304. Predictive ability of genomic prediction in layers when including CADD scores as genome function information

M.C.A.M. Bink[1*], M.P.L. Calus[2], M.F.L. Derks[2], J. Visscher[1] and B.C. Perez[1]

[1]Hendrix Genetics, P.O. Box 114, 5830 AC Boxmeer, the Netherlands; [2]Wageningen University & Research, Animal Breeding and Genomics, P.O. Box 338, 6700 AH Wageningen, the Netherlands; *marco.bink@hendrix-genetics.com*

Abstract

Accuracy of genomic prediction is key for genetic progress in breeding programs and including genome functional annotation may help. Here we report on the added value of Combined Annotation Dependent Depletion (CADD) scores to weigh SNPs in GBLUP analyses. Multiple transformations of CADD scores were considered and empirically validated on a layer dataset including 5 traits and 18K animals with 27K SNP genotypes. This initial analysis revealed that the use of (squared) CADD scores yielded marginally higher accuracy for 3 traits. We anticipate that the added value of CADD scores will increase by using a higher number of SNPs and ultimately whole genome sequence SNPs.

Introduction

Since the introduction of genomic prediction (Meuwissen *et al.*, 2001), animal breeding has been transitioning towards approaches that increasingly exploit genomic information. An emerging strategy involves weighting of SNPs that have a higher probability of being causal or being tightly linked to causal variants (Xiang *et al.*, 2021).

The combined annotation dependent depletion (CADD) (Rentzsch *et al.*, 2019) model that was developed to investigate SNPs in human populations, can score variants at coding and non-coding locations in the genome. Similar models have been deployed for livestock species. For chicken and pig, so-called cCADD and pCADD scores were created, as a tool to prioritize variants and evaluate sequences to highlight new sites of interest to explain biological functions that are relevant to animal breeding (Groβ *et al.*, 2020).

The objective of this study is to validate the added value of CADD scores to weight SNPs in genomic prediction on five traits in commercial layers. We explore multiple weighting strategies, i.e. transformations of CADD scores, in GBLUP models. This project is part of EuroFAANG (https://eurofaang.eu), a synergy of five Horizon 2020 projects that share the common goal to discover links between genotype to phenotype in farmed animals and meet global Functional Annotation of ANimal Genomes (FAANG) objectives.

Materials & methods

Chicken CADD scores. The chicken CADD scores were retrieved from Groβ *et al.* (2020). CADD is a machine learning tool originally developed to identify deleterious variants in the human genome. CADD uses a wide range of annotations to predict impact from sequence variation. Important annotations include conservation scores, gene model annotations, epigenomic data, and functional predictions. CADD combines these annotations to provide an impact score for any possible SNP variant in the genome. The higher the CADD score, the more likely the variant has impact. The scores can be used to prioritize genetic variants in genetic evaluation.

Weighting SNPs in genomic prediction. A typical genomic relationship matrix G is computed as: $G=XX^T/c$, equivalently written as: $G=XIX^T/c$, where X contains centred (and possibly scaled) SNP genotypes, c is a constant, and I is an identity matrix (VanRaden, 2007). This implies that all SNPs have

R.F. Veerkamp and Y. de Haas (eds)

DOI: 10.3920/978-90-8686-940-4_304, © M.C.A.M. Bink *et al.* 2022

an equal weight of 1. The computation can be generalized as: $G_w=XDX^T/c$, where D is a diagonal matrix, with weights for each SNP on the diagonal that are proportional to the variance that this SNP explains. The calc_grm program (Vandenplas and Calus 2020) was used to compute G and G_w.

Scenarios of weights. In this approach, the weight represents the *a priori* variance associated with a particular SNP. So, if one SNP has a weight of 1, and another one has a weight of 2, this means that the second SNP explains twice as much variance as the first one. The CADD scores do not represent variances and therefore three options are considered to translate CADD scores into weights, i.e. original (CADD), squared (CADDsq), and square root (CADDrt) (Figure 1).

To develop a benchmark for these scenarios, two options were considered: [A] all SNPs having equal weights (Equal); [B] random values (Random). For option [B] 10 replicates were generated to provide standard error estimates. The random values were generated by squaring random draws from $N(0,1)$. Within each scenario, all weights are scaled by dividing by the average weight, such that after scaling the mean weight is equal to 1.

Layers dataset. The dataset was provided by the layer division (ISA BV) of Hendrix Genetics and pertains to the pure line WA of the commercial layer breeding program. A total of 18,419 animals was genotyped with a 60K SNP array. These genotypes were curated on minor allele frequency (>0.05) and missing values (<10%). SNPs with unknown position or located at sex chromosomes were also discarded. After quality control, 26,841 autosomal SNPs were available for further analyses. Including ancestors of the genotyped animals yielded a pedigree of 21,060 animals. The dataset included five traits, i.e. shell breaking strength (BS), egg number (EN), egg weight (EW), hatchability (HT), and survival (SV). The number of records varied between 2,361 and 5,802 (Table 1). The dataset has been pre-corrected for all non-genetic effects to ease analyses. The pre-correction was executed using the models as used in the routine genetic evaluations. That is, the feature 'yielddev' in the MiXBLUP software (www.mixblup.eu) was used to produce yield deviations, i.e. the sum of estimated animal & residual effects. Variance components were estimated using the average information restricted maximum likelihood algorithm implemented in the AIREMLF90 software (Misztal *et al.*, 2014), and yielded heritability estimates ranging from 0.06 to 0.75 (Table 1). When omitting SNP genotypes, the pedigree-based accuracy of EBV was between 0.06 and 0.37 for the validation animals.

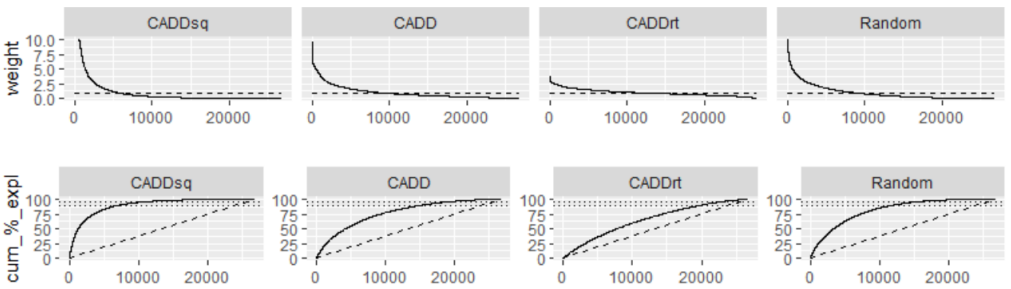

Figure 1. Distribution of weights (upper), and cumulative % of *a priori* explained genetic variances (lower) for CADD and random scores. Dashed lines pertain to scenario with equal weights, and dotted lines pertain to 90 and 95% explained variance.

Table 1. Numbers of records, heritability estimates, split dates for validation, and pedigree-based accuracy for validation animals, on five layers traits.

Trait	No. records	Heritability	CV hatch date	Accuracy$_{(ped)}$
Breaking strength (BS)	2,522	0.26	2014-11-04	0.25
Egg number (EN)	3,400	0.12	2014-11-04	0.11
Egg weight (EW)	5,802	0.75	2016-06-21	0.37
Hatchability (HT)	2,361	0.30	2016-06-21	0.20
Survival (SV)	3,536	0.06	2015-03-24	0.06

Validation set & accuracy. In animal breeding programs the main interest is in accurate breeding values on selection candidates. The forward validation scheme aligns best with this interest. Here, the dataset was split into a training set of 75% oldest animals and a validation set of 25% youngest animals. The patterns and numbers of records on the five traits differ substantially and therefore different dates were used to assign animals to subsets for each trait (Table 1). The validation set allowed a dissection in birthyears to assess impact of connectedness between training and validation animals with youngest animals least connected. Measure for accuracy was the correlation between estimated breeding values and pre-corrected phenotypes (on validation set). All data manipulation was done with dedicated R scripts (R Core Team, 2020).

Results

The weighting scenarios on CADD scores yielded accuracy estimates that were similar to those from Equal scenario (Figure 2) with slightly higher values for 3 traits (BS, HT, and SV). For these traits, the squared CADD scores yielded highest accuracies. Note that for the hatchability trait accuracies were not different from the Random scenario. For other traits the Random scenario always yielded the lowest accuracies. As expected, accuracies dropped for younger animals in the validation set (Figure 3) and these patterns were very similar for all scenarios (some not shown). The rate of drop in accuracy with age was trait dependent.

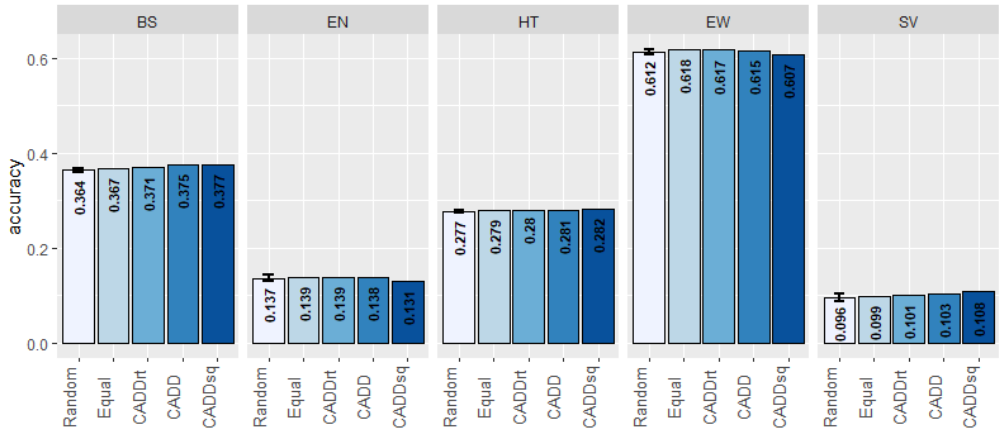

Figure 2. Accuracy of genomic prediction on validation animals for five traits and five weight scenarios (standard error bars are included for Random).

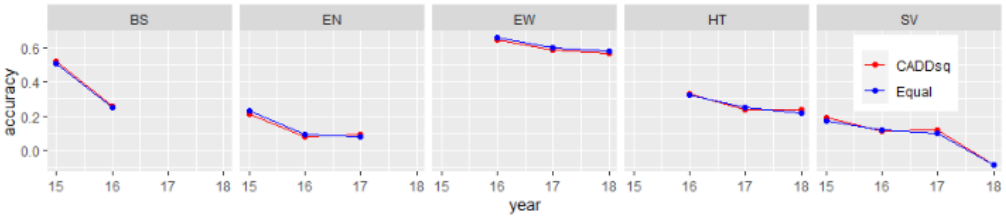

Figure 3. Accuracy estimates versus birthyear of validation animals.

Discussion

We reported on the use of CADD scores to weight SNPs in genomic prediction when using data from a 60K SNP array. These results did not show a consistent added value in predictive ability when using CADD scores to weight SNPs in GBLUP models for genomic prediction. Among the weighting scenarios some trends were observed. The square root CADD scores yielded accuracy closest to the Equal scenario while the squared CADD scores were always most different. This is consistent with the patterns of cumulative percentages explained genetic variance (Figure 1), where the square root CADD scores are most close to the Equal scenario. The trait-dependency of results points to potential differences in trait architecture with some traits highly polygenic and others pre-dominantly driven by fewer genes. This deserves further investigation. With whole genome sequence SNPs, it is known to be important to use a subset for genomic prediction. CADD enables selecting variants with impact, independent from the training data.

Acknowledgements

The GENE-SWitCH project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 817998.

References

Groβ C., Bortoluzzi C., de Ridder D., Megens H.J., Groenen M.A.M. *et al.* (2020) PLoS Genet 1:22. https://doi.org/10.1371/journal.pgen.1009027

Meuwissen T.H.E., Hayes B.J., and Goddard M.E. (2001) Genetics. 157(4):1819–29. https://doi.org/10.1093/genetics/157.4.1819

Misztal I., Legarra A., and Aguilar I. (2014) J Dairy Sci. 97(6):3943–3952. https://doi.org/10.3168/jds.2013-7752

Rentzsch P., Witten D., Cooper G.M., Shendure J., and Kircher M. (2019) Nucleic Acids Res. 47(D1):D886–D894. https://doi.org/10.1093/nar/gky1016

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (www.R-project.org)

VanRaden P.M., (2008) J Dairy Sci 91(11):4414-4423. https://doi.org/10.3168/jds.2007-0980

Vandenplas, J., and Calus M.P.L. (2020). Calc_grm – a program to compute pedigree, genomic, and combined relationship matrices. WUR-ABG, Wageningen Livestock Research.

Xiang R., MacLeod I.M., Daetwyler H.D., de Jong G., O'Conner E. *et al.* (2021) Nature Commun 12(1):860. https://doi.org/10.1038/s41467-021-21001-0