Accuracy of genomic prediction of dry matter intake in Dutch Holsteins using sequence variants from meta-analyses

B. Gredler-Grandl[1*], B. Raymond[1], G.C.B. Schopen[1], P.K. Chitneedi[2], Z. Cai[3], C.I.V. Manzanilla-Pech[3], T. Iso-Touru[4], D. Fischer[4], S. Bolormaa[5], T.S. Chud[6,7], F.S. Schenkel[6], Y. Wang[8,9], C. Li[8,9], D. Hailemariam[8,9], B. Villanueva[10], A. Fernandez[10], C. Kuehn[2], G. Sahana[3], M.H. Lidauer[4], J.E. Pryce[5,11], O. González-Recio[10], G. Plastow[8], C.F. Baes[6,12], N. Charfeddine[13], Y. de Haas[1], R.F. Veerkamp[1] and A.C. Bouwman[1]

[1]Wageningen University & Research, P.O. Box 338, 6700 AH Wageningen, the Netherlands; [2]Institute for Genome Biology, Research Institute for Farm Animal Biology (FBN), Wilhelm-Stahl-Allee 2, 18196 Dummerstorf, Germany; [3]Center for Quantitative Genetics and Genomics, Aarhus University, Blichers Alle 20, 8830 Tjele, Denmark; [4]Natural Resources Institute Finland (Luke), Myllytie 1, FI-31600 Jokioinen, Finland; [5]Agriculture Victoria, AgriBio, Centre for AgriBioscience, 5 Ring Road, Bundoora, Victoria, 3083, Australia; [6]University of Guelph, Centre for Genetic Improvement of Livestock, Guelph, ON N1G 2W1, Canada; [7]Fast Genetics, Saskatoon, SK S7K 2K6, Canada; [8]Department of Agriculture, Food & Nutritional Science, University of Alberta, Edmonton, Alberta, Canada; [9]Lacombe Research and Development Centre, Agriculture and Agri-food Canada, Lacombe, Alberta, Canada; [10]Departamento de Mejora Genética Animal, Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria, INIA-CSIC Crta. de la Coruña km 7.5, 28040 Madrid, Spain; [11]School of Applied Systems Biology, La Trobe University, Bundoora, Victoria, 3083, Australia; [12]University of Bern, Institute of Genetics, Vetsuisse Faculty, Bern 3001, Bern, Switzerland; [13]Spanish Holstein Association (CONAFE), Ctra. De Andalucía km 23600 Valdemoro, 28340 Madrid, Spain; [*]birgit.gredler-grandl@wur.nl

Abstract

We evaluated the accuracy of biology informed genomic prediction for dry matter intake in 2,162 Dutch Holstein cows. Sequence variants were selected from meta-analyses including GWAS summary statistics for QTL and metabolomic QTL in several dairy and crossbred beef populations. Selected variants were prioritized in GBLUP models in a five-fold cross-validation. The accuracies were compared to genomic prediction based on routine 50k genotype data. The average accuracy for the 50k scenario was 0.683. Adding selected sequence variants in the GBLUP model did not improve the accuracies for dry matter intake. Next steps will include testing Bayesian variable selection methods to prioritize variants in genomic prediction for dry matter intake.

Introduction

Feed intake is an important trait, but large scale recording schemes on individual animals are costly and hamper implementations of genetic evaluations for dry matter intake (DMI). High prediction accuracies are difficult to achieve for traits like DMI, because the size of the reference population is often limited. Raymond et al. (2018) have shown a way to increase accuracy of genomic prediction (GP) in a multibreed evaluation when prioritising sequence variants from a meta-GWAS separate from the routine array genotypes. Additionally, functional information to pinpoint QTL regions may improve predictions. Metabolites in blood plasma represent intermediate phenotypes between the genome and transcriptome and final expressed phenotypes. Here, we evaluate the accuracy of biology informed GP for DMI in Dutch Holstein cows by prioritising variants from various meta-GWAS analyses.

Material & Methods

Meta-GWAS for QTL.

To increase power of QTL detection for DMI, a meta-GWAS based on local GWAS results of eight populations was carried out using the METAL software package (Willer et al., 2010). Local GWAS results for imputed sequence variants were available for five different Holstein populations in Australia (584 cows), Canada (588 cows), Denmark (495 cows), Germany (140 cows) and Spain (561 cows), one Finnish Red population from Finland (366 cows) and two crossbred beef populations in Germany (253 bulls of a Charolais x Holstein cross) and Canada (7,552 heifers and steers of Angus, Charolais, Kinsella Composite and crossbred animals thereof). Phenotypes used in the local GWAS were either de-regressed breeding values (DRBV) or raw phenotypes for DMI corrected for fixed effects across lactation. Variant effects and standard errors of the effect from the local GWAS were standardized based on the genetic standard deviation for DMI for each population. Variants with an imputation accuracy $r^2<0.6$, a minor allele frequency (MAF) <0.001 or an effect size >5 standard deviations apart from the mean were not considered in the meta-analysis. Three different meta-GWAS were carried out: a meta-analysis including all populations (ALL), Holstein populations only (HOL) and beef populations only (BEEF). The total number of sequence variants considered were 30,216,688, 19,647,876 and 27,839,929 for ALL, HOL and BEEF, respectively.

Meta-GWAS for mQTL.

GWAS summary statistics of imputed sequence variants from three cattle populations (Holstein, Charolais x Holstein, mixed beef breed composite) were available for segregating blood plasma metabolomic QTL. Metabolites considered were amino acids, short and long-chain fatty acids and compounds from energy and protein metabolism functions. The METAL software package (Willer et al., 2010) was used for the meta-analysis. Variant effects were standardized by the genetic standard deviation of plasma metabolite concentration. Variants with MAF < 0.01 and/or imputation r^2 <0.6 were discarded. The number of animals contributing to the meta-GWAS varied across metabolites between 241 and 1,103. The number of variants considered across metabolites were between 14,343,591 and 19,467,841.

SNP selection scenarios.

To filter out non-causal and select independently associated variants from the QTL meta-analysis, a forward selection of variants based on the conditional and joint effect method (COJO) described in Yang et al. (2012) was conducted. Variants were selected using the following COJO model parameters: conditional and joint p-value threshold to declare a genome-wide significant variant of p=5e-3 (scencarios ALL3, HOL3, BEEF3) or p=5e-5 (scenarios ALL5, HOL5, BEEF5), collinearity between selected markers of 0.9 and window size of 10 Mb. P-value significance thresholds were chosen to achieve a reasonable number of selected variants for GP. The mQTL meta-GWAS resulted in 20,426 trait associated SNPs with $p < 10^{-6}$ for 27 metabolites. Of those, all variants segregating in the target population for GP (see below) were used in the mQTL scenario. For the base scenario (50k) variants available on the Illumina Bovine snp50 v3 beadchip (Illumina Inc., San Diego, CA, USA) were selected from the imputed sequence genotypes of the target population for GP. The number of selected variants per scenario is shown in Table 1. Selected variants were used to derive the genomic relationship matrices (GRM) for estimation of heritability and genomic breeding values (GEBV) for DMI.

Genomic prediction.

An independent data-set of 2,162 Dutch Holstein cows imputed to sequence, and with DRBV

for DMI with reliability ≥0.30 were used in GP models. For the base scenario (50k) and all scenarios where only one GRM was fitted the following model was run in the mtg2 software (Lee, 2016):

$$\mathbf{y} = \mathbf{1}\boldsymbol{\mu} + \mathbf{W}\mathbf{g} + \mathbf{e} \quad (1)$$

where **y** is a vector of DRBV for DMI for the reference set (missing for the validation set), $\mathbf{1}\boldsymbol{\mu}$ is the overall mean, **g** is a vector of additive genetic effects for all cows, **W** is a design matrix linking **g** to DRBV in **y** and **e** is a vector containing residuals. **g** and **e** are assumed to be normally distributed with $\boldsymbol{g} = N(0, \boldsymbol{GRM}\sigma_g^2)$ and $\boldsymbol{e} = N(0, \boldsymbol{I}\sigma_e^2)$.

For scenarios with two different GRM (50k variants and selected sequence variants) the following model was used:

$$\mathbf{y} = \mathbf{1}\boldsymbol{\mu} + \mathbf{W_1}\mathbf{g_1} + \mathbf{W_2}\mathbf{g_2} + \mathbf{e} \quad (2)$$

where subscripts 1 and 2 represent the first (base 50k) and second (selected sequence variants) GRM fitted, respectively, $\mathbf{W_1}$ and $\mathbf{W_2}$ are identical design matrices linking DRBV to the two genetic effects $\mathbf{g_1}$ and $\mathbf{g_2}$. As above, **g** and **e** are assumed to be normally distributed.

A five-fold cross validation was carried out. 400 of the 2,162 cows were randomly selected five times as validation and the remainder (1,762 cows) comprised the reference set. Accuracy of GP was measured as the correlation between the GEBV and DRBV in the validation set. The prediction bias was assessed by regressing DRBV of the validation cows on their GEBV.

Table 1. Number of variants to build the GRM and estimated heritability of DRBV in all scenarios (average across 5 cross-validations).

Scenario	Number of variants	Heritability	SE
50k	37,179	0.849	0.024
ALL3	1,810	0.456	0.032
ALL5	166	0.142	0.026
HOL3	1,834	0.442	0.032
HOL5	63	0.086	0.026
BEEF3	1,746	0.473	0.032
BEEF5	321	0.240	0.030
mQTL	17,056	0.675	0.034
50k + ALL3	37,179 + 1,810	0.796[1] + 0.060[1]	0.037 + 0.030
50k + ALL5	37,179 + 166	0.853 + <0.001	0.025 + 0.007
50k + HOL3	37,179 + 1,834	0.825 + 0.026	0.034 + 0.026
50k + HOL5	37,179 + 63	0.849 + <0.001	0.025 + 0.005
50k + BEEF3	37,179 + 1,746	0.854 + 0.004	0.036 + 0.028
50k + BEEF5	37,179 + 321	0.858 + <0.001	0.026 + 0.012
50k + mQTL	37,179 + 17,056	0.844 + 0.008	0.025 + 0.008

[1] The first and second number represent the heritability estimates for 50k and the respective sequence variant set, respectively.

Results

The heritability estimated in this study is based on DRBV and can therefore be interpreted as the explained genetic variance of DRBV. The estimates shown in Table 1 are similar between the 50k and all scenarios fitting two GRMs. When using selected sequence variants only, heritability ranged between 0.086 (HOL5) and 0.675 (mQTL).

The accuracy of GP for all scenarios is presented in Table 2. All scenarios fitting two GRMs resulted in similar accuracy as in the 50k scenario only. When using selected variants only, the highest accuracy was observed when using variants from the beef meta-GWAS (BEEF3) and using all populations (ALL3). GEBV were overestimated in all scenarios fitting two GRMs. GEBVs in ALL5 and HOL5 were least biased.

Table 2. Accuracy and slope of the regression of DRBV on GEBV for all genomic prediction scenarios (average across 5 cross-validations).

Scenario	Accuracy	SE	Slope
50k	0.683	0.027	0.909
ALL3 only	0.557	0.035	0.957
ALL5 only	0.363	0.043	1.014
HOL3 only	0.549	0.035	0.973
HOL5 only	0.204	0.048	0.992
BEEF3 only	0.568	0.034	0.983
BEEF5 only	0.469	0.039	1.068
mQTL only	0.475	0.039	0.914
50k + ALL3	0.686	0.027	0.936
50k + ALL5	0.682	0.027	0.932
50k + HOL3	0.683	0.027	0.933
50k + HOL5	0.683	0.028	0.932
50k + BEEF3	0.653	0.029	0.935
50k + BEEF5	0.682	0.027	0.932
50k + mQTL	0.682	0.027	0.931

Discussion

We present accuracies of GEBV for DMI for different sequence variant selection scenarios. No clear advantage in accuracy could be found when adding relevant sequence variants in GBLUP models fitting two GRMs. Reasons may be related to the genetic architecture of the trait DMI in Holstein. Furthermore, variants selected in the different scenarios have generally a very low MAF and imputation errors and linkage disequilibrium can influence the result of GP. Here, we used GBLUP for estimation of GEBV. Bayesian variable selection methods may potentially result in higher accuracy. Further studies will be undertaken in the near future testing Bayesian variable selection methods in a larger Dutch Holstein population and across dairy and beef breeds with DMI phenotypes.

Acknowledgments

This work is carried out within the BovReg project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815668, and the global dry matter initiative gDMI. Canadian authors acknowledge additional funding support from Genome Alberta and Alberta Agriculture and Forestry, project GAB-A3GP37, as well as funders, partners and collaborators involved in the Resilient Dairy Genome Project (http://www.resilientdairy.ca).

References

Lee S.H, and Van der Werf J.H (2016) Bioinformatics 32(9):1420-1422. https://doi.org/10.1093/bioinformatics/btw012

Raymond B, Bouwman A.C., Wientjes Y.C.J., Schrooten C., Houwing-Duistermaat J. et al. (2018) Genet. Sel. Evol. 50:49. https://doi.org/10.1186/s12711-018-0419-5

Willer C.J., Li Y., and Abecasis G.R. (2010) Bioinformatics 26(17):2190-2191. https://doi.org/10.1093/bioinformatics/btq340

Yang J., Ferreira T., Morris A.P., Medland S.E., Madden P.A., et al. (2012) Nat. Genet. 44(4):369-375. https://doi.org/10.1038/ng.2213