282. On the use of SNPs of large effect to improve prediction accuracy in pigs

M.S. Lopes[1,2*], M. Derks[1,3], M. van Son[4], A.B. Gjuvsland[4], C.A. Sevillano[1], E. Grindflek[4] and E.F. Knol[1]

[1]Topigs Norsvin Research Center, P.O. Box 43, 6640 AA Beuningen, the Netherlands; [2]Topigs Norsvin, Visconde do Rio Branco 1310, 80.420-210 Curitiba, Brazil; [3]Wageningen University & Research, P.O. Box 338, 6700 AH Wageningen, the Netherlands; [4]Norsvin, Storhamargata 44, 2317 Hamar, Norway; marcos.lopes@topigsnorsvin.com

Abstract

The aim of this study was to develop a strategy for selecting SNPs that could receive higher weight than other SNPs in a GBLUP approach due to their expected association with important pig phenotypes. In addition, we aimed to investigate if such a strategy yields improved prediction accuracy compared to a traditional GBLUP approach. Four prediction accuracy scenarios were evaluated using three production traits in two pig populations. Our results show that adding extra weight to SNPs that are expected to be (close to) causal variants in the G matrix increases the prediction accuracy of genomic prediction. The advantage of weighted G matrix compared to a traditional one is not very large, but the added value is consistent.

Introduction

The first practical use of genomic information in animal prediction models was the application of marker-assisted selection (MAS) where markers of large effect were identified in linkage-based studies using microsatellite markers and then applied in selection. The true benefits of genomics, however, became more pronounced only after the development of dense single nucleotide polymorphism (SNP) panels. With SNPs, the association studies moved from linkage-based to genome-wide association studies (GWAS), and MAS was replaced by genomic selection (Meuwissen *et al.*, 2001), which consists of the estimation of genomic breeding values using a large number of SNPs spread across the whole genome. With GWAS, the power to detect new QTL has increased and SNPs tightly-linked to the causal variants have been identified (van Son *et al.*, 2019). However, GWAS findings have not been extensively exploited for selection purposes as the focus has been mainly on genomic selection. Different methods for genomic selection have been developed, including the extensive Bayesian alphabet (Habier *et al.*, 2011) which allows to put more emphasis on SNPs with large effects. For practical application, however, the most used method in pig breeding is the so-called 'single-step' genomic evaluation (Misztal *et al.*, 2009), which can be considered as an extension of the genomic BLUP (GBLUP) assuming that all SNPs contribute equally to the trait (infinitesimal model). Nevertheless, it has been shown that a finite number of genes control quantitative traits (Hayes and Goddard, 2001) and an approach to give different weights to SNPs when building a genomic relationship (G) matrix has been proposed (Zhang *et al.*, 2014). Therefore, further improvement of traditional genomic prediction models could be achieved by weighting SNPs differently in the G matrix if they are, for example, significant in GWAS (Zhang *et al.*, 2014) or are expected to be a causal variant identified using a porcine Combined Annotation Dependent-Depletion (pCADD) tool (Derks *et al.*, 2021). The aim of this study was to develop a strategy for selecting SNPs that could receive higher weight than other SNPs in the G matrix due to their expected association with important pig phenotypes. In addition, we aimed to investigate if such a strategy yields improved prediction accuracy compared to a traditional GBLUP approach.

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

DOI: 10.3920/978-90-8686-940-4_282, © M.S. Lopes *et al.* 2022

Materials & methods

GWAS. Our first step was to perform a GWAS using all traits available in the Topigs Norsvin breeding program. In total, we evaluated 113 traits using five pig populations: Large White, Landrace, Pietrain, Duroc and a Synthetic line. The single-SNP GWAS was ran within population using a single-trait linear animal model in GCTA (Yang *et al.*, 2011). The response variables for most traits were pre-corrected phenotypes for all non-genetic effects using all data available in the routine genetic evaluation of Topigs Norsvin using MiXBLUP (ten Napel *et al.*, 2017). For binary traits, maternal and indirect genetic effects and traits that are an index (e.g. selection index), the response variables were breeding values with a reliability greater than 0.50. Breeding values and reliability estimates were also extracted from the routine evaluation of Topigs Norsvin using MiXBLUP (ten Napel *et al.*, 2017). The number of animals per trait and per line ranged from ~1,300 to ~40,000. Significant association was declared when the SNP presented a *P*-value $<1.0\times10^{-10}$. All significant SNPs located within 5 Mb from another significant SNP were considered to belong to the same QTL region and only the most significant SNP of each QTL region was selected for the further steps.

pCADD. The second step was to select whole-genome sequence variants based on pCADD scores as described in Derks *et al.* (2021). The pCADD score estimates a probability of having a functional (deleterious) impact for each variant. This probability is based on a machine-learning procedure comparing sequence data across species and including annotation data of known functional elements. The score is trait-independent but depends on the quality of the annotation in the region of interest. We selected variants based on their high pCADD score across different annotations as well as variants that are underlying known QTL regions (Derks *et al.*, 2021).

Genotypes. Animals used in the GWAS were genotyped using Illumina Neogen medium-density SNP chips (50K or 80K) (Lincoln, NE, USA). Part of these animals (most influential boars) were also genotyped using the Affymetrix 660K high-density SNP chip (Santa Clara, CA, USA). All animals had their genotypes imputed towards the medium density chip using Fimpute v3 (Sargolzaei *et al.*, 2014) before performing the GWAS. In addition, ~2,000 SNPs were selected from the GWAS and pCADD analyses (2K set) and placed in an Illumina Neogen custom 25K SNP chip together with ~23,000 SNPs (23K set) from the 50K chip that were segregating across all Topigs Norsvin populations or were present on X or Y chromosome. After genotyping over 100,000 animals across all populations with the 25K SNP chip, another round of imputation was performed towards a panel of SNPs that combines all SNPs from the 50K and 25K SNP chips. This new imputed dataset was used for the validation step. Quality control was performed at each level of imputation as described in van Son *et al.* (2019).

Validation. We selected significant SNPs from the GWAS for 113 traits in five populations as well as SNPs based on high pCADD scores which are trait independent. The reason for this is that our final goal is to have a set of SNPs that can be used in multi-trait and multi-population scheme. However, for the validation in this study, we evaluated only the traits backfat (BF) at the end of the test period (~120 kg), average daily gain (DG) and feed intake (FI) during the test period (~25-120 kg) from two populations (Large White and Landrace). From each population we selected the youngest 6,000 animals that had pre-corrected phenotypes for all three traits. The 1000 (very) youngest animals from each population, all genotyped with the 25K SNP chip, were taken as our validation set. The remaining 5,000 animals, genotyped with both 25K and 50K, were taken as our reference (training) set. Four prediction accuracy scenarios were evaluated: 1. 23K, which used a standard G matrix using the same SNPs as the routine genetic evaluation (23K set); 2. 2K, which uses a standard G matrix using only the GWAS and pCADD SNPs (2K set); 3. 25K, which uses a standard G matrix built from a combined set with SNPs from the 23K and 2K sets; 4. w25K, which uses a weighted G matrix using the same SNPs as the 25K scenario. Accuracy of prediction of each scenario was assessed as the correlation between the pre-corrected phenotype and estimated breeding values of the

1000 validation animals from each population. The G matrices of all scenarios were built using calc_grm (Calus and Vandenplas, 2016). To define the weights used in the w25K scenario, we assumed that the 2K set and the 23K set of SNPs can explain the same amount of variance. Therefore, the sum of the weights of the 2K set was equal to the sum of the weights of the 23K set, which means that each 2K SNP has about 20 times heavier weight than the 23K SNPs in the G matrix. We applied the same weights for each SNP independently of the evaluated trait and population. In a standard G matrix, all SNPs have the same weight. Different levels of weights were evaluated (e.g. the sum of the weights of all 2K set was half of the sum of the weights of the 23K set), but the results were similar to those obtained with equal weights and will therefore not be presented.

Results

We selected 260 significant SNPs from the GWAS and 876 with high pCADD scores. The GWAS SNPs explained up to 16% of the phenotypic and 40% of the genetic variance of the evaluated traits. After including them in the custom 25K SNP chip, we obtained 220 GWAS and 784 pCADD SNPs working properly and segregating in both the Large White and the Landrace populations. Therefore, the final 2K set was composed of 1,004 SNPs, while the final 23K set was composed of 21,117 autosomal SNPs also segregating in both populations. The highest prediction accuracies were observed for scenario w25K, except for FI in the Landrace population (Table 1). However, the relative increase in prediction accuracy of the w25K scenario compared to the traditional 23K was not that high, ranging from 0.5 to 4%.

Discussion

Our results show that selecting SNPs that are expected to be (close to) causal variants, including these SNPs in a custom SNP chip and weighting them heavier than standard SNPs in the G matrix might be beneficial to increase the prediction accuracy in the evaluated populations. Although the advantage of w25K scenario compared to the traditional one (23K) is not very large, the added value is consistent. A similar situation has been shown by Khansefid *et al.* (2020) who showed that a custom panel enriched with, or close to, causal mutations yield higher prediction accuracy than their traditional scenario. It is also interesting to observe that using only 1,004 SNP selected from the GWAS or based on pCADD scores (2K scenario), we already obtain high accuracies that are not too far from those obtained with more than 21,000 SNPs (23K scenario). Further, this is a work in progress, and we expect to improve it by investigating other strategies for selecting causal variants and improving the strategy of weighting the SNPs in the G matrix. For practical reasons, we will be looking for strategies that allow improvements in a multi-trait and multi-population scheme. However, working within trait and within population, the added value of a weighted G matrix might be larger due to the differences in allele frequency and linkage disequilibrium across populations.

Table 1. Accuracies of prediction.

Population	Trait	Scenarios[1]			
		23K	**2K**	**25K**	**w25K**
Large White	Backfat	0.474	0.387	0.481	**0.492**
	Average daily gain	0.330	0.228	0.334	**0.340**
	Average feed intake	0.470	0.371	0.474	**0.480**
Landrace	Backfat	0.453	0.412	0.460	**0.473**
	Average daily gain	0.440	0.386	0.444	**0.452**
	Average feed intake	0.522	0.423	**0.525**	0.524

[1] Scenarios: 23K= standard G matrix used in the routine genetic evaluation; 2K= standard G matrix using only the GWAS and pCADD SNPs; 25K= standard G matrix using a combined set with SNPs from 23K and 2K; w25K= weighted G matrix using a combined set with SNPs from 23K and 2K. Numbers in **bold** indicate the highest accuracy of each line.

References

Calus M., and Vandenplas J. (2016) Animal Breeding and Genomics Centre, Wageningen UR Livestock Research.

Derks M.F., Gross C., Lopes M.S., Reinders M.J., Bosse M., *et al.* (2021) Genomics 113(4):2229-2239. https://doi.org/10.1016/j.ygeno.2021.05.017

Habier D., Fernando R.L., Kizilkaya K., and Garrick D.J. (2011). BMC bioinformatics 12(1):186. https://doi.org/10.1186/1471-2105-12-186

Hayes B., and Goddard M.E. (2001). Genetics Selection Evolution 33(3):1-21. https://doi.org/10.1186/1297-9686-33-3-209

Khansefid M., Goddard M.E., Haile-Mariam M., Konstantinov K.V., Schrooten C., *et al.* (2020). Frontiers in genetics 11. https://doi.org/10.3389/fgene.2020.598580

Meuwissen T., Hayes B., and Goddard M. (2001). Genetics 157(4):1819-1829. https://doi.org/10.1093/genetics/157.4.1819

Misztal I., Legarra A., and Aguilar I. (2009). Journal of dairy science 92(9):4648-4655. https://doi.org/10.3168/jds.2009-2064

Sargolzaei M., Chesnais J.P., and Schenkel F.S. (2014). BMC Genomics 15:478. https://doi.org/10.1186/1471-2164-15-478

ten Napel J., Vandenplas J., Lidauer M., Stranden I., Taskinen M., *et al.* (2017) MiXBLUP Manual 202. Available at: https://www.mixblup.eu/download.html

van Son M., Lopes M.S., Martell H.J., Derks M.F., Gangsei L.E., *et al.* (2019) Frontiers in Genetics 10, 272. https://doi.org/10.3389/fgene.2019.00272

Yang J., Lee S.H., Goddard M.E., and Visscher P.M. (2011). American Journal of Human Genetics 88(1):76-82. https://doi.org/10.1016/j.ajhg.2010.11.011

Zhang Z., Ober U., Erbe M., Zhang H., Gao N., *et al.* (2014). PloS one 9(3):e93017. https://doi.org/10.1371/journal.pone.0093017