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

Citation for published version:
Ros-Freixedes, R, Johnsson, M, Chen, C-Y, Valente, BD, Herring, WO, Gorjanc, G & Hickey, JM 2022 'Genomic prediction with whole-genome sequence data in intensely selected pig lines' bioRxiv, bioRxiv. https://doi.org/10.1101/2022.02.02.478838

Digital Object Identifier (DOI):
10.1101/2022.02.02.478838

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

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

Roger Ros-Freixedes[1,2§], Martin Johnsson[1,3], Andrew Whalen[1], Ching-Yi Chen[4], Bruno D Valente[4], William O Herring[4], Gregor Gorjanc[1], John 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.

[§]Corresponding author: RRF roger.ros@roslin.ed.ac.uk

Abstract

Background

Early simulations indicated that whole-genome sequence data (WGS) could improve prediction accuracy and its persistence across generations and breeds. However, results in real datasets have been ambiguous so far. Large data sets that capture most of the genome diversity in a population must be assembled so that allele substitution effects are estimated with higher accuracy. The objectives of this study were to use a large pig dataset to assess the benefits of using WGS for genomic prediction compared to using commercial marker arrays, to identify scenarios in which WGS provides the largest advantage, and to identify potential pitfalls for its effective implementation.

Methods

We sequenced 6,931 individuals from seven commercial pig lines with different numerical size. Genotypes of 32.8 million variants were imputed for 396,100 individuals (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. Genomic predictions were performed using either data from a marker array or variants preselected from WGS based on linkage disequilibrium, functional annotation, or association tests. Both single and multi-line training sets were explored.

Results

Using WGS improved prediction accuracy relative to the marker array, provided that training sets were sufficiently large, especially for traits with high heritability and low number of quantitative trait nucleotides. The performance of each set of predictor variants was not robust across traits and lines. The most robust results were obtained when preselected variants with statistically significant associations were added to the marker array. Under this method, average improvements of prediction accuracy of 2.5 and 4.2 percentage points were observed in within-line and multi-line scenarios, respectively, with training sets of around 80k individuals.

Conclusions

Our results evidenced the potential for WGS to improve genomic prediction accuracy in intensely selected pig lines. Although the prediction accuracy improvements achieved so far were modest at best, we would expect that more robust improvements could be attained with a combination of larger training sets and optimised pipelines.

Introduction

Whole-genome sequence data (WGS) has the potential to empower the identification of causal variants that underlie quantitative traits or diseases [1–4], increase the precision and scope of population genetic studies [5,6], and enhance livestock breeding. Genomic prediction has been successfully implemented in the main livestock species and it has increased the rate of genetic gain [7]. Genomic prediction has provided many benefits such as greater accuracies of genetic evaluations and the reduction of the generational interval in dairy cattle. However, since its early implementations, genomic prediction is typically performed using marker arrays that capture the effects of the (usually unknown) causal variants via linkage disequilibrium. Alternatively, WGS are assumed to contain the causal variants themselves. For this reason, it was hypothesized that such data could further improve prediction accuracy and its persistence across generations and breeds. Early simulations indicated that causal mutations from WGS could increase prediction accuracy [8–13]. One simulation study indicated that the magnitude of prediction accuracy improvement relative to dense marker arrays ranged from 2.5 to 3.7%, with a persistence of over 10 generations [11]. Another one reached improvements of 30% if causal variants with low minor allele frequency could be captured by the WGS [9]. However, benefits could be on the lower end of that range in standard livestock populations due to small effective population sizes and long-term negative selection [10].

During the last few years, there have been several attempts at improving the accuracy of genomic prediction with the use of WGS in the main livestock species. Results have been ambiguous so far. When predicting genomic breeding values within breed or line, some studies found no relevant improvement of prediction accuracy for

WGS compared to marker arrays [14–18]. Other studies found small, and often unstable, improvements (e.g., from 1 to 5% or no improvement depending on prediction method [19–21], or trait-dependent results [21,22]). When predicting genomic breeding values across populations, the identification of causal variants from WGS can improve prediction accuracy [23–26], especially for small populations where initial prediction accuracy was low or that were not included in the training population [23,25–28].

One of the most successful strategies to exploit WGS consists in augmenting available marker arrays with preselected variants from WGS based on their association with the trait of interest [29–32]. In some cases, this strategy improved prediction accuracy by up to 9% [31] and 11% [32]. However, it did not improve prediction accuracies in other within-line scenarios [16]. Nevertheless, this shows how identifying causal variants could enhance genomic prediction with WGS. Whole-genome sequence data has already been applied in genome-wide association studies (GWAS) to identify variants associated to a variety of traits in livestock [2,33–35], including pigs [36,37]. However, the fine-mapping of causal variants remains challenging due to the pervasive long-range linkage disequilibrium across extremely dense variation.

High accuracy in estimating allele substitution effects and, ideally, the identification of causal variants amongst millions of other variants are important for the usefulness of WGS in research and breeding. This requires large data sets able to capture most of the genome diversity in a population. Despite that low-cost sequencing strategies have been developed, which typically involve sequencing a subset of the individuals in a population at low coverage and then imputing WGS for the remaining individuals [38–40], the cost of generating accurate WGS at this scale, as well as the large computational requirements for the analyses of such datasets, have limited the population sizes or number of populations tested in some of the previous studies. This hinders the interpretation of results across studies, which are very diverse in population structures, sequencing strategies and prediction methodologies used. The largest studies on the use of WGS for genomic prediction to date have been performed in cattle, for which large multi-breed reference panels are available from the 1000 Bull Genomes Project [2,19,33]. This has enabled the imputation of WGS for cattle populations. The lack of such available reference panels has been cited as an important limiting factor for performing similar studies in other species, such as pigs [36].

We have previously described our approach to impute WGS in large pedigreed populations without the need for haplotype phased reference panels [41]. Following that strategy, we generated WGS for 396,100 pigs from seven intensely selected lines with diverse genetic backgrounds and numerical size. The objectives of this study were to use this large pig dataset to assess the benefits of using WGS for genomic prediction compared to using commercial marker arrays, to identify scenarios in which WGS provides the largest advantage, and to identify potential pitfalls for its effective implementation.

Materials and 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 total coverage of approximately 27,243x. Sequencing effort in each of the seven lines was proportional to population size. Approximately 1.5% (0.9 to 2.1% in each line) of the pigs in each line were sequenced. Most pigs were sequenced at low coverage, with target coverage of 1 or 2x, but a subset of pigs were sequenced at higher coverage of 5, 15, or 30x. Thus, the average individual coverage was 3.9x, but the median coverage was 1.5x. The number of pigs sequenced and at which coverage for each line is summarized in Table 1.

The sequenced pigs and their coverage were selected following a three-part sequencing strategy developed to represent the haplotype diversity in each line. First (1), sires and dams with the highest number of genotyped progeny were sequenced at 2x and 1x, respectively. Sires were sequenced at a greater coverage because they contributed with more progeny than dams. Then (2), the individuals with the greatest genetic footprint on the population (i.e., those that carry more of the most common haplotypes) and their immediate ancestors were sequenced at a coverage between 1x and 30x (AlphaSeqOpt part 1; [42]). The sequencing coverage was allocated with an algorithm that maximises the expected phasing accuracy of the common haplotypes from the cumulated family information. Finally (3), pigs that carried haplotypes with low cumulated coverage (below 10x) were sequenced at 1x (AlphaSeqOpt part 2; [43]). Sets (2) and (3) were based on haplotypes inferred from marker array genotypes (GGP-Porcine HD BeadChip; GeneSeek, Lincoln, NE), which were phased and imputed using AlphaPhase [44] and AlphaImpute [45].

Most sequenced pigs and their relatives were also genotyped either at low density (15k markers) using the GGP-Porcine LD BeadChip (GeneSeek) or at high density (80k markers) using the GGP-Porcine HD BeadChip (GeneSeek). Quality control of the marker array data was based on the individuals genotyped at high density. Markers with minor allele frequency below 0.01, call rate below 0.80, or that failed the Hardy-Weinberg equilibrium test were removed. After quality control, 38,634 to 43,966 markers remained in each line.

Sequencing and data processing

Tissue samples were collected from ear punches or tail clippings. Genomic DNA was extracted using Qiagen DNeasy 96 Blood & Tissue kits (Qiagen Ltd., Mississauga, ON, Canada). Paired-end library preparation was conducted using the TruSeq DNA PCR-free protocol (Illumina, San Diego, CA). Libraries for resequencing at low coverage (1 to 5x) were produced with an average insert size of 350 bp and sequenced on a HiSeq 4000 instrument (Illumina). Libraries for resequencing at high coverage (15 or 30x) were produced with an average insert size of 550 bp and sequenced on a HiSeq X instrument (Illumina). All libraries were sequenced at Edinburgh Genomics (Edinburgh Genomics, University of Edinburgh, Edinburgh, UK).

DNA sequence reads were pre-processed using Trimmomatic [46] to remove adapter sequences from the reads. The reads were then aligned to the reference genome *Sscrofa11.1* (GenBank accession: GCA_000003025.6) using the BWA-MEM algorithm [47]. Duplicates were marked with Picard (http://broadinstitute.github.io/picard). Single nucleotide polymorphisms (SNPs) and short insertions and deletions (indels) were identified with the variant caller GATK HaplotypeCaller (GATK 3.8.0) [48,49] using default settings. Variant discovery with GATK HaplotypeCaller was performed separately for each individual and then a joint variant set for all the individuals in each population was obtained by extracting the variant positions from all the individuals.

We extracted the read counts supporting each allele directly from the aligned reads stored in the BAM files using a pile-up function to avoid biases towards the reference allele introduced by GATK when applied on low-coverage WGS [50]. That pipeline uses the tool pysam (version 0.13.0; https://github.com/pysam-developers/pysam), which is a wrapper around htslib and the samtools package [51]. We extracted the read counts for all biallelic variant positions, after filtering variants in potential repetitive regions (defined as variants that had mean depth values 3 times greater than the average realized coverage) with VCFtools [52]. This amounted to a total of 55.6 million SNP (19.6 to 31.1 million within each line) and 10.2 million indels (4.1 to 5.6 million within each line). A more complete description of the variation across the lines is provided in [53].

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 [54,55]. This method used all the available marker array and WGS. Imputation was performed separately for each line using complete multi-generational pedigrees, which encompassed from 21,129 to 122,753 individuals each (Table 1). We have previously published reports on the accuracy of imputation in the same populations using this method [41]. The estimated average individual-wise dosage correlation was 0.94 (median: 0.97). Individuals with low predicted imputation accuracy were removed before further analyses. An individual was predicted to have low imputation accuracy if itself or all of its grandparents were not genotyped with a marker array or if it had a low degree of connectedness to the rest of the population. These criteria were based on the analysis of simulated and real data on imputation accuracy [41]. A total of 396,100 individuals remained, with each line comprising between 17,224 and 104,661 individuals (Table 1). The expected average individual-wise dosage correlation of the remaining individuals was 0.97 (median: 0.98) according to our previous estimates. We also excluded from the analyses variants with a minor allele frequency lower than 0.023, as their estimated variant-wise dosage correlations was lower than 0.90 [41]. After imputation, 32.8 million variants (14.5 to 19.9 million within each line) remained for downstream analyses, out of which 9.9 million segregated across all seven lines.

Traits

We analysed data of 8 traits that are commonly included in selection objectives of pig breeding programmes: average daily gain (ADG, g), backfat thickness (BFT, mm), loin depth (LD, mm), average daily feed intake (ADFI, kg), feed conversion ratio (FCR), total number of piglets born (TNB), litter weight at weaning (LWW, kg), and return to oestrus 7 days after weaning (RET, binary trait). Most pigs with records were born during the 2008–2020 period. Breeding values were estimated by line with a linear mixed model that included polygenic and non-genetic (as relevant for each trait) effects. Deregressed breeding values (dEBV) were obtained following the method by VanRaden and Wiggans [56]. Only individuals in which the trait was directly measured were retained for further analyses. The number of records for each trait used in the analyses of each line is detailed in Table 2.

Simulated traits

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) and heritability levels (h^2; 0.10, 0.25 or 0.50). Positions of the QTN were sampled randomly amongst all variants called across all lines. Because QTN were sampled from all variants, some QTN were fixed in some of the lines while segregating in others. There were only negligible differences in the number of segregating QTN per line (53 to 61, 531 to 583, or 5375 to 6058, respectively). Marker effects of the QTN were sampled from a gamma distribution with shape=2 and scale=5. After a polygenic term was calculated for each individual using these marker effects, residual terms were sampled from a normal distribution with a variance parameter adjusted to produce the desired heritability level. The number of records for the simulated traits is detailed in Table 2. In these simulations, we used the imputed genotypes as real genotypes and, therefore, implicitly cancelled any errors that might arise from the processing of the sequencing reads and genotype imputation.

Training and testing sets

We split the individuals in each population into training and testing sets. The testing sets were defined as those individuals from full-sib families from the last generation of the pedigree (i.e., individuals that did not have any progeny of their own). Only families with a minimum of 5 full-sibs were considered. 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 any given time.

Genome-wide association study

To assess whether variants from the WGS could provide a finer mapping of causal variants than marker array data, and to provide an association-based criterion to preselect variants for the genomic prediction tests, we performed a GWAS for each trait and line. This step included only the individuals in the training set. We fitted a univariate linear mixed model that accounted for the genomic relationship matrix as:

$$\mathbf{y} = \mathbf{x}_i \beta_i + \mathbf{u} + \mathbf{e},$$

where $\mathbf{y}$ is the vector of dEBV, $\mathbf{x}_i$ is the vector of genotypes for the ith SNP coded as 0 and 2 if homozygous for either allele or 1 if heterozygous, β_i is the additive effect of the ith SNP on the trait, $\mathbf{u} \sim N(0, \sigma_{\mathrm{u}}^2 \mathbf{K})$ is the vector of polygenic effects with the covariance matrix equal to the product of the polygenic additive variance σ_{u}^2 and a genomic relationship matrix $\mathbf{K}$, and $\mathbf{e}$ is a vector of uncorrelated residuals. Due to computational limitations, the genomic relationship matrix $\mathbf{K}$ was calculated using only imputed SNP genotypes in the marker array regardless of whether the association study involves the SNPs in the marker array or the variants in WGS. We used the FastLMM software [57,58] to fit the model.

We used the same p-value threshold ($p<10^{-6}$) for both marker array and for sequence associations, because while the WGS contains many more variants, they are also expected to be in higher linkage disequilibrium. This threshold was based on Bonferroni's multiple test correction assuming that the markers from the marker array were independent. For the simulated traits, we defined genomic regions that contained significant associations and assessed whether or not they contained a QTN. These regions were defined by overlapping 500-kb segments centered on the significant markers.

Genomic prediction in within-line scenarios

To test whether variants from the WGS could provide greater prediction accuracy than the marker array, we tested genomic prediction using variants from the marker array, from the WGS, or combining them. The marker array data (referred to as ‘Chip’) was set as the benchmark for prediction accuracy. It contained all ~40k variants in the marker array. For the sequence-based predictors, we preselected sets of variants because currently available methods for genomic prediction are not yet capable of handling datasets as large as the complete WGS. We tested different alternative strategies for preselecting the predictor variants:

- *LDTags.* Tag variants retained after pruning based on linkage disequilibrium. Variants were removed so that no pairs of SNPs with $r^2>0.1$ remained in any 10-Mb window (windows slid by 2,000 variants) using Plink 1.9 [59]. The number of predictor variants preselected by this method was on average of 30k variants (range: 5k to 80k).
- *Top40k.* Variants preselected based on GWAS analyses. To mimic the number of variants in Chip, we preselected the variants with the lowest p-value (not necessarily below the significance threshold) in each of consecutive non-overlapping 55-kb windows along the genome. In addition, to test the impact of variant density on prediction accuracy, we preselected 10k, 25k, 75k, or 100k predictor variants following the same criterion.
- *ChipPlusSign.* Variants preselected based on GWAS analyses as in Top40k, but only significant variants ($p\leq10^{-6}$) were preselected and merged with those in Chip. When a 55-kb window contained more than one significant variant, only that with the lowest p-value was selected as a proxy, in order to reduce the

preselection of multiple significant SNPs tagging the same causal variant. On average, 309 significant variants were identified per trait and line (range: 23 to 1083; Table 3). These significant variants were merged with those in Chip.

- *Functional.* Variants that were annotated as loss-of-function or missense according to Ensembl Variant Effect Predictor (Ensembl VEP; version 97, July 2019) [60]. The most severe predicted consequence type for each variant was retrieved. The number of predictor variants preselected by this method was on average of 35k variants (range: 27k to 40k).
- *Rand40k.* The same number of predictor variants as in Chip, chosen randomly.

Genomic prediction was performed by fitting a univariate model with BayesR [61,62], with a mixture of normal distributions as the prior for variant effects, including one distribution that sets the variant effects to zero. The model was:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

where $\mathbf{y}$ is the vector of dEBV, $\mathbf{1}$ is a vector of ones, μ is the general mean, $\mathbf{X}$ is a matrix of genotypes, $\boldsymbol{\beta}$ is a vector of variant effects, and $\mathbf{e}$ is a vector of uncorrelated residuals. The prior variance of the variant effects in $\boldsymbol{\beta}$ had four components with variances $\sigma_1^2 = 0$, $\sigma_2^2 = 0.0001\sigma_g^2$, $\sigma_3^2 = 0.001\sigma_g^2$, or $\sigma_4^2 = 0.01\sigma_g^2$, where σ_g^2 is the total genetic variance. We used a uniform and almost uninformative prior for the mixture distribution. We used a publicly available implementation of BayesR (https://github.com/syntheke/bayesR; accessed on 30 April 2021), with default settings. Prediction accuracy was calculated in the testing set as the correlation between the genomic estimated breeding value and the dEBV. Bias of the prediction accuracy was calculated as the regression coefficient of the dEBV on the genomic estimated breeding values.

It has been noted that using the same reference individuals for preselecting variants through GWAS and for training the predictive equation can reduce prediction accuracy and bias the predicted breeding values [16,63]. To account for that, we reanalysed some of the scenarios after splitting the training set into two exclusive subsets, one for GWAS to preselect the predictor variants and one for training the predictive equation. The GWAS subset was defined by randomly selecting either 10% or 50% of the individuals in the original training set. Those individuals were excluded from the subset used for training the predictive equation afterwards.

Genomic prediction in multi-line scenarios

We considered multi-line scenarios in which the training set consisted of merging the training sets that had been defined for each line. All analyses were performed as for the within-line scenarios but with a line effect. In the multi-line scenarios, all SNPs from the marker array that passed quality control and were imputed for at least one line were included in the baseline (referred to as 'ML-Chip'). For ease of computation, the strategies for preselection of predictor variants from WGS were applied only to the subset of 9.9 million variants that had been called and imputed in all seven lines. Thus, we defined the predictor sets 'ML-Top40k' and 'ML-ChipPlusSign' by preselecting variants following the same criteria as in within-line scenarios, but using a multi-line GWAS analyses with line effect instead. For ML-ChipPlusSign, 60 to 7247 significant variants were identified per trait (Table 3) and merged with those in ML-Chip. For comparison purposes, prediction accuracy was calculated for the testing set of each individual line.

Results

Prediction accuracy within line

Whole-genome sequence data can improve prediction accuracy of marker array data when there is a sufficiently large training set and if an appropriate set of predictor variants is preselected. Figure 1 shows the prediction accuracy for the case with the largest training set using different sets of predictor variants. In this case, all tested sets of variants from the WGS, except for LDTags, yielded increases of prediction accuracy that ranged from +2.0% to +9.2%. Using WGS also reduced bias relative to Chip in some scenarios. However, the performance across predictor variants set was not robust for the most part, and differed for each trait and line (Additional File 1), often leading to no improvements of prediction accuracy or even reduced prediction accuracy relative to Chip. One stable feature of the results was LDTags showing a noticeable decrease in prediction accuracy in most traits and lines.

The size of the training set was one of the main factors that determined the capacity of predictor variants from the WGS to improve the baseline prediction accuracy of Chip. Figures 2 and 3 show the difference in prediction accuracy of Top40k and ChipPlusSign with respect to the baseline of Chip against the number of phenotypic records available in the training set. We observed large variability for the difference in prediction accuracy, especially when the training set was small. This variability was larger in Top40k than in ChipPlusSign, in a way that shrinkage of variation as the training set was larger was more noticeable in ChipPlusSign. Gains in prediction accuracy were low-to-moderate in the most favourable cases. In the most unfavourable ones we observed large losses in prediction accuracy for Top40k but more restrained losses for ChipPlusSign with moderate training set sizes. For both sets of predictor variants, there was a positive trend that supported the need for large training sets. This trend was clearer in ChipPlusSign than in Top40k, because of the apparent lower robustness of the latter. Results for the other sets of predictor variants are provided in Additional File 2.

The genetic architecture of the traits was also related to the success of WGS for improving prediction accuracy. As the true genetic architecture of real complex traits is mostly unknown, we used simulated traits to show that traits with high heritability and low number of QTN were more likely to show larger improvements in predictive performances. With Top40k (Figure 4), heritability seemed to be the main factor that affected the expected improvement with large training sets (from null improvements when h^2=0.1 to improvements of approximately 0.05 when h^2=0.5, regardless of number of QTN, with a training set of 92k individuals). With ChipPlusSign (Figure 5), the expected improvements with the same training set (92k individuals) were not only greater in magnitude but depended on both heritability and number of QTN (from null improvements when h^2=0.1 to improvements of approximately 0.03 to 0.10 when h^2=0.5 with a number of 100 to 10k QTN, respectively). Results confirmed the trends observed for the real traits (Figures 4 and 5); for instance, the higher robustness of ChipPlusSign compared to Top40k.

We observed diminishing returns when we increased the density of the predictor variants. Increasing the number of predictor variants from the 40k in Top40k to 75k selected in the same way yielded small improvements in prediction accuracy compared to Top40k, but increases up to 100k variants provided smaller or null additional gains (Additional File 3).

Splitting the original training set into two exclusive subsets, one for the GWAS-based preselection of the variants and one for the training of the predictive equation did not improve the prediction accuracy (Additional File 4). For

ChipPlusSign, this strategy reduced the bias but prediction accuracy decreased too, probably because of the smaller subset available for training the predictive equation.

Prediction accuracy in multi-line scenarios

The performance of genomic predictors trained with multi-line datasets was systematically lower than in the within-line scenarios (Additional File 5). Nonetheless, the ML-ChipPlusSign predictor variants in general increased prediction accuracy relative to ML-Chip (Figure 6). The increase in genomic prediction accuracy for each line was largely dependent on the number of individuals of each line in the training set. Therefore, the greatest improvements were achieved for the largest lines. However, in the multi-line scenarios we observed increases of prediction accuracy for some traits and lines for which no improvements were observed in the within-line scenarios (Figure 7). In contrast, results for ML-Top40k were not robust (Additional File 6).

Association tests

First, we assessed the performance of GWAS using the simulated traits. Table 4 shows the number of regions with significant associations that were detected using either Chip or WGS, and whether they contained zero, one or multiple true QTN. The WGS allowed the detection of a much larger proportion of true QTN than the Chip, especially for the traits with high heritability and with large population sizes. The most favourable scenarios for identifying regions that contained unequivocally a single QTN with WGS were those in which the trait was controlled by a low number of true QTN. However, even though the genetic architecture was very simple and consisted of additive effects alone, the regions with significant associations only captured a small fraction of the QTN that segregated within each line. Moreover, using WGS also increased the number of regions with significant associations that contained no QTN, which could therefore be considered as false positives. Some of the selected regions contained multiple QTN, which could indicate either a 'hit by chance' or an inability to disentangle multiple causal variants. While false positives also occur with Chip, their incidence was more severe with the WGS, especially for traits with a large number of QTN. Large population sizes further aggravated the inflation of genome-wide p-values.

Despite this, with the real traits we found that GWAS using WGS can contribute to a better understanding of the genetic mechanisms that underlie the traits of interest. To illustrate this, we examined the GWAS results for BFT in line A, for which a large number of phenotypic records were available. Figure 8 shows the results for chromosome 1 as an example, while Additional File 7 shows the results for six genomic regions of interest. The main genomic regions and candidate genes associated to BFT detected with Chip in the same genetic lines studied here were reported elsewhere [64]. We will use the candidate genes reported there to refer to the genomic regions with significant associations. Using Chip, we identified 6 genomic regions ($p<10^{-6}$). Using WGS (with a more stringent significance threshold of $p<10^{-9}$ to focus on the most significant associations), we confirmed 3 of these genomic regions that co-located to candidate genes *MC4R*, *DOLK*, and *DGKI* or *PTN*. However, the most associated variants in each of these genomic regions located outside the coding region of these putative causal genes. These signals sometimes had very strong evidence of association for some variants that were relatively distant from our candidate functional gene, which could cast doubts about the fine-mapping of the causal mutation. The region at SSC18, 9–13 Mb, contained two candidate genes

DGKI and *PTN*, but the WGS revealed significantly associated variants within *DGKI* and none within *PTN*, despite that the strongest associations were away from both genes at 10.5-11 Mb. Using the WGS we also detected 24 additional genomic regions that contained candidate genes such as *CYB5R4*, *IGF2*, and *LEPR*. These genes were previously detected in other lines using the Chip but not in this one [64], sometimes because there were no markers for the associated region in Chip (SSC2, 0–4 Mb). The region at SSC1, 52.5–53.5 Mb, showed many significant variants that encompassed not only the previously identified candidate gene *CYB5R4*, but also *MRAP2* (annotated with functions on feeding behavior and energy homeostasis). In contrast, candidate gene *LEPR* was located within the region at SSC6, 146.5–147.0 Mb, where many significant variants were located, although the most significant variants were not in the coding regions of the gene. Using the WGS we also identified additional candidate genes that had not been previously detected in any of the lines, such as *CYP24A1* (annotated with functions on fatty acid omega-oxidation and vitamin D metabolism; not shown). For many of the other genomic regions, it was difficult to pinpoint a candidate gene with the available information or there were no annotated genes.

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 at best. On one hand, these modest 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 results from several traits and lines with different training set sizes and the use of simulated phenotypes allowed us to identify the most favourable scenarios for genomic prediction with WGS. We will discuss (1) the prediction accuracy that we achieved with WGS compared to commercial marker array data and the scenarios in which WGS may become beneficial, and (2) the potential pitfalls for its effective implementation and the need for an optimised strategy.

Prediction accuracy with whole-genome sequence data

We compared the genomic prediction accuracy of the current marker array (Chip) with sets of preselected sequence variants in a way that the number of variants remained similar across sets. Improvements of prediction accuracy can be limited if current marker arrays are already sufficiently dense to capture a large proportion of the genetic variance in intensely selected livestock populations. These populations typically have small effective population size [10,19]. Nevertheless, modest improvements have been achieved under certain scenarios. In our study, the most robust results were obtained for the ChipPlusSign set, where variants that showed statistically significant associations to the trait were preselected and added to the information from the marker array. This is consistent with previous reports that showed an improvement of prediction accuracy under similar approaches [29–32]. We added 23 to 1083 significant variants to those in Chip in different scenarios. In the most successful ones, at least around 200 significant variants were added and average improvements of prediction accuracy of 2.5 percentage points were observed with training sets of around 80k individuals. In other instances, however, additions of a larger number of variants have been proposed. The addition of 1623 variants (preselected as the combination of 3-5 variants for each of the top QTL per trait and breed) to a 50k array increased prediction reliability (accuracy squared) by up to 5 percentage points in Nordic cattle [29]. Adding the 16k SNPs with largest estimated effects to a 60k array increased prediction reliability on average by 2.7 (up to 4.8) percentage points in Holstein cattle [30]. For the custom 50k array for Hanwoo cattle, it has been reported that adding at least around 12k SNPs (3k for each of four traits) improved prediction accuracy by up to ~6 percentage points [32]. The addition of ~400 variants preselected by GWAS with regional heritability mapping to a 50k array increased prediction accuracy by 9 percentage points in sheep [31]. In other cases in Nordic cattle, however, the addition of ~1500 variants preselected by GWAS to a 54k panel produced negligible improvements in the prediction of traits with low heritability [65].

Preselecting an entirely new set of predictor variants from WGS, as in Top40k, proved more challenging than ChipPlusSign. In Top40k, we preselected the variants with the lowest p-value in each of consecutive non-overlapping 55-kb windows along the genome. This strategy did not perform much differently from just taking random variants from these windows, as in Rand40k. One possible reason for these results is that at this variant density, random variants effectively tag the same associations as Top40k thanks to linkage disequilibrium. Denser sets of predictor variants provided only small further improvements of prediction accuracy with diminishing returns.

The modest performance of ChipPlusSign and Top40k could also be a consequence of the difficulty for fine-mapping causal variants through GWAS with WGS. Theoretically, the identification of all causal variants associated with a trait should enable the improvement of prediction accuracy [12]. 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 detected signal. For instance, results in cattle showed that GWAS with WGS did not detect clearer associated regions relative to marker arrays and failed to capture QTL for genomic prediction [14], as the effect of potential QTL were spread across multiple variants. Therefore, WGS performed better with simple genetic architectures (i.e., traits with low number of QTN). This is consistent with expectations and simulation results [8] that indicated that the benefit of using WGS for genomic prediction would be limited by the number and size of QTN. When there are many QTN with small effects it becomes much more difficult to properly estimate their effects accurately. Therefore, for largely polygenic traits (as most traits of interest in livestock production), training sets need to be very large before WGS can increase prediction accuracy [8].

The advantage of using WGS might be limited by the current training set sizes, especially in scenarios where marker arrays are already yielding high prediction accuracy [14,20]. Multi-line training sets could be particularly beneficial with the use of WGS because they allow a larger training set with low pairwise relationship degree among individuals. Previous simulations suggested that WGS might be the most beneficial with multi-breed reference panels [66], especially for numerically small populations. Our results with a multi-line training set indicated that WGS can improve prediction accuracy in scenarios that are less optimised than within-line genomic prediction. The average improvements of prediction accuracy of 4.2 percentage points were observed for the populations that contributed around 80k individuals to the training set. However, in general those predictions were still less accurate than using variants preselected under within-line training sets. In our multi-line scenarios we only used variation that segregated across all seven lines. We observed that population-specific variation accounted only for small fractions of genetic variance [53] and it seems unlikely that they would contribute much to prediction accuracy across breeds. Another possible obstacle is the differences in the allele substitution effects of the causal mutations across breeds. This can be caused by differences in allele frequency, contributions of non-additive effects and different genetic backgrounds, or even gene-by-environment interactions among others [24,67].

We observed low robustness of genomic prediction with WGS across traits and lines, and drops in prediction accuracy in those scenarios where genomic prediction with WGS failed. Regarding bias, we did not observe a systematic increase for ChipPlusSign despite using the same individuals for variant preselection and for training the predictors [16,63]. When we split the training set into two subsets, one for GWAS-based variant preselection and the other for training of the predictive equations, we did not observe any improvement in accuracy or bias. One hypothesis is that both subsets belong to the same population and therefore retained similar inter-relationship degrees (i.e., they are not strictly independent sets of individuals). Moreover, the reduction in individuals available for training the predictors negatively affected prediction accuracy.

We did not directly test persistence of prediction accuracy, but previous studies with real data found no higher persistence of prediction accuracy for WGS, not even with low degree of relationship between training and testing sets [14]. We would expect such obstacles to persistence of accuracy until causal variants can be successfully identified.

Suboptimal strategy and pitfalls

The use of WGS for genomic prediction can only be reached after many other steps are completed to produce the genotypes at whole-genome level. Each of these steps has its pitfalls. It is unavoidable that the success of using of WGS is sensitive not only to the prediction methodology itself but also to the strategy followed until genotyping. This strategy includes the choice of which individuals to sequence, the bioinformatics pipeline to call variants, the imputation of the WGS and choice of variant filters. When combined with the multiplicity of prediction methods and the preselection of predictor variants (which is unavoidable with current datasets, predictive methodologies and computational capacities), there are many options and variables in the whole process that can affect the final result and that are not yet well understood. Therefore, a much greater effort for optimising such strategies is required. Here we tested relatively simple approaches to see how they performed with large WGS datasets. We have discussed what in our opinion are the main pitfalls of our approach for selection of the individuals to sequence [55] and the biases that may appear during processing of sequencing reads [50] elsewhere, and therefore here we will focus on imputation of WGS and its use for genomic prediction.

Imputation accuracy

It is widely recognized that imputation from marker arrays to WGS from very few sequenced individuals can introduce genotyping errors and that genotype uncertainty can be high [17,21,68,69]. The accuracy of the imputed WGS is one of the main factors that may limit its performance for genomic prediction. In a simulation study, van den Berg et al. [17] quantified the impact of imputation errors on prediction accuracy and showed that prediction accuracy decreases as errors accumulate, especially in the testing set.

Imputation of WGS is particularly challenging because typically we have to impute a very large number of variants for a very large number of individuals from very few sequenced individuals. We assessed the imputation accuracy of our approach elsewhere [41,55] and recommended that ~2% of the population should be sequenced. In our study, line D was the line where prediction accuracy with Top40k performed the worst, mostly performing below Chip predictors. In this line, only 0.9% of the individuals in the population had been sequenced and therefore lower imputation accuracy could be expected. Although there was not enough evidence for establishing a link between these two features (sequencing effort and prediction accuracy), we recommend cautious design of a sequencing strategy that is suited to the intended imputation method [55].

Prediction accuracy could be improved by accounting for genotype uncertainty of the imputed WGS. For that, it could be advantageous to use allele dosages rather than best-guess genotypes [69], although most current implementations cannot handle such information.

Preselection of predictor variants

Simply using WGS to increase the number of markers does not improve prediction accuracy [18,21,24]. Due to the large dimensionality of WGS, there is a need to remove uninformative variants [24,31,66,68,70]. Predictor variants must be causal or at least informative of the causal variants, which depends on the distance between the markers and the causal variants [13]. For this reason, variants that are in weak linkage disequilibrium with causal mutations have a ‘dilution’ effect, i.e., they add noise and limit prediction accuracy [24,31,70]. However, if too stringent filters are applied during preselection of predictor variants, there is a risk of removing true causal variants, and that would debilitate persistence of accuracy across generations and across populations [66,71]. For instance, the impact of removing predictor variants with low minor allele frequency can vary depending on the minor allele frequency of the causal variants as well as the distance between predictor and causal variants [13]. Losing causal or informative variants would negatively affect multi-line or multi-breed prediction.

A popular strategy to preselect the predictor variants is based on association tests. Genome-wide association studies on WGS are expected to confirm associations that were already detected with marker arrays and identify novel associations (e.g., [36,72]). However, preliminary inspection of our GWAS results for the real traits showed that the added noise could easily offset the added information and fine-mapping remains challenging. Multi-breed GWAS [4] and meta-analyses [73] are suitable alternatives for GWAS to accommodate much larger population sizes and for combining results of populations with diverse genetic backgrounds. Multi-breed GWAS can be more efficient to identify informative variants than single-breed GWAS, which may benefit even prediction within lines [74]. Because the signal of some variants may go undetected for some traits but not for other correlated traits, combining GWAS information of several traits can also help identifying weak or moderate associations [25]. We did not test whether combining the significant markers from the different single-trait GWAS yielded greater improvements in prediction accuracy [29,32]. Multi-trait GWAS models could be more suited for that purpose [72,75]. To improve fine-mapping, other GWAS models that incorporate biological information have been proposed (e.g., functional annotation [76] or metabolomics [77]).

There have been other suggested methods that may improve variant preselection for genomic prediction. VanRaden et al. [30] suggested that preselecting variants based on the genetic variance that they contribute rather than the significance of the association could be more advantageous, as the former would indirectly preselect variance with higher minor allele frequency. Other authors proposed preselection of variants using statistics that do not depend on GWAS, such as the fixation index (F_{ST}) score between groups of individuals with high and low phenotype values [70], as an alternative to avoid the negative impact of spurious associations.

Preselecting predictor variants based on functional annotation was not useful, as it reduced prediction accuracy in several traits and lines. Previous studies showed that subsets of variants based on functionality either did not improve or reduced prediction accuracy [20] and that adding preselected variants from coding regions to marker arrays produced lower prediction accuracy than just adding the same number of variants without considering functional classification [32]. A plausible explanation is that functional variants are enriched for lower minor allele frequency, which can be less informative for prediction [13]. Furthermore, functional annotation does not necessarily capture true effects, and the method we used is biased towards protein-coding variants, which may lead to an underrepresentation of functional non-coding variants that may explain a large fraction of quantitative trait variance. Xiang et al. [78] found that expression QTL and non-coding variants explained more variation in quantitative traits in cattle than protein-coding functional variants. When functional annotation is not considered, intergenic variants are more likely to be preselected by chance. Such variants tend to be more common and widespread across populations, and therefore can act as tag variants and capture much larger fractions of trait variance [53].

Another popular strategy to reduce the number of variants is to prune variants based on linkage disequilibrium (LDTags). This strategy performed very poorly in our populations. Other studies reported different outcomes, where pruning for r^2>0.9 provided positive results [18,21]. It is possible that this was in part due to the stringent threshold (r^2>0.1) that we used in order to retain only a small number of variants.

New models and methods

It is also likely that models, methods, and their implementations need to be improved to handle the complexity of WGS and to efficiently estimate marker effects of so many variants with high accuracy, among other features. This is a very active area of research and multiple novel methodologies have been proposed over the last years. Some examples are a combination of subsampling and Gibbs sampling [79], and a model that simultaneously fits a GBLUP term for a polygenic effect and a BayesC term for variants with large effects selected by the model (BayesGC) [26]. Testing alternative models and methods for genomic prediction was out of the scope of this report. However, together with refinements in the preselection of predictor variants, it remains an interesting avenue for further optimisation of the analysis pipeline.

Some of the most promising methods are designed to incorporate prior biological information into the models. One of such methods is BayesRC [23], which extends BayesR by assigning flatter prior distributions to classes of variants that are more likely to be causal [19,22]. Similarly, GFBLUP [80] could be used to incorporate prior biological information from either QTL databases or GWAS as genomic features [21,35,68]. The model MBMG [27], which fits two genomic relationship matrices according to prior biological information, has also been proposed for multi-breed scenarios to improve genomic prediction in small populations. Haplotype-based prediction methods could provide greater prediction accuracy with WGS than SNP-based methods in pigs [81] and cattle [82]. These methods reduce the number of model dimensions. However, the uptake of such methods has been limited so far due to their greater complexity, for example, to define haplotype blocks.

Conclusion

Our results evidenced the potential for WGS to improve genomic prediction accuracy in intensely selected pig lines. The performance of each set of predictor variants was not robust across traits and lines and the improvements that we achieved so far were modest at best. The most robust results were obtained when variants that showed statistically significant associations to the trait were preselected and added to the marker array. With this method, average improvements of prediction accuracy of 2.5 and 4.2 percentage points were observed in within-line and multi-line scenarios, respectively, with training sets of around 80k individuals. We would expect that a combination of larger training sets and improved pipelines could help achieve greater improvements of prediction accuracy. The robustness of the whole strategy for generating WGS at the population level must be carefully stress-tested and further optimised.

Ethics approval and consent to participate

The samples used in this study were derived from the routine breeding activities of PIC.

Consent for publication

Not applicable.

Availability of data and material

The software packages AlphaSeqOpt, AlphaPhase, AlphaImpute and AlphaPeel are available from the AlphaGenes website (http://www.alphagenes.roslin.ed.ac.uk). The datasets generated and analysed in this study are derived from the PIC breeding programme and not publicly available.

Competing interests

The authors declare that they have no competing interests. BDV, CYC, and WOH are employees of Genus PIC.

Funding

The authors acknowledge the financial support from the BBSRC ISPG to The Roslin Institute (BBS/E/D/30002275), from Genus plc, Innovate UK (grant 102271), and from grant numbers BB/N004736/1, BB/N015339/1, BB/L020467/1, and BB/M009254/1. MJ acknowledges financial support from the Swedish Research Council for Sustainable Development Formas Dnr 2016-01386.

Authors' contributions

RRF, GG and JMH designed the study; CYC assisted in preparing the datasets; RRF, AW and MJ performed the analyses; RRF wrote the first draft; AW, CYC, BDV, WHO, GG and JMH assisted in the interpretation of the results and provided comments on the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work has made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF) (http://www.ecdf.ed.ac.uk/).

References

1. Pasaniuc B, Rohland N, McLaren PJ, Garimella K, Zaitlen N, Li H, et al. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. Nat Genet. 2012;44:631–5.

2. Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brondum RF, et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet. 2014;46:858–65.

3. Nicod J, Davies RW, Cai N, Hassett C, Goodstadt L, Cosgrove C, et al. Genome-wide association of multiple complex traits in outbred mice by ultra-low-coverage sequencing. Nat Genet. 2016;48:912–8.

4. Sanchez M-P, Govignon-Gion A, Croiseau P, Fritz S, Hozé C, Miranda G, et al. Within-breed and multi-breed GWAS on imputed whole-genome sequence variants reveal candidate mutations affecting milk protein composition in dairy cattle. Genet Sel Evol. 2017;49:68.

5. Das A, Panitz F, Gregersen VR, Bendixen C, Holm L-E. Deep sequencing of Danish Holstein dairy cattle for variant detection and insight into potential loss-of-function variants in protein coding genes. BMC Genomics. 2015;16:1043.

6. Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, et al. Large-scale whole-genome sequencing of the Icelandic population. Nat Genet. 2015;47:435–44.

7. VanRaden PM. Symposium review: How to implement genomic selection. J Dairy Sci. 2020;103:5291–301.

8. Clark SA, Hickey JM, van der Werf JH. Different models of genetic variation and their effect on genomic evaluation. Genet Sel Evol. 2011;43:18.

9. Druet T, Macleod IM, Hayes BJ. Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. Heredity. 2014;112:39–47.

10. MacLeod IM, Hayes BJ, Goddard ME. The Effects of Demography and Long-Term Selection on the Accuracy of Genomic Prediction with Sequence Data. Genetics. 2014;198:1671–84.

11. Meuwissen T, Goddard M. Accurate Prediction of Genetic Values for Complex Traits by Whole-Genome Resequencing. Genetics. 2010;185:623–31.

12. Pérez-Enciso M, Rincón JC, Legarra A. Sequence- vs. chip-assisted genomic selection: accurate biological information is advised. Genet Sel Evol. 2015;47:43.

13. van den Berg I, Boichard D, Guldbrandtsen B, Lund MS. Using Sequence Variants in Linkage Disequilibrium with Causative Mutations to Improve Across-Breed Prediction in Dairy Cattle: A Simulation Study. G3 GenesGenomesGenetics. 2016;6:2553–61.

14. van Binsbergen R, Calus MPL, Bink MCAM, van Eeuwijk FA, Schrooten C, Veerkamp RF. Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. Genet Sel Evol. 2015;47:71.

15. Calus MPL, Bouwman AC, Schrooten C, Veerkamp RF. Efficient genomic prediction based on whole-genome sequence data using split-and-merge Bayesian variable selection. Genet Sel Evol. 2016;48:49.

16. Veerkamp RF, Bouwman AC, Schrooten C, Calus MPL. Genomic prediction using preselected DNA variants from a GWAS with whole-genome sequence data in Holstein–Friesian cattle. Genet Sel Evol. 2016;48:95.

17. van den Berg I, Bowman PJ, MacLeod IM, Hayes BJ, Wang T, Bolormaa S, et al. Multi-breed genomic prediction using Bayes R with sequence data and dropping variants with a small effect. Genet Sel Evol. 2017;49:70.

18. Frischknecht M, Meuwissen THE, Bapst B, Seefried FR, Flury C, Garrick D, et al. Short communication: Genomic prediction using imputed whole-genome sequence variants in Brown Swiss Cattle. J Dairy Sci. 2018;101:1292–6.

19. Hayes BJ, MacLeod IM, Daetwyler HD, Bowman PJ, Chamberlain AJ, Vander Jagt CJ, et al. Genomic prediction from whole genome sequence in livestock: the 1000 Bull Genomes Project. Proc 10th World Congr Genet Appl Livest Prod WCGALP. Vancouver, BC, Canada; 2014. p. 183.

20. Heidaritabar M, Calus MPL, Megens H-J, Vereijken A, Groenen MAM, Bastiaansen JWM. Accuracy of genomic prediction using imputed whole-genome sequence data in white layers. J Anim Breed Genet. 2016;133:167–79.

21. Song H, Ye S, Jiang Y, Zhang Z, Zhang Q, Ding X. Using imputation-based whole-genome sequencing data to improve the accuracy of genomic prediction for combined populations in pigs. Genet Sel Evol. 2019;51:58.

22. Zhang C, Kemp RA, Stothard P, Wang Z, Boddicker N, Krivushin K, et al. Genomic evaluation of feed efficiency component traits in Duroc pigs using 80K, 650K and whole-genome sequence variants. Genet Sel Evol. 2018;50:14.

23. MacLeod IM, Bowman PJ, Vander Jagt CJ, Haile-Mariam M, Kemper KE, Chamberlain AJ, et al. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genomics. 2016;17:144.

24. Raymond B, Bouwman AC, Schrooten C, Houwing-Duistermaat J, Veerkamp RF. Utility of whole-genome sequence data for across-breed genomic prediction. Genet Sel Evol. 2018;50:27.

25. Xiang R, MacLeod IM, Daetwyler HD, de Jong G, O'Connor E, Schrooten C, et al. Genome-wide fine-mapping identifies pleiotropic and functional variants that predict many traits across global cattle populations. Nat Commun. 2021;12:860.

26. Meuwissen T, van den Berg I, Goddard M. On the use of whole-genome sequence data for across-breed genomic prediction and fine-scale mapping of QTL. Genet Sel Evol. 2021;53:19.

27. Raymond B, Bouwman AC, Wientjes YCJ, Schrooten C, Houwing-Duistermaat J, Veerkamp RF. Genomic prediction for numerically small breeds, using models with pre-selected and differentially weighted markers. Genet Sel Evol. 2018;50:49.

28. Moghaddar N, Brown DJ, Swan AA, Gurman PM, Li L, Werf JH. Genomic prediction in a numerically small breed population using prioritized genetic markers from whole-genome sequence data. J Anim Breed Genet. 2021;

29. Brøndum RF, Su G, Janss L, Sahana G, Guldbrandtsen B, Boichard D, et al. Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. J Dairy Sci. 2015;98:4107–16.

30. VanRaden PM, Tooker ME, O'Connell JR, Cole JB, Bickhart DM. Selecting sequence variants to improve genomic predictions for dairy cattle. Genet Sel Evol. 2017;49:32.

31. Al Kalaldeh M, Gibson J, Duijvesteijn N, Daetwyler HD, MacLeod I, Moghaddar N, et al. Using imputed whole-genome sequence data to improve the accuracy of genomic prediction for parasite resistance in Australian sheep. Genet Sel Evol. 2019;51:32.

32. Lopez BIM, An N, Srikanth K, Lee S, Oh J-D, Shin D-H, et al. Genomic Prediction Based on SNP Functional Annotation Using Imputed Whole-Genome Sequence Data in Korean Hanwoo Cattle. Front Genet. 2021;11:603822.

33. Hayes BJ, Daetwyler HD. 1000 Bull Genomes Project to Map Simple and Complex Genetic Traits in Cattle: Applications and Outcomes. Annu Rev Anim Biosci. 2019;7:89–102.

34. Sanchez M-P, Guatteo R, Davergne A, Saout J, Grohs C, Deloche M-C, et al. Identification of the ABCC4, IER3, and CBFA2T2 candidate genes for resistance to paratuberculosis from sequence-based GWAS in Holstein and Normande dairy cattle. Genet Sel Evol. 2020;52:14.

35. Yang R, Xu Z, Wang Q, Zhu D, Bian C, Ren J, et al. Genome-wide association study and genomic prediction for growth traits in yellow-plumage chicken using genotyping-by-sequencing. Genet Sel Evol. 2021;53:82.

36. Yan G, Liu X, Xiao S, Xin W, Xu W, Li Y, et al. An imputed whole-genome sequence-based GWAS approach pinpoints causal mutations for complex traits in a specific swine population. Sci China Life Sci. 2021;

37. Yang R, Guo X, Zhu D, Tan C, Bian C, Ren J, et al. Accelerated deciphering of the genetic architecture of agricultural economic traits in pigs using a low-coverage whole-genome sequencing strategy. GigaScience. 2021;10:giab048.

38. Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR. Low-coverage sequencing: Implications for design of complex trait association studies. Genome Res. 2011;21:940–51.

39. Hickey JM. Sequencing millions of animals for genomic selection 2.0. J Anim Breed Genet. 2013;130:331–2.

40. Hickey JM, Gorjanc G, Cleveland MA, Kranis A, Jenko J, Mésázros G, et al. Sequencing Millions of Animals for Genomic Selection 2.0. Proc 10th World Congr Genet Appl Livest Prod WCGALP. Vancouver, BC, Canada; 2014. p. 377.

41. Ros-Freixedes R, Whalen A, Chen C-Y, Gorjanc G, Herring WO, Mileham AJ, et al. Accuracy of whole-genome sequence imputation using hybrid peeling in large pedigreed livestock populations. Genet Sel Evol. 2020;52:17.

42. Gonen S, Ros-Freixedes R, Battagin M, Gorjanc G, Hickey JM. A method for the allocation of sequencing resources in genotyped livestock populations. Genet Sel Evol. 2017;49:47.

43. Ros-Freixedes R, Gonen S, Gorjanc G, Hickey JM. A method for allocating low-coverage sequencing resources by targeting haplotypes rather than individuals. Genet Sel Evol. 2017;49:78.

44. Hickey JM, Kinghorn BP, Tier B, Wilson JF, Dunstan N, van der Werf JH. A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. Genet Sel Evol. 2011;43:12.

45. Hickey JM, Kinghorn BP, Tier B, van der Werf JH, Cleveland MA. A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. Genet Sel Evol. 2012;44:9.

46. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.

47. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013;1303.3997v1 [q – bio.GN].

48. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43:491–8.

49. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. 2018;10.1101/201178.

50. Ros-Freixedes R, Battagin M, Johnsson M, Gorjanc G, Mileham AJ, Rounsley SD, et al. Impact of index hopping and bias towards the reference allele on accuracy of genotype calls from low-coverage sequencing. Genet Sel Evol. 2018;50:64.

51. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

52. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.

53. Ros-Freixedes R, Valente B, Chen C-Y, Herring WO, Gorjanc G, Hickey JM, et al. Rare and population-specific functional variants across pig lines. bioRxiv [Internet]. 2022; Available from: https://doi.org/10.1101/2022.02.01.478603

54. Whalen A, Ros-Freixedes R, Wilson DL, Gorjanc G, Hickey JM. Hybrid peeling for fast and accurate calling, phasing, and imputation with sequence data of any coverage in pedigrees. Genet Sel Evol. 2018;50:67.

55. Ros-Freixedes R, Whalen A, Gorjanc G, Mileham AJ, Hickey JM. Evaluation of sequencing strategies for whole-genome imputation with hybrid peeling. Genet Sel Evol. 2020;52:18.

56. VanRaden PM, Wiggans GR. Derivation, Calculation, and Use of National Animal Model Information. J Dairy Sci. 1991;74:2737–46.

57. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. FaST linear mixed models for genome-wide association studies. Nat Methods. 2011;8:833–5.

58. Widmer C, Lippert C, Weissbrod O, Fusi N, Kadie C, Davidson R, et al. Further Improvements to Linear Mixed Models for Genome-Wide Association Studies. Sci Rep. 2015;4:6874.

59. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015;4.

60. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17:122.

61. Erbe M, Hayes BJ, Matukumalli LK, Goswami S, Bowman PJ, Reich CM, et al. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci. 2012;95:4114–29.

62. Moser G, Lee SH, Hayes BJ, Goddard ME, Wray NR, Visscher PM. Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using a Bayesian Mixture Model. Haley C, editor. PLOS Genet. 2015;11:e1004969.

63. MacLeod IM, Bolormaa S, Schrooten C, Goddard ME, Daetwyler H. Pitfalls of pre-selecting subsets of sequence variants for genomic prediction. Proc 22nd Conf Assoc Adv Anim Breed Genet AAABG. Townsville, Queensland, Australia; 2017. p. 141–4.

64. Gozalo-Marcilla M, Buntjer J, Johnsson M, Batista L, Diez F, Werner CR, et al. Genetic architecture and major genes for backfat thickness in pig lines of diverse genetic backgrounds. Genet Sel Evol. 2021;53:76.

65. Gebreyesus G, Lund MS, Sahana G, Su G. Reliabilities of Genomic Prediction for Young Stock Survival Traits Using 54K SNP Chip Augmented With Additional Single-Nucleotide Polymorphisms Selected From Imputed Whole-Genome Sequencing Data. Front Genet. 2021;12:667300.

66. Iheshiulor OOM, Woolliams JA, Yu X, Wellmann R, Meuwissen THE. Within- and across-breed genomic prediction using whole-genome sequence and single nucleotide polymorphism panels. Genet Sel Evol. 2016;48:15.

67. Legarra A, Garcia-Baccino CA, Wientjes YCJ, Vitezica ZG. The correlation of substitution effects across populations and generations in the presence of non-additive functional gene action. PREPRINT. 2021;

68. Sarup P, Jensen J, Ostersen T, Henryon M, Sørensen P. Increased prediction accuracy using a genomic feature model including prior information on quantitative trait locus regions in purebred Danish Duroc pigs. BMC Genet. 2016;17:11.

69. Pausch H, MacLeod IM, Fries R, Emmerling R, Bowman PJ, Daetwyler HD, et al. Evaluation of the accuracy of imputed sequence variant genotypes and their utility for causal variant detection in cattle. Genet Sel Evol. 2017;49:24.

70. Ling AS, Hay EH, Aggrey SE, Rekaya R. Dissection of the impact of prioritized QTL-linked and -unlinked SNP markers on the accuracy of genomic selection. BMC Genomic Data. 2021;22:26.

71. Fragomeni BO, Lourenco DAL, Masuda Y, Legarra A, Misztal I. Incorporation of causative quantitative trait nucleotides in single-step GBLUP. Genet Sel Evol. 2017;49:59.

72. Bolormaa S, Swan AA, Stothard P, Khansefid M, Moghaddar N, Duijvesteijn N, et al. A conditional multi-trait sequence GWAS discovers pleiotropic candidate genes and variants for sheep wool, skin wrinkle and breech cover traits. Genet Sel Evol. 2021;53:58.

73. van den Berg I, Xiang R, Jenko J, Pausch H, Boussaha M, Schrooten C, et al. Meta-analysis for milk fat and protein percentage using imputed sequence variant genotypes in 94,321 cattle from eight cattle breeds. Genet Sel Evol. 2020;52:37.

74. van den Berg I, Boichard D, Lund MS. Sequence variants selected from a multi-breed GWAS can improve the reliability of genomic predictions in dairy cattle. Genet Sel Evol. 2016;48:83.

75. Yoshida GM, Yáñez JM. Multi-trait GWAS using imputed high-density genotypes from whole-genome sequencing identifies genes associated with body traits in Nile tilapia. BMC Genomics. 2021;22:57.

76. Yang J, Fritsche LG, Zhou X, Abecasis G. A Scalable Bayesian Method for Integrating Functional Information in Genome-wide Association Studies. Am J Hum Genet. 2017;101:404–16.

77. Li J, Mukiibi R, Wang Y, Plastow GS, Li C. Identification of candidate genes and enriched biological functions for feed efficiency traits by integrating plasma metabolites and imputed whole genome sequence variants in beef cattle. BMC Genomics. 2021;22:823.

78. Xiang R, Berg I van den, MacLeod IM, Hayes BJ, Prowse-Wilkins CP, Wang M, et al. Quantifying the contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits. Proc Natl Acad Sci. 2019;116:19398–408.

79. Xavier A, Xu S, Muir W, Rainey KM. Genomic prediction using subsampling. BMC Bioinformatics. 2017;18:191.

80. Edwards SM, Sørensen IF, Sarup P, Mackay TFC, Sørensen P. Genomic Prediction for Quantitative Traits Is Improved by Mapping Variants to Gene Ontology Categories in *Drosophila melanogaster*. Genetics. 2016;203:1871–83.

81. Bian C, Prakapenka D, Tan C, Yang R, Zhu D, Guo X, et al. Haplotype genomic prediction of phenotypic values based on chromosome distance and gene boundaries using low-coverage sequencing in Duroc pigs. Genet Sel Evol. 2021;53:78.

82. Li H, Zhu B, Xu L, Wang Z, Xu L, Zhou P, et al. Genomic Prediction Using LD-Based Haplotypes Inferred From High-Density Chip and Imputed Sequence Variants in Chinese Simmental Beef Cattle. Front Genet. 2021;12:665382.

Figures

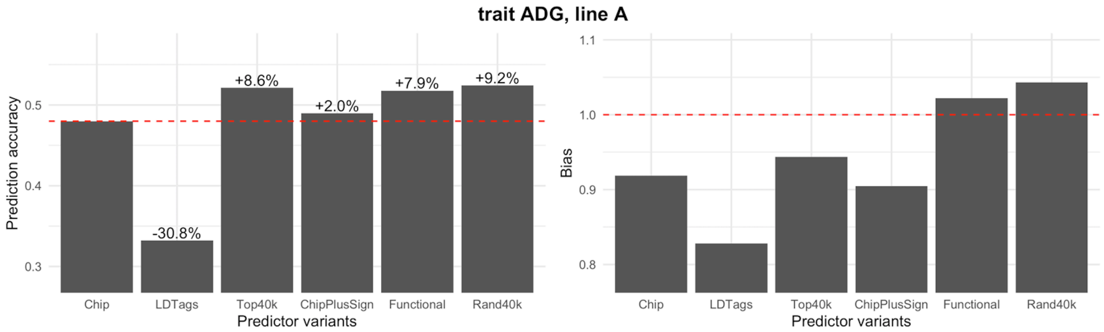
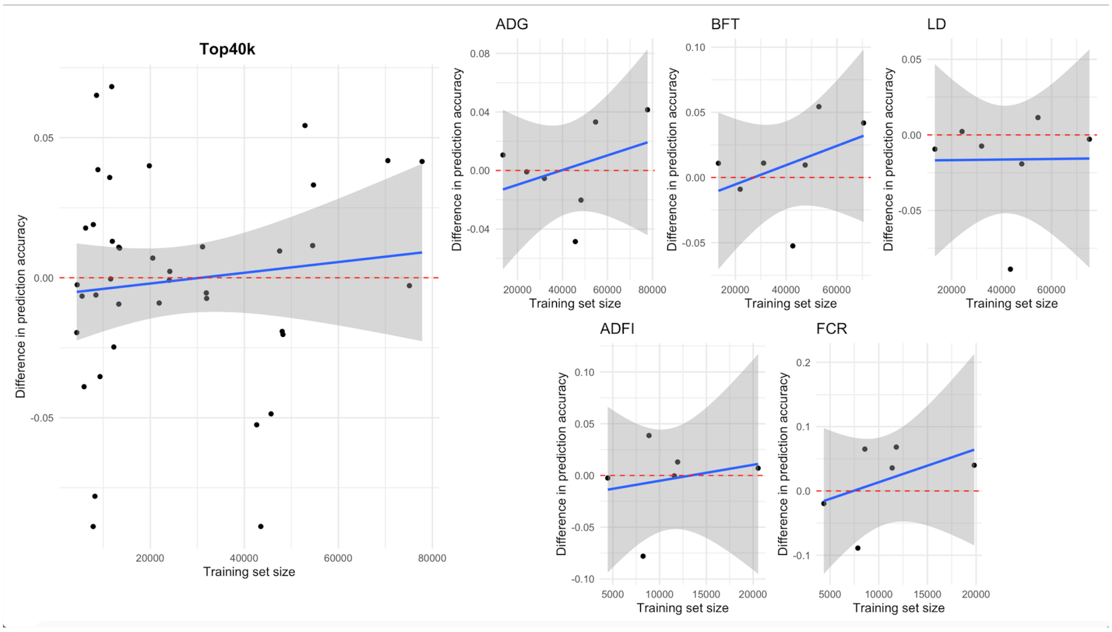


Figure 1. Prediction accuracy for each set of predictor variants for trait ADG in line A. Left: Correlation (left). Dashed line at value of Chip as a reference. Values indicate relative difference to Chip. Right: Bias. Dashed line at the ideal value.



966

967 **Figure 2.** Genomic prediction accuracy with the Top40k predictor variants for the real traits. The difference of prediction accuracy between Top40k and Chip is shown, for all traits and lines (left) or by trait (right). Red dashed line at ‘no difference’.

970

971

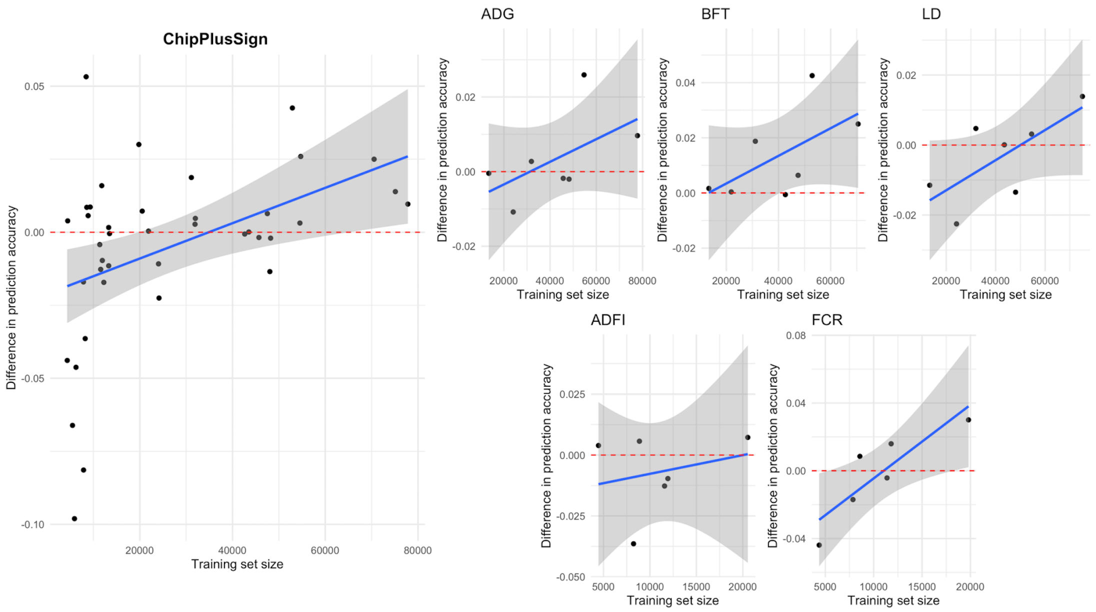


Figure 3. Genomic prediction accuracy with the ChipPlusSign predictor variants for the real traits. The difference of prediction accuracy between ChipPlusSign and Chip is shown, for all traits and lines (left) or by trait (right). Red dashed line at 'no difference'.

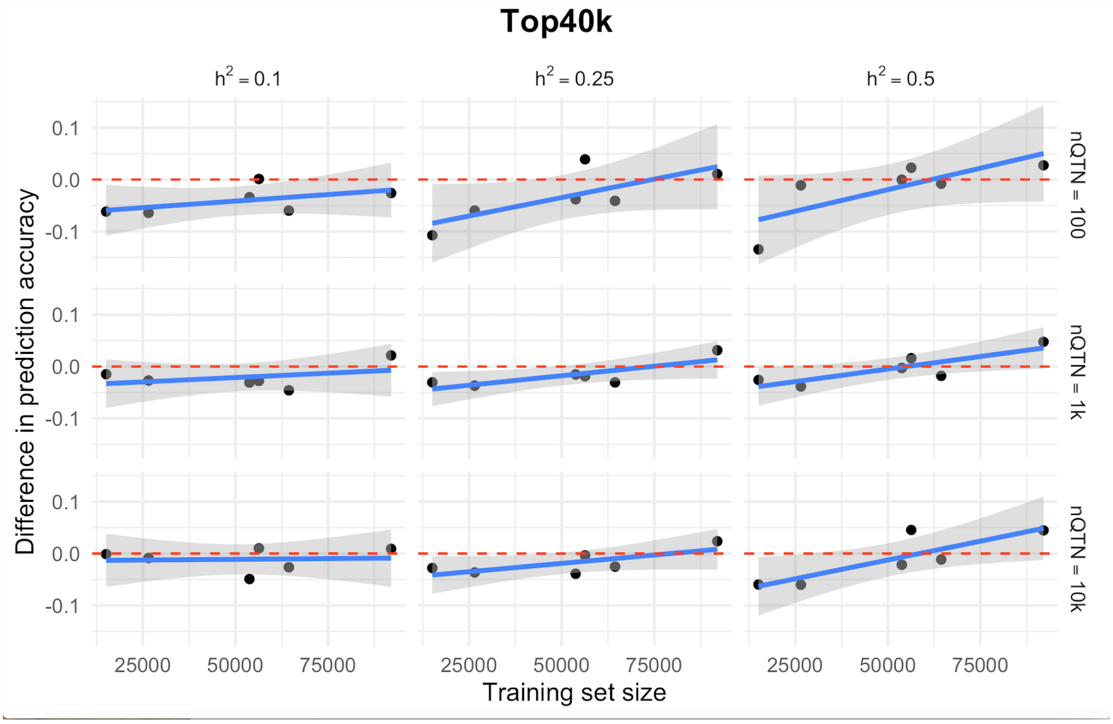


Figure 4. Genomic prediction accuracy with the Top40k predictor variants for the simulated traits. The difference of prediction accuracy between Top40k and Chip is shown by heritability (h^2) and number of quantitative trait nucleotides (nQTN) of the simulated traits. Red dashed line at ‘no difference’.

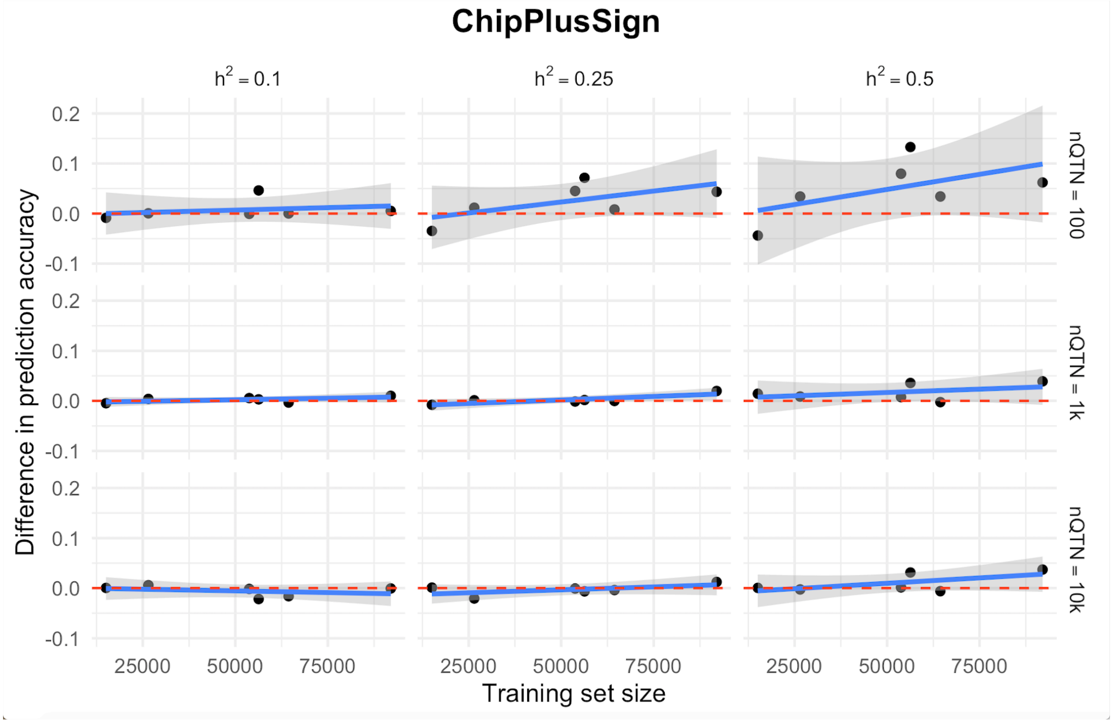


Figure 5. Genomic prediction accuracy with the ChipPlusSign predictor variants for the simulated traits. The difference of prediction accuracy between ChipPlusSign and Chip is shown by heritability (h^2) and number of quantitative trait nucleotides (nQTN) of the simulated traits. Red dashed line at 'no difference'.

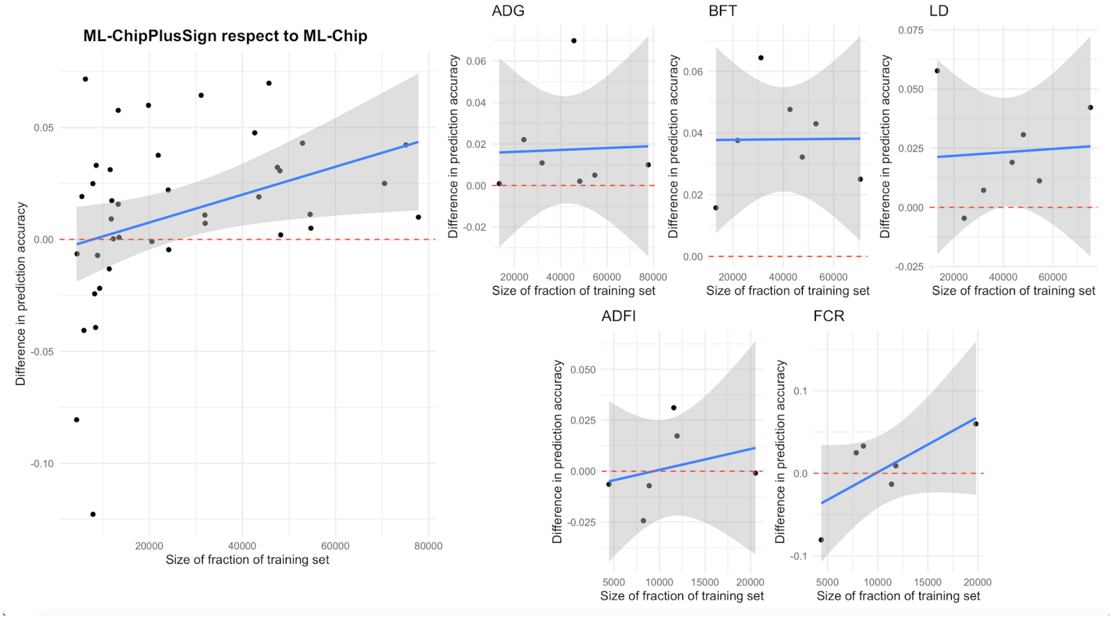


Figure 6. Genomic prediction accuracy with the ML-ChipPlusSign predictor variants for the real traits. The difference of prediction accuracy between ML-ChipPlusSign and ML-Chip is shown, for all traits and lines (left) or by trait (right). Red dashed line at ‘no difference’.

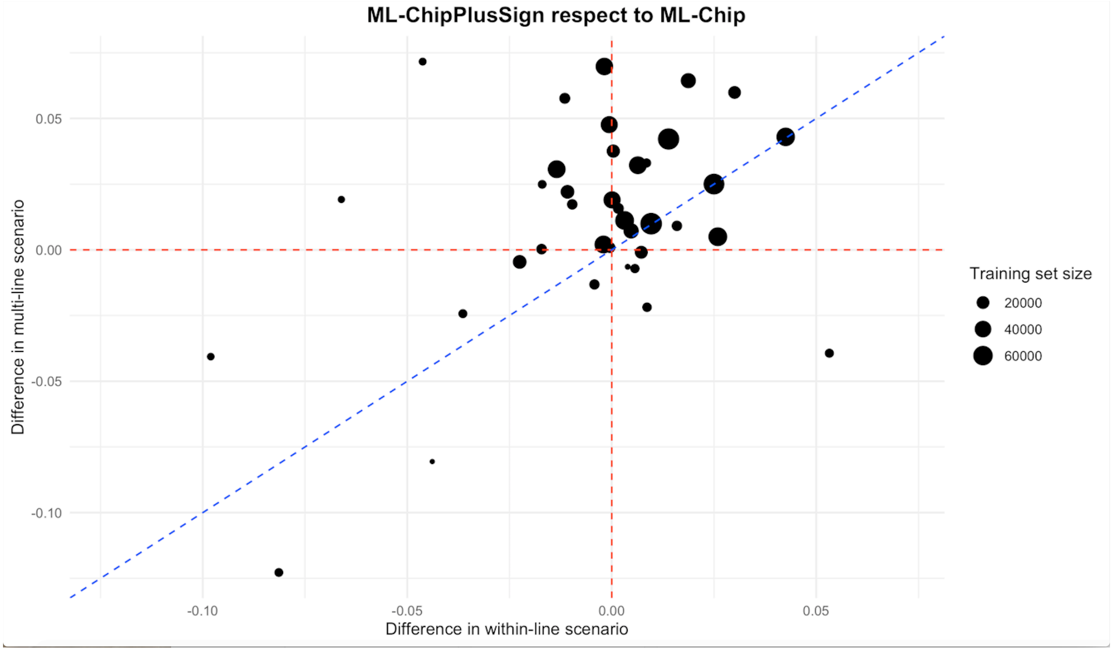


Figure 7. Comparison of the difference in genomic prediction accuracy in the multi-line scenarios (between ML-ChipPlusSign and ML-Chip) and in the within-line scenarios (between ChipPlusSign and Chip). Red dashed line at ‘no difference’. Blue dashed line is the bisector.

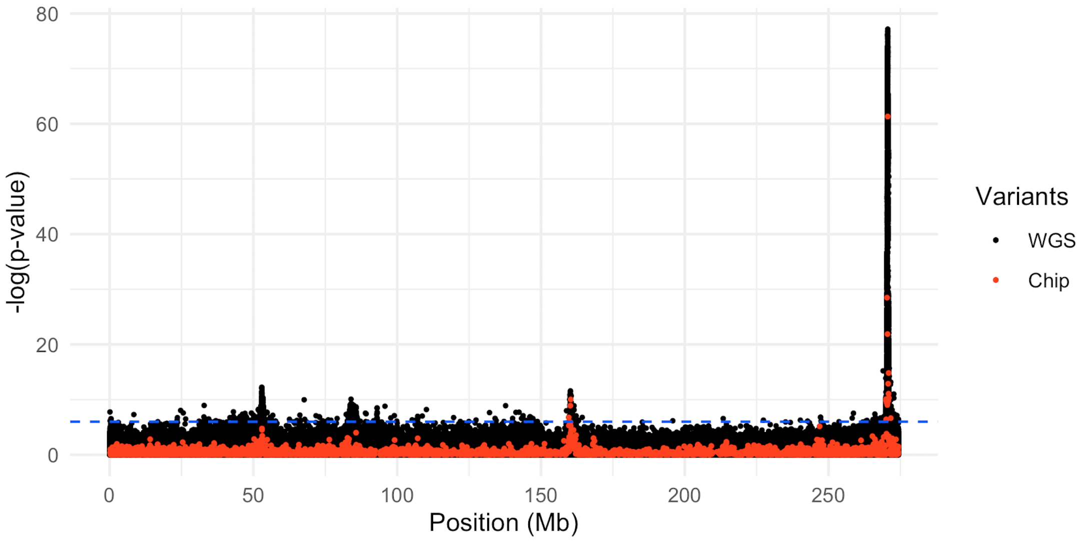


Figure 8. Genome-wide association study results for trait BFT in line A. Only chromosome 1 is displayed as an example. In red, results for the variants in the marker array (Chip); in black, results for the whole-genome sequence data (WGS). The blue dashed line indicates significance threshold with Bonferroni's multiple test correction assuming that the markers from the marker arrays were independent (p-value ≤ 10^{-6}).

Tables

1015 **Table 1.** Number of sequenced pigs and pigs with imputed data.

Line	Individuals sequenced	Individuals sequenced by coverage				Individuals used in analyses	
		1x	2x	5x	15–30x	Pedigree	Imputed
A	1,856	1,044	649	73	90	122,753	104,661
B	1,366	685	545	44	92	88,964	76,230
C	1,491	628	728	54	81	84,420	66,608
D	731	362	311	16	42	79,981	60,474
E	760	394	274	27	65	50,797	41,573
F	381	193	137	16	35	35,309	29,330
G	445	217	176	15	37	21,129	17,224

1016

Table 2. Number of phenotypic records per trait and line.

Trait	A	B	C	D	E	F	G
ADG	88,342	64,285	56,173	51,061	35,423	26,335	15,452
BFT	80,146	62,027	55,233	47,509	34,527	23,872	15,268
LD	85,233	64,141	56,026	48,509	35,495	26,453	15,274
ADFI	21,960	9,525	9,062	12,256	12,444	4,105*	4,851
FCR	21,200	9,217	8,654	12,044	12,316	4,016*	4,754
TNB	13,581	10,721	9,626	7,729*	6,506*	-	3,230*
LWW	-	9,112	7,251	-	-	-	2,813*
RET	-	6,978	6,327	-	-	-	1,669*
Simulated	104,661	76,230	66,608	60,474	41,573	29,330	17,224

ADG average daily gain, *BFT* backfat thickness, *LD* loin depth, *ADFI* average daily feed intake, *FCR* feed conversion ratio, *TNB* total number of piglets born, *LWW* litter weight at weaning, *RET* return to oestrus 7 days after weaning.

*Included in multi-line scenarios, but excluded in within-line scenarios because of the limited size of the testing set.

Table 3. Number of significant variants from the whole-genome sequence data that were added to the marker array in ChipPlusSign.

Trait	A	B	C	D	E	F	G	Multi-line
ADG	646	581	424	498	279	219	143	4731
BFT	1083	758	664	518	1030	218	237	6149
LD	633	579	458	518	222	215	43	7247
ADFI	145	224	169	23	183	-	119	767
FCR	198	224	162	95	56	-	134	1369
TNB	71	117	161	-	-	-	-	248
LWW	-	32	73	-	-	-	-	480
RET	-	184	31	-	-	-	-	60

ADG average daily gain, *BFT* backfat thickness, *LD* loin depth, *ADFI* average daily feed intake, *FCR* feed conversion ratio, *TNB* total number of piglets born, *LWW* litter weight at weaning, *RET* return to oestrus 7 days after weaning.

Table 4. Number of significantly associated genomic regions in the genome-wide association study for the simulated phenotypes that contained 0, 1 or 2 or more quantitative trait nucleotides (QTN).

h^2	nQTN	Line	Chip		Whole-genome sequence		
		size	0 QTN	1 QTN	0 QTN	1 QTN	≥2 QTN
0.10	100	27k	4	1	8	6	0
		56k	11	3	19	19	0
		92k	10	7	44	19	0
	1k	27k	1	0	4	0	1
		56k	1	0	16	3	1
		92k	1	0	283	9	0
	10k	27k	1	0	1	0	0
		56k	0	0	16	2	1
		92k	2	0	186	17	12
0.25	100	27k	11	6	26	15	1
		56k	22	8	44	28	3
		92k	20	7	90	34	1
	1k	27k	0	0	8	1	3
		56k	3	0	34	15	6
		92k	6	0	692	49	16
	10k	27k	0	0	2	0	0
		56k	0	0	90	9	22
		92k	4	0	564	56	164
0.50	100	27k	18	9	24	24	1
		56k	30	13	116	41	3
		92k	17	9	425	44	1
	1k	27k	6	0	22	9	6
		56k	5	1	238	59	32
		92k	11	1	903	169	120
	10k	27k	0	0	4	0	0
		56k	0	0	360	77	172
		92k	10	0	379	116	508