J. Dairy Sci. 101:1292–1296
https://doi.org/10.3168/jds.2017-12890

Short communication: Genomic prediction using imputed whole-genome sequence variants in Brown Swiss Cattle

Mirjam Frischknecht,*†[1] Theodorus H. E. Meuwissen,‡ Beat Bapst,* Franz R. Seefried,* Christine Flury,† Dorian Garrick,§ Heidi Signer-Hasler,† Christian Stricker,# Intergenomics Consortium,‖ Anna Bieber,¶ Ruedi Fries, Ingolf Russ,†† Johann Sölkner,‡‡ Alessandro Bagnato,§§ and Birgit Gredler-Grandl***

*Qualitas AG, Zug 6300, Switzerland
†School of Agricultural, Forest and Food Sciences (HAFL), Bern University of Applied Sciences, Zollikofen 3052, Switzerland
‡Department of Animal and Aquacultural Sciences, Norwegian University of Life Science, Ås 1432, Norway
§Institute of Veterinary, Animal & Biomedical Sciences, Massey University, Palmerston North 4442, New Zealand
#agn Genetics, Davos 7260, Switzerland
‖Interbull Center, Uppsala 75007, Sweden
¶Department of Animal Sciences, Research Institute of Organic Agriculture (FiBL), Frick 5070, Switzerland
**Chair of Animal Breeding, Technische Universität München, Freising-Weihenstephan 85354, Germany
††Tierzuchtforschung e.V., Poing-Grub 85586, Germany
‡‡Department of Sustainable Agricultural Systems, Division of Livestock Sciences, University of Natural Resources and Life Sciences, Wien 1180, Austria
§§Department of Veterinary Sciences and Technologies for Food Safety, University of Milan, Milano 20133, Italy

ABSTRACT

The accuracy of genomic prediction determines response to selection. It has been hypothesized that accuracy of genomic breeding values can be increased by a higher density of variants. We used imputed whole-genome sequence data and various single nucleotide polymorphism (SNP) selection criteria to estimate genomic breeding values in Brown Swiss cattle. The extreme scenarios were 50K SNP chip data and whole-genome sequence data with intermediate scenarios using linkage disequilibrium-pruned whole-genome sequence variants, only variants predicted to be missense, or the top 50K variants from genome-wide association studies. We estimated genomic breeding values for 3 traits (somatic cell score, nonreturn rate in heifers, and stature) and found differences in accuracy levels between traits. However, among different SNP sets, accuracy was very similar. In our analyses, sequence data led to a marginal increase in accuracy for 1 trait and was lower than 50K for the other traits. We concluded that the inclusion of imputed whole-genome sequence data does not lead to increased accuracy of genomic prediction with the methods.

Key words: genomic prediction, Brown Swiss, whole-genome sequence data

Short Communication

Genomic prediction has had a great effect worldwide, especially on dairy breeding programs. Currently, routine genomic evaluations in dairy cattle are often based on 50K SNP chip data. However, it has been shown in simulation studies that using higher-density SNP information could increase accuracy of genomic breeding values (e.g., Meuwissen and Goddard, 2010; Druet et al., 2014; Iheshiulor et al., 2016). It has been hypothesized that the use of whole-genome sequence data should in particular increase accuracy of genomic estimated breeding values (**GEBV**), as sequence data includes the causal variants. Thanks to the 1000 Bull Genomes Project an unprecedented amount of sequence data became available to all project partners (Daetwyler et al., 2014). In 2015, the fifth run of the project was released, including sequences of 1,682 bulls and cows. However, these individuals still represent only a small fraction of all individuals of the global cattle population. An alternative to sequencing more individuals is to impute sequence data into target individuals genotyped for a smaller amount of SNP. Using this approach, sequence information of a large number of individuals becomes accessible. It has been shown that imputation using the reference panel from the 1000 Bull Genomes Project is highly accurate (e.g., Daetwyler et al., 2014; Frischknecht et al., 2016); however, Druet et al. (2014) showed, in a simulation study, that especially for traits influenced mainly by QTL with low minor allele frequency (**MAF**), the increase in accuracy compared with 50K scenarios is limited. A few studies using real data to evaluate accuracy of genomic prediction have

Received March 20, 2017.
Accepted September 28, 2017.

[1] Corresponding author: mirjam.frischknecht@qualitasag.ch

been published [e.g., in Holstein (van Binsbergen et al., 2015; Veerkamp et al., 2016) or Fleckvieh (Erbe et al., 2016)]; those studies found no advantage in the accuracy of genomic prediction using sequencing data over 50K SNP chip data.

The objective of the current study was to examine the effect of different SNP selection strategies on the accuracy of genomic prediction in Brown Swiss cattle. The main goal was to investigate the effect of imputed whole-genome sequence data on accuracy of genomic prediction. We estimated genomic breeding values in the Brown Swiss population using different densities of SNP data from 50K SNP data to whole-genome sequence level. Deregressed breeding values (**DRBV**; Garrick et al., 2009) were used as input phenotypes. Sequence genotypes were derived from a 2-step imputation: 50K to HD (800K) with FImpute (Sargolzaei et al., 2011), and from HD to full sequence with Beagle (Browning and Browning, 2009) and Minimac (Fuchsberger et al., 2015) using 123 Brown Swiss and Original Braunvieh animals from the 1000 Bull Genomes Project data set (Daetwyler et al., 2014) as reference individuals (Frischknecht et al., 2016). For the estimation of SNP effects, we used imputed allele dosage. For imputation, our data set, including sequenced and imputed individuals, comprised 23,001 animals with 16,184,800 SNP and small insertions or deletions. For further analyses, we excluded the small insertions or deletions and SNP with MAF <1% within the whole population and an imputation R^2 <0.1 (value provided by Minimac).

We evaluated the effect of SNP panel density on 3 traits: a reproduction trait [nonreturn rate in heifers (**NRH**); reference individuals (**ref**): n = 2,018, validation individuals (**val**): n = 240], a conformation trait [stature (**STA**); ref: n = 5,294, val: n = 596], and SCS (ref: n = 4,786, val: n = 560; Table 1). We calculated the proportion of variance that can be attributed to the SNP ($\sigma^2_{\mathrm{SNP}} / \sigma^2_{\mathrm{P}}$, where σ_{SNP} is the genetic variance attributed to SNP and σ_{P} is the phenotypic variance) in the data set with gcta using the reml function (Yang et al., 2010, 2011). Individuals for genomic prediction were chosen according to the reliability of the breeding value and, among these, the 10% youngest individuals were selected as validation individuals.

We estimated SNP effects for 5 different SNP selection scenarios: (1) 50K SNP chip data was used (**50K**; 38,436 SNP; average MAF: 0.247); (2) the full sequence panel was used (**SEQ**; 12,413,067 SNP; average MAF: 0.191); (3) variants with annotation missense from the Variant Effect Predictor (McLaren et al., 2016) were used (**MISS**; 33,037 SNP; average MAF: 0.182); (4) we randomly removed 1 SNP in SNP pairs from the full sequence panel that were located within a window of 10,000 SNP and showed almost perfect linkage disequilibrium (**LD**; r^2 ≥0.999999; 5,353,086 SNPs; average MAF: 0.203); (5) we performed a genome-wide association study (**GWAS**) with the full sequence panel using the bulls of the reference population and selected the 50,000 SNP with the lowest *P*-values [**TOP**; 50,000 SNP; average MAF: 0.268 (NRH), 0.241 (SCS), 0.265 (STA)]. Consequently, the number of SNP in the TOP scenario was similar to the 50K scenario. For GWAS, a mixed linear model was fitted in EMMAX (Kang et al., 2010) with allele dosage as input data using a G-Matrix from GCTA (Yang et al., 2011) and proportion of Original Braunvieh genes calculated from pedigree data as covariate. We estimated genomic breeding values using the program gbcpp (Iheshiulor et al., 2015). Using gbcpp we fitted marker effects as in BayesC (captures variants with larger effects) and a polygenic effect as a genomic BLUP term (captures genomic relationships due to polygenes; BayesC-L), which in addition to SNP makes use of the genomic relationship matrix. The model for BayesC-L can be described as

$$\mathbf{y} = 1'\mu + \mathbf{g} + \mathbf{X}\boldsymbol{\beta} + e, \qquad [1]$$

where $\mathbf{y}$ is a vector of DRBV; $\mathbf{g} \sim N\left(0, \mathbf{G}\sigma^2_{\mathrm{g}}\right)$ is a vector of random polygenic effects ($\mathbf{G}$ = the genomic relationship matrix and $\sigma_{\mathbf{g}}$ = variance of the polygenic effect); and $\boldsymbol{\beta}$ is a vector of SNP effects with elements, which are distributed $N\left(0, \sigma^2_{\mathrm{SNP}}\right)$ with a probability π and with probability $(1 - \pi)$ equal to zero (σ_{SNP} = variance of SNP effects); $\mathbf{X}$ is a matrix of marker genotypes; and $e \sim N\left(0, \sigma^2_{\mathrm{e}} / \mathrm{w}_{\mathrm{i}}\right)$ is the residual variance (σ_{e} = variance of residuals), with w_{i} being the weight of the y_{i}, which is in our analysis the reliability of the DRBV. For π we used a fixed value per trait, which we scaled according to the number of SNP (Supplemental Table S1; https://doi.org/10.3168/jds.2017-12890). Accuracy of genomic

Table 1. Accuracy of genomic breeding values for all scenarios[1]

Scenario[2]	SCS	STA	NRH
50K	0.556	0.538	0.397
TOP	0.502	0.478	0.324
MISS	0.504	0.495	0.347
LD	0.542	0.530	0.401
SEQ	0.548	0.527	0.403

[1] STA = stature; NRH = nonreturn rate in heifers.

[2] Scenario: 50K = SNP from 50K SNP chip; TOP = top associated variants from GWAS; MISS = variants with annotation Missense; LD = linkage disequilibrium pruned sequence data; SEQ = whole-genome sequence data.

breeding values was calculated as the Pearson correlation of the DRBV and the GEBV.

We compared the accuracies of the 5 different SNP density scenarios. For the 50K scenario, across all traits we observed accuracies from 0.397 for NRH to 0.556 for SCS (Table 1). Variation in accuracy between traits was expected due to differences in the number of reference individuals per trait and the different heritabilities of the traits. For SCS and STA, similar levels of GEBV accuracy were found, likely due to a similar number of individuals in the experiment and a similar proportion of the variance explained by the SNP. Within-trait differences observed for the different SNP selection scenarios were much smaller than the differences between traits. We found the lowest accuracies with the TOP scenario, which indicates that including only the top associated variants in the prediction will not increase accuracy. It is likely that with this method we included a large number of SNP in high linkage disequilibrium, which may have adversely affected the analysis. Additionally, those SNP were no longer distributed across the whole genome, but were rather concentrated in some regions (Figure 1 for STA). Therefore, relationships between individuals might be captured less precisely, which might explain part of the decrease in accuracy. The scenario MISS was included, as missense variants should be more likely to be causal because they are variants that have a direct effect on proteins. Including more causal variants would be expected to explain a larger fraction of the genetic variation. However, looking at the σ^2_{SNP} / σ^2_{P} ratio (Supplemental Table S1; https://doi.org/10.316/jds.2017-12890), we found the opposite: less variance was explained by missense variants than other choices of SNP. This could be associated to the lower average MAF compared with 50K and a low number of SNP compared with SEQ, for example. For all traits, MISS leads to slightly lower accuracies than 50K; compared with 50K, LD and SEQ led only to very moderate increase for NRH and to a very moderate decrease for SCS and STA (Table 1). For STA, LD gives slightly higher accuracies, indicating that filtering of SNP can be beneficial to the accuracy. This may be due to the fact that we excluded highly redundant information, which may not be properly accounted for by the algorithm. The QTL effects distributed across a large number of redundant SNP in high linkage disequilibrium are harder to distinguish from residual error, because the SNP effects of the QTL can be smeared across multiple SNP and, therefore, the effect of a single SNP can be reduced to a similar level as SNP with no effect (Veerkamp et al., 2016).

Our results support the findings of published studies performed in other breeds that the increase of accuracy of genomic prediction is marginal, if anything, by increasing marker density from 50K SNP chip to sequence data. In a recent study, a small increase in accuracy was found by adding selected sequence variants to an HD panel (VanRaden et al., 2017). We ran our analyses only with selected variants and did not include all 50K data. A combination of 50K data and selected variants might have led to a higher accuracy in our data. Further possible reasons for this small increase in accuracy probably include sequence genotypes being imputed, thus losing recombination events occurring in the genome of imputed individuals; additionally, imputation accuracy is lower for rare variants (van Binsbergen et al., 2014; Pausch et al., 2017). In our study, imputation was based on a low number of reference individuals (n = 123), which means we could only capture the variance covered by those 123 individuals. As the individuals were chosen to explain a major fraction of the genetic variance, we expected to cover common haplotypes relatively well. However, rare haplotypes could also have a large effect on the trait (Gonzalez-Recio et al., 2015), and they cannot be detected if not present in the reference population from imputation. In a simulation study by Druet et al. (2014), it has been shown that when using imputed data the increase in accuracy is decreased depending on the MAF distribution of the QTL influencing the trait. Additionally, increasing the number of parameters to estimate without increasing the number of observations in general leads to a decrease in accuracy (Berger et al., 2015). In the present study, this was somewhat compensated by using a variable selection model (Bayes C-L); however, only limited data are available, which hinders optimization within the extremely high-dimensional parameter space. Additionally the used data has a large influence on the outcome of the study; the number of individuals was relatively small and changes in the training or the validation population could have had a large effect on the results. Overall, currently it is not recommended to implement SNP effect estimation based on sequence data for routine genomic prediction.

ACKNOWLEDGMENTS

This study was conducted within the project Swiss-LowInputGenetics (SLIG). We thank the Swiss Commission for Technology and Innovation for financial support of this project. We acknowledge Braunvieh Schweiz (Switzerland) for providing phenotypic data. The partners of the Low-InputBreed project (FP7-project no. KBBE 222 623) are acknowledged for giving the permission to use their genotypic data. The 1000 Bull Genomes Project is acknowledged for providing

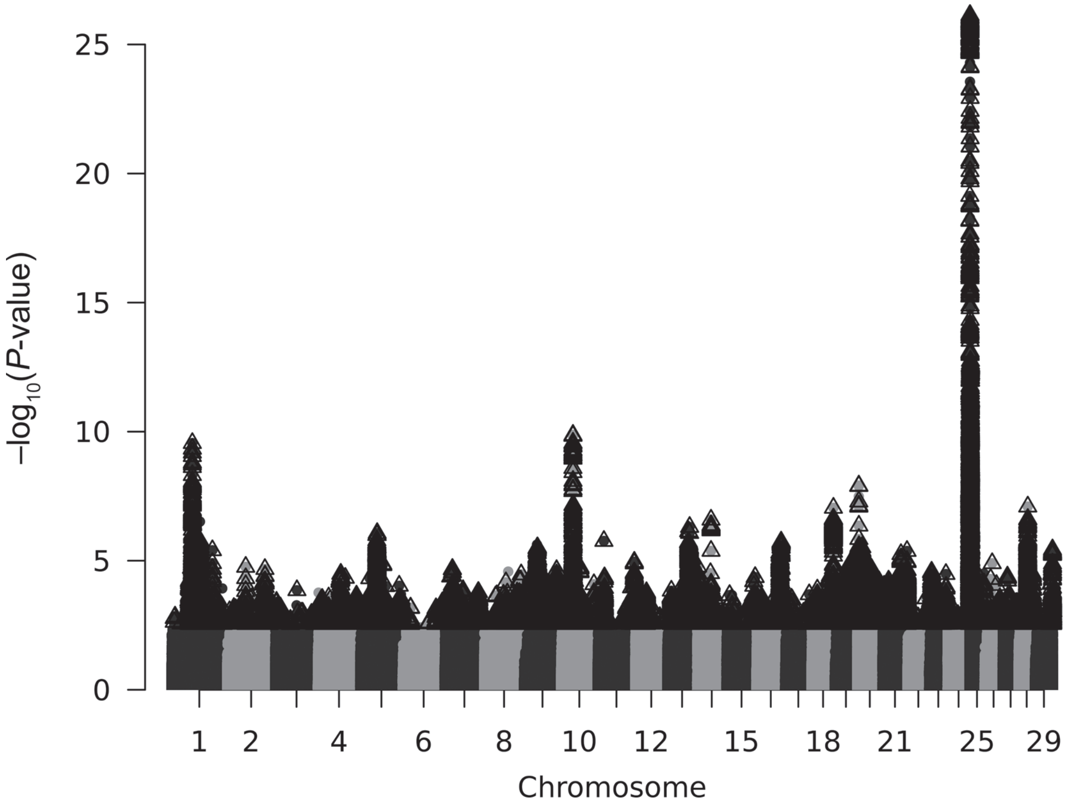

Figure 1. Manhattan plot for the genome-wide association study (GWAS) in stature (STA); SNP in a triangle are the SNP used for estimation of genomic breeding values for STA in the top scenario.

sequence genotypes of the reference panel for imputation. Computations were done using server facilities of Qualitas AG, Zug, Switzerland.

REFERENCES

Berger, S., P. Pérez-Rodríguez, Y. Veturi, H. Simianer, and G. de los Campos. 2015. Effectiveness of shrinkage and variable selection methods for the prediction of complex human traits using data from distantly related individuals. Ann. Hum. Genet. 79:122–135. https://doi.org/10.1111/ahg.12099.

Browning, B. L., and S. R. Browning. 2009. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am. J. Hum. Genet. 84:210–223. https://doi.org/10.1016/j.ajhg.2009.01.005.

Daetwyler, H. D., A. Capitan, H. Pausch, P. Stothard, R. van Binsbergen, R. F. Brøndum, X. Liao, A. Djari, S. C. Rodriguez, C. Grohs, D. Esquerré, O. Bouchez, M.-N. Rossignol, C. Klopp, D. Rocha, S. Fritz, A. Eggen, P. J. Bowman, D. Coote, A. J. Chamberlain, C. Anderson, C. P. VanTassell, I. Hulsegge, M. E. Goddard, B. Guldbrandtsen, M. S. Lund, R. F. Veerkamp, D. Boichard, R. Fries, and B. J. Hayes. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat. Genet. 46:858–865. https://doi.org/10.1038/ng.3034.

Druet, T., I. M. Macleod, and B. J. Hayes. 2014. Toward genomic prediction from whole-genome sequence data: Impact of sequencing design on genotype imputation and accuracy of predictions. Heredity (Edinb.) 112:39–47. https://doi.org/10.1038/hdy.2013.13.

Erbe, M., G. Ni, H. Pausch, R. Emmerling, T. Meuwissen, D. Cavero, K.-U. Götz, and H. Simianer. 2016. Experiences from genomic prediction using sequence data in different species. Page 103 in Book of Abstracts of the 67th Annual Meeting of the European Federation of Animal Science. Wageningen Academic Publishers, Wageningen, the Netherlands.

Frischknecht, M., H. Pausch, B. Bapst, F. R. Seefried, C. Flury, H. Signer-Hasler, R. Fries, D. J. Garrick, C. Stricker, and B. Gredler. 2016. Accurate sequence imputation enables precise QTL mapping in Brown Swiss cattle. Page 104 in Book of Abstracts of the 67th Annual Meeting of the European Association for Animal Production. Wageningen Academic Publishers, Wageningen, the Netherlands.

Fuchsberger, C., G. R. Abecasis, and D. A. Hinds. 2015. Minimac2: Faster genotype imputation. Bioinformatics 31:782–784. https://doi.org/10.1093/bioinformatics/btu704.

Garrick, D. J., J. F. Taylor, and R. L. Fernando. 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet. Sel. Evol. 41:55 https://doi.org/10.1186/1297-9686-41-55.

Gonzalez-Recio, O., H. D. Daetwyler, I. M. MacLeod, J. E. Pryce, P. J. Bowman, B. J. Hayes, and M. E. Goddard. 2015. Rare variants in transcript and potential regulatory regions explain a small percentage of the missing heritability of complex traits in cattle. PLoS One 10:e0143945. https://doi.org/10.1371/journal.pone.0143945.

Iheshiulor, O. O. M., J. A. Woolliams, X. Yu, and T. H. E. Meuwissen. 2015. Genomic predictions in Norwegian Red Cattle: comparison of methods. Page 353 in Book of Abstracts of the 66th Annual Meeting of the European Federation of Animal Science. Wageningen Academic Publishers, Wageningen, the Netherlands.

Iheshiulor, O. O. M., J. A. Woolliams, X. Yu, R. Wellmann, and T. H. E. Meuwissen. 2016. Within- and across-breed genomic prediction using whole-genome sequence and single nucleotide polymorphism panels. Genet. Sel. Evol. 48:15. https://doi.org/10.1186/s12711-016-0193-1.

Kang, H. M., J. H. Sul, S. K. Service, N. A. Zaitlen, S.-Y. Kong, N. B. Freimer, C. Sabatti, and E. Eskin. 2010. Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. 42:348–354. https://doi.org/10.1038/ng.548.

McLaren, W., L. Gil, S. E. Hunt, H. S. Riat, G. R. S. Ritchie, A. Thormann, P. Flicek, and F. Cunningham. 2016. The Ensembl variant effect predictor. Genome Biol. 17:122 https://doi.org/10.1186/s13059-016-0974-4.

Meuwissen, T., and M. Goddard. 2010. Accurate prediction of genetic values for complex traits by whole-genome resequencing. Genetics 185:623–631. https://doi.org/10.1534/genetics.110.116590.

Pausch, H., I. M. Macleod, R. Fries, R. Emmerling, and J. Phil. 2017. Evaluation of the accuracy of imputed sequence variants and their utility for causal variant detection in cattle. Genet. Sel. Evol. 49:24. https://doi.org/10.1186/s12711-017-0301-x.

Sargolzaei, M., J. Chesnais, and F. Schenkel. 2011. FImpute—An efficient imputation algorithm for dairy cattle populations. J. Dairy Sci. 94:421.

van Binsbergen, R., M. C. A. M. Bink, M. P. L. Calus, F. A. van Eeuwijk, B. J. Hayes, I. Hulsegge, and R. F. Veerkamp. 2014. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. Genet. Sel. Evol. 46:41. https://doi.org/10.1186/1297-9686-46-41.

van Binsbergen, R., M. P. L. Calus, M. C. A. M. Bink, F. A. van Eeuwijk, C. Schrooten, and R. F. Veerkamp. 2015. Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. Genet. Sel. Evol. 47:71. https://doi.org/10.1186/s12711-015-0149-x.

VanRaden, P. M., M. E. Tooker, J. R. O'Connell, J. B. Cole, and D. M. Bickhart. 2017. Selecting sequence variants to improve genomic predictions for dairy cattle. Genet. Sel. Evol. 49:32. https://doi.org/10.1186/s12711-017-0307-4.

Veerkamp, R. F., A. C. Bouwman, C. Schrooten, and M. P. L. Calus. 2016. Genomic prediction using preselected DNA variants from a GWAS with whole-genome sequence data in Holstein–Friesian cattle. Genet. Sel. Evol. 48:95. https://doi.org/10.1186/s12711-016-0274-1.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher. 2010. Common SNPs explain a large proportion of heritability for human height. Nat. Genet. 42:565–569. https://doi.org/10.1038/ng.608.

Yang, J., S. H. Lee, M. E. Goddard, and P. M. Visscher. 2011. GCTA: A tool for genome-wide complex trait analysis. Am. J. Hum. Genet. 88:76–82. https://doi.org/10.1016/j.ajhg.2010.11.011.