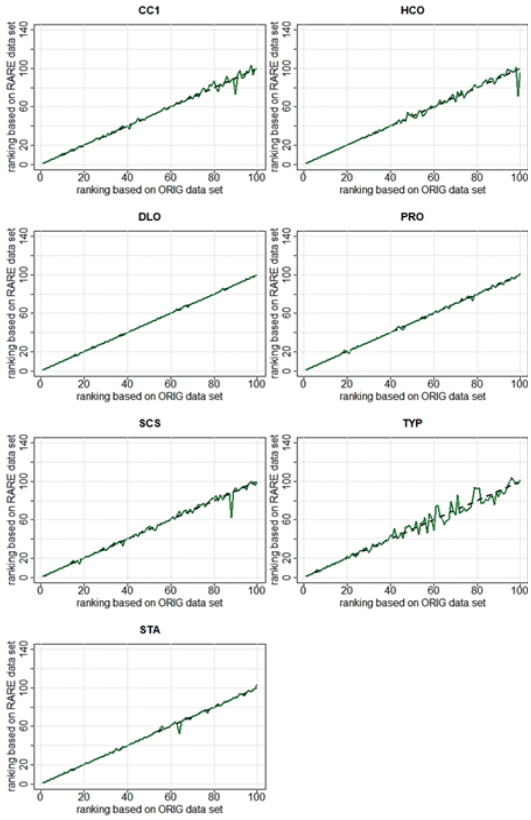

Fig. 4. Re-ranking of 100 top reference bulls based on the RARE data set as compared to the ORIG data set.

and cows born after 2010 as well as the top 100 reference bulls born after 2000. The greatest drop in the ranking of the top 100 candidate bulls was 35 for a bull evaluated for HCO, while the greatest increase in the ranking was 18 for bulls evaluated for both PRO and SCS. Moreover, in the evaluation for HCO and PRO seven bulls that were outside the top 100 ranking based on the ORIG data set were found in the top 100 ranking in the evaluation based on the RARE data set. For cows the maximum drop/increase in the top 100 ranking was found for TYP (28/30). Note that the SNP effects for this trait were estimated based on the national data, thus the ranking was the least stable. For the reference bulls the ranking rearrangements were lower. The maximum drop in the top 100 ranking was 15 for TYP, while among the traits evaluated based on the EuroGenomics reference population it ranged between 2 (DLO and PRO) and 7 (CC1). The maximum increase in the ranking was 28 for HCO. Also, the number of reference bulls that dropped out of the top 100 group defined by the ORIG evaluation was very low, between 0 (SCS) and 2 (TYP). Detailed information regarding changes

in ranking for candidate and reference animals is presented in Figures 2-4. Less re-ranking among the top ranked reference individuals proves that the genomic evaluation is stable and reliable.

For the results of the Interbull validation test there were no considerable differences between the RARE and ORIG data sets. Regardless of the data set used, all the traits evaluated based on the EuroGenomics reference population (CC1, HCO, DLO, PRO, SCS, STA) passed the validation test and regardless of the data set the estimated intercepts were very close to the expected ones, with the greatest difference of 0.190 observed for DLO with the RARE data set. Conversely, the trait evaluated based on the national reference population (TYP) failed the test regardless of the data set. The quality of EBV prediction expressed by the coefficient of determination (R2) varied markedly between the traits, ranging from 61.2% for STA with ORIG data to 10.2% for HCO with RARE data. However, it was not influenced by the inclusion of rare variants, since differences in R2 between predictions based on the ORIG and RARE data sets were always less than 1%. The results of the Interbull validation test are summarised in Table 2.

Table 2. A comparison of summary statistics of the Interbull genetic trend validation test based on the ORIG data set (upper line) and the RARE data set (bottom line)

Trait	$\hat{\beta}_1$	$SE(\hat{\beta}_1)$	$R^2_{model(3)}$	$R^2_{model(5)}$	$\lvert\hat{\beta}_1 - E(\beta_1)\rvert$	Result of Interbull test
Stature	1.065	0.012	61.2	13.1	0.065	passed
	1.089	0.012	60.6		0.089	passed
Type	0.724	0.059	12.2	1.9	0.276	did not pass
	0.730	0.060	11.8		0.270	did not pass
Protein yield	0.996	0.015	42.1	2.3	0.004	passed
	1.014	0.015	41.5		0.014	passed
Heifer conception rate	1.022	0.039	10.4	0.9	0.022	passed
	1.075	0.042	10.2		0.075	passed
Cow conception rate	1.081	0.027	20.4	2.4	0.081	passed
	1.131	0.028	20.1		0.131	passed
Somatic cell score	0.939	0.012	47.9	10.1	0.062	passed
	0.955	0.013	47.4		0.045	passed
Longevity	1.122	0.064	22.5	3.4	0.122	passed
	1.190	0.067	22.7		0.190	passed

The traits in the analysis were selected to represent a range of heritabilities (e.g. STA vs HCO) and the size of the reference population (e.g. PRO vs TYP). Although we did not have any a priori expectations, the selection was made to enable the detection of an eventual different impact of including rare variants depending on the trait. However, this was not the case.

The functional consequences of enrichment of rare variants had already been demonstrated by the 1000 Bulls Genome Project Consortium [Hayes and Daetwyler 2019]. For instance, rare variants (i.e. MAF<0.005) representing non-synonymous mutations amounted from 0.09% to 1.33% of all rare SNPs, while among common SNPs (i.e. MAF>0.05) non-synonymous mutations it was only between 0.06% and

0.09%. Therefore, in the presented study we investigated the influence of rare SNP variants on the genomic evaluation of Polish Holstein-Friesian cattle. For this purpose we used two data sets, one with rare variants and the other without them. None of the rare SNP effects was significant, although we noted changes in the ranking of the top 100 candidate bulls. The obvious drawback of using the 50K Illumina SNP chip to track rare variants is that commercial microarray platforms were designed to harbour common variations. Still both in our data and in other national Holstein-Friesian populations genotyped using the chip, one can well track what is called "low frequency variants" in human genetic application, i.e. variants with MAF ranging between 0.005 and 0.01. Even such polymorphisms show an excess of non-synonymous variants in human genomes (0.04%-0.76%) as compared to common SNPs.

The use of rare variants in genomic selection has some disadvantages. It lengthens computational time. The accuracy of genotyping rare SNPs is lower compared to SNPs with more balanced genotype counts, as is the accuracy of such SNP effect estimation. Moreover, including low frequency SNPs does not affect the outcome of the Interbull validation test, since the effects of rare SNPs are less accurately estimated, which undermines the advantage of having them in the EBV prediction model. However, on the basis of individual animals the use of rare SNP information can provide a more accurate ranking of selection candidates, which is due to the fact that the differentiation among top ranked individuals with high GEBVs can be made more accurate by including the extra information from additional SNPs. This has further implications for genomic evaluation based on whole-genome sequence data [Druet *et al*. 2014, O'Connell *et al*. 2016], although not in a rare SNP context, in which the amount of low frequency SNPs will be much greater than in our study. It is also worth noting that the use of rare variants may vary depending on the methodology used for SNP prioritizing, potentially yielding results with different degrees of accuracy. Such a comparison is presented between the BayesB, BayesC and Fst methods [Chang *et al.* 2018], indicating greater genomic and phenotypic accuracy of the latter (in most cases), providing a more appropriate tool for analyses that include rare variants. In genomic selection based on whole-genome sequences, rare variants could well have a stronger impact on selection, the Interbull validation process and evaluation reliability. Still it has to be kept in mind that for the purpose of predicting of genomic breeding values, defined as the cumulative additive genetic effect of all possible causal mutations, the addition of rare variants does not necessarily provide an improvement. In artificially bred populations remaining under a directional selection for many decades, such as dairy cattle, high linkage disequilibrium allows for the estimation of genomic effects accurately even with a moderate number of highly polymorphic markers (i.e. with moderate MAF) markers. The most important advantage of inclusion of rare SNPs lies in the fact that they facilitate a more accurate assessment of effects of all SNPs from the genomic evaluation model, since the higher SNP density more precisely fits the polygenic nature of additive genetic variation. We can further hypothesize that this fact has implications for the greater accuracy of direct genomic values in individuals with rare alleles.

REFERENCES

1. BENJAMINI Y., HOHBERG Y., 1995 – Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society, Series B (Statistical Methodology), 57, 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x.
2. BLOOM J.S., BOOCOCK J., TREUSCH S., SADHU M.J., DAY L., OATES-BARKER H., KRUGLYAK L., 2019 – Rare variants contribute disproportionately to quantitative trait variation in yeast. eLife, 2019, 8, e49212. doi:10.7554/eLife.49212.
3. BOMBA L., WALTER K., SORANZO N., 2017 – The impact of rare and low-frequency genetic variants in common disease. ***Genome Biology*** 18, 77. doi:10.1186/s13059-017-1212-4.
4. BROWNING B.L., BROWNING S.R., 2009 – A unified approach to genotype imputation and haplotype phase inference for large data sets of trios and unrelated individuals. ***American Journal of Human Genetics*** 84, 210-223. doi:10.1016/j.ajhg.2009.01.005.
5. CHANG L., TOGHIANI S., LING A., AGGREY S.E., REKAYA R., 2018 – High density marker panels, SNPs prioritizing and accuracy of genomic selection. ***BMC Genetics*** 19, 4 doi.org/10.1186/s12863-017-0595-2.
6. DUNNETT C.W., 1955 – A multiple comparisons procedure for comparing several treatments with a control. ***Journal of the American Statistical Association*** 50, 1096-1121. doi:10.1080/01621459.1955.10501294.
7. DRUET T., MACLEOD I.M., HAYES B.J., 2014 – Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. Heredity, 112, 39-47. doi:10.1038/hdy.2013.13.
8. FRAZER K.A., MURRAY S.S., SCHORK N.J., TOPOL E.J., 2009 – Human genetic variation and its contribution to complex traits. ***Nature Reviews Genetics*** 10, 241251. doi:10.1038/nrg2554.
9. HAYES B.J., DAETWYLER H.D., 2019 – 1000 Bull Genomes project to map simple and complex genetic traits in cattle: Applications and outcomes. ***Annual Review of Animal Biosciences*** 7, 89-102. doi.org/10.1146/annurev-animal-020518-115024.
10. HENDERSON C.R., 1984 – Applications of Linear Models in Animal Breeding. Guelph University of Guelph.
11. LI H., DURBIN R., 2009 - Fast and accurate short read alignment with Burrows Wheeler Transform. ***Bioinformatics*** 25, 1754-1760. doi: 10.1093/bioinformatics/btp324.
12. LI H., HANDSAKER B., WYSOKER A., FENNELL T., RUAN J., HOMER N., MARTH G., ABECASIS G., DURBIN R., 1000 Genome Project Data Processing Subgroup, 2009 – The Sequence alignment/map (SAM) format and SAMtools. ***Bioinformatics*** 25, 2078–2079. doi: 10.1093/bioinformatics/btp352.
13. LIU Z., GODDARD M.E., HAYES B.J., REINHARDT F., REENTS R., 2016 – Technical note: Equivalent genomic models with a residual polygenic effect. ***Journal of Dairy Science*** 99, 2016-2025. doi:10.3168/jds.2015-10394.
14. MCKENNA A., HANNA M., BANKS E., SIVACHENKO A., CIBULSKIS K., KERNYTSKY A., GARIMELLA K., ALTSHULER D., GABRIEL S., DALY M., DEPRISTO M.A., 2010 – The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. ***Genome Research*** 20, 1297-1303.
15. KOSIŃSKA-SELBI B., SUCHOCKI T., EGGER-DANNER CH., SCHWARZENBACHER H., FRĄSZCZAK M., SZYDA J., 2020 – Exploring the potential genetic heterogeneity in the incidence of hoof disorders in Austrian Fleckvieh and Braunvieh cattle. ***Frontiers in Genetics*** 11, 577116. doi: 10.3389/fgene.2020.577116.
16. MANOLIO T.A., COLLINS F.S., COX N.J., GOLDSTEIN D.B., HINDORFF L.A., HUNTER D.J., MCCARTHY M.I., RAMOS E.M., CARDON L.R., CHAKRAVARTI A., CHO J.H., GUTTMACHER A.E., KONG A., KRUGLYAK L., MARDIS E., ROTIMI C.N., SLATKIN M.,

VALLE D., WHITTEMORE A.S., BOEHNKE M., CLARK A.G., EICHLER E.E., GIBSON G., HAINES J.L., MACKAY T.F.C., MCCARROLL S.A., VISSCHER P.M., 2009 – Finding the missing heritability of complex diseases. ***Nature*** 461, 747-753. doi:10.1038/nature08494.

17. MANTYSAARI E., LIU Z., VANRADEN P., 2010 – Interbull validation test for genomic evaluation. Interbull Bulletin, 41, Paris France, March 4-5 2010.
18. O'CONNELL J.R., TOOKER M.E., BICKHART D.M., VANRADEN P.M., 2016 – Selection of Sequence Variants to Improve Genomic Predictions. Interbull Bulletin, 50, 58-66.
19. SCHAID D.J., CHEN W., LARSON N.B., 2018 – From genome-wide associations to candidate causal variants by statistical fine-mapping. ***Nature Reviews Genetics*** 19, 491-504, doi:10.1038/s41576-018-0016-z.
20. STYRKARSDOTTIR U., THORLEIFSSON G., SULEM P., GUDBJARTSSON D.F., SIGURDSSON A., JONASDOTTIR A., ODDSSON A., HELGASON A., MAGNUSSON O.T., WALTERS G.B., FRIGGE M.L., HELGADOTTIR H.T., JOHANNSDOTTIR H., BERGSTEINSDOTTIR K., OGMUNDSDOTTIR M.H., CENTER J.R., NGUYEN T.V., EISMAN J.A., CHRISTIANSEN C., STEINGRIMSSON E., JONASSON J.G., TRYGGVADOTTIR L., EYJOLFSSON G.I., THEODORS A., JONSSON T., INGVARSSON T., OLAFSSON I., RAFNAR T., KONG A., SIGURDSSON G., MASSON G., THORSTEINSDOTTIR U., STEFANSSON K., 2013 – Nonsense mutation in the LGR4 gene is associated with several human diseases and other traits. ***Nature*** 497, 517–520. doi:10.1038/nature12124.
21. SUCHOCKI T., EGGER-DANNER CH., SCHWARZENBACHER H., SZYDA J., 2020 – Two-stage GWAS for the identification of causal variants underlying hoof disorders in cattle. ***Journal of Dairy Science*** 103, 4483-4494. doi.org/10.3168/jds.2019-17542.
22. SULEM P., GUDBJARTSSON D.F., WALTERS G.B., HELGADOTTIR H.T., HELGASON A., GUDJONSSON S.A., ZANON C., BESENBACHER S., BJORNSDOTTIR G., MAGNUSSON O.T., MAGNUSSON G., HJARTARSON E., SAEMUNDSDOTTIR J., GYLFASON A., JONASDOTTIR A., HOLM H., KARASON A., RAFNAR T., STEFANSSON H., ANDREASSEN O.A., PEDERSEN J.H., PACK A.I., DE VISSER M.C.H., KIEMENEY L.A., GEIRSSON A.J., EYJOLFSSON G.I., OLAFSSON I., KONG A., MASSON G., JONSSON H., THORSTEINSDOTTIR U., JONSDOTTIR I., STEFANSSON K., 2011 – Identification of low-frequency variants associated with gout and serum uric acid levels. ***Nature Genetics*** 43, 1127–1130. doi: 10.1038/ng.972.
23. 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. dx.doi.org/10.1007/s13353-011-0047-z.