Proceedings, 10th World Congress of Genetics Applied to Livestock Production

Do rare variants contribute to the genomic prediction accuracy?

T. Suchocki[1], **A. Żarnecki**[2] **and J. Szyda**[1]
[1]Wrocław University of Environmental and Life Sciences, Wrocław, Poland,
[2] National Research Institute of Animal Production, Cracow-Balice, Poland.

ABSTRACT: The major goal of this study is identification of SNPs with rare allelic variants i.e. with minor allele frequency lower than 1%, in data set of bulls from Polish Holstein-Friesian breed, and comparison of accuracy of breeding value prediction for data sets with and without rare alleles. Data set consisted of 3,100 proven and 1,968 young bulls. Each bull was genotyped using 50K Illumina BeadChip. In our analysis production, fertility and udder health traits were considered. Using SNP and SNP-BLUP model two evaluations were carried out: (1) with all available SNPs, including rare variants (53,862 SNPs); (2) with common SNPs only, for which minor allele frequency exceeds 1% (46,267 SNPs). Finally, statistical significance of SNP estimates and reliability of predicted breeding values were compared between two data sets. Results showed that including rare variants into analysis increase the accuracy of DGV and GEBV.

Keywords: dairy cattle; rare variants; genetics

Introduction

Predicting phenotypes from genotype data is important for plant and animal breeding, and evolutionary biology. Genomic-based phenotype prediction has been applied using data from single-nucleotide polymorphism (SNP) genotyping platforms. Usually, a set of markers included in the final analysis is edited based on a minor allele frequency (MAF) and a call rate. Such filtering leads to the fact that additive effects of SNPs with rare genotypes are not considered in the analysis and impact of such markers on estimated breeding values is unknown. Recently, rare genetic variants, i.e. polymorphisms with low minor allele frequency have been brought into focus in the context of genetic determination of complex traits. 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 of the underlying genetic variation of such traits (Manolio et al. (2009)). Consequently, it is expected that those are rare variants which represent functional mutations, exhibit large effects on complex phenotypes and are thus responsible for a considerable proportion of the observed genetic variation. The biological explanation is that since a mutation is functional, it is subjected to selection, which as a consequence, affects the population allele frequency stronger than in the case of a neutral mutation (Frazer et al. (2009)). Dairy cattle poses an ideal population to verify this hypothesis. It has undergone a directional selection on production traits for many generations and it has very good records of complex traits and familial relationship. Moreover, recent success of genomic selection provided extensive information on genotypes of single nucleotide polymorphisms distributed all over the genome and available for many individuals.

The main goal of our study was to verify whether including rare variants into a genomic selection model allows for capturing a considerable part of missing heritability underlying traits under selection in dairy cattle, by comparing the accuracy of breeding value prediction using only common variants and a mixture of common and rare variants.

Materials and Methods

Animals. The core data set represented the status quo of the evaluation from April 2009 with traits represented by EBVs for three production traits: milk- (MY), protein- (PY) and fat- (FY) yields, an udder health trait represented by somatic cell score (SCS) and two fertility traits: non-return rate of heifers (HCO) and non return rate of cows (CC1). The trait values were deregressed using the method described by Jairath et al. (1998). The training part used for the estimation of additive effects of SNPs, consisted of 3,100 Polish Holstein-Friesian proven bulls with the average number of effective daughters (EDC) equal to 275 (11-14,403) for production traits, 338 (11-17,311) for udder health trait and 202 (11-8,765) for fertility traits. The validation part consisted of 1,968 young bulls without daughters.

Genotypes. SNP genotypes were detected by the use of the Illumina BovineSNP50 Genotyping BeadChip, which consists of 54,001 SNPs (version 1) or 54,609 SNPs (version 2). The original set of SNPs used for the estimation of Direct Genomic Values (DGV) consisted of 46,267 polymorphisms resulting from filtering based on minor allele frequency (MAF), with a minimum MAF of 0.01, and technical SNP quality expressed by the minimum call rate of 90%. The data set including rare variants was selected without SNP filtering on MAF and consisted of 53,862 polymorphisms.

DGV estimation. The following model was used to estimate additive effects of SNPs: $\mathbf{y} = \mathbf{X}\mathrm{b} + \mathbf{Zg} + \mathbf{e}$, where $\mathbf{y}$ represents a vector of deregressed EBVs, $\mathbf{X}$ is a design matrix for fixed effects, b is a vector of fixed effects, which in the current model comprise only a general mean, $\mathbf{Z}$ is a design matrix for SNP genotypes, which is parameterized as 0, 1, and 2 for a homozygous, a heterozygous, and an alternative homozygous SNP genotype respectively, $\mathbf{g}$ is a vector of random additive SNP effects assuming $\mathbf{g}\sim\mathrm{N}\left(\mathbf{0},\mathbf{I}\frac{\hat{\sigma}_a^2}{\mathrm{N_{snp}}}\right)$ with $\mathbf{I}$ being an identity matrix and $\hat{\sigma}_a^2$ representing the estimate of the additive genetic variance of a given trait, and $\boldsymbol{e}$ is a vector of residuals with $\mathbf{e}\sim\mathrm{N}(\mathbf{0},\mathbf{D}\hat{\sigma}_e^2)$ with $\mathbf{D}$ being a diagonal matrix containing the reciprocal of EDC on the diagonal. A direct Genomic Value (DGV) is the sum of additive effects of SNPs: $\widehat{\mathbf{DGV}} = \mathbf{X}\hat{\mathrm{b}} + \mathbf{Z}\hat{\mathbf{g}}$.

GEBV estimation. Genomically Enhanced Breeding Values (GEBV) are the combination of own genomic information contained in SNP estimates represented by DGV and polygenic information contained in phenotypes of relatives represented by EBV. Consequently, GEBV of proven bulls is given by:

$$GEBV = \begin{bmatrix} r_{DGV} & r_{EBV} \end{bmatrix} \begin{bmatrix} r_{DGV} & r_{DGV}r_{EBV} \\ r_{DGV}r_{EBV} & r_{EBV} \end{bmatrix}^{-1} \begin{bmatrix} DGV \\ EBV \end{bmatrix}$$

and of selection candidates by:

$$GEBV = \begin{bmatrix} r_{DGV} & r_{PI} \end{bmatrix} \begin{bmatrix} r_{DGV} & r_{DGV}r_{PI} \\ r_{DGV}r_{PI} & r_{PI} \end{bmatrix}^{-1} \begin{bmatrix} DGV \\ PI \end{bmatrix},$$

where PI represents pedigree index and r_x is a reliability of a given source of information. In the current analysis PI was calculated using values corresponding to the national genetic evaluation from April 2009. The reliability of DGV was estimated by the approach of Strandén and Garrick (2009) with SNP allele frequencies in the base population estimated following VanRaden (2008).

Validation. The assessment of the predictive ability of the model was carried out using the standard procedure recommended by the Interbull organization (Mäntysaari et al. (2010)) with the historical data set represented by information from and active dairy cattle population from April 2009 and the current data set represented by information from April 2013. The validation test was calculated for two data sets – ORIG and RARE.

Results and Discussion

Figure 1 presents comparison of common SNP effects between ORIG and RARE data sets. There are considerable differences in effects of SNPs across the whole genome. It shows that SNPs with rare variants can influence on genomic breeding value estimates.

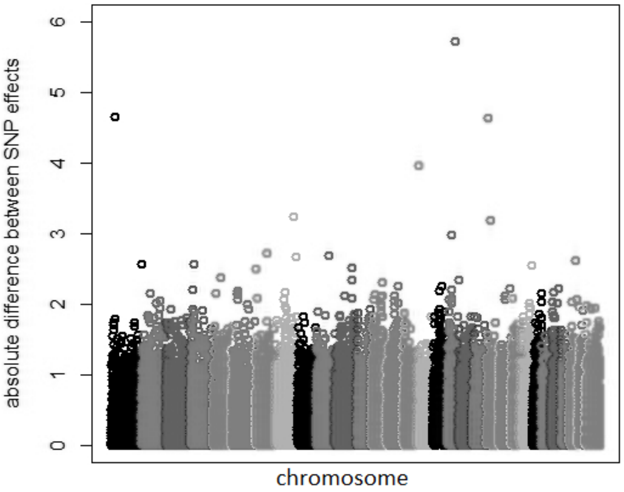

Figure 1: A Manhattan plot of absolute differences in common SNP effects estimated based on RARE and ORIG data sets.

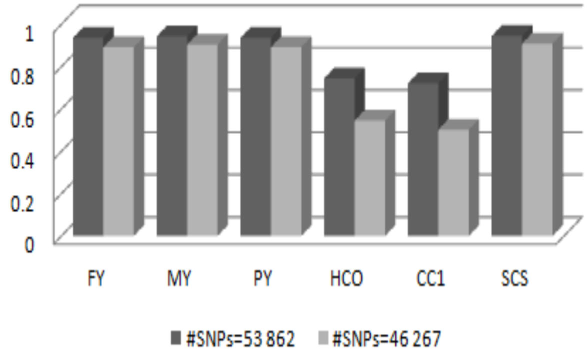

Figure 2: Comparison of accuracy of Direct Genomic Values (DGV) for training animals calculated based on ORIG and RARE data sets.

Figure 2 depicts the accuracy of DGV. For each trait the accuracy obtained for the RARE data set is higher than for the ORIG data. The largest differences are observed for fertility traits, HCO - 21% and CC1 - 19%. For the other traits the difference is considerably lower and does not exceed 5%.

Figure 3 presents the accuracy of GEBV. We can observe a similar situation as for DGV - the largest differences in accuracy for GEBV occur for fertility traits (HCO - 17% and CC1 - 16%) while for the other traits the difference is in maximum 2%.

The comparison of slope is presented on figure 4. For production and udder health traits there are no significant differences. For CC1 the difference is very large and it shows that including rare variants results in the overestimation of slope coefficient in this test.

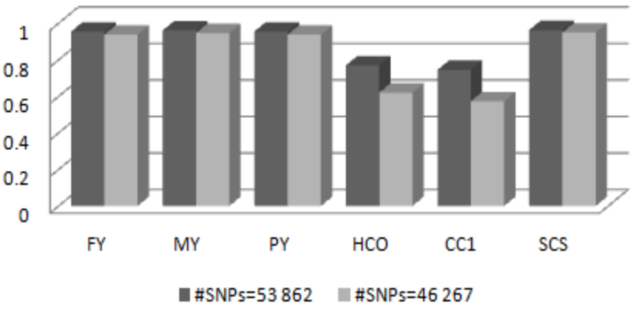

Figure 3: Comparison of accuracy of Genomically Enhanced Breeding Values (GEBV) for training animals calculated based on ORIG and RARE data sets.

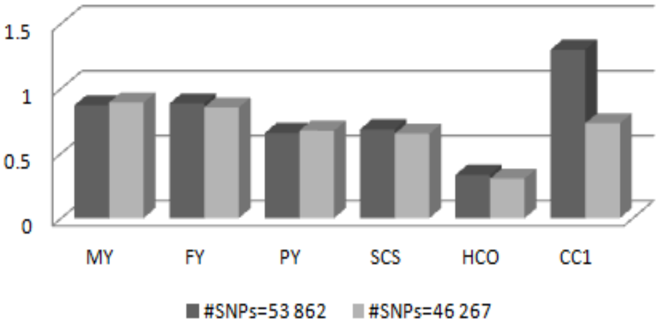

Figure 4: Comparison of slope coefficient for Interbull model

Conclusions

The results showed that including the rare variants into analysis increase the accuracy of DGV and GEBV for each kind of trait. Especially huge increasing we can observe for traits with low heritability (fertility traits). Additionally including the rare variants into analysis could be helpful for capturing a considerable part of missing heritability underlying traits under selection in dairy cattle - the markers from RARE data set can explain more missing heritability and markers from ORIG data set can change its effects.

Instead of many advantages of including the rare variants into analysis we should be more carefully in interpreting the results because especially for low-heritable traits can we observe the effect of overestimation of the slope coefficient in the Interbull test.

Literature Cited

Frazer, K.A., Murray, S.S., Schork, N.J., et al. (2009). Nat. Rev. Genet. 10: 241–251.

Jairath L., Dekkers, J.C.M., Schaeffer, L.R., et al. (1998). J. Dairy Sci. 81: 550–562.

Manolio, T.A., Collins, F.S., Cox, N.J., et al. (2009). Nature 461: 747–753.

Mäntysaari, E.A., Liu, Z., VanRaden, P. (2010). Interbull Bulletin 41: 17-22.

Strandén, I., Garrick, D.J. (2009). J. Dairy Sci. 92: 2971–2975.

VanRaden, P.M. (2008). J. Dairy Sci. 91: 4414–4423.