 doi:10.20944/preprints202202.0164.v1

*Article*

# The Impact of Rare Single Nucleotide Polymorphism Variants on the Genomic Evaluation of Dairy Cattle

**Tomasz Suchocki [1,2]*, Michalina Jakimowicz[1], Andrzej Zarnecki [2], Arkadiusz Dziech[3] and 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
* Correspondence: tomasz.suchocki@upwr.edu.pl; Tel.: +48-71-3205758

**Simple Summary:** Our research focused on incorporating rare genetic variants into model used in routine genomic selection in dairy cattle in Poland. While most scientists have based their research on common genetic variants, our results indicate that rare variants could increase the accuracy of rankings for bulls and cows. More stable rankings could increase the success of selection and improve the economics of milk production. A better understanding of the contribution of rare variants to the estimation of GEBVs and bull ranking is also important for the future utilization of the whole genome DNA sequence information in genomic selection, in which an even higher number of rare variants is expected.

**Abstract:** The experiments described in this research article were designed to test the effect of rare variants into genomic prediction in dairy cattle. Common polymorphisms are able to explain 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 (humans) or artificial (livestock) selection pressure. Therefore, it is important to check whether the use of rare variants could increase the accuracy of ranking of animals by providing the tool for more precise differentiation among the bulls with high additive genetic merit. The goal of our study was to verify whether including rare variants in a genomic selection model allows for a more accurate description of the additive genetic background of traits under selection in dairy cattle. We used the linear mixed model for comparison SNP estimates for Holstein-Friesian cattle of the two data sets – a set containing only single nucleotide polymorphisms defined by minor allele frequency ≥ 0.01, 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 DGV and GEBV and compared them between both data sets. In all the analyses we used production, fertility, conformation and udder health traits. We also assessed the time required for the two most computationally demanding components of genomic selection: preparing 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, but had no effect on the outcome of the quality of EBV prediction as expressed by the Interbull validation test.

**Keywords:** rare variants, genome-wide association study, validation test, SNP chip , genomic selection

## 1. Introduction

 doi:10.20944/preprints202202.0164.v1

Predicting phenotypes from genotype data is important for plant and animal breeding, as well as for 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 the result that 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, that is, polymorphisms with low, typically below 1%, minor allele frequency, have been brought into focus in the context of genetic determination of complex traits [1, 2]. The main reason for this is the phenomenon of so-called "missing heritability" indicated for most of the complex phenotypes measured in humans, which denotes that common polymorphisms are able to explain only a small proportion, ranging between 1.5% and 50%, of the underlying genetic variation of such traits [3]. Variants representing functional mutations with large effects on complex phenotypes are expected to be rare because of natural (humans) or artificial (livestock) selection pressure against an unfavourable allele [4]. The biological explanation is that since a mutation is functional, it is subjected to selection, which as a consequence affects population allele frequency stronger than in the case of a neutral mutation [5]. 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 [6, 7]. Also, in yeasts (Saccharomyces cerevisiae) the importance of rare variants in phenotypes of quantitative traits was higher than might be 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 (QTL) rich in rare variants had larger substitution effects, while the one with an abundance of common variants was less influential [8]. Dairy cattle is an ideal population to verify this hypothesis. It has undergone directional selection for production traits over many generations and has very good records of complex traits and familial relationship. 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 allows for a more accurate description of the additive genetic background of traits under selection in dairy cattle. The analysis involved comparisons of two data sets – a set containing only SNPs defined by MAF ≥ 0.01 and call rate higher than 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 February 2022 doi:10.20944/preprints202202.0164.v1

## 2. Materials and Methods

The analyzed data set originated from the EuroGenomics Cooperative U.A. Holstein-Friesian dairy cattle population. For each bull born before 2010, pseudophenotypes in the form of deregressed breeding values (DRP) corresponding to the Interbull evaluation from April 2020 were available (based on pedigree model). 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 traits except 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.

**Table 1.** Summary of analyzed sub-sets of individuals and traits' 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% |

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 February 2022 doi:10.20944/preprints202202.0164.v1

| | | | | | |
|---|---|---|---|---|---|
| Longevity | 21,795 | 25,998 | 43,392 | 0.173 | 40% |
| Somatic cell score | 34,168 | 23,090 | 43,392 | 0.320 | 20% |

Most of the reference (87%) individuals were genotyped using Illumina BovineSNP50 BeadChip Version 2. All individuals genotyped with other platforms were imputed to the above microarray using Beagle software [9] Almost all imputed animals were genotyped using EuroG10K BeadChip v2-5. In the final analysis, two data sets of SNP genotypes were used: (i) ORIG consisted of 46,216 SNPs representing the standard common SNP set used for the routine genomic evaluation in Poland, and (ii) RARE consisted of 54,378 SNPs without preselection on MAF, including common and rare polymorphisms. The SNP selection criterion for ORIG comprised MAF of at least 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.

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 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 bulls in validation data set were originally genotyped using Illumina BovineSNP50 BeadChip Version 2, while 93% of cows using EuroG10K BeadChip v2-5.

The following mixed model [10] was used to estimate the additive effects of SNPs $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z_1 g} + \boldsymbol{Z_2 a} + \boldsymbol{\varepsilon}$ (**1**), where $\boldsymbol{y}$ represents the vector of deregressed breeding values of the reference bulls (for all traits except TYP we used deregressed MACE EBVs calculated for the Polish scale and for TYP we used deregressed national EBVs), $\boldsymbol{\beta}$ is the general mean, $\boldsymbol{X}$ is vector of ones, $\boldsymbol{Z_1}$ is the design matrix for SNP genotypes, which is parameterized as -1, 0, or 1 for a homozygous, heterozygous and an alternative homozygous SNP genotype, respectively, $\boldsymbol{g}$ is the vector of random additive SNP effects, $\boldsymbol{Z_2}$ is the design matrix for a polygenic effect, $\boldsymbol{a}$ 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 analyzed phenotypes [11]. $\boldsymbol{\varepsilon}$ is the vector of error terms with $\boldsymbol{\varepsilon} \sim \boldsymbol{N}(\mathbf{0}, \boldsymbol{D}\hat{\boldsymbol{\sigma}}_{\boldsymbol{e}}^{\mathbf{2}})$, where $\boldsymbol{D}$ is the diagonal matrix containing the reciprocal of bulls' effective daughter contributions (EDC; for all traits except TYP we used EDC from MACE evaluation calculated on the Polish scale, for TYP we used EDC from Poland) on the diagonal and $\hat{\boldsymbol{\sigma}}_{\boldsymbol{\varepsilon}}^{\mathbf{2}}$ representing the error variance. The covariance structure of $\boldsymbol{g}$ was assumed to be $\boldsymbol{g} \sim \boldsymbol{N}\left(\mathbf{0}, \boldsymbol{I}\frac{\hat{\sigma}_{\mathbf{a}}^{\mathbf{2}}}{\mathbf{N_{snp}}}\right)$, with I being the identity matrix, $\hat{\boldsymbol{\sigma}}_{\mathbf{a}}^{\mathbf{2}}$ representing the additive genetic variance of a given trait and $N_{SNP}$ being the number of SNPs used (here 46,216 for ORIG and 54,495 for RARE data set) assigning the same small fraction of the polygenic variance to each of the $N_{SNP}$ polymorphisms. $\boldsymbol{a} \sim \boldsymbol{N}(\mathbf{0}, \boldsymbol{A}\hat{\boldsymbol{\sigma}}_{\mathbf{a}*}^{\mathbf{2}})$, where $\boldsymbol{A}$ is the numerator relationship matrix and $\hat{\boldsymbol{\sigma}}_{\mathbf{a}*}^{\mathbf{2}}$ 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 of the above model was based on solving the mixed model equations [12] $\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{g}} \\ \hat{\boldsymbol{a}} \end{bmatrix} =$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 February 2022 doi:10.20944/preprints202202.0164.v1

$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z_1 & X^T R^{-1} Z_2 \\ Z_1^T R^{-1} X & Z_1^T R^{-1} Z_1 + G_1^{-1} & Z_1^T R^{-1} Z_2 \\ Z_2^T R^{-1} X & Z_2^T R^{-1} Z_1 & Z_2^T R^{-1} Z_2 + G_2^{-1} \end{bmatrix}^{-1} \begin{bmatrix} X^T R^{-1} y \\ Z_1^T R^{-1} y \\ Z_2^T R^{-1} y \end{bmatrix}$ (2), where $R = D\hat{\sigma}_e^2$, $G_1 = I\frac{\hat{\sigma}_a^2}{N_{snp}}$ and $G_2 = A\hat{\sigma}_{a*}^2$. Consequently, the variance of $y$ is then given by $Z_1 G_1 Z_1^{\mathrm{T}} + Z_2 G_2 Z_2^{\mathrm{T}} + R$. 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. Run under the Suse Linux Bourn shell environment.

The effects of particular SNPs ($g_i$) were tested for significance i.e. $H_0: g_i = 0$ 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 and 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 effect are assumed to be normally distributed it is common to use *W* statistics to test significance of random SNP effects [13, 14].

In the validation we used two data sets: full and truncated. The full data set consisted of all available bulls with daughter information, and 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 [15]. The bias of the genomic evaluation was estimated using the following weighted linear regression model: $y = \beta_0 + \beta_1 \cdot GEBV_r + e$ (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 truncated data set. The weights used in covariance of residual vector in Model (3) were expressed by: $w_i = \frac{EDC_i}{EDC_i + k}$ (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: $y = \beta_0 + \beta_1 \cdot PI_r + e$ (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 $H_0: \beta_1 = E(\beta_1)$, tested based on the $t$ statistics $t = \frac{|\hat{\beta}_1 - E(\beta_1)|}{SE(\hat{\beta}_1)}$, is accepted on the 5% significance level; note that $\hat{\beta}_1$ and $SE(\hat{\beta}_1)$ are estimated with model (3) and in the case of our data where all validation bulls are genotyped $E(\beta_1) = 1$, (ii) $R^2$ from model (3) is higher than $R^2$ from the model (5).

**3. Results**

We observed the differences in minimum MAF between the ORIG and the RARE data sets. For ORIG the minimum MAF was equal to 0.011, while for RARE 0.002. We didn't observe differences between 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 of the estimation of SNP effects by 14.9% on average, varying from 6.8% for SCS to 28.0% for CC1. However, for the second most computationally burdened component of genomic selection, that is, preparation of pedigree

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 February 2022 doi:10.20944/preprints202202.0164.v1

and genomic data files, there was no significant difference in computational time. The computational time of the analyses is presented in Figure 1.

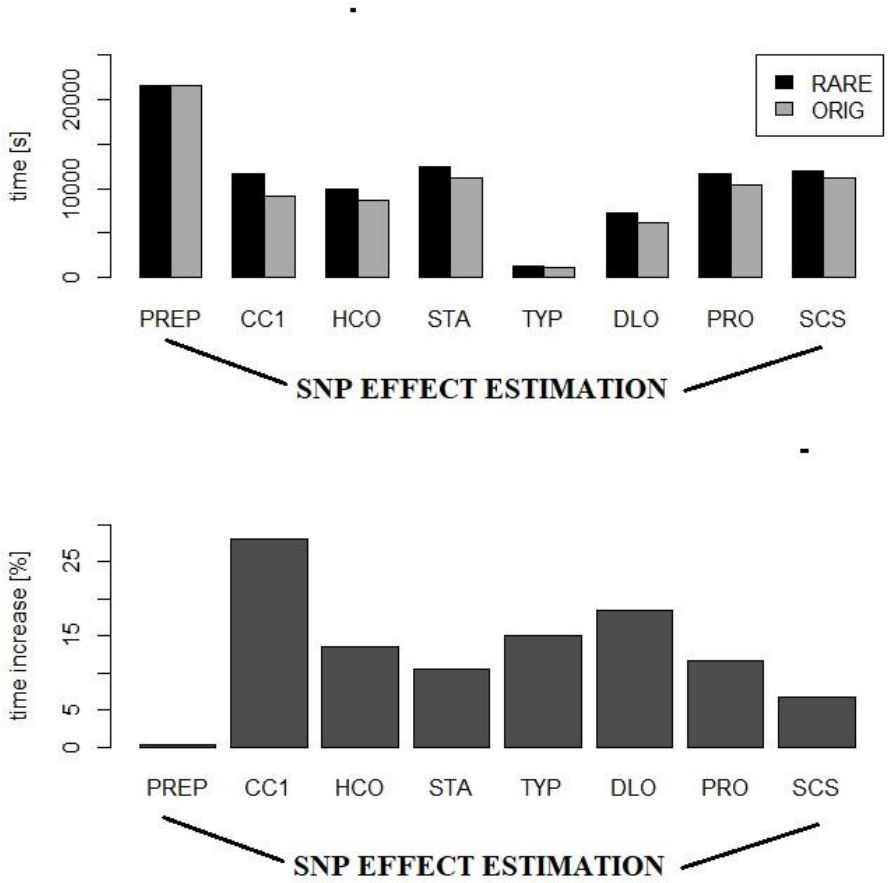


**Figure 1.** Time in seconds [s], and time increase [%] for using the RARE data set in two main time-consuming elements of the genomic evaluation system. PREP = preparation of data sets for all traits, animals and SNPs. SNP ESTIMATION = SNP effect estimations.

The estimators of SNP effects were consistent between evaluations based on common SNP present in the ORIG and RARE data sets. The Pearson correlation coefficients between SNP effects common to both data sets were at least 0.999 for all 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 and cows born after 2010 and the top 100 reference bulls born after 2000. The biggest drop in the top 100 ranking for candidate bulls was 35 for a bull evaluated for HCO, and the biggest 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 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, and 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 traits evaluated based on the EuroGenomics reference population it varied 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,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 February 2022 doi:10.20944/preprints202202.0164.v1

varying 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.

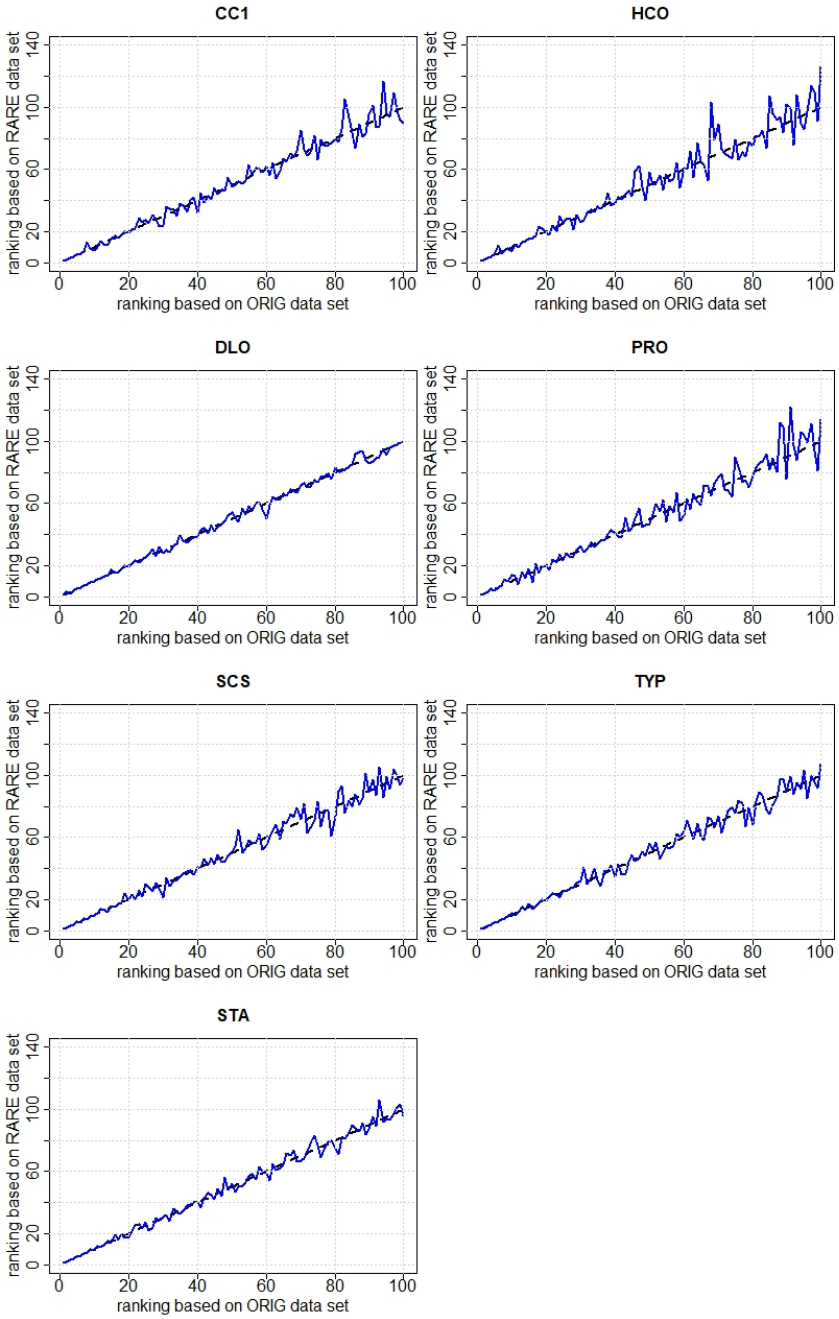


**Figure 2.** Re-ranking of the 100 top candidate bulls based on the RARE data set as compared to the ORIG data set.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 February 2022 doi:10.20944/preprints202202.0164.v1

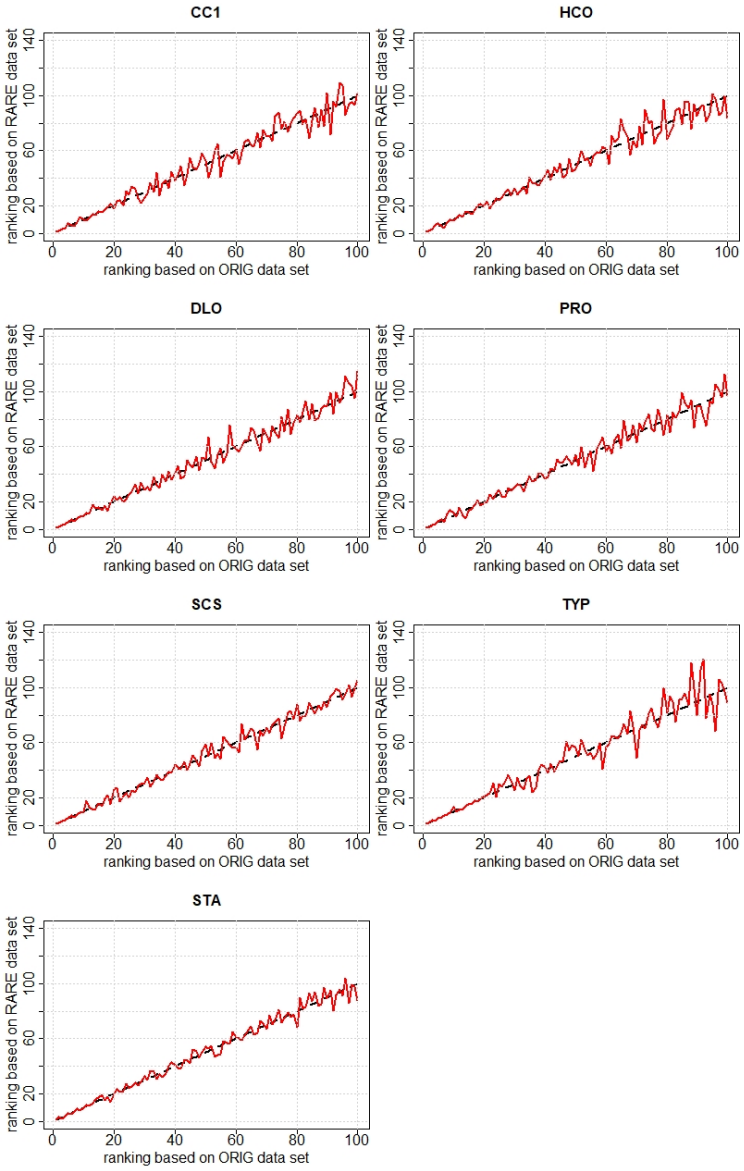


**Figure 3.** Re-ranking of the 100 top candidate cows based on the RARE data set as compared to the ORIG data set.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 February 2022 doi:10.20944/preprints202202.0164.v1

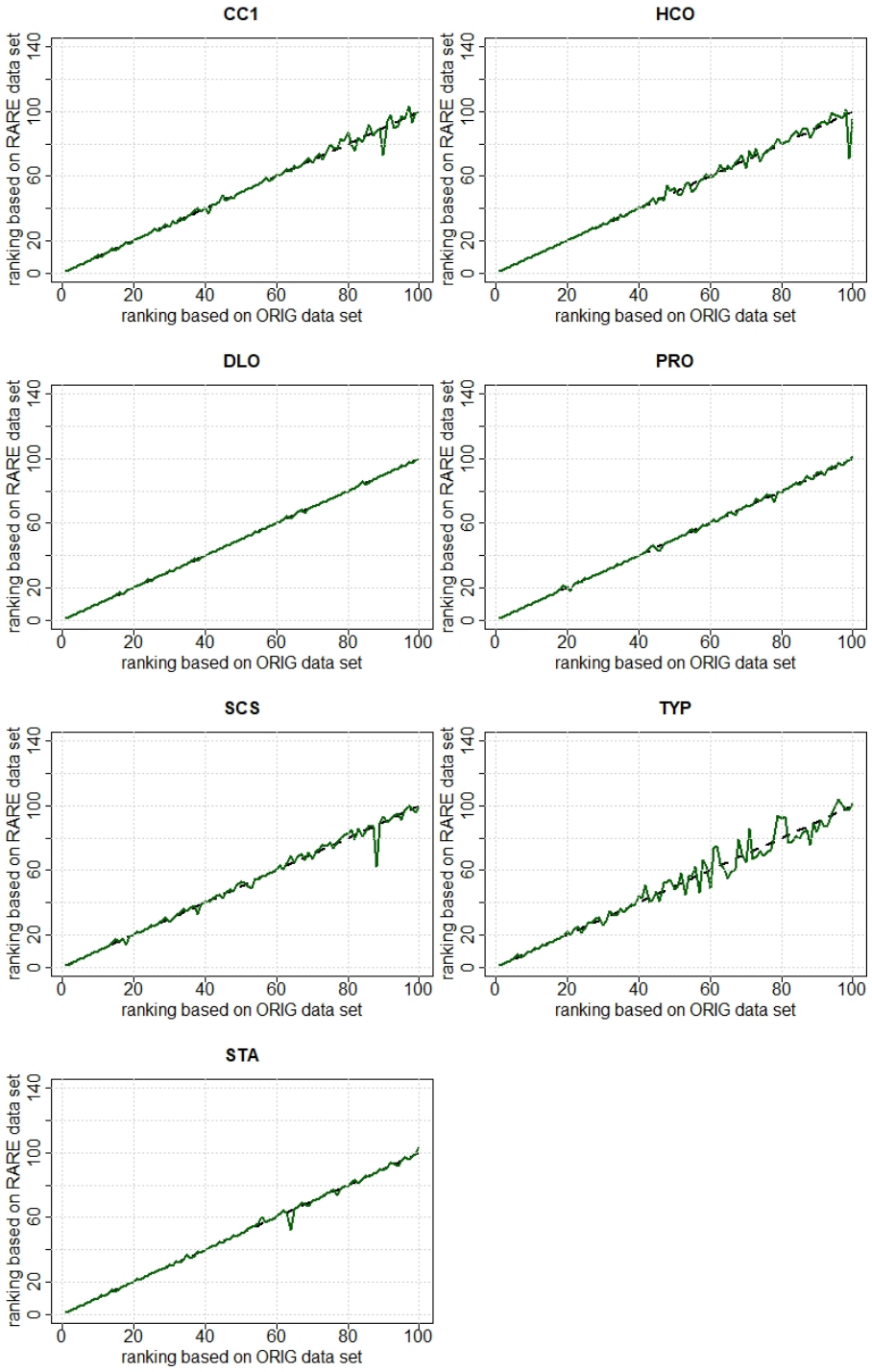


**Figure 4.** Re-ranking of the 100 top reference bulls based on the RARE data set as compared to the ORIG data set.

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 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 highest 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 ($R^2$) varied markedly among traits, between 61.2 % for STA with ORIG data and 10.2% for HCO with RARE data, but was not

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 February 2022 doi:10.20944/preprints202202.0164.v1

influenced by the inclusion of rare variants since differences in $R^2$ between predictions based on the ORIG and RARE data sets were always less than 1%. The results of the Interbull validation test were summarised in Table 2 .

**Table 2.** The comparison of the 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 |

## 4. Discussion

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 eventual different impact of including rare variants, depending on trait. This was however not the case.

 doi:10.20944/preprints202202.0164.v1

The functional consequences of enrichment of rare variants has already been demonstrated by the 1000 Bulls Genome Project Consortium [4]. For instance, rare variants (i.e. MAF<0.005) representing non-synonymous mutations amounted between 0.09% and 1.33% of all rare SNPs, while among common SNPs (i.e. MAF>0.05) non-synonymous mutations 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. To achieve that, we used two data sets, one with rare variants and the second one without them. None of the rare SNP effect was significant, but we noted changes in the ranking of the top 100 candidate bulls. The obvious drawback of using the 50K Illumina SNP chip for tracking rare variants is that commercial microarray platforms were designed to harbour common variations. Still in our data, as well as in other national Holstein-Friesian populations genotyped by the chip, one can well track what is called "low frequency variants" in human genetic application, that is, 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 effect of rare SNPs are less accurately estimated what overshadows 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 [16, 17], although not in a rare SNP context, in which the amount of low frequency SNPs will be much higher 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, which might yield results with different degrees of accuracy. Such a comparison is presented between the BayesB, BayesC and Fst methods [18], indicating higher 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 which are 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 (i.e. with moderate MAF) markers.

## 5. Conclusions

The most important advantage of inclusion of rare SNPs lies in the fact that they allow for 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 on the higher accuracy of direct genomic values of individuals with rare alleles.

**Author Contributions:** Conceptualization, Tomasz Suchocki and Joanna Szyda; Data curation, Andrzej Zarnecki; Formal analysis, Tomasz Suchocki, Michalina Jakimowicz, Andrzej Zarnecki, Arkadiusz Dziech and Joanna Szyda; Investigation, Andrzej Zarnecki; Methodology, Tomasz Suchocki and Joanna Szyda; Resources, Andrzej Zarnecki; Visualization, Michalina Jakimowicz and Arkadiusz Dziech; Writing – original draft, Tomasz Suchocki, Michalina Jakimowicz and Arkadiusz Dziech; Writing – review & editing, Tomasz Suchocki, Michalina Jakimowicz and Joanna Szyda.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable

**Informed Consent Statement:** Not applicable

**Data Availability Statement:** The datasets analysed for this study are available upon a formal request to the EuroGenomics 247 cooperative and for TYP – upon a formal request to the Polish Federation of Cattle Breeders and Dairy 248 Farmers.

**Acknowledgments:** Calculations have been carried out using resources provided by Wroclaw Centre for Networking and Supercomputing (http://wcss.pl), grant No. 509.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Bomba, L.; Walter, K.; Soranzo, N. The Impact of Rare and Low-Frequency Genetic Variants in Common Disease. *Genome Biol* **2017**, *18*, 77, doi:10.1186/s13059-017-1212-4.
2. Schaid, D.J.; Chen, W.; Larson, N.B. From Genome-Wide Associations to Candidate Causal Variants by Statistical Fine-Mapping. *Nat Rev Genet* **2018**, *19*, 491–504, doi:10.1038/s41576-018-0016-z.
3. 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.; et al. Finding the Missing Heritability of Complex Diseases. *Nature* **2009**, *461*, 747–753, doi:10.1038/nature08494.
4. Hayes, B.J.; Daetwyler, H.D. 1000 Bull Genomes Project to Map Simple and Complex Genetic Traits in Cattle: Applications and Outcomes. *Annu. Rev. Anim. Biosci.* **2019**, *7*, 89–102, doi:10.1146/annurev-animal-020518-115024.
5. Frazer, K.A.; Murray, S.S.; Schork, N.J.; Topol, E.J. Human Genetic Variation and Its Contribution to Complex Traits. *Nat Rev Genet* **2009**, *10*, 241–251, doi:10.1038/nrg2554.
6. 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.; et al. Identification of Low-Frequency Variants Associated with Gout and Serum Uric Acid Levels. *Nat Genet* **2011**, *43*, 1127–1130, doi:10.1038/ng.972.
7. Styrkarsdottir, U.; Thorleifsson, G.; Sulem, P.; Gudbjartsson, D.F.; Sigurdsson, A.; Jonasdottir, A.; Jonasdottir, A.; Oddsson, A.; Helgason, A.; Magnusson, O.T.; et al. Nonsense Mutation in the LGR4 Gene Is Associated with Several Human Diseases and Other Traits. *Nature* **2013**, *497*, 517–520, doi:10.1038/nature12124.
8. Bloom, J.S.; Boocock, J.; Treusch, S.; Sadhu, M.J.; Day, L.; Oates-Barker, H.; Kruglyak, L. Rare Variants Contribute Disproportionately to Quantitative Trait Variation in Yeast. *eLife* **2019**, *8*, e49212, doi:10.7554/eLife.49212.
9. Browning, B.L.; Browning, S.R. A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. *The American Journal of Human Genetics* **2009**, *84*, 210–223, doi:10.1016/j.ajhg.2009.01.005.
10. Szyda, J.; Żarnecki, A.; Suchocki, T.; Kamiński, S. Fitting and Validating the Genomic Evaluation Model to Polish Holstein-Friesian Cattle. *J Appl Genetics* **2011**, *52*, 363–366, doi:10.1007/s13353-011-0047-z.
11. Liu, Z.; Goddard, M.E.; Hayes, B.J.; Reinhardt, F.; Reents, R. Technical Note: Equivalent Genomic Models with a Residual Polygenic Effect. *Journal of Dairy Science* **2016**, *99*, 2016–2025, doi:10.3168/jds.2015-10394.
12. Henderson, C.R. *Applications of Linear Models in Animal Breeding*; Univ. of Guelph: Guelph, 1984; ISBN 9780889550308.
13. Suchocki, T.; Egger-Danner, Ch.; Schwarzenbacher, H.; Szyda, J. Two-Stage Genome-Wide Association Study for the Identification of Causal Variants Underlying Hoof Disorders in Cattle. *Journal of Dairy Science* **2020**, *103*, 4483–4494, doi:10.3168/jds.2019-17542.
14. Kosińska-Selbi, B.; Suchocki, T.; Egger-Danner, C.; Schwarzenbacher, H.; Frąszczak, M.; Szyda, J. Exploring the Potential Genetic Heterogeneity in the Incidence of Hoof Disorders in Austrian Fleckvieh and Braunvieh Cattle. *Front. Genet.* **2020**, *11*, 577116, doi:10.3389/fgene.2020.577116.
15. Mäntysaari, E.A.; Liu, Z.; VanRaden, P. Interbull Validation Test for Genomic Evaluations. *Interbull Bulletin* **2010**, 17–17.
16. Druet, T.; Macleod, I.M.; Hayes, B.J. Toward Genomic Prediction from Whole-Genome Sequence Data: Impact of Sequencing Design on Genotype Imputation and Accuracy of Predictions. Heredity 2014, 112, 39–47, doi:10.1038/hdy.2013.13.
17. VanRaden, P.M.; Tooker, M.E.; O'Connell, J.R.; Cole, J.B.; Bickhart, D.M. Selecting Sequence Variants to Improve Genomic Predictions for Dairy Cattle. Genetics Selection Evolution 2017, 49, 32, doi:10.1186/s12711-017-0307-4.
18. Chang, L.-Y.; Toghiani, S.; Ling, A.; Aggrey, S.E.; Rekaya, R. High Density Marker Panels, SNPs Prioritizing and Accuracy of Genomic Selection. BMC Genet 2018, 19, 4, doi:10.1186/s12863-017-0595-2.