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 genomic prediction accuracy and its persistence across generations and breeds. However, empirical results 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 high 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 eight complex traits. Genomic predictions were performed using either data from a marker array or variants preselected from WGS based on association tests.

Results

The prediction accuracy with each set of preselected WGS variants was not robust across traits and lines and the improvements in prediction accuracy that we achieved so far with WGS compared to marker arrays were generally small. The most favourable results for WGS were obtained when the largest training sets were available and used to preselect variants with statistically significant associations to the trait for augmenting the established marker array. With this method and training sets

of around 80k individuals, average improvements of genomic prediction accuracy of 0.025 were observed in within-line scenarios.

Conclusions

Our results showed that WGS has a small potential to improve genomic prediction accuracy compared to marker arrays in intensely selected pig lines in some settings. Thus, although we expect that more robust improvements could be attained with a combination of larger training sets and optimised pipelines, the use of WGS in the current implementations of genomic prediction should be carefully evaluated on a case-by-case basis against the cost of generating WGS at a large scale.

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 in livestock populations, such as cattle and pig, and the reduction of the generational interval, most notably in dairy cattle. Since its early implementations, genomic prediction is typically performed using marker arrays that capture the effects of the (usually unknown) causal variants via linkage and linkage disequilibrium [8,9]. In contrast, WGS are assumed to contain the causal variants. For this reason, it was hypothesized that WGS could further improve genomic prediction accuracy and its persistence across generations and breeds. Early simulations indicated that causal mutations from WGS could increase prediction accuracy. 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 [10]. Another study reported improvements in prediction accuracy of up to 30% if causal variants with low minor allele frequency could be captured by the WGS [11]. However, benefits could be low in typical livestock populations due to small effective population sizes and recent directional selection [12].

During the last few years, there have been several attempts at improving the accuracy of genomic prediction with WGS in the main livestock species. Empirical results have been ambiguous so far. When predicting genomic breeding values within a population, some studies found no relevant improvement in genomic prediction

accuracy with WGS compared to marker arrays [13–16]. Other studies found small, and often unstable, improvements (e.g., from 1 to 5% or no improvement depending on prediction method [17–19], or trait-dependent results [19,20]). With genomic prediction across populations, the identification of causal variants from WGS can improve prediction accuracy [21–24], especially for numerically small populations or for populations that are not represented in the training [21,23–27].

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 [28–31]. In some cases, this strategy improved genomic prediction accuracy by up to 9% [30] and 11% [31], but this strategy did not improve prediction accuracy in other within-line settings [15]. Nevertheless, these examples indicate 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 [2,32–34], including pigs [35,36]. However, the fine-mapping of causal variants remains challenging due to the pervasive long-range linkage disequilibrium across extremely dense variation [37].

The estimation of allele substitution effects with high accuracy 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. 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. However, despite this, the cost of generating accurate WGS at such a large 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 in livestock on the use of WGS for genomic prediction to date have been performed in cattle, for which a large multi-breed reference panel is available from the 1000 Bull Genomes Project [2,17,32]. This reference panel has enabled the imputation of WGS in many cattle populations. The lack of such reference panels hampers the potential of WGS in other species, such as pigs [35].

We have previously described our approach to impute WGS in large pedigreed populations without external reference panels [38]. 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 re-sequenced the whole genome of 6,931 individuals from seven commercial pig lines (Genus PIC, Hendersonville, TN) with a total coverage of approximately 27,243x. Breeds of origin of the nine lines included Large White, Landrace, Pietrain, Hampshire, Duroc, and synthetic lines. Sequencing effort in each of the seven lines was proportional to population size. The number of pigs that were

117 available in the pedigree of each line and the number of sequenced pigs, by coverage,
118 is summarized in Table 1. Approximately 1.5% (0.9 to 2.1% in each line) of the pigs
119 in each line were sequenced. Most pigs were sequenced at low coverage, with target
120 coverage of 1 or 2x, but a subset of pigs was sequenced at a higher coverage of 5, 15,
121 or 30x. Thus, the average individual coverage was 3.9x, but the median coverage was
122 1.5x. Most of the sequenced pigs were born during the 2008–2016 period. The
123 population structure across the seven lines was assessed with a principal component
124 analysis using the sequenced pigs and is shown in Additional file 1.

125 The sequenced pigs and their coverage were selected following a three-part
126 sequencing strategy developed to represent the haplotype diversity in each line. First
127 (1), sires and dams with the highest number of genotyped progeny were sequenced at
128 2x and 1x, respectively. Sires were sequenced at a greater coverage because they
129 contributed with more progeny than dams. Then (2), the individuals with the greatest
130 genetic footprint on the population (i.e., those that carry more of the most common
131 haplotypes) and their immediate ancestors were sequenced at a coverage between 1x
132 and 30x (AlphaSeqOpt part 1; [39]). The sequencing coverage was allocated with an
133 algorithm that maximises the expected phasing accuracy of the common haplotypes
134 from the accumulated family information. Finally (3), pigs that carried haplotypes
135 with low accumulated coverage (below 10x) were sequenced at 1x (AlphaSeqOpt part
136 2; [40]). Sets (2) and (3) were based on haplotypes inferred from marker array
137 genotypes (GGP-Porcine HD BeadChip; GeneSeek, Lincoln, NE), which were phased
138 with AlphaPhase [41] and imputed with AlphaImpute [42].

139 Most sequenced pigs and their relatives were also genotyped with marker
140 arrays either at low density (15k markers) using the GGP-Porcine LD BeadChip
141 (GeneSeek) or at high density (50k or 80k markers) using different versions of the

142 GGP-Porcine HD BeadChip (GeneSeek). In our study we only used markers included
143 in the 50k array, which is the latest version of the high-density array. Markers in the
144 15k array were nested within the 50k array and markers from the 80k array that were
145 not included in the 50k array were discarded. The number of pigs genotyped at each
146 density is summarized in Table 1. Quality control of the marker array data was based
147 on the individuals genotyped at high density. Markers with minor allele frequency
148 below 0.01, call rate below 0.80, or a significant deviation from the Hardy-Weinberg
149 equilibrium were removed. After quality control, 38,634 to 43,966 markers remained
150 in each line.
151

Sequencing and data processing

152 Tissue samples were collected from ear punches or tail clippings. Genomic
153 DNA was extracted using Qiagen DNeasy 96 Blood & Tissue kits (Qiagen Ltd.,
154 Mississauga, ON, Canada). Paired-end library preparation was conducted using the
155 TruSeq DNA PCR-free protocol (Illumina, San Diego, CA). Libraries for
156 resequencing at low coverage (1 to 5x) were produced with an average insert size of
157 350 bp and sequenced on a HiSeq 4000 instrument (Illumina). Libraries for
158 resequencing at high coverage (15 or 30x) were produced with an average insert size
159 of 550 bp and sequenced on a HiSeq X instrument (Illumina). All libraries were
160 sequenced at Edinburgh Genomics (Edinburgh Genomics, University of Edinburgh,
161 Edinburgh, UK).
162 DNA sequence reads were pre-processed using Trimmomatic [43] to remove
163 adapter sequences from the reads. The reads were then aligned to the reference
164 genome *Sscrofa11.1* (GenBank accession: GCA_000003025.6) using the BWA-MEM
165 algorithm [44]. Duplicates were marked with Picard

166 (http://broadinstitute.github.io/picard). Single nucleotide polymorphisms (SNPs) and
167 short insertions and deletions (indels) were identified with the variant caller GATK
168 HaplotypeCaller (GATK 3.8.0) [45,46] using default settings. Variant discovery with
169 GATK HaplotypeCaller was performed separately for each individual and then a joint
170 variant set for all the individuals in each population was obtained by extracting the
171 variant positions from all the individuals.

172 We extracted the read counts supporting each allele directly from the aligned
173 reads stored in the BAM files using a pile-up function to avoid biases towards the
174 reference allele introduced by GATK when applied on low-coverage WGS [47]. That
175 pipeline uses pysam (version 0.13.0; https://github.com/pysam-developers/pysam),
176 which is a wrapper around htslib and the samtools package [48]. We extracted the
177 read counts for all biallelic variants, after filtering out variants observed in less than
178 three sequenced individuals and variants in potential repetitive regions (defined as
179 variants that had mean depth values 3 times greater than the average realized
180 coverage) with VCFtools [49]. This pipeline delivered a total of 55.6 million SNP
181 (19.6 to 31.1 million within each line) and 10.2 million indels (4.1 to 5.6 million
182 within each line). A more complete description of the variation across the lines is
183 provided in [50].
184

Genotype imputation

185 Genotypes were jointly called, phased and imputed for a total of 483,353
186 pedigree-related individuals using the 'hybrid peeling' method implemented in
187 AlphaPeel [51,52]. This method used all the available marker array and WGS.
188 Imputation was performed separately for each line using complete multi-generational
189 pedigrees, with 21,129 to 122,753 individuals per line (Table 1). We have previously

published reports on the accuracy of imputation in the same populations using this method [38]. 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 (defined as the sum of coefficients of pedigree-based relationship between the individual and the rest of individuals). These criteria were based on the analysis of imputation accuracy in simulated and empirical data [38]. A total of 396,100 individuals remained, with 17,224 and 104,661 individuals per line (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, because their estimated variant-wise dosage correlations was lower than 0.90 [38]. 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 eight complex 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 of VanRaden et al. [53]. 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.

Training and testing sets

We split the individuals in each line into training and testing sets. The testing sets were defined as 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 in the testing set. This design was chosen to mimic a realistic situation in which breeding programmes evaluate selection candidates available in a selection nucleus at any given time.

To assess the effect of the size of the training set on prediction accuracy, we created training sets with a reduced number of phenotype records for the three largest lines and the three traits with the largest number of records. We did this by removing the oldest animals in a way that approximately the most recent 10, 20, or 35 to 45 thousand phenotype records remained in each of the reduced training sets.

Due to the computational requirements of the analyses, we could not perform repetitions for every analysis. However, we estimated variability of the results across repetitions in the largest, an intermediate, and the smallest lines for two traits with a large and small number of phenotype records. For doing this, we randomly split the test sets into five subsets, with each full-sib family represented exclusively in one of the subsets. Training sets for each repetition were defined as for the general case.

238

Genome-wide association study

239 To provide an association-based criterion to preselect variants for genomic
240 prediction, we performed a GWAS for each trait and line. This step included only the
241 individuals in the training set. We fitted a univariate linear mixed model that
242 accounted for the genomic relationships as:

243 $$\mathbf{y} = \mathbf{x}_i\beta_i + \mathbf{u} + \mathbf{e},$$

244 where $\mathbf{y}$ is the vector of dEBV, $\mathbf{x}_i$ is the vector of genotypes for the ith variant coded
245 as 0 and 2 if homozygous for either allele or 1 if heterozygous, β_i is the allele
246 substitution effect of the ith variant on the trait, $\mathbf{u}\sim N(0, \sigma^2_u\mathbf{K})$ is the vector of
247 polygenic effects with the covariance matrix equal to the product of the polygenic
248 additive genetic variance σ^2_u and a genomic relationship matrix $\mathbf{K}$, and $\mathbf{e}$ is a vector of
249 uncorrelated residuals. Due to computational limitations, the genomic relationship
250 matrix $\mathbf{K}$ was calculated using only imputed genotypes in the marker array. We used
251 the FastLMM software [54,55] to fit the model.

252

Within-line genomic prediction

253 To test whether variants from the WGS could provide greater genomic
254 prediction accuracy than the marker array, we tested genomic prediction using
255 variants from the marker array, from the WGS, or combining them. The marker array
256 data (also referred to as 'Chip') was set as the benchmark for prediction accuracy. It
257 contained all ~40k variants in the marker array. For WGS, we preselected sets of
258 variants because currently available methods for genomic prediction are not yet
259 capable of handling datasets as large as the complete WGS without exorbitant

260 computational resources. We tested different alternative strategies for preselecting
261 variants for the prediction model based on the GWAS results:

262 • *Top40k.* To mimic the number of variants in Chip, we preselected the variants
263 with the lowest p-value (not necessarily below the significance threshold) in each
264 of consecutive non-overlapping 55-kb windows along the genome. In addition, to
265 test the impact of variant density on prediction accuracy, we preselected 10k,
266 25k, 75k, or 100k variants following the same criterion.

267 • *ChipPlusSign.* Variants preselected as in Top40k, but only significant variants
268 ($p \leq 10^{-6}$) were preselected and merged with those in Chip. When a 55-kb window
269 contained more than one significant variant, only that with the lowest p-value was
270 selected as a proxy to reduce the preselection of multiple significant variants
271 tagging the same causal variant. When the most significant variant from WGS
272 was already included in the marker array, the variant was considered only once
273 and in the rare cases of genotype discordance, the genotype was replaced with the
274 mean genotype value in that line. On average, 309 significant variants were
275 identified per trait and line (range: 23 to 1083; Table 3) and merged with those in
276 Chip.

277

278 Genomic prediction was performed by fitting a univariate model with BayesR
279 [56,57], which uses a mixture of normal distributions as the prior for variant effects,
280 including one distribution that sets the variant effects to zero. The model was:

281 $$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

282 where $\mathbf{y}$ is the vector of dEBV, $\mathbf{1}$ is a vector of ones, μ is the general mean, $\mathbf{X}$ is a
283 matrix of variant genotypes, $\boldsymbol{\beta}$ is a vector of variant effects, and $\mathbf{e}$ is a vector of
284 uncorrelated residuals. The prior variance of the variant effects in $\boldsymbol{\beta}$ had four

components with mean zero and variances $\sigma_1^2 = 0$, $\sigma_2^2 = 0.0001\sigma_g^2$, $\sigma_3^2 = 0.001\sigma_g^2$, and $\sigma_4^2 = 0.01\sigma_g^2$, where σ_g^2 is the total genetic variance. We used a uniform and almost uninformative prior for the mixture distribution with the total genetic variance re-estimated in every iteration. 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 predicted genomic breeding values and the dEBV. Bias of the prediction accuracy was calculated as the regression coefficient of the dEBV on the predicted genomic breeding values. For ease of comparison between traits and lines, the difference between prediction accuracy of WGS and the marker array was calculated. The difference in prediction accuracy was analysed by fitting linear models with the size of the training set as a covariate and trait and line as fixed effects when appropriate.

Multi-line genomic prediction

We considered multi-line scenarios in which the training set was formed by merging the training sets that had been defined for each line. All analyses were performed as for the within-line scenarios but with an additional effect of the line in the prediction model. In the multi-line scenarios, all variants 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 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 variant sets 'ML-Top40k' and 'ML-ChipPlusSign' by preselecting variants following the same criteria as in within-line scenarios, but using a multi-line GWAS with an

309 additional effect of the line. For ML-ChipPlusSign, 60 to 7247 significant variants
310 were identified per trait (Table 3) and merged with those in ML-Chip. For comparison
311 purposes, genomic prediction accuracy was calculated for the testing set of each
312 individual line.
313

Results

Within-line genomic prediction accuracy

314 Whole-genome sequence data improved genomic prediction accuracy
315 compared to marker array data in some scenarios, especially when there was a
316 sufficiently large training set and if an appropriate set of variants was preselected.
317 Figure 1 shows the prediction accuracy for the three traits and three lines with the
318 largest training sets using the two different sets of WGS variants. Results for the rest
319 of traits and lines, as well as results for the bias, are provided in Additional File 2. For
320 BFT in line B, the two tested sets of variants from the WGS increased prediction
321 accuracy by 0.054 (+9.8%), for Top40k, and by 0.043 (+7.7%), for ChipPlusSign.
322 However, the performance of WGS was not robust and differed for each trait and line,
323 and even across repetitions within trait and line (Additional File 3), often leading to
324 no improvements of prediction accuracy or even reduced prediction accuracy relative
325 to the marker array. For instance, Top40k reduced prediction accuracy by 0.020 (–
326 3.4%), for ADG in line C, and ChipPlusSign by 0.020 (–2.2%), for LD in line C.
327 Using WGS reduced bias compared to the marker array in some, but not all,
328 scenarios.

329 There was a trend that the capacity of WGS variants to improve the genomic
330 prediction accuracy compared to the marker arrays was larger for the traits and lines
331 with larger training sets. Figures 2 and 3 show the difference in prediction accuracy of

Top40k and ChipPlusSign compared to the marker array against the training set size. 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. Within trait and line, the variability across repetitions was also larger in Top40k than in ChipPlusSign (Additional File 3). Gains in prediction accuracy were low-to-moderate in the most favourable cases. In the most unfavourable cases we observed large losses in prediction accuracy for Top40k but more limited losses for ChipPlusSign with moderate training set sizes. The regression coefficient between the difference in prediction accuracy and the size of the training set was positive but had stronger statistical evidence for ChipPlusSign ($b=0.5\cdot10^{-6}$ individual^{-1}; $p=0.032$; R^2 for each trait between 0.06 and 0.75) than for Top40k ($b=0.5\cdot10^{-6}$ individual^{-1}; $p=0.24$; R^2 for each trait between 0.00 and 0.20), because of the apparent lower robustness with Top40k.

Results within trait and line (Figure 4) confirmed that the impact of WGS on genomic prediction accuracy depended on line, but also that in general WGS yielded higher prediction accuracy compared to the marker array when the training set was the largest. Under this setting, the regression coefficient between the difference in prediction accuracy and the size of training set was $0.6\cdot10^{-6}$ individual^{-1} ($p<0.001$), for Top40k, and $0.3\cdot10^{-6}$ individual^{-1} ($p=0.017$) for ChipPlusSign. This was at least partly driven by the lower number of significant associations that were detected with smaller training sets. With a training set of 20k individuals or less, 118 to 287 significant variants were added to the marker array; with a training set of 35k to 45k individuals, 288 to 709 significant variants; and with all available individuals in the training set, 424 to 1083 significant variants. Thus, if the marker array was augmented with the

357 significant variants detected with all available individuals (ChipPlusSign*), then
358 WGS yielded the same prediction accuracy than the marker array or higher in most
359 scenarios even when the set for training the predictive equation was smaller.

360 Results from simulated traits (Additional File 4) confirmed the trends
361 observed for the empirical traits; for instance, the higher robustness of ChipPlusSign
362 compared to Top40k. Results from the simulated traits also showed the impact of the
363 genetic architecture of the traits on the success of WGS in improving genomic
364 prediction accuracy. Traits with high heritability and low number of QTN were more
365 likely to show larger improvements in prediction accuracy.

366 We observed diminishing returns when we increased the density of the
367 variants used in prediction. Increasing the number of variants from the 40k in Top40k
368 to 75k selected in the same way yielded small improvements in genomic prediction
369 accuracy compared to Top40k, but increases up to 100k variants provided smaller or
370 null additional gains (Additional File 5).

371

Multi-line genomic prediction accuracy

372 The accuracy of genomic prediction trained across multi-line datasets was
373 systematically lower than in the within-line datasets (Additional File 6). Nonetheless,
374 when using multi-line training sets, the ML-ChipPlusSign variants in general
375 increased genomic prediction accuracy relative to the marker array (ML-Chip; Figure
376 5). For the traits that accumulated the largest multi-line training sets (i.e., ADG, BFT,
377 and LD), the improvements of prediction accuracy in each individual line seemed
378 unrelated to the number of individuals that each line contributed to the multi-line
379 training set. However, for the traits that accumulated smaller multi-line training sets
380 (i.e., ADFI and FCR), ML-ChipPlusSign only improved prediction accuracy in the

lines that contributed more individuals to the multi-line training set, and reduced prediction accuracy in the lines that contributed less individuals to the multi-line training set. Therefore, as happened in the within-line scenarios, the greatest improvements of prediction accuracy with WGS were achieved for the largest individual lines, although ML-ChipPlusSign in the multi-line scenarios also improved prediction accuracy compared to ML-Chip for some traits and lines for which no improvements were observed in the within-line scenarios, including numerically small lines (Figure 6). In contrast, results for ML-Top40k were not robust across traits (Additional File 7).

Preselection of variants through genome-wide association study

Although GWAS with WGS has the potential to detect associations that are not captured by marker arrays, the fine-mapping of the associated regions and the preselection of variants through GWAS with WGS was limited due to the pervasiveness of linkage disequilibrium (Additional File 8) and was affected by false positives in a more severe way than GWAS with marker arrays, especially for highly polygenic traits (Additional File 4).

Discussion

Our results showed that WGS has some potential to improve genomic prediction accuracy compared to marker arrays in intensely selected pig lines, but the use of WGS in current implementations should be carefully evaluated. On one hand, the small and non-robust 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

404 to leverage the potential of WGS, especially in scenarios where genomic prediction
405 with marker arrays is already yielding high accuracy. The results from several traits
406 and lines with different training set sizes allowed us to identify the most favourable
407 scenarios for genomic prediction with WGS. We will discuss (1) the prediction
408 accuracy that we achieved with WGS compared to commercial marker array data and
409 the scenarios in which WGS may become beneficial, (2) the potential pitfalls for its
410 effective implementation and the need for an optimised strategy, and (3) the
411 suitability of WGS for genomic prediction.
412

Prediction accuracy with whole-genome sequence data

413 We compared the genomic prediction accuracy of the current marker array
414 (Chip) with sets of preselected WGS variants in a way that the number of variants
415 remained similar across sets. Improvements in prediction accuracy can be limited if
416 current marker arrays are already sufficiently dense to capture a large proportion of
417 the genetic diversity in intensely selected livestock populations. These populations
418 typically have small effective population size due to intense selective breeding
419 [12,17]. Nevertheless, modest improvements have been achieved under certain
420 scenarios across several studies. In our study, the most robust results were obtained
421 with the ChipPlusSign variant sets, where the marker array was augmented with WGS
422 variants that had statistically significant associations to the trait. This is consistent
423 with previous reports that showed an improvement in prediction accuracy under
424 similar approaches [28–31]. We augmented the marker array with 23 to 1083
425 significant variants in different scenarios. In the most successful scenarios, a
426 minimum of around 200 significant variants were added and prediction accuracy
427 improved by 0.025 on average with training sets of around 80k individuals. Other

428 studies suggested additions of a larger number of variants. In Nordic cattle, adding
429 1623 variants (preselected as the combination of 3-5 variants for each of the top QTL
430 per trait and breed) to a 50k marker array increased reliability (accuracy squared) by
431 up to 0.05 [28], but a similar approach produced negligible improvements for low
432 heritability traits [58]. In Holstein cattle, adding around 16k variants (preselected as
433 the largest allele substitution effects) to a 60k marker array increased reliability on
434 average by 0.027 (up to 0.048) [29]. In Hanwoo cattle, adding around 12k variants (3k
435 for each of four traits) to a custom 50k marker array improved accuracy by up to
436 ~0.06 [31]. In sheep, adding around 400 variants (preselected by GWAS with regional
437 heritability mapping) to a 50k marker array increased accuracy by 0.09 [30].

438 The modest performance of ChipPlusSign and Top40k could also be a
439 consequence of the difficulty for fine-mapping causal variants through GWAS on
440 WGS. Theoretically, the identification of all causal variants associated with a trait
441 should improve genomic prediction accuracy [59]. Even though WGS allows the
442 detection of a very large number of associations, problems such as false positives or
443 p-value inflation also become more severe, so that the added noise might offset the
444 detected signal. For instance, results in cattle showed that GWAS on WGS did not
445 detect clearer associated regions relative to marker arrays and failed to capture QTL
446 for genomic prediction [13], because the effect of potential QTL were spread across
447 multiple variants. Therefore, WGS performed better with simple genetic architectures
448 (i.e., traits with a low number of QTN). This is consistent with expectations and
449 simulation results [60] that indicated that the benefit of WGS for genomic prediction
450 would be limited by the number and size of QTN. Therefore, for largely polygenic
451 traits (as most traits of interest in livestock production), training sets need to be very
452 large before WGS can increase genomic prediction accuracy [60].

453 The advantage of using WGS might be limited by the small effective
454 population size of livestock populations under selection [61] and by the current
455 training set sizes, especially in scenarios where marker arrays are already yielding
456 high genomic prediction accuracy [13,18]. Multi-line training sets could be
457 particularly beneficial with the use of WGS because they allow a larger training set
458 with low pairwise relationship degree among individuals. Previous simulations
459 suggested that WGS might be the most beneficial with multi-breed reference panels
460 [62], especially for numerically small populations. Our results with a multi-line
461 training set indicated that WGS can improve prediction accuracy in scenarios that are
462 less optimised than within-line genomic prediction by up to 0.04. However, in general
463 those predictions were still less accurate than in within-line scenarios. In our multi-
464 line scenarios, we only used variation that segregated across all seven lines. We
465 observed that population-specific variation accounted only for small fractions of
466 genetic variance [50] and it seems unlikely that they would contribute much to
467 genomic prediction accuracy across breeds. Another possible obstacle is the
468 differences in the allele substitution effects of the causal mutations across breeds. This
469 can be caused by differences in allele frequency, contributions of non-additive effects
470 and different genetic backgrounds, or even gene-by-environment interactions among
471 others [22,63].

472 We observed low robustness of genomic prediction with WGS across traits
473 and lines, and drops in prediction accuracy in some scenarios. Regarding bias, it has
474 been noted that using the same reference individuals for preselecting variants through
475 GWAS and for training the predictive equation can reduce genomic prediction
476 accuracy and bias the predicted genomic breeding values [15,64]. In complementary
477 tests, we observed no systematic increase in accuracy or bias after splitting the

478 training set into two exclusive subsets, one for GWAS to preselect the predictor
479 variants and the other for training the predictive equation (Additional File 9). One
480 hypothesis is that both subsets belonged to the same population and therefore retained
481 similar inter-relationships (i.e., they are not strictly independent sets of individuals).
482 Moreover, the reduction in individuals available for training the predictors negatively
483 affected genomic prediction accuracy.

484 We did not directly test persistence of genomic prediction accuracy across
485 generations, but previous studies with empirical data found no higher persistence of
486 prediction accuracy with WGS, not even with low degree of relationship between
487 training and testing sets [13]. We expect such obstacles to persistence of accuracy
488 until causal variants can be successfully identified and their non-additive effects are
489 understood.

490

Suboptimal strategy and pitfalls

491 The use of WGS for genomic prediction can only be reached after many other
492 steps are completed to produce genotype data at the whole-genome level. Each of
493 these steps has potential pitfalls to which the success of using WGS is sensitive. This
494 strategy includes the choice of which individuals to sequence, the bioinformatics
495 pipeline to call variants, the imputation of the WGS, and filtering of variants. When
496 combined with the multiplicity of methods for preselecting variants for genomic
497 prediction (which is unavoidable with current datasets, genomic prediction methods,
498 and computational resources), there are many variables in the whole process that can
499 affect the final result and that are not yet well understood. Therefore, a much greater
500 effort for optimising such pipelines is required. Here we tested relatively simple
501 approaches to evaluate 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 [52] and the biases that may appear during processing of sequencing reads [47] elsewhere. Here we will focus discussion on imputation of WGS and its use for genomic prediction.

Imputation accuracy

Imputation of WGS is particularly challenging because typically we must impute a very large number of variants for a very large number of individuals from few sequenced individuals. As a consequence, genotype uncertainty can be high [19,25,65,66]. The accuracy of the imputed WGS is one of the main factors that may limit its potential for genomic prediction. In a simulation study, van den Berg et al. [25] quantified the impact of imputation errors on genomic prediction accuracy and showed that prediction accuracy decreases as errors accumulate, especially in the testing set.

We assessed the imputation accuracy of our approach elsewhere [38,52] and recommended that ~2% of the population should be sequenced in intensely selected populations. In our study, line D was the line where genomic prediction accuracy with Top40k performed the worst, mostly performing worse than with the marker array. 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 genomic prediction accuracy), we recommend cautious design of a sequencing strategy that is suited to the intended imputation method [52].

Genomic prediction accuracy could be improved by accounting for genotype uncertainty of the imputed WGS. For that, it could be advantageous to use allele

527 dosages rather than best-guess genotypes [66], although most current implementations
528 of genomic prediction methods cannot handle such information.
529
530 ***Preselection of predictor variants***
531 Using WGS to simply increase the number of variants does not improve
532 genomic prediction accuracy [16,19,22]. Due to the large number of variants in WGS,
533 there is a need to remove uninformative variants [22,30,62,65,67]. We can expect
534 variants that are causal or at least informative about the causal variants, which
535 depends on their distance to the causal variants, to be the most predictive [68]. For
536 this reason, variants that are in weak linkage disequilibrium with causal mutations
537 have a 'dilution' effect, i.e., they add noise and limit prediction accuracy [22,30,67].
538 However, if too stringent filters are applied during preselection of predictor variants,
539 there is a risk of removing true causal variants, and that would debilitate persistence
540 of accuracy across generations and across populations [62,69]. For instance, the
541 impact of removing variants with low minor allele frequency can vary depending on
542 the minor allele frequency of the causal variants as well as the distance between
543 preselected and causal variants [68]. Losing causal or informative variants would
544 negatively affect multi-line or multi-breed prediction.
545 A popular strategy to preselect variants for the prediction model is based on
546 association tests. Genome-wide association studies on WGS are expected to confirm
547 associations that were already detected with marker arrays and identify novel
548 associations (e.g., [35,70]). However, preliminary inspection of our empirical GWAS
549 results showed that the added noise could easily offset the added information and
550 fine-mapping remains challenging. Multi-breed GWAS [4] and meta-analyses [71] are
551 suitable alternatives for GWAS to accommodate much larger population sizes and for

552 combining results of populations with diverse genetic backgrounds. Multi-breed
553 GWAS can be more efficient to identify informative variants than single-breed
554 GWAS, which may benefit even prediction within lines [72]. Because the signal of
555 some variants may go undetected for some traits but not for other correlated traits,
556 combining GWAS information of several traits can also help identifying weak or
557 moderate associations [23]. We did not test whether combining the significant
558 markers from the different single-trait GWAS yielded greater improvements in
559 prediction accuracy [28,31]. Multi-trait GWAS could be more suited for that purpose
560 [70,73]. To improve fine-mapping, other GWAS models that incorporate biological
561 information have been proposed (e.g., functional annotation [74] or metabolomics
562 [75]).

563 Other methods were suggested to improve variant preselection for genomic
564 prediction. VanRaden et al. [29] suggested that preselecting variants based on the
565 genetic variance that they contribute rather than the significance of the association
566 could be more advantageous, because the former would indirectly preselect variants
567 with higher minor allele frequency. Other authors proposed preselection of variants
568 using others statistics, such as the fixation index (F_{ST}) between groups of individuals
569 with high and low phenotype values to avoid the negative impact of spurious
570 associations [67].

571

572 ***New models and methods***

573 It is also likely that genomic prediction models, estimation methods, and their
574 implementations need to be improved to leverage the potential of WGS. This is an
575 active area of research and multiple novel methodologies have been proposed over the
576 last years. Some examples are a combination of subsampling and Gibbs sampling

577 [76], and a model that simultaneously fits a GBLUP term for a polygenic effect and a
578 BayesC term for variants with large effects selected by the model (BayesGC) [24].
579 Testing alternative models and methods for genomic prediction was out of the scope
580 of this study. However, together with refinements in the preselection of variants, it
581 remains an interesting avenue for further optimisation of the analysis pipeline.

582 Some of the most promising methods are designed to incorporate prior
583 biological information into the models. One of such methods is BayesRC [21], which
584 extends BayesR by assigning flatter prior distributions to classes of variants that are
585 more likely to be causal [17,20]. Similarly, GFBLUP [77] could be used to
586 incorporate prior biological information from either QTL databases or GWAS as
587 genomic features [19,34,65]. The model MBMG [26], which fits two genomic
588 relationship matrices according to prior biological information, has also been
589 proposed for multi-breed scenarios to improve genomic prediction in small
590 populations. Haplotype-based models have been shown to provide greater prediction
591 accuracy with WGS than variant-based models in pigs [78] and cattle [79]. However,
592 the uptake of such models has been limited so far due to additional complexity, for
593 example, to define haplotype blocks.

594

Suitability of whole-genome sequence data for genomic prediction

595 The small improvements in genomic prediction accuracy that we achieved
596 with WGS reflect the limited dimensionality of genomic information [61]. The WGS
597 variants only produce small increases in prediction accuracy compared to marker
598 arrays because the effective population size of intensely selected livestock populations
599 is typically small and marker arrays already capture a large proportion of their
600 independent chromosome segments. Thus, the use of WGS in current

601 implementations of genomic prediction should be carefully evaluated against the cost
602 of generating the WGS, especially given the large size of the datasets that are
603 required. Sequencing costs are expected to continue to decrease and therefore large
604 datasets of WGS will become more affordable in time, while efforts to develop and
605 optimise scalable and accurate pipelines for WGS-based data generation, storage, and
606 analysis are on-going (e.g., [80,81]). These advances, together with a finer knowledge
607 of the genetic architecture of traits empowered by WGS, could allow a case-by-case
608 refinement of genomic prediction. However, to date, the low robustness of the results
609 for complex traits discourage the generalised use of WGS for traits that are already
610 accurately predicted by conventional means.
611

Conclusion

612 Our results showed that WGS has some potential to improve genomic
613 prediction accuracy compared to marker arrays in intensely selected pig lines.
614 However, the prediction accuracy with each set of preselected WGS variants was not
615 robust across traits and lines and the improvements in prediction accuracy that we
616 achieved so far with WGS compared to marker arrays were generally small. The most
617 favourable results for WGS were obtained when the largest training sets were
618 available and used to preselect variants with statistically significant associations to the
619 trait for augmenting the established marker array. With this method and training sets
620 of around 80k individuals, average improvements of genomic prediction accuracy of
621 0.025 were observed in within-line scenarios. A combination of larger training sets
622 and improved pipelines could further improve genomic prediction accuracy. The
623 robustness of the whole strategy for generating WGS at the population level must be
624 carefully stress-tested and further optimised. However, with the current

625 implementations of genomic prediction, the use of WGS should be carefully evaluated
626 on a case-by-case basis against the cost of generating the WGS at a large scale.
627

Ethics approval and consent to participate

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

Consent for publication

630 Not applicable.

Availability of data and material

631 The software packages AlphaPhase, AlphaImpute, and AlphaPeel are available from
632 https://github.com/AlphaGenes. The software package AlphaSeqOpt is available from
633 the AlphaGenes website (http://www.alphagenes.roslin.ed.ac.uk). The datasets
634 generated and analysed in this study are derived from the PIC breeding programme
635 and not publicly available.

Competing interests

636 CYC, BDV, and WOH are employed by Genus PIC. The remaining authors declare
637 that the research was conducted in the absence of potential conflicts of interest.

Funding

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

Authors' contributions

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

Acknowledgements

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

651

References

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

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

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

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

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

668 6. Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, et
669 al. Large-scale whole-genome sequencing of the Icelandic population. Nat Genet.
670 2015;47:435–44.

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

673 8. Habier D, Fernando RL, Dekkers JCM. The impact of genetic relationship
674 information on genome-assisted breeding values. Genetics. 2007;177:2389–97.

675 9. Clark SA, Hickey JM, Daetwyler HD, Werf JH van der. The importance of
676 information on relatives for the prediction of genomic breeding values and the
677 implications for the makeup of reference data sets in livestock breeding schemes.
678 Genet Sel Evol. 2012;44:4.

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

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

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

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

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

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

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

699 17. Hayes BJ, MacLeod IM, Daetwyler HD, Bowman PJ, Chamberlain AJ, Vander
700 Jagt CJ, et al. Genomic prediction from whole genome sequence in livestock: the
701 1000 Bull Genomes Project. Proc 10th World Congr Genet Appl Livest Prod
702 WCGALP. Vancouver, BC, Canada; 2014. p. 183.

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

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

20. 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.

21. 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.

22. 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.

23. 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.

24. 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.

25. 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.

26. 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.

27. 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;

28. 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.

29. 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.

30. 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.

31. 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.

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

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

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

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

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

763 37. Johnsson M, Jungnickel MK. Evidence for and localization of proposed causative
764 variants in cattle and pig genomes. Genet Sel Evol GSE. 2021;53:67.

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

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

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

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

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

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

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

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

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

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

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

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

797 50. Ros-Freixedes R, Valente B, Chen C-Y, Herring WO, Gorjanc G, Hickey JM, et
798 al. Rare and population-specific functional variants across pig lines. Genet Sel Evol.
799 2022;54:39.

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

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

806 53. VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD,
807 Taylor JF, et al. Invited review: reliability of genomic predictions for North American
808 Holstein bulls. J Dairy Sci. 2009;92:16–24.

809 54. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. FaST
810 linear mixed models for genome-wide association studies. Nat Methods. 2011;8:833–
811 5.

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

815 56. Erbe M, Hayes BJ, Matukumalli LK, Goswami S, Bowman PJ, Reich CM, et al.
816 Improving accuracy of genomic predictions within and between dairy cattle breeds
817 with imputed high-density single nucleotide polymorphism panels. J Dairy Sci.
818 2012;95:4114–29.

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

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

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

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

830 61. Pocrnic I, Lourenco DAL, Masuda Y, Misztal I. Accuracy of genomic BLUP
831 when considering a genomic relationship matrix based on the number of the largest
832 eigenvalues: a simulation study. Genet Sel Evol GSE. 2019;51:75.

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

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

839 64. MacLeod IM, Bolormaa S, Schrooten C, Goddard ME, Daetwyler H. Pitfalls of
840 pre-selecting subsets of sequence variants for genomic prediction. Proc 22nd Conf
841 Assoc Adv Anim Breed Genet AAABG. Townsville, Queensland, Australia; 2017. p.
842 141–4.

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

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

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

852 68. van den Berg I, Boichard D, Guldbrandtsen B, Lund MS. Using Sequence
853 Variants in Linkage Disequilibrium with Causative Mutations to Improve Across-
854 Breed Prediction in Dairy Cattle: A Simulation Study. G3 GenesGenomesGenetics.
855 2016;6:2553–61.

856 69. Fragomeni BO, Lourenco DAL, Masuda Y, Legarra A, Misztal I. Incorporation of
857 causative quantitative trait nucleotides in single-step GBLUP. Genet Sel Evol.
858 2017;49:59.

859 70. Bolormaa S, Swan AA, Stothard P, Khansefid M, Moghaddar N, Duijvesteijn N,
860 et al. A conditional multi-trait sequence GWAS discovers pleiotropic candidate genes

861 and variants for sheep wool, skin wrinkle and breech cover traits. Genet Sel Evol.
862 2021;53:58.

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

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

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

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

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

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

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

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

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

890 80. Eggertsson HP, Jonsson H, Kristmundsdottir S, Hjartarson E, Kehr B, Masson G,
891 et al. Graphtyper enables population-scale genotyping using pangenome graphs. Nat
892 Genet. 2017;49:1654–60.

893 81. Talenti A, Powell J, Hemmink JD, Cook E a. J, Wragg D, Jayaraman S, et al. A
894 cattle graph genome incorporating global breed diversity. Nat Commun. 2022;13:910.

895

896

Figures

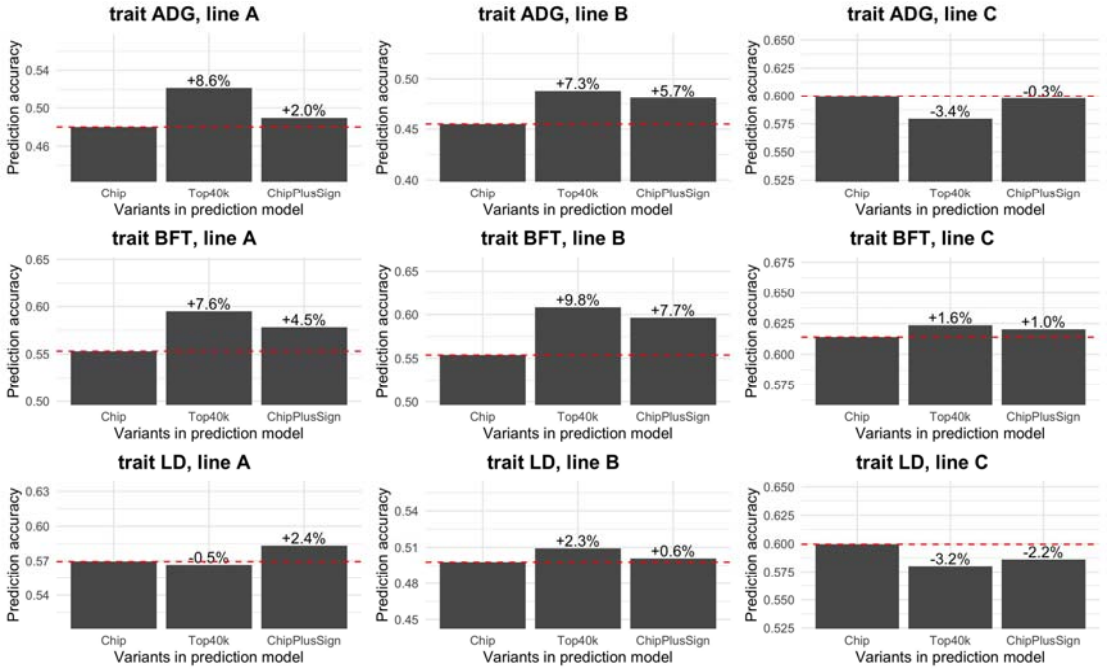
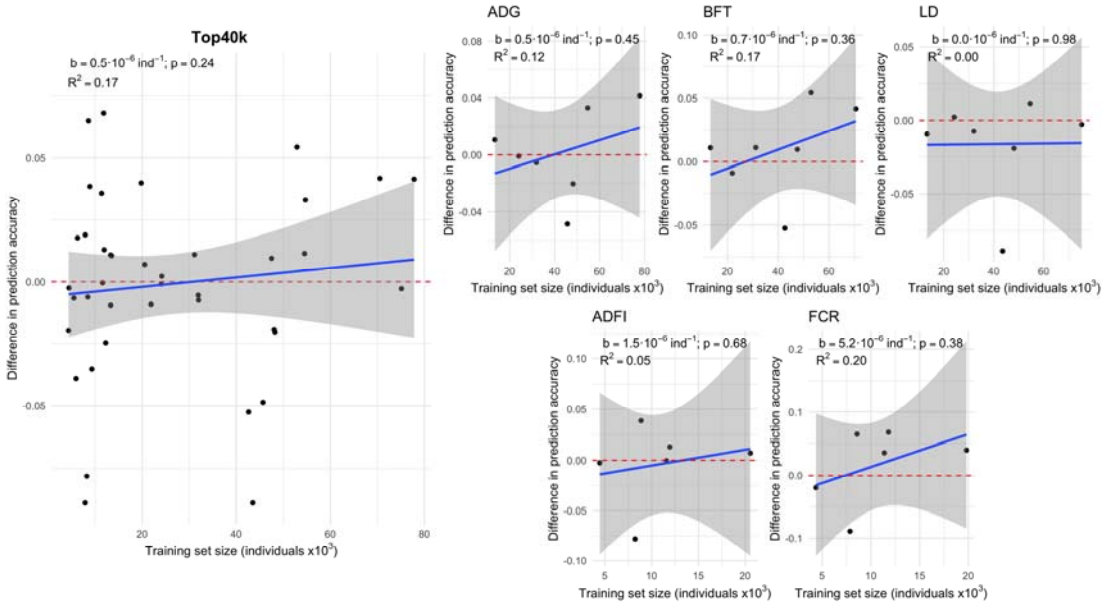


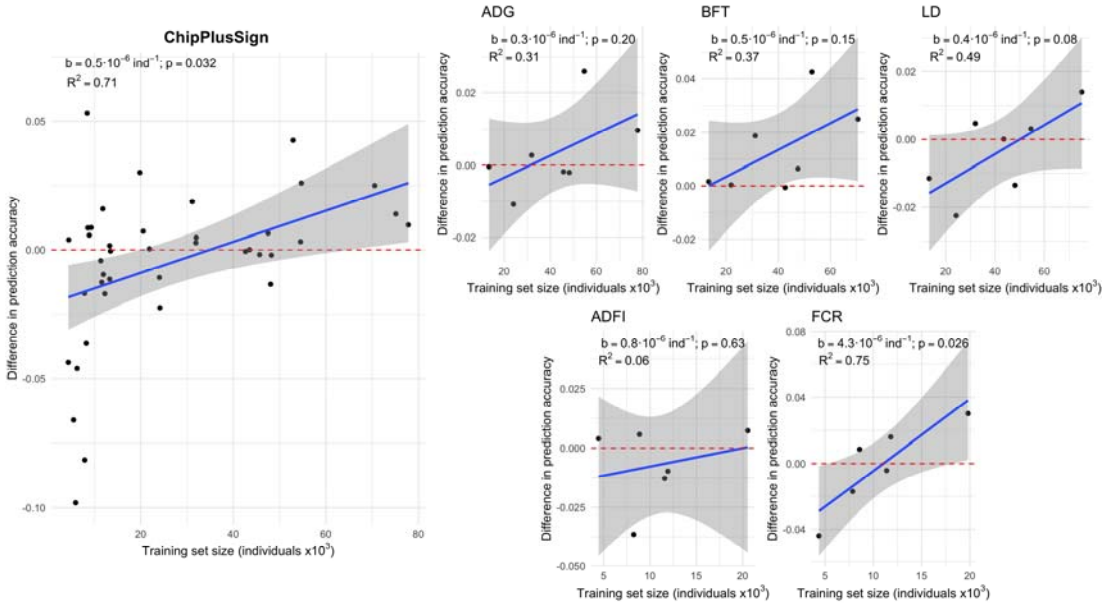
Figure 1. Genomic prediction accuracy for each set of variants for traits ADG, BFT, and LD in the three largest lines. Dashed line at value of marker array (Chip) as a reference. Values indicate relative difference to marker array (Chip).



903

904 **Figure 2.** Genomic prediction accuracy with the Top40k variants for the complex
905 traits. The difference between the Top40k and marker array is shown for all traits and
906 lines (left) or by trait (right). Red dashed line at 'no difference'. Regression
907 coefficient (b) and p-value of training set size is provided, as well as the coefficient of
908 determination (R^2) of the model. The linear model for the joint analyses included the
909 trait effect.

910



911

912 **Figure 3.** Genomic prediction accuracy with the ChipPlusSign variants for the
913 complex traits. The difference between the ChipPlusSign and marker array is shown
914 for all traits and lines (left) or by trait (right). Red dashed line at 'no difference'.
915 Regression coefficient (b) and p-value of training set size is provided, as well as the
916 coefficient of determination (R^2) of the model. The linear model for the joint analyses
917 included the trait effect.

918

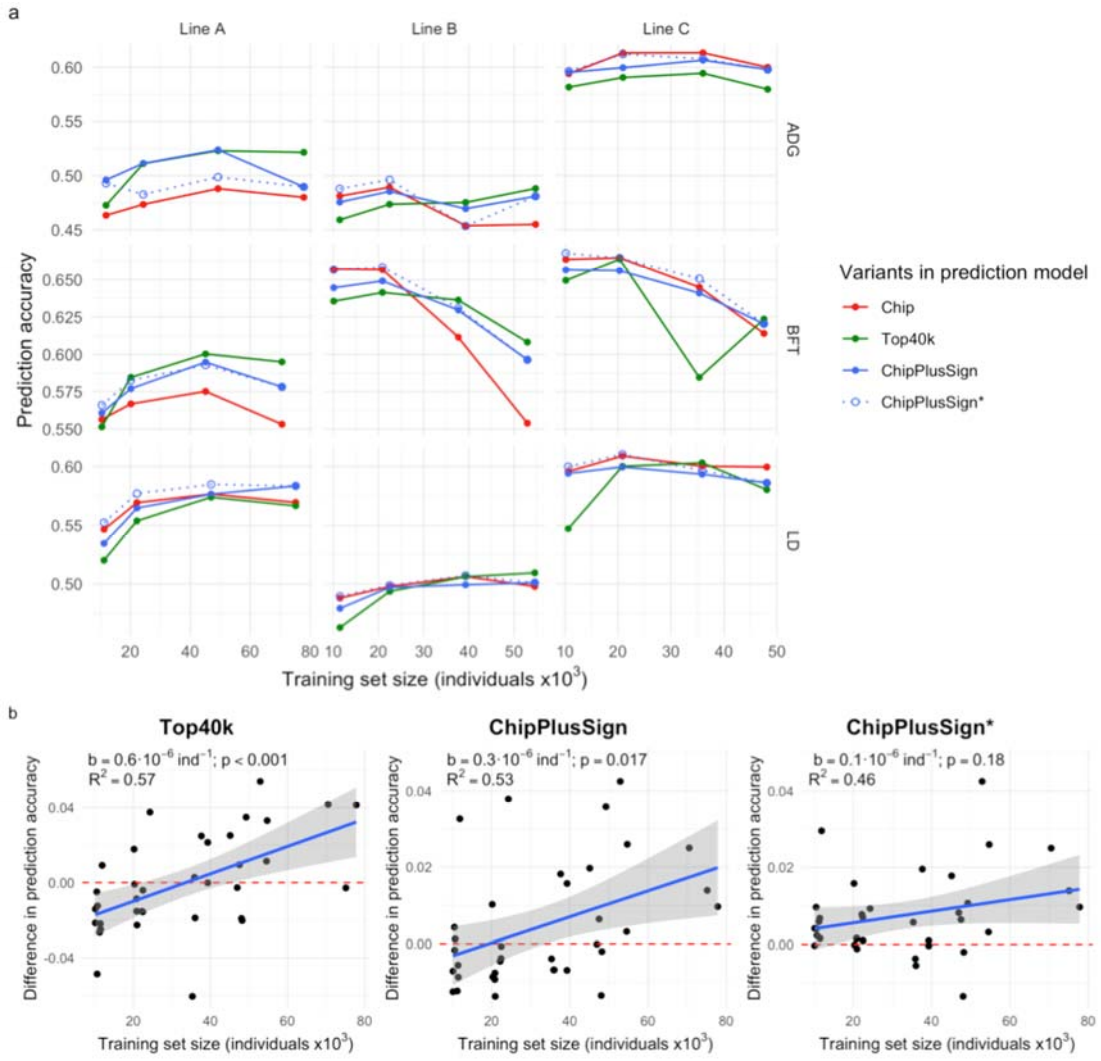
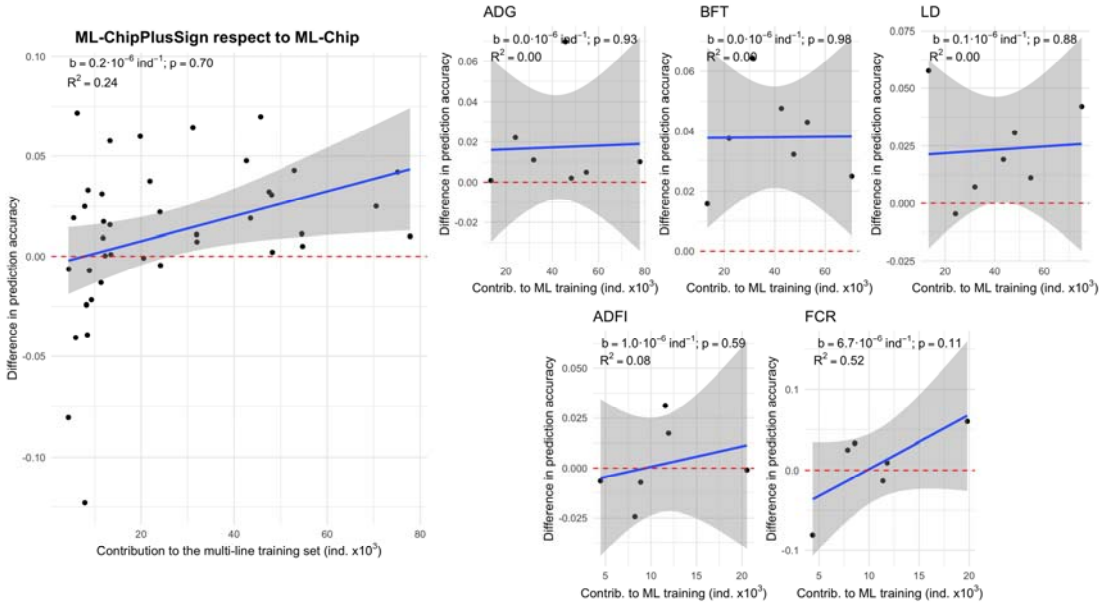


Figure 4. Effect of training set size on the genomic prediction accuracy for each set of variants for traits ADG, BFT, and LD in the three largest lines. **(a)** Genomic prediction accuracy with the marker array (Chip) or with preselected WGS data (Top40k, ChipPlusSign, and ChipPlusSign*). In ChipPlusSign* variants are preselected based on associations tested using the largest training set available. **(b)** The difference between the ChipPlusSign and Chip is shown for all traits and lines. Red dashed line at 'no difference'. Regression coefficient (b) and p-value of training set size is provided, as well as the coefficient of determination (R^2) of the model. The linear model for the joint analyses included the trait and line effects.



930

931 **Figure 5.** Genomic prediction accuracy with the ML-ChipPlusSign variants for the
932 complex traits. The difference between ML-ChipPlusSign and marker array (ML-
933 Chip) is shown for all traits and lines (left) or by trait (right). Red dashed line at 'no
934 difference'. Regression coefficient (b) and p-value of training set size is provided, as
935 well as the coefficient of determination (R^2) of the model. The linear model for the
936 joint analyses included the trait effect.

937

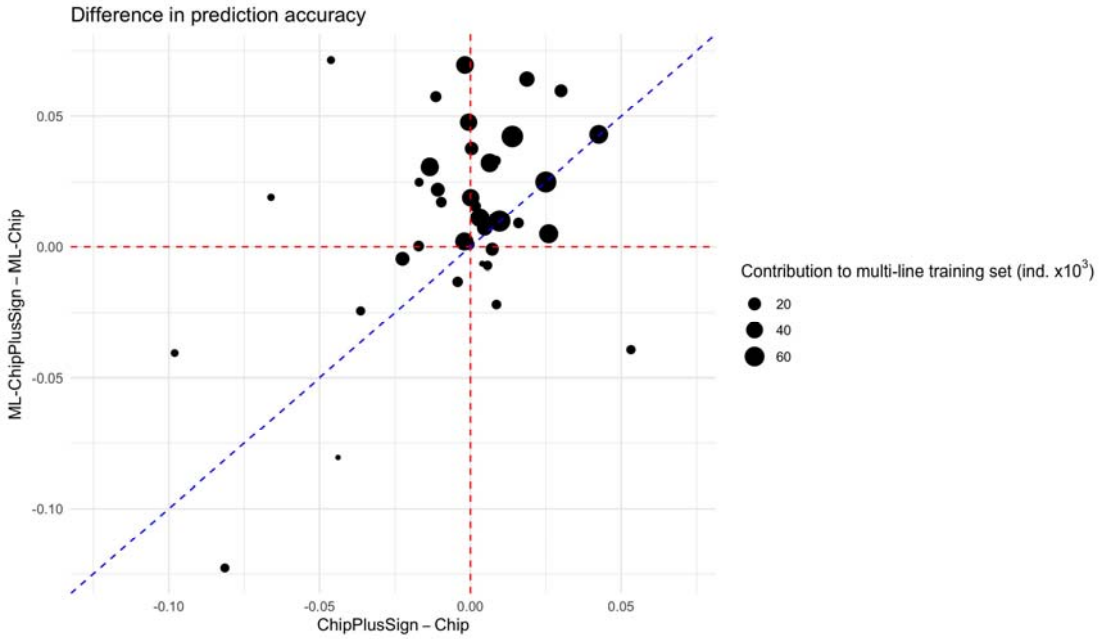


Figure 6. Comparison of the difference in genomic prediction accuracy in the multi-line scenarios (between ML-ChipPlusSign and ML-Chip) and in the within-line scenarios (between ChipPlusSign and Chip) for all traits and lines. Red dashed line at 'no difference'. Blue dashed line is the bisector.

Tables

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

Line	Individuals sequenced	Individuals sequenced by coverage				Individuals used in the analyses			
		1x	2x	5x	15–30x	Pedigree	LD	HD	Imputed
A	1,856	1,044	649	73	90	122,753	39,485	66,763	104,661
B	1,366	685	545	44	92	88,964	39,110	38,763	76,230
C	1,491	628	728	54	81	84,420	35,343	34,358	66,608
D	731	362	311	16	42	79,981	16,650	54,297	60,474
E	760	394	274	27	65	50,797	22,768	20,685	41,573
F	381	193	137	16	35	35,309	11,747	17,758	29,330
G	445	217	176	15	37	21,129	11,472	6,661	17,224

Pedigree number of individuals included in the pedigree used for imputation, *LD* number of individuals genotyped with the low-density marker array, *HD* number of individuals genotyped with high-density marker arrays, *Imputed* number of individuals with imputed genotypes that remain after filtering out individuals with low predicted imputation accuracy.

951 **Table 2.** Number of phenotype records per trait and line.

Trait	Set	A	B	C	D	E	F	G
ADG	Training	77,811	54,709	48,219	45,693	31,918	24,046	13,479
	Test	9,435	8,387	6,977	4,789	3,019	1,808	1,572
BFT	Training	70,529	52,910	47,512	42,636	31,127	21,892	13,300
	Test	8,560	7,957	6,747	4,301	2,936	1,602	1,568
LD	Training	75,117	54,537	48,054	43,517	31,987	24,154	13,303
	Test	9,021	8,415	6,995	4,411	3,024	1,807	1,570
ADFI	Training	20,535	8,866	8,235	11,573	11,930	4,000	4,457
	Test	1,358	638	802	641	478	97*	364
FCR	Training	19,805	8,572	7,857	11,378	11,804	3,900	4,364
	Test	1,328	624	775	624	477	97*	360
TNB	Training	12,250	9,315	8,438	7,700	5,834	-	2,865
	Test	254	428	400	23*	125*	-	98*
LWW	Training	-	7,884	6,251	-	-	-	2,505
	Test	-	246	220	-	-	-	47*
RET	Training	-	5,928	5,496	-	-	-	1,481
	Test	-	332	282	-	-	-	70*

952 *ADG* average daily gain, *BFT* backfat thickness, *LD* loin depth, *ADFI* average daily

953 feed intake, *FCR* feed conversion ratio, *TNB* total number of piglets born, *LWW* litter

954 weight at weaning, *RET* return to oestrus 7 days after weaning.

955 *Included in multi-line scenarios, but excluded in within-line scenarios because of the

956 limited size of the testing set.

957

958 **Table 3.** Number of significant variants from the whole-genome sequence data that
959 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

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

963