Genomic prediction using whole-genome sequence data in intensely selected pig lines

R. Ros-Freixedes[1,2*], M. Johnsson[1,3], A. Whalen[1], C.Y. Chen[4], B.D. Valente[4], W.O. Herring[4], G. Gorjanc[1] and J.M. Hickey[1]

[1]The Roslin Institute and Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Easter Bush, Midlothian, Scotland, UK; [2]Departament de Ciència Animal, Universitat de Lleida - Agrotecnio-CERCA Center, Lleida, Spain; [3]Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden; [4]The Pig Improvement Company, Genus plc, Hendersonville, TN, USA; [*]roger.ros@udl.cat

Abstract

We used a large pig dataset to assess the utility of whole-genome sequence data for genomic prediction. We imputed genotypes of 32.8 million variants for 396,100 individuals from seven commercial pig lines (17,224 to 104,661 per line). We used BayesR to perform genomic prediction for 8 real traits and 9 simulated traits with different genetic architectures. Sequence data improved prediction accuracy relative to the marker array, provided that the training set was sufficiently large and especially for traits with high heritability and low number of quantitative trait nucleotides. The most robust results were obtained when variants with statistically significant associations to the trait were preselected and added to the marker array, which yielded average improvements of prediction accuracy of 2.5 and 4.2 percentage points in within-line and multi-line scenarios, respectively, with training sets of around 80k individuals. More robust improvements could be achieved with larger training sets and optimised pipelines.

Introduction

Whole-genome sequence data (WGS) could improve genomic prediction accuracy and its persistence across generations and breeds because it is assumed to contain the causal variants. Early simulations indicated that using causal mutations from such data could increase prediction accuracy by up to 30% if causal variants with low minor allele frequency were captured (Druet *et al.*, 2014). Results with real data have been mixed so far, comprising either no relevant improvement of prediction accuracy compared to marker array data or small, and often non-robust, improvements (e.g., Zhang *et al.*, 2018; Song *et al.*, 2019). One of the strategies to exploit WGS that has been more successful to date consists in augmenting the available marker arrays with preselected variants from WGS based on their association with the trait of interest. In some cases, this strategy improved prediction accuracy by up to 9% (Al Kalaldeh *et al.*, 2019) or 11% (Lopez *et al.*, 2021). Thus, large data sets that capture most of the genome diversity in a population must be assembled so that the allele substitution effects are estimated with high accuracy for millions of variants. We generated WGS for 396,100 pigs from seven different intensely selected lines from diverse genetic backgrounds and with different numerical size. The objectives of this study were to use this dataset to assess the utility of WGS for genomic prediction compared to commercial marker array data and to identify scenarios in which WGS may become beneficial.

Materials & Methods

Populations and sequencing strategy. We performed whole-genome re-sequencing of 6,931 individuals from seven commercial pig lines (Genus PIC, Hendersonville, TN) with a median individual coverage of 1.5x. Approximately 1.5% (0.9 to 2.1% in each line) of the pigs in each line were sequenced. Most pigs in each line were also genotyped with commercial marker arrays (GeneSeek, Lincoln, NE, USA).

Genotype imputation. Genotypes were jointly called, phased, and imputed for a total of 483,353 pedigree-related individuals using the 'hybrid peeling' method implemented in AlphaPeel (Whalen *et al.*, 2018), which used all marker array and WGS that was available. Imputation was performed separately for each line using its complete multi-generational pedigree, which encompassed from 21,129 to 122,753 individuals. Individuals with low predicted imputation accuracy were removed before further analyses according to Ros-Freixedes *et al.* (2020). A total of 396,100 individuals remained, from 17,224 to 104,661 individuals for each line, with an average predicted imputation accuracy (individual-wise dosage correlation) of 0.97 (median: 0.98). We also excluded from the analyses variants with a minor allele frequency lower than 0.023, because their average predicted imputation accuracy (variant-wise dosage correlation) was lower than 0.90. After imputation, 32.8 million variants (14.5 to 19.9 million within each line) remained for downstream analyses, of which 9.9 million segregated across all seven lines.

Traits. We analysed data of 8 traits: average daily gain (ADG), backfat thickness (BFT), loin depth (LD), average daily feed intake (ADFI), feed conversion ratio (FCR), total number of piglets born, litter weight at weaning, and return to oestrus 7 days after weaning. Most pigs with records were born during the 2008–2020 period. Deregressed breeding values (dEBV) were used. To assist in the interpretation of results, we also created 9 simulated traits with different numbers of quantitative trait nucleotides (QTN; 100, 1,000 or 10,000 QTN; sampled randomly across the called variants) and heritability levels (h^2; 0.10, 0.25 or 0.50).

Genomic prediction. For each line, we defined a testing set as the full-sib families from the last generation of the pedigree. The training set was defined as all those individuals that had a pedigree coefficient of relationship lower than 0.5 with any individual of the testing set. This design was chosen to mimic a realistic situation in which breeding companies evaluate the selection candidates available in the selection nucleus at the given time. We tested genomic prediction using variants from the marker array (referred to as 'Chip'; ~40k variants) or for sets of preselected variants from the WGS of a similar size: (i) LDTags: tag variants retained after pruning based on linkage disequilibrium; (ii) Top40k: 40k variants preselected based on genome-wide single-marker regression analyses in the training sets (variants with the lowest p-value in each consecutive non-overlapping 55-kb windows along the genome); (iii) ChipPlusSign: significant variants with ($p \leq 10^{-6}$; only the lowest p-value within each 55-kb window; average per trait: 309, range: 23 to 1083) were preselected and added to Chip; (iv) Functional: variants that were annotated as loss-of-function or missense; and (v) Rand40k: 40k variants chosen randomly. Genomic prediction was performed by fitting a univariate mixed model with BayesR (Erbe *et al.*, 2012). Prediction accuracy was calculated in the testing set as the correlation between the genomic estimated breeding value and the dEBV. Additionally, we considered multi-line (ML) training sets that were conformed by merging the training sets of each line and by performing all analyses adding the line as an effect.

Results

The performance of each set of predictor variants was not robust and differed for each trait and line, sometimes leading to no improvements or losses of prediction accuracy. The most robust results were obtained for ChipPlusSign (Figure 1). The size of the training set was one of the main factors that determined the capacity of variants from the WGS to improve prediction accuracy compared to Chip. The simulation results suggested that genetic architecture of the traits also conditioned the utility of WGS for improving prediction accuracy compared to Chip, as greater improvements were achieved for traits with high heritability and low number of QTN (Figure 2). However, for empirical polygenic traits, average improvements of prediction accuracy of up to 2.5 percentage points were observed with training sets of around 80k individuals. In the most successful scenarios, at least around 200 significant variants were added to Chip. With multi-line training sets, average

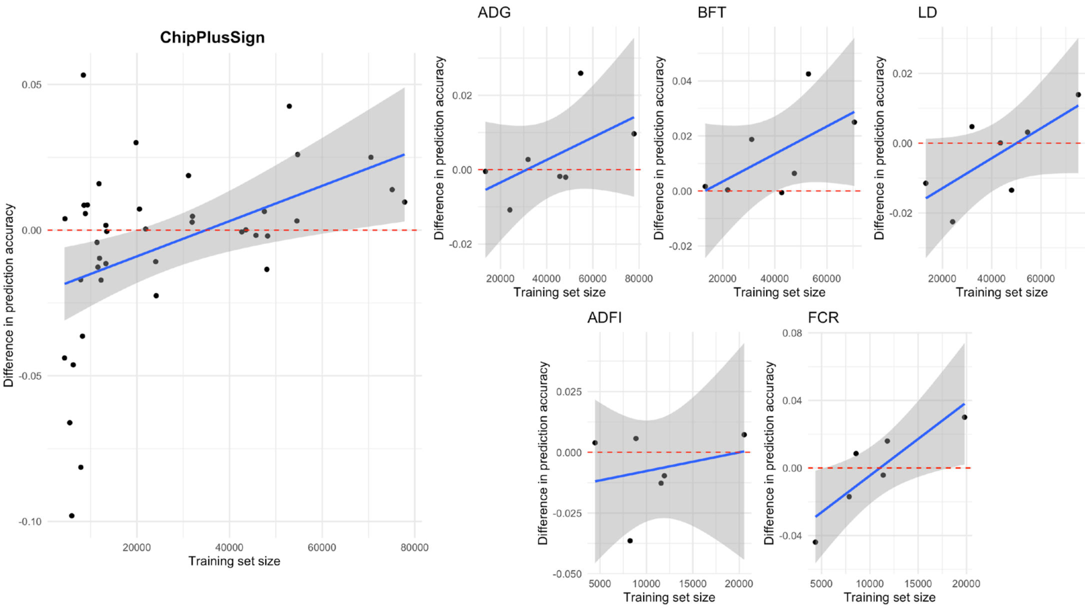

Figure 1. Genomic prediction accuracy with ChipPlusSign for the empirical traits. The difference of prediction accuracy between ChipPlusSign and Chip is shown, for all traits and lines (left) or by trait (right).

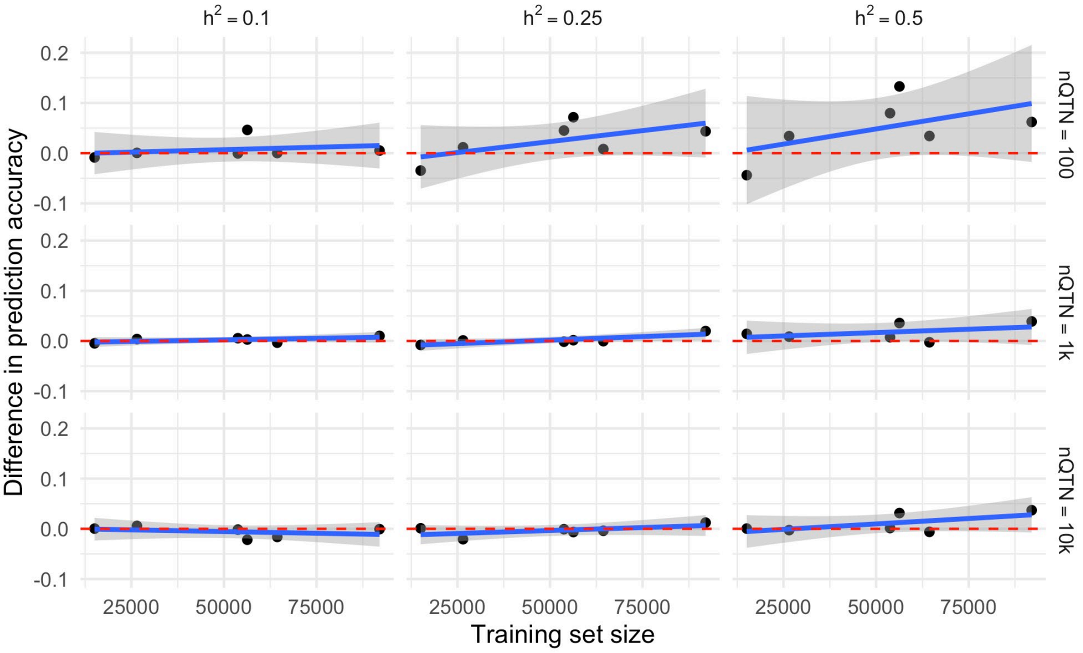

Figure 2. Genomic prediction accuracy with ChipPlusSign for the simulated traits. The difference of prediction accuracy between ChipPlusSign and Chip is shown.

improvements of prediction accuracy with ML-ChipPlusSign were of up to 4.2 percentage points (not shown). Preselecting an entirely new set of predictor variants from WGS in Top40k proved more challenging than ChipPlusSign and did not perform much differently from just taking random variants from these windows, as in Rand40k. Selection based on functional variants and, more severely, LDTags reduced prediction accuracy in many cases.

Discussion

Our results evidenced the potential for WGS to improve genomic prediction accuracy in intensely selected pig lines, provided that the training sets are large enough. Improvements achieved so far were modest. On one hand, these limited improvements indicated that the strategies that we tested were likely suboptimal. On the other hand, the positive trend for the largest training sets indicated that we might have not reached the critical mass of data that is needed to leverage the potential of WGS, especially in scenarios where genomic prediction with marker arrays is already yielding high accuracy.

The modest performance of ChipPlusSign is also a consequence of the difficulty for fine-mapping causal variants through genome-wide association studies with WGS. Even though WGS allows the detection of a very large number of associations, problems such as false positives or p-value inflation also become more severe in a way that added noise might offset the added information. Therefore, WGS performed better with simple genetic architectures (i.e., traits with low number of QTN). This is consistent with expectations and simulation results (Clark *et al.*, 2011).

Multi-line training sets could be particularly beneficial for the use of WGS because they allow a larger training set with low relationship degree between the individuals. Previous simulations suggested that WGS might be the most beneficial with multi-breed reference panels (Iheshiulor *et al.*, 2016), especially for numerically small populations.

References

Al Kalaldeh M., Gibson J., Duijvesteijn N., Daetwyler H.D., MacLeod I., *et al.* (2019) Genet Sel Evol 51:32. 10.1186/s12711-019-0476-4

Clark S.A., Hickey J.M., and van der Werf J.H. (2011) Genet Sel Evol 43:18. 10.1186/1297-9686-43-18

Druet T., MacLeod I.M., and Hayes B.J. (2014) Heredity 112:39-47. 10.1038/hdy.2013.13

Erbe M., Hayes B.J., Matukumalli L.K., Goswami S., Bowman P.J., *et al.* (2012). J Dairy Sci 95:4114-29. 10.3168/jds.2011-5019

Iheshiulor O.O.M., Woolliams J.A., Yu X., Wellmann R., and Meuwissen T.H.E. (2016). Genet Sel Evol 48:15. 10.1186/s12711-016-0193-1

Lopez B.I.M., An N., Srikanth K., Lee S., Oh J.D., *et al.* (2021). Front Genet 11:603822. 10.3389/fgene.2020.603822

Ros-Freixedes R., Whalen A., Chen C.Y., Gorjanc G., Herring W.O., *et al.* (2020) Genet Sel Evol 52:17. 10.1186/s12711-020-00536-8

Song H., Ye S., Jiang Y, Zhang Z., Zhang Q., *et al.* (2019) Genet Sel Evol 51:58. 10.1186/s12711-019-0500-8

Whalen A., Ros-Freixedes R., Wilson D.L., Gorjanc G. and Hickey J.M. (2018) Genet Sel Evol 50:67. 10.1186/s12711-018-0438-2

Zhang C., Kemp R.A., Stothard P., Wang Z., Boddicker N., *et al.* (2018). Genet Sel Evol 50:14. 10.1186/s12711-018-0387-9