# INTEGRATING WHOLE GENOME SEQUENCE VARIANTS SELECTED FROM LARGE-SCALE ASSOCIATION ANALYSIS INTO THE SINGLE-STEP MARKER MODEL IN NEW ZEALAND DAIRY CATTLE

**Y. Wang, M.A. Nilforooshan, R.G. Sherlock, A.M. Winkelman and B.L. Harris**

Research and Development, Livestock Improvement Corporation, Hamilton 3240, New Zealand

## SUMMARY

Genome-wide association studies (GWAS) of whole-genome sequence data are used to identify causative variants or single nucleotide polymorphisms (SNPs) closely linked to causative variants that are associated with traits of interest. SNP associated with causative variants are often not included in the standard SNP chips used for routine genomic evaluation, e.g., the Illumina Bovine SNP50 chip. Several studies have shown the potential of including pre-selected sequence variants (SEQ) in genomic prediction. However, studies integrating such SNPs using a single-step genomic prediction model remain limited. A large-scale genome-wide association study of 42 routinely evaluated traits was conducted for New Zealand dairy cattle, and 4431 sequence variants were selected using an iterative GWAS approach. The current study evaluated the performance of genomic prediction by combining pre-selected sequence variants with the standard Illumina 50k SNP panel used in the NZ dairy industry in a single-step marker model. Various levels of increased ratio of prediction accuracy, decreased dispersion, and increased bias were observed in focal cows and bulls for live weight. A clear advantage of including these sequence variants in the model was not observed.

## INTRODUCTION

Genomic prediction using single nucleotide polymorphism data relies on the linkage disequilibrium (LD) between SNPs and causative variants. In the dairy industry, genomic predictions of breeding values have primarily used genotypes from the Illumina Bovine 50k SNP panel. However, standard Illumina 50k SNP arrays may not capture causal variants effectively because they consist of randomly selected markers chosen mainly for their polymorphism across breeds. Rare causal variants may not be in strong LD with the SNPs on these arrays and would have little, if any, impact on genomic predictions. In contrast, whole-genome sequence data is expected to capture many rare causal variants. Pre-selected sequence variants, either causal or closely linked to causal variants, are advantageous for genomic prediction and have been reported in several studies (VanRaden *et al.* 2017; Moghaddar *et al.* 2019). This study aimed to assess the impact of incorporating pre-selected sequence variants into the single-step genomic prediction of live weight in the New Zealand dairy population.

## MATERIALS AND METHODS

**Data.** The population included Holstein-Friesian (HF), Jersey (JER), European Red Dairy (RDC) purebred, and Holstein-Friesian Jersey (HFJ) crossbred animals. Animals that didn't fit the above-mentioned groups were categorised as "other" (OTH). Table 1 shows the number of animals with live weight phenotypes and genotypes in each breed category.

**Sequence variants selection.** The sequence variants were selected from large-scale genome-wide association studies in New Zealand dairy cattle, which included 42 routinely evaluated traits (Wang, unpublished results). A total of 292,667 animals genotyped on various genotype panels were imputed to the whole genome sequence level using Beagle 5.4 (B.L. Browning 2018). After imputation, variants with a minor allele frequency lower than 0.5% or dosage R-squared lower than

0.9 were removed. A total of 16,453,913 variants were retained for the GWAS analysis. In this study, we conducted an iterative GWAS analysis in which the most significant variants identified in the current iteration were then added as covariates in the next iteration. Such iterations continued until no variants that passed the p-value threshold ($5\times10^{-8}$) remained. Using this approach a list of 4799 variants that were significantly associated with the phenotype was generated. Multiple variants selected from different traits have the same locations. Thus, 4431 unique positions were added to the current genomic prediction model.

**Table 1. Number of animals with live weight phenotype and genotype by breed included in the model**

| Breed | Number of animals with phenotype | Number of animals with genotype |
|---|---|---|
| Holstein-Friesian | 466,513 | 71,244 |
| Jersey | 208,411 | 41,150 |
| Holstein-Friesian Jersey cross | 638,445 | 157,618 |
| European Red | 10,255 | 757 |
| Other | 86,302 | 15,172 |
| Total | 1,409,926 | 285,941 |

**Statistical analysis.** The single-step marker model is described in Harris (2022):

$$\mathrm{y = Xb + Z_g M_g m + Z_n u_n + Za + Zp + e}$$

where $\mathbf{y}$ is the phenotype vector; $\mathbf{X}$, $\mathbf{Z_g}$, $\mathbf{Z_n}$, and $\mathbf{Z}$ are the incidence matrices; $\mathbf{b}$ is the vector of fixed effects, including hybrid vigor, genetic groups, and breed covariates; $\mathbf{M_g}$ is the SNP marker matrix, where g refers to genotyped animals; $\mathbf{m}$ is the vector of SNP marker effects with each column centered to have a mean of zero; $\mathbf{u_n}$ is the vector of genomic breeding values for non-genotyped animals and n refers to non-genotyped animals; $\mathbf{a}$ is the vector of polygenic effects; $\mathbf{p}$ is the vector of permanent effects and $\mathbf{e}$ is the random residual effect.

In this study, we used two sets of SNPs for genomic prediction: (i) filtered Illumina 50k and (ii) filtered Illumina 50k and 4431 pre-selected sequence variants. The Linear Regression (LR) validation method (Legarra and Reverter 2018) was used to compare the prediction performance before and after including the pre-selected sequence variants in the model. The impact of the two marker sets on the genomic predictions was assessed using "*whole*" and "*partial*" datasets. The whole dataset included all available pedigree, genotype, and phenotype information, resulting in genomic evaluations that we will refer to as $u_w$. In the partial data set, phenotypes obtained after 2020-05-31 were removed, and the resulting evaluations will be referred as $u_p$). The whole dataset contained 1,995,603 live weight records from 1,409,927 animals, whereas the partial dataset contained 1,676,349 records from 1,193,138 animals.

LR validation statistics were obtained for two focal populations - validation bulls and validation cows. The focal population of validation bulls included genotyped bulls with at least 20 daughters and records in the whole but not the partial data. The number of daughters per sire ranges from 20 to 559, with a mean of 97.08. The focal population of validation cows included genotyped cows with at least one record in the whole, but no record in the partial data. The number of focal bulls and cows in each breed category can be found in Table 2. Due to their poor population representation, animals in the RDC and OTH groups were not considered when assessing the genomic prediction performance. For both focal populations, prediction performance was evaluated by bias, dispersion, and ratio of accuracies calculated from GEBVs obtained from the whole and partial dataset described in the LR method. The bias $\hat{\Delta}_p$ was calculated as the difference between the mean $u_p$ and $u_w$. The expected value of the bias is zero (i.e. unbiased). The dispersion $\hat{b_p}$ was calculated as the regression coefficient of $u_w$ on $u_p$, which was expected to be one when there is no over- or under-dispersion. Values greater than one indicate under-dispersion, and values less than one indicate over-dispersion. The correlation between $u_w$ and $u_p$ is an estimator of the ratio of prediction accuracies ($\rho_{w,p}$).

**Table 2. Number of live weight records for the whole and partial dataset, and the number of validation animals per breed**

| Breed | **Number of records** (whole) | **Number of records** (partial) | **Focal** | |
|---|---|---|---|---|
| | | | Bulls | Cows |
| Holstein-Friesian | 693,439 | 626,823 | 327 | 8,675 |
| Jersey | 298,783 | 262,074 | 176 | 5,968 |
| Holstein-Friesian Jersey cross | 877,050 | 680,383 | 307 | 37,123 |
| European Red | 12,627 | 11,557 | - | - |
| Other | 113,703 | 95,511 | - | - |
| Total | 1,995,602 | 1,676,348 | 810 | 51,766 |

## RESULTS AND DISCUSSION

**Ratio of accuracy.** Overall, the ratio of accuracy increased to varied degrees in both focal bull and cow populations when pre-selected sequence variants were added to the model. JER animals had a relatively lower ratio of accuracy compared to HF and HFJ. The ratio of accuracy increased from 0.769 to 0.791 for bulls and from 0.844 to 0.854 for cows. After including SEQ variants, the ratio of accuracy increased from 0.818 to 0.830 in HOL bulls and 0.854 to 0.861 in HOL cows. For HFJ animals, the ratio of accuracy increased from 0.826 to 0.846 in bulls and 0.892 to 0.899 in cows, the highest among all breed groups. Standard errors were larger for bulls than for cows using both marker sets (Table 3).

**Table 3. The ratio of the prediction accuracies ($\rho_{w,p}$) for the live weight from validation bulls and cows. Standard errors within brackets**

| Breed | **Focal Bulls** | | **Focal Cows** | |
|---|---|---|---|---|
| | Filtered 50k | Filtered 50k+SEQ | Filtered 50k | Filtered 50k+SEQ |
| Holstein-Friesian | 0.818 (0.032) | 0.830 (0.031) | 0.854 (0.006) | 0.861 (0.005) |
| Jersey | 0.769 (0.048) | 0.791 (0.046) | 0.844 (0.007) | 0.854 (0.007) |
| Holstein-Friesian Jersey cross | 0.826 (0.032) | 0.846 (0.031) | 0.892 (0.002) | 0.899 (0.002) |

**Dispersion.** In general, adding the sequence variants resulted in inconsistent changes in the dispersion across the breed and focal groups (Table 4). Among all breed groups of focal bulls, HFJ had the most unbiased prediction; also, after including SEQ variants, bias decreased from 1.040 to 1.030. HF had a slight overdispersion with a 0.6% increase in bias after adding SEQ variants. JER had a higher overdispersion level, with 0.866 before and 0.868 after SEQ variants were included. In contrast, focal cows showed a much smaller dispersion than bulls, which was close to 1 across all breeds. Adding sequence variants in the model only resulted in around a 1% change in all breed groups.

**Table 4. Dispersion ($\hat{b}_p \frown$ ) for the live weight from validation bulls and cows. Standard errors within brackets**

| **Breed** | **Focal Bulls** | | **Focal Cows** | |
|---|---|---|---|---|
| | Filtered 50k | Filtered 50k+SEQ | Filtered 50k | Filtered 50k+SEQ |
| Holstein-Friesian | 0.927 (0.036) | 0.921 (0.034) | 1.028 (0.007) | 1.020 (0.006) |
| Jersey | 0.866 (0.055) | 0.868 (0.051) | 0.990 (0.008) | 0.979 (0.008) |
| Holstein-Friesian Jersey cross | 1.040 (0.041) | 1.030 (0.037) | 1.061 (0.003) | 1.053 (0.003) |

**Bias.** Overall, bias was slightly increased after including sequence variants in all scenarios aside of HF cows (Table 5). However, in general, the increased level was negligible compared to the standard deviation of genetic variance in both bulls (40kg in the "*whole*" and 37kg in the "*partial*" evaluation) and cows (37kg in the "*whole*" and 32kg in the "*partial*" evaluations). Among all breeds, the highest bias was observed in JER, where the bias increased from 5.20 to 5.26 for bulls after adding sequence variants in the model. For cows, the value increased from 4.04 to 4.27. For all the other breed groups, the bias was all under 3.

**Table 5. Bias ($\hat{\Delta}_p$ ) for the live weight from validation bulls and cows. Standard errors between brackets**

| **Breed** | **Focal Bulls** | | **Focal Cows** | |
|---|---|---|---|---|
| | Filtered 50k | Filtered 50k+SEQ | Filtered 50k | Filtered 50k+SEQ |
| Holstein-Friesian | 1.800 (0.886) | 1.860 (0.860) | 0.407 (0.163) | 0.352 (0.161) |
| Jersey | 5.200 (1.010) | 5.260 (0.965) | 4.044 (0.155) | 4.275 (0.153) |
| Holstein-Friesian Jersey cross | 1.970 (0.920)) | 2.340 (0.872) | 1.191 (0.076) | 1.195 (0.075) |

## CONCLUSIONS

Enhancing the standard SNP chip with pre-selected sequence variants for genomic prediction of live weight was assessed using changes in the ratio of the accuracies, bias, and dispersion of the resulting GEBVs. The changes differed across breeds within the two focus populations but were generally small. Overall, the improvements in the estimates were minor, and no clear benefit of the sequence variants was found.

## ACKNOWLEDGMENTS

The NZ Ministry of Primary Industries financially supported this study through the SFF Futures Program: Resilient Dairy–Innovative Breeding for a Sustainable Dairy Future (grant number: PGP06-17006). The authors wish to acknowledge the use of New Zealand eScience Infrastructure (NeSI) high-performance computing facilities as part of this research.

## REFERENCES

Browning B.L., Zhou Y. and Browning S.R. (2018) *Am. J. Hum. Genet.* **103**: 338.

Harris B.L. (2022) *JDS communications* **3**: 152.

Legarra A. and Reverter A. (2018) *Genet. Sel. Evol.* **50**: 53.

Moghaddar N., Khansefid M., van Der Werf J.H.J., Bolormaa S., Duijvesteijn N., Clark S.A., Swan A.A., Daetwyler H.D. and MacLeod I.M. (2019) *Genet. Sel. Evol.* **51**: 72.

VanRaden P.M., Tooker M.E., O'Connell J.R., Cole J.B. and Bickhart D.M. (2017) *Genet. Sel. Evol.* **49**: 32.