


RESEARCH ARTICLE

Open Access



Genomic evaluation of feed efficiency component traits in Duroc pigs using 80K, 650K and whole-genome sequence variants

Chunyan Zhang¹, Robert Alan Kemp², Paul Stothard¹, Zhiquan Wang¹, Nicholas Boddicker², Kirill Krivushin¹, Jack Dekkers³ and Graham Plastow^{1*} 

Abstract

Background: Increasing marker density was proposed to have potential to improve the accuracy of genomic prediction for quantitative traits; whole-sequence data is expected to give the best accuracy of prediction, since all causal mutations that underlie a trait are expected to be included. However, in cattle and chicken, this assumption is not supported by empirical studies. Our objective was to compare the accuracy of genomic prediction of feed efficiency component traits in Duroc pigs using single nucleotide polymorphism (SNP) panels of 80K, imputed 650K, and whole-genome sequence variants using GBLUP, BayesB and BayesRC methods, with the ultimate purpose to determine the optimal method to increase genetic gain for feed efficiency in pigs.

Results: Phenotypes of average daily feed intake (ADFI), average daily gain (ADG), ultrasound backfat depth (FAT), and loin muscle depth (LMD) were available for 1363 Duroc boars from a commercial breeding program. Genotype imputation accuracies reached 92.1% from 80K to 650K and 85.6% from 650K to whole-genome sequence variants. Average accuracies across methods and marker densities of genomic prediction of ADFI, FAT, LMD and ADG were 0.40, 0.65, 0.30 and 0.15, respectively. For ADFI and FAT, BayesB outperformed GBLUP, but increasing marker density had little advantage for genomic prediction. For ADG and LMD, GBLUP outperformed BayesB, while BayesRC based on whole-genome sequence data gave the best accuracies and reached up to 0.35 for LMD and 0.25 for ADG.

Conclusions: Use of genomic information was beneficial for prediction of ADFI and FAT but not for that of ADG and LMD compared to pedigree-based estimates. BayesB based on 80K SNPs gave the best genomic prediction accuracy for ADFI and FAT, while BayesRC based on whole-genome sequence data performed best for ADG and LMD. We suggest that these differences between traits in the effect of marker density and method on accuracy of genomic prediction are mainly due to the underlying genetic architecture of the traits.

Background

Feed is of major economic importance in pig production, accounting for 60 to 70% of total costs. The grow-finish phase accounts for the largest proportion of total feed, at about 75% [1]. Thus, improving grow-finish feed efficiency will significantly reduce production cost and increase profitability. Although intense selection for lean growth has improved feed efficiency dramatically in the

past decades, with feed conversion ratio (FCR) values of 2.0 or less currently achievable [1], further improvements require direct measurement and selection on feed intake (FI) and other components of feed efficiency. This is especially the case for high-quality products with increased marbling, since fat deposition has a high genetic correlation with FI (0.37 [2]). However, the expense of recording FI on large numbers of selection candidates limits the opportunities of using this approach. Genomic selection (GS) or prediction is a promising approach to address this issue, since it allows for early selection among candidates without FI records, higher rates of genetic gain, and

*Correspondence: plastow@ualberta.ca

¹ Department of Agricultural, Food and Nutritional Science, University of Alberta, Edmonton, AB T6G 2R3, Canada

Full list of author information is available at the end of the article



better management of inbreeding, compared with traditional selection based on pedigree and phenotype [3, 4].

GS has been widely applied in livestock breeding programs, using medium-to-high density single nucleotide polymorphism (SNP) panels [5]. The most successful implementation of GS is in dairy cattle, which has made it possible to reduce generation intervals and costs by eliminating progeny testing [6, 7]. Unlike dairy cattle, where the biggest impact is on reducing generation interval [8], the largest benefit for pigs is in increasing the accuracy of selection for traits such as feed intake. However, implementation of GS in pigs is still very limited [9–12], which might be due to the low monetary value of a boar compared to a dairy bull and the relatively low genomic prediction power for pigs in most breeding programs, due to not having access to large numbers of animals that have the necessary phenotype and genotype records compared to dairy cattle (primarily for Holsteins). It was anticipated that these limitations could be addressed by increasing the numbers of animals with quality phenotypes that are genotyped and the number of markers used (especially for markers that are in linkage disequilibrium (LD) with the underlying causative mutations) or by using the causative mutations themselves [7]. Using whole-genome sequence data is also expected to increase the accuracy of genomic prediction, since all or most of the causal mutations that underlie quantitative traits loci (QTL) are expected to be included in the data. Inclusion of the causal mutations is expected to increase the accuracy of genomic prediction across generations and even across breeds [7]. This was confirmed using simulated data [13–16] but, in practice, the use of imputed sequence data in cattle and chicken has shown little increase (0–3%) in the accuracy of genomic prediction [17–21]. Many factors can influence the accuracy of genomic prediction, including the genetic architecture of the traits, the statistical method applied [13, 22], marker density, LD between QTL and SNPs [16], effective population size [19, 23, 24], size of the reference population, relatedness of selection candidates with individuals in the training data [13, 22, 25], and imputation accuracy of marker genotypes [14]. The availability of higher density SNP panels and sequence information for pigs provided the opportunity to examine this for feed efficiency in a commercial Duroc breeding population.

Therefore, this study aimed at evaluating the accuracy of genomic prediction of feed efficiency component traits of average daily feed intake, average daily gain, ultrasound backfat depth, and loin muscle depth, using 80K and imputed 650K SNPs, as well as imputed whole-genome sequence variants. Three methods, GBLUP [26], BayesB [27, 28] and BayesRC [19], were compared to determine the best method and marker density for each

trait. Possible factors that influence the accuracy of genotype imputation and genomic prediction were also discussed. The ultimate aim was to investigate the feasibility and optimal approach for using genomic information to increase genetic gain for feed efficiency in pigs.

Methods

Ethics statement

Data were collected at the Prairie Sun Research and Development Facility (Genesis Inc., Oakville, MB). All animals used in this study were raised under commercial production-like conditions and fed standard diets designed to exceed the pig's requirements, as described previously [29]. The proposed work was reviewed by the University of Alberta Animal Care and Use Committee. No other specific permissions were required for the work, since the animals were cared for according to the Canadian Quality Assurance Program, which includes attention to animal health and well-being and is in line with the Canadian Council on Animal Care guidelines.

Animals and data collection

A total of 1363 Duroc boars (from 63 sires and 439 dams) tested in 2014 were used for this study. At weaning, on average, two boars per litter were selected to create a group of 24 or 48 boars, depending on the number of litters weaned in a given week. The average genetic relationship among these 1363 individuals was about 0.12 based on pedigree information. The boars were placed in nursery pens at a stocking density of 24 per pen, with littermates split between the two pens when groups of 48 were stocked. At completion of the nursery phase (approximately 9 weeks of age), each group of boars was put into a single test pen (22 to 24 boars per pen) that was fitted with two electronic feeders per pen (IVOG, Insentec BV, Marknesse, the Netherlands). Boars from a nursery pen were kept together in the test pen. Following a 7-d acclimation period, feed intake was recorded in a test period of 14 weeks. Body weights were recorded at the beginning (~45 kg) and end (~110 kg) of the test, with an intermediate weight of ~80 kg. In addition, when average weight in the pen was near 110 kg (actual weight 112 ± 11.05 kg, actual age 155 ± 7.27 d), boars were individually weighed and depths of backfat (FAT) and *longissimus* muscle (LMD) were measured approximately 7 cm off the midline over the last three ribs using ultrasound (Aloka 500, Imagemedical Inc., QC) and Biotronics Toolbox Software (Biotronics Inc., Ames, IA).

Individual meal events were edited to remove outliers and obvious errors using adapted procedures recommended by Casey et al. [30], as described in [29]. All boars had to have a minimum of 63 valid feed intake days to pass the edits, along with a minimum of two valid feed

intake days per week while on test. Following these edits, daily feed intake was calculated as the sum of individual feed intake events per day. Average daily feed intake (ADFI) was calculated as the predicted feed intake at the midpoint age on test for each boar based on intra-pig linear regression of daily feed intake on age. Average daily gain (ADG) was calculated using linear regression of weight on age using the weights recorded at the start and end of test, along with one or two intermediate weights, with a minimum of two weeks between any two weight records. All phenotypic records (ADFI, ADG, FAT, and LMD) were further edited by removing observations that were more than three standard deviations from their respective means. After editing, all traits followed a normal distribution and were used for further analysis.

Variant genotyping and imputation

Genomic DNA was isolated from tail tissue samples following the DNA Extraction instruction manual (Thermo Fisher Scientific Ltd., Ottawa, ON, Canada). Samples from all animals (1363) with phenotypic records were genotyped using the Geneseek-Neogen GPPHD 80K SNP chip. A deep pedigree for these animals was traced back ~8 generations. The common ancestors and their genetic contribution to the studied population (1363) were calculated using the PEDIG program [31]. On the basis of “the proportion of genetic diversity” strategy, as suggested by Druet et al. [14], the top 29 ancestors (22 boars and 7 sows) based on their genetic contributions to the 1363 evaluated animals that had available tissue samples, were selected for next-generation sequencing (with an average 12-fold coverage). These ancestors cumulatively contributed about 70% of the genetics of the studied population. To improve imputation accuracy, 171 animals were genotyped with the Affymetrix Axiom® 650K SNP Array, including: (1) the 94 sires, maternal grand-sires/great-grand-sires of the 1363 animals, (2) the 29 sequenced animals, (3) 19 sons of the sequenced animals, and (4) the next 29 ancestors (19 boars and 10 dams), which cumulatively contributed about 20% of the genetics of the studied population. In order to test the accuracy of imputed genotypes across three different genotyping platforms, the 29 sequenced animals and 67 of the animals with 650K genotypes were also genotyped with the 80K SNP chip. All genotyping and sequencing analyses were conducted by Delta Genomics (Edmonton, AB, Canada). Library construction for next-generation sequencing was performed with 1 µg of genomic DNA according to library preparation protocols (Bio-O Scientific NEXTflex™ DNA Sequencing Kit). The Illumina 100 paired-end sequencing kit was used for sequencing on an Illumina HiSeq 2000 PE100. Variant calling was performed according to GATK Best Practices work flow [32,

33]. More specifically, Illumina reads were aligned to the reference genome (*Sscrofa 10.2*) using BWA [34]. Then, duplicates were marked and GATK INDEL realignment [35] and base quality score recalibration were applied. After that, we performed variant calling with Haplotype-Caller and joint genotyping on all samples. Finally, SNPs and Indels were filtered using parameters recommended by GATK Best Practices [32, 33].

A total of 16,560,854 autosomal variants were detected in the 29 sequenced animals, including 2,576,543 Indels and 13,984,543 SNPs. Before imputation, alleles for all SNPs on the 80K and 650K panels were converted to the standard reference (*Sscrofa 10.2*), with the reference-based allele denoted 0 and the alternate allele denoted 1. SNPs or variants for each genotyping platform were filtered for analysis according to the following criteria: SNP or variant call rate higher than 95%, SNP or variant with map information on autosomes (*Sscrofa 10.2*), Chi square of Hardy–Weinberg equilibrium test less than 600, and minor allele frequency (MAF) in the genotyped animals higher than 5%. Stepwise imputation from 80K to 650K and then to the whole-genome sequence was performed by Fimpute v2.2 [36] with inclusion of pedigree information. Leave-one-out cross-validation using the 96 animals that had both 80K and 650K genotypes and the 29 animals that had both 650K and whole-genome sequence genotypes was used to evaluate the imputation accuracy in each step. Only SNPs or variants with an imputation accuracy higher than 95% were used for further analysis. Genotype imputation accuracy was defined as the percentage of correctly imputed genotypes among the animals. Finally, 38,440 SNPs remained from the 80K panel, 429,130 SNPs remained from the 650K panel, and 4,844,535 variants were contained in the imputed whole-genome sequence.

Genomic evaluation

Phenotype correction and estimation of breeding values

Significance of all possible systematic effects on phenotype, including the fixed effects of contemporary group (78 levels) consisting of ultrasonic test date and grow-finish pen, ultrasonic test machine (two levels, for FAT and LMD only), and the covariate of animal age at the end of the test (140 to 170 days), were tested using the following univariate animal model in ASREML [37]:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e}, \quad (1)$$

where \mathbf{y} is the vector of observations for the trait, \mathbf{b} is a vector of fixed effects (contemporary group and machine) and covariate (age), \mathbf{a} is a vector of random additive genetic effects [$\mathbf{a} \sim N(\mathbf{0}, \mathbf{A} \times \sigma_a^2)$], where \mathbf{A} is the additive genetic relationship matrix constructed using pedigree and σ_a^2 is the additive genetic variance, \mathbf{e} is a vector

of random residuals [$\mathbf{e} \sim N(\mathbf{0}, \mathbf{I} \times \sigma_e^2)$], where \mathbf{I} is the identity matrix and σ_e^2 is the residual variance, and \mathbf{X} and \mathbf{Z} are incidence matrices associating \mathbf{b} and \mathbf{a} with \mathbf{y} . Only significant ($P < 0.01$) fixed effects were included in the final model to estimate the variance components and residuals of the traits. The effects of contemporary group and animal age were significant for all traits, and ultrasonic test machine was significant for FAT and LMD. The interaction between contemporary group and ultrasonic test machine was not significant. Corrected phenotypes were calculated as the sum of the estimated breeding value and the estimated residuals from the above univariate pedigree-based animal model.

Then, the 1363 Duroc boars were split into training ($n=1167$) and prediction datasets ($n=196$) based on birthdate, before and after June 10, 2014, respectively. The 196 youngest animals for prediction were from 19 sires and 88 dams, and almost all had half-sibs in the training dataset. The genetic relationship between individuals in the training and prediction datasets averaged 0.11 based on pedigree data. First, a full animal model (all available phenotypes) was used to obtain estimated breeding values (EBV), i.e. EBV1, and corrected phenotypes (y_{c1}) for the validation animals. These y_{c1} were used to measure the accuracy and bias of all prediction models. Second, a reduced animal model (masking the phenotypes of validation animals) was used to calculate the EBV (EBV2) of the validation animals and corrected phenotypes (y_{c2}) of training animals. These y_{c2} of training animals were used as pseudo-phenotypes in the BayesRC method (see below) to estimate the effect of SNPs. The resulting EBV2 of validation animals were then used to evaluate the pedigree-based prediction ability (BLUP method below).

Pre-selection, biological priors and classification of whole-genome sequence variants

The top variants (SNPs and Indels) were selected from the imputed whole-genome sequence data based on their effects on phenotype, as estimated in the training dataset ($n=1167$) using method BayesB in GenSel [27, 28]. The following model was used:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \sum_j^k \mathbf{z}_j \alpha_j \delta_j + \boldsymbol{\varepsilon}, \quad (2)$$

where \mathbf{y} is the vector of observations for the traits, \mathbf{b} is a vector of the significant fixed effects and covariate, as described in Eq. (1), \mathbf{z}_j is the vector of genotype covariates ($-10/0/10$) across animals for SNP j ($j=1$ to k), α_j is the allele substitution effect for SNP j , and δ_j is an indicator for whether SNP j was included ($\delta_j=1$) or excluded ($\delta_j=0$) in the model for a given Markov chain Monte

Carlo (MCMC) iteration. A total of 50,000 iterations were run for each analysis, with the first 5000 iterations used as burn-in. The prior probability of a SNP to have no effect was set equal to $\pi=0.9995$ based on the posterior value obtained from BayesC π . Due to the computational demands associated with testing the very large number of sequence variants ($\sim 4.8 \times 10^6$) simultaneously, an alternative split-and-merge method was used, similar to Calus et al. [17]. Briefly, on each chromosome, the sequence-based variants were extracted and merged with the SNPs from the 80K SNP panel on the other chromosomes to generate sub-datasets ($n=18$). The association analysis was then conducted separately on each sub-dataset using BayesB. Subsequently, results from all sequence variants across all chromosomes were combined and ordered according to the absolute value of the estimated marker effect from highest to lowest for each trait. The top 0.05% (equal to $1-\pi$) of variants were considered to have an important effect on the trait and selected as markers that were given a different prior for sequence variant classification (see below). Finally, 7855 markers were selected for the four traits (2025 for each trait, with 245 shared between at least two traits).

The imputed whole-genome variants were annotated based on the *Sscrofa 10.2* assembly of the swine genome using NGS-SNP [38]. All variants were then defined as belonging to one of three broad categories, as suggested by MacLeod et al. [19]. The first category, which will be referred to as “NSC”, comprised variants that were statistically associated with the traits (preselected from genome-wide association analyses (GWAS), as described above) and variants predicted to cause a non-synonymous coding change, including missense variants, splice site variants, in-frame Indels, frame shift variants, and stop gained/lost mutations. The second category, referred to as “REG”, included variants in regions that were predicted to have potential regulatory roles, mainly those within 5000 bp upstream and downstream of genes, variants in the 3' or 5' untranslated genic regions, and non-coding exon variants. All other variants were allocated to the third category, referred to as “CHIP”. These were mainly intergenic but included some intronic and synonymous coding variants. Then, the imputed whole-genome sequence variants were further filtered based on LD using PLINK [39] by excluding a random variant of a pair of variants that were in complete LD ($r^2 > 0.99$) in a 5000-kb sliding-window with 50 variants. LD pruning was carried out first independently within each category (NSC, REG and CHIP) and then any REG or CHIP variant that was in complete LD with an NSC variant was removed. Finally, all CHIP variants that were in complete LD with a REG variant were removed. The remaining 2,154,844 variants, henceforth referred to as “SEQ”, were

used for genomic prediction. They included 13,642 NSC, 157,809 REG and 1,983,393 CHIP variants.

Genomic prediction

Genomic predictions for the validation animals were estimated based on their genotypes (38,440 from 80K, 429,130 from imputed 650K and 2,154,844 from SEQ) and the marker effects estimated in the training dataset using three methods: GBLUP [26], BayesB [27, 28] and BayesRC [19] (the latter was only used for “SEQ”). Accuracy of prediction was evaluated by correlating the genomic breeding value of the validation animals with their corrected phenotype and dividing by the square root of the heritability of the trait. Bias of genomic predictions was estimated as the linear regression of predictions on corrected phenotypes for the validation animals, with a regression coefficient equal to 1 indicating no bias. Corrected phenotypes used for validation were obtained from analysis of the full dataset using the model of Eq. (1), as the sum of the pedigree-based EBV1 and residuals. The accuracy of genomic predictions was compared to the accuracy of pedigree-based predictions of the validation animals, which were obtained by fitting the model of Eq. (1) to the dataset with phenotypes for validation animals masked.

GBLUP

The genomic relationship matrix (\mathbf{G}) based on each of the three sets of genotypes was calculated using PLINK. The GBLUP approach was applied to the model of Eq. (1), but using the genomic relationship matrix \mathbf{G} , instead of the pedigree-based relationship matrix, and with the phenotypes of validation animals masked.

BayesB

In the Bayesian approach, first the fraction of loci with no effect, π , was estimated using method BayesC π in GenSel, using the full dataset. The posterior mean of π was similar for all traits, at approximately 0.99, 0.999, 0.9995 for the 80K, 650K and SEQ genotypes, respectively. Then, the BayesB method using the model of Eq. (2) was applied to genotypes and phenotypes of the training dataset with the corresponding estimates of π to simultaneously estimate effects of SNPs across the entire genome for the 80K, 650K and SEQ genotypes. The total number of iterations was 80,000, with 10,000 discarded as burn-in. Then the genomic prediction for the animals in the validation dataset were computed as in Eq. (3):

$$\text{GEBV}_i = \sum_{j=1}^k z_{ij} \hat{\alpha}_j, \quad (3)$$

where GEBV_i is the genomic EBV for validation animal i , $j=1$ to k is the number of SNPs in the respective genotype datasets, z_{ij} is the SNP genotype code ($-10/0/10$) for validation animal i for SNP j , and $\hat{\alpha}_j$ is the effect estimate for SNP j obtained from BayesB according to Eq. (2).

BayesRC

BayesRC was applied to the SEQ variants only, following MacLeod et al. [19]. Briefly, BayesRC uses an MCMC approach to estimate variant effects that are modelled as a mixture of four normal distributions, including a null distribution, $N(0, 0.0 \times \sigma_g^2)$, and three others: $N(0, 0.0001 \times \sigma_g^2)$, $N(0, 0.001 \times \sigma_g^2)$, $N(0, 0.01 \times \sigma_g^2)$, where σ_g^2 is the additive genetic variance for the trait based on whole-sequence genotypes. The first distribution accommodates the likelihood that many variants have no effect on the trait, thus reducing the complexity of the model. The model fitted to the datasets was:

$$y_{c2} = \mathbf{1}\mu + \mathbf{Za} + \mathbf{Wv} + \mathbf{e}, \quad (4)$$

where y_{c2} is the corrected phenotype for the trait, \mathbf{Z} is the design matrix allocating phenotypes to polygenic breeding values, \mathbf{a} is the vector of polygenic breeding values [$N(\mathbf{0}, \mathbf{A} \times \sigma_a^2)$], with \mathbf{A} as the genetic relationships calculated from pedigree and σ_a^2 as the additive genetic variance not explained by the variants, \mathbf{W} is the design matrix of variant genotypes (0/1/2), centred and standardized to have unit variance, \mathbf{v} is the vector of estimated variant effects based on a mixture of the four distributions as listed above, and \mathbf{e} is the vector of random residuals.

Prior independent biological information was used to allocate each variant to a “class” c ($c=3$), as described above, where the purpose is to provide one or more classes that are expected to be enriched for QTL or for variants linked to the QTL. As described by MacLeod et al. [19], within each class c , a uniform *Dirichlet* prior was used for the proportion of effects in each of the four normal distributions of SNP effects.

For all traits, we implemented five replicate chains of 80,000 iterations of the Gibbs sampler, with 10,000 iterations discarded as burn-in. Very good agreement was found in the final results across the five replicate chains (correlation of posterior estimates of marker effects equal to 0.999). Final estimates were derived from the means of the five replicate chains. Using the resulting posterior means of marker effects, the genomic breeding value for the validation animals were calculated using Eq. (3).

Results

Genotype imputation accuracy

The average genotype imputation accuracy for individual SNPs was 92.1% from 80K to 650K and 85.6% from 650K to whole-genome sequence, with the complete range from 0 to 100% across SNPs. Most SNPs had an imputation accuracy higher than 90%, 77% of SNPs for imputation from 80K to 650K and 57% for imputation from 650K to sequence. About 12.6 and 25.9% of SNPs had an imputation accuracy lower than 80% for imputation of 80K–650K and of 650K to sequence imputation, respectively (Table 1). Only variants with an imputation accuracy higher than 95% were kept for final genomic prediction.

Genomic prediction accuracy

Genomic prediction versus pedigree-based prediction

The accuracy and bias of (G)EBV for the studied traits are in Table 2. Generally, the average accuracy of GEBV was moderate to high for ADFI (0.40) and FAT (0.65), and relatively low for LMD (0.30) and ADG (0.15). Compared with the pedigree-based evaluation, the use of genomics was beneficial for ADFI and FAT, with smaller bias and an accuracy that was improved by on average 42.9 and

32.7%, respectively. However, for ADG and LMD, pedigree-based prediction gave better accuracy and smaller bias, and no improvement was observed from using genomic data.

Bayesian methods versus GBLUP

Improvement in the accuracy of genomic predictions based on BayesB compared with GBLUP is shown in Fig. 1. Generally, BayesB performed better than GBLUP for ADFI and FAT for all three sets of genotypes (positive in Fig. 1). For ADG and LMD, GBLUP gave higher accuracy using 80K and 650K SNPs (negative in Fig. 1), but little difference in accuracy was observed between the two methods when using SEQ data. When applied to the SEQ data, BayesRC resulted in higher accuracy than BayesB and GBLUP for both ADG and LMD. For ADFI and FAT, the accuracy from BayesRC was between those from GBLUP and BayesB (Table 2).

Table 1 Percentage of SNPs in different ranges of imputation accuracy from 80K to 650K and 650K to sequence

Range of imputation accuracy (%)	80K to 650K	650K to sequence
< 80	12.6	25.9
80–85	4.2	5.3
85–90	6.2	11.4
90–95	11.8	10.5
> 95	65.2	46.9

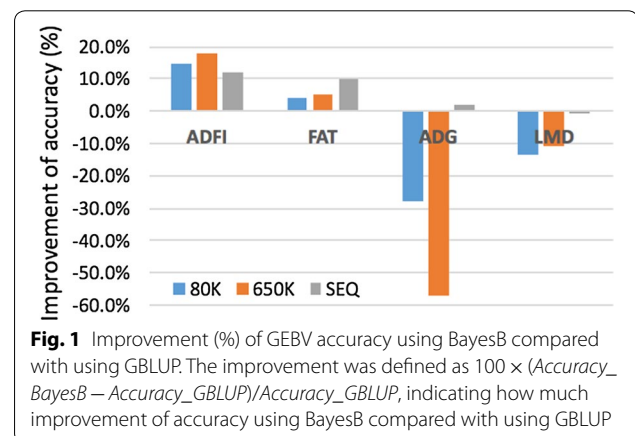


Fig. 1 Improvement (%) of GEBV accuracy using BayesB compared with using GBLUP. The improvement was defined as $100 \times (Accuracy_{BayesB} - Accuracy_{GBLUP}) / Accuracy_{GBLUP}$, indicating how much improvement of accuracy using BayesB compared with using GBLUP

Table 2 Accuracy and bias of (G)EBV evaluated using pedigree, 80K, 650K and SEQ data using different prediction methods

Resource	Method	ADFI		FAT		ADG		LMD	
		Accuracy	Bias	Accuracy	Bias	Accuracy	Bias	Accuracy	Bias
Pedigree	BLUP	0.28	0.83	0.49	0.91	0.28	0.53	0.42	1.17
80K	GBLUP	0.38	0.96	0.66	0.98	0.17	0.31	0.29	0.63
	BayesB	0.44	1.14	0.68	1.12	0.12	0.23	0.25	0.59
650K	GBLUP	0.38	0.99	0.64	0.95	0.20	0.38	0.29	0.6
	BayesB	0.45	1.17	0.68	1.17	0.09	0.15	0.26	0.59
SEQ	GBLUP	0.37	0.95	0.59	0.96	0.12	0.28	0.32	0.69
	BayesB	0.41	1.07	0.65	1.27	0.12	0.21	0.32	0.77
	BayesRC	0.40	0.64	0.62	0.81	0.25	0.32	0.35	0.64
Average accuracy of using genomic data		0.40	0.97	0.65	1.02	0.15	0.30	0.30	0.71

ADFI average daily feed intake, FAT ultrasound backfat depth, ADG average daily gain, LMD ultrasound loin muscle depth

Accuracy from different marker densities

The change in the accuracy of genomic predictions with increasing SNP density is in Fig. 2. Increasing the SNP density slightly decreased the prediction accuracy for FAT, for which use of 80K SNPs gave the best accuracy regardless of the statistical method used. For ADFI, use of SEQ data decreased the accuracy compared with the SNP panels, and little difference in accuracy was observed between 80K and 650K. For LMD, increasing the number of SNPs resulted in similar or greater accuracy for both GBLUP and BayesB. For ADG, almost no improvement in accuracy was observed with increasing marker density. In conclusion, SEQ data with the BayesRC method gave the best accuracy for ADG (0.25) and LMD (0.35), while use of 80K SNPs with the BayesB method gave the best accuracy for ADFI (0.44) and FAT (0.68).

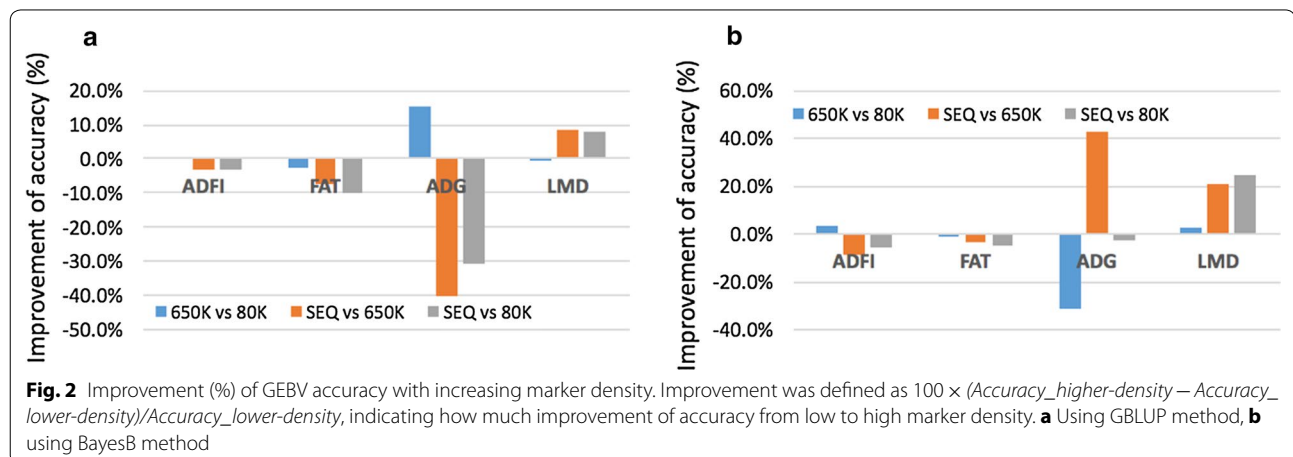
Discussion

Genotype imputation

In the current study, we obtained high imputation accuracies, which reached 0.92 for imputation from 80K to 650K SNPs. In pigs, there are several reports on the accuracy of genotype imputation from lower densities to 60K SNPs, with correlations between observed and imputed genotypes ranging from 0.952 to 0.995 for imputation from 6K to 60K [40], from 0.879 to 0.991 for imputation from 3K or 6K to 60K [41, 42], and from 0.88 to 0.95 for imputation from 9K to 60K in different scenarios [9]. However, imputation using high-density and whole-genome sequence using pig data has not been reported to date. In beef cattle, the percentage of correctly imputed genotypes was on average 95% for imputation from 3K to 50K [43] and ranged from 84 to 99% for imputation from 50K to 777K [24, 44]. The squared correlation (R^2) between imputed and observed genotypes ranged from 0.80 to 0.96 for imputation from different low densities

to 50K [45] and from 0.90 to 0.96 for imputation from 50K to 777K [46]. For sequence imputation, research in Holstein–Friesian bulls has shown that stepwise imputation (50K–777K to sequence) yields higher accuracy (correlation between observed and imputed genotypes) (0.77 to 0.83) than using a one-step method, which had accuracies ranging from 0.37 to 0.46 for imputation from 50K to sequence and from 0.77 to 0.83 for imputation from 777K to sequence [47]. Imputation accuracy measured as the percentage of variants correctly imputed was on average 85.6% to whole-genome sequence data in the current study, which was lower than that obtained in Holstein cattle (97%) with a large multi-breed reference population ($n = 444$) [48].

Many factors can influence the accuracy of genotype imputation. MAF is one important factor, especially when imputing to sequence data, since the number of SNPs with a very low MAF is usually limited in SNP panels but large in sequence data (see Additional file 1: Fig. S1). The effect of MAF on accuracy was even greater when the accuracy was measured as the percentage of correctly imputed variants (the measurement applied here) than measured by other statistics. In cattle, SNPs with a very low MAF had very poor imputation accuracy, which heavily influenced the overall imputation accuracy [47], especially when the reference population (founders) was small [16, 49], as is the case in this study, where only 29 common ancestors were sequenced as the reference population. Relationships between individuals is another factor that affects imputation accuracy. In general, imputation accuracy increases with lower relatedness within the reference population and larger relationships between reference and imputed individuals. As reported previously, a multi-breed reference population generated higher imputation accuracy for a given breed than using the same breed as a reference [23, 50]. In the present



study, in order to maximize relationships between reference and imputed animals, 29 common ancestors that contributed about 70% of the alleles present in the imputed animals were selected for sequencing. However, as we were restricted to one breed and in availability of tissue samples, the selected animals (29) in the reference population were related to each other, with genetic relationships ranging from 0.03 to 0.49 and averaging 0.06.

Genomic prediction

The average accuracies of genomic predictions for ADFI, FAT, ADG and LMD obtained in this study were 0.40, 0.65, 0.15 and 0.30, respectively. Limited literature is available on genomic prediction for feed efficiency and component traits in pigs (Table 3). Studies for Duroc pigs using 60K SNPs showed accuracies of genomic predictions for ADFI, FAT, ADG and LMD of about 0.15, 0.37–0.56, 0.24–0.58 and 0.30, respectively [10–12]. A study using imputed 60K SNPs in Yorkshire pigs [9] reported accuracies of 0.69–0.86 for FAT and of 0.66–0.88 for growth rate. Accuracies of genomic predictions obtained in this study were much higher for ADFI than accuracies obtained in these previous reports but much lower for ADG, while accuracies obtained for FAT and LMD were in the range of previous reports (Table 3). These differences can be explained by the many factors that influence the accuracy of genomic prediction, which will be discussed later.

The advantage of using genomic information for breeding value prediction over using pedigree information (BLUP method) was not uniform across traits. Compared

to pedigree-BLUP, using genomic data increased prediction accuracy and decreased prediction bias for ADFI and FAT but not for ADG and LMD. Similar results were reported in cattle [51] and sheep [23], where the use of genomic data for genomic prediction was not beneficial for all traits. Use of genomic information is generally expected to increase prediction accuracy, such as the reports in chicken [18, 52] and pigs [53], since genomic data can consider the Mendelian sampling terms better compared with pedigree information, and can produce more accurate genetic relationships among animals. However, this is not always true, as discussed above. The other two main factors that affect genomic prediction accuracy are the ability of markers to capture the total genetic variance of the traits (so-called “genomic heritability”) and the accuracy of the estimates of marker effects [54]. In most cases, heritability estimates obtained from dense markers were lower than estimates obtained from pedigree-based animal models (see Additional file 2: Table S1), which indicates that “missing heritability” exists, and this has been reported to be an issue in human genetics [55, 56]. Missing heritability mainly results from incomplete LD between causal variants and genotyped SNPs, which can be exacerbated by causal variants having lower MAF than the genotyped SNPs [55, 56]. Missing heritability can also be related to the genetic architecture of the traits, epistatic effects, genotype-by-environment interactions, and others [57]. For example, if the SNPs used are causal variants or are closely-linked to causal variants for the traits, they can capture a large proportion of the genetic variance and give high genomic prediction accuracies, such as for ADFI and FAT in this study, for which QTL with relatively large effects have been detected (data not shown). If the SNPs used do not capture all the genetic variation for the trait, prediction accuracy is limited, such as the low prediction accuracy found for ADG and LMD, for which the SNP panels only captured 53 to 83% of the genetic variance based on pedigree- and genotype-based estimates of heritability (see Additional file 2: Table S1). A similar trend was also reported in sheep [23], where no significant regions or markers were detected for the two traits for which prediction accuracy was not increased by using genomic data compared with using pedigree information.

Genetic architecture of traits and genomic prediction method

Genetic architecture and the statistical method used for genomic prediction are two interrelated factors that have a large influence on the accuracy of genomic prediction. Usually, higher accuracy can be achieved when the model assumptions more closely represent the underlying genetic architecture of the traits. We found that BayesB outperformed GBLUP in the accuracy of genomic

Table 3 Literature estimates of the accuracy of genomic predictions of feed efficiency component traits in pigs

Trait	Accuracy ^a	Breed and reference
Days to 250 lbs	0.66–0.84	Yorkshire [9]
ADG	0.50–0.58 ^b	Danish Duroc [12]
	0.40–0.43 ^b	Danish Duroc [10]
	0.24	Duroc [11]
Feed conversion ratio	0.39–0.45 ^b	Danish Duroc [12]
	0.11	Duroc [11]
FAT	0.69–0.86	Yorkshire [9]
	0.55–0.56 ^b	Danish Duroc [10]
	0.37	Duroc [11]
ADFI	0.15	Duroc [11]
Residual feed intake	0.09	
LMD	0.30	

^a Correlation of genomic predictions and corrected phenotype divided by square root of heritability, which was also used in our study; ^b converted from the reliability reported in the literatures

ADFI average daily feed intake, FAT ultrasound backfat depth, ADG average daily gain, LMD ultrasound loin muscle depth

prediction for ADFI and FAT. BayesB assumes that only a small proportion of SNPs have a large effect on the trait, which is in agreement with our GWAS results, where relatively large QTL were detected on *Sus scrofa* chromosomes (SSC) SSC1 and SSC18 for ADFI and FAT, using BayesB in the full dataset (data not shown). With this method, the effects of SNPs surrounding large QTL, such as those on SSC1 and SSC18 for ADFI and FAT, are easier to detect and more accurately estimated. This could be the main reason why BayesB gave higher accuracy for ADFI and FAT than GBLUP. The advantage of BayesB for genomic prediction for traits that are, at least in part, determined by QTL of large effect was also recognized by Meuwissen et al. [58] and demonstrated by other empirical studies [5, 22, 46, 59, 60]. For ADG and LMD, GBLUP performed better and increased the accuracy by 3 to 11% compared to BayesB. GBLUP assumes an infinitesimal model and, thereby, that all markers have the same contribution to the trait (e.g. no major QTL control the trait). Compared with FAT, few QTL were detected for ADG (data not shown), indicating that ADG may be determined by many loci with very small individual effects.

We also implemented the BayesRC method for the imputed whole-genome sequence data. Compared to BayesB and GBLUP, BayesRC gave higher accuracy for ADG and LMD, improving accuracy from 0.12 to 0.25 for ADG and from 0.32 to 0.35 for LMD. For ADFI and FAT, the accuracy from BayesRC was between those obtained with GBLUP and BayesB. The advantage of BayesRC compared with GBLUP and BayesB is that it can incorporate prior biological information by defining classes of variants that are likely enriched for causal mutations and by fitting a mixture distribution for the effects of variants in each class [15, 19, 61], which is more precise and sensitive to the genetic architecture of the traits. BayesRC resulted in the most accurate genomic predictions for ADG and LMD but also introduced greater bias for ADFI and LMD (Table 2). Both simulation and empirical studies have also shown that BayesRC can increase the power of detection of causal variants and improve the accuracy of genomic prediction compared to GBLUP [19, 46], in agreement with this study. BayesRC is also able to detect a larger proportion of variance when there is a large number of QTL with small individual effects [62], as was the case for ADG in this study. The posterior π values for class NSC that was obtained for ADG (0.413) was much smaller than the posterior π for FAT (0.641), which indicates that a larger proportion of variants were in class NSC for ADG and these variants may have small individual effects on the trait. Therefore, for traits with such a genetic architecture (e.g. ADG), the advantage of BayesRC for genomic prediction is greater. For traits with known large QTL, such as ADFI and FAT, accuracies

obtained with BayesB and BayesRC were similar or higher than those obtained using GBLUP. However, the advantage of BayesRC for sequence data depends on the completeness and accuracy of the prior biological information. With a better understanding of the functional annotation of genes and variants in the future [63], the benefit of using whole-sequence data for genomic prediction is anticipated to be further improved.

Impact of marker density on genomic prediction

Increasing marker density has the potential to improve the accuracy of genomic prediction and the use of whole-genome sequence data is expected to give the best accuracy, as the causative mutations are expected to be included in the genotype data [16, 64]. However, this was not found to be always the case in our study. An increase in marker density did not improve prediction accuracy for some traits, such as FAT, for which 80K SNPs gave the highest accuracy, regardless of the statistical method used. This result was also reported for backfat thickness in pigs by Pérez-Enciso et al. [62]. Similar results were also obtained in cattle, for which imputed 777K SNPs resulted in no or very little increase in the accuracy of genomic prediction for some traits [17, 65–67] compared with using 50K SNPs. SNPs on commercially available low-density SNP chips (e.g. pig 60K and 80K, bovine 50K) were selected to have a high MAF and can, thus, capture a relatively large amount of the variance for traits that are determined by a relatively small number of QTL (e.g. backfat in pigs). Increasing marker density has little effect on capturing the remaining proportion of genetic variance for such traits. Furthermore, with GBLUP and BayesB, the QTL effects and opportunities for their detection become smaller with increasing density [16], thus resulting in less accurate genomic predictions. Therefore, we suggest that the 80K SNP panel is sufficient for within-breed genomic prediction for FAT and yields acceptable accuracy (0.68). In contrast, when some of the QTL mutations or the linked SNPs are not in the SNP panel, a higher density may include more SNPs that are in high LD with the QTL for the traits, resulting in an increase in the genetic variance captured and more accurate genomic predictions. This appeared to be the case for ADG in this study. As discussed above, when considering the best method for each trait, the imputed 650K SNPs increased the accuracy of genomic prediction by 3.4% for ADFI (BayesB) and 15.2% for ADG (GBLUP).

Results from using sequence data to improve the accuracy of genomic prediction have been inconsistent. Simulation studies suggested that including whole-sequence data could improve the accuracy of genomic prediction by as much as 40%, depending on the trait, statistical method, and MAF of the causal mutations affecting the

trait [14, 16, 25, 68]. However, empirical studies in cattle and chickens have reported either no or a very small increase in accuracy when using imputed whole-genome sequence data compared to using the available low- or high-density SNP chips [17, 18, 21, 64, 69]. In pigs, simulation based on whole-genome sequence showed an increase in accuracy of ~3.8% over 60K and ~2.8% over 650K SNPs [62]. We found that using SEQ data and BayesRC gave the highest prediction accuracies for LMD and ADG. For LMD, using SEQ data increased the accuracy from 8 (GBLUP) to ~20% (BayesB). Using SEQ data, however, resulted in a decrease in accuracy for FAT and ADFI compared to using 80K and 650K SNP chips. Druet et al. [14] explained that the advantage of using imputed sequence data for genomic prediction is affected by the accuracy of imputation and, more importantly, by the allele frequency distribution of the QTL. When the MAF of QTL is very low, genomic predictions from imputed sequence data can result in up to 30% improvement in accuracy. However, for rare variants, imputation accuracy is usually poorer than for variants with a high MAF [14, 47, 50]. Thus, a large reference population must be sequenced to improve the results. In this study, the small number of sequenced animals may have influenced the accuracy of imputed sequence variants and, thus, may have limited the potential of whole-genome sequence data to improve prediction accuracy.

Pre-selection and prior biology of sequence variants

A significant challenge for genomic prediction using whole-genome sequence data is the computational requirement due to the large number of markers. Pre-selecting the most important markers and/or filtering out the uninformative ones can address this problem. The split-and-merge approach, which splits one large computational task into many smaller ones, was first proposed by Calus et al. [17] to pre-select the most important markers from whole-genome sequence data. Some studies [48, 70–72] showed that using preselected markers from sequence data through GWAS and adding them to the 50K SNP panel can increase the accuracy of genomic prediction by up to 5 percentage points. However, Veerkamp et al. [73] and Calus et al. [17] found no improvement in accuracy using a similar approach. In our study, first a modified split-and-merge approach was used by integrating the 80K SNPs into the split association analyses for each chromosome, in order to better account for polygenic effects and to improve the accuracy of estimates of marker effects. Second, all SEQ markers, not only the pre-selected ones, were considered simultaneously in the genomic prediction model (BayesRC), which may avoid the loss of the marginal genetic variance contributed by the non-selected sequence variants

and the possible bias derived from strict pre-selection, as discussed by Calus et al. [17] and Veerkamp et al. [73]. Pruning SNPs that are in complete and high LD with other SNPs is also an efficient way to reduce the number of uninformative markers, which was shown to be important for the application of Bayesian models that explicitly estimate a SNP variance component using sequence data, since performance of these models may be poorer without pruning [17].

Other factors affecting genomic prediction accuracy

According to Goddard [54], prediction accuracy depends on both the proportion of genetic variance that can be captured by markers (so-called “genomic heritability”) [74] and the accuracy of estimates of marker effects. However, there are important trade-offs between these two factors. Usually the estimate of genomic heritability increases when more markers are used, especially when the added markers are in high LD with the QTL [55, 56]. A similar trend was found in our study, where the use of SEQ variants increased the genomic heritability for all traits compared with using the commonly available SNP panels (see Additional file 2: Table S1). However, the accuracy of estimates of marker effects was impaired as the number of effects to be estimated increased, which is mainly due to the relatively small size of the training population ($n \ll p$, where n is the number of animals in training and p is the number of markers). Additional issues can also arise as a result of the small sample size, including (1) causal mutations (usually with a small MAF) may be missed, are more easily filtered out during quality control, or are more poorly imputed to the whole population [47, 50], which decreases the value of such causal variants in the prediction process, thus negatively influencing prediction accuracy; and (2) a small amount of phenotypic data is not sufficient to detect causative mutations and to distinguish their effects from random noise. Therefore, as Meuwissen et al. [75] highlighted, a large training dataset is needed to take full advantage of high-density markers (especially for whole-genome sequence data) for accurate genomic prediction.

Relatedness between individuals is also very important for both genotype imputation and genomic prediction. Ideally, having less related animals in a large reference population is helpful to break down high levels of LD, thus making it easier to identify the causal mutations and to capture all the genetic variance. For example, use of multiple breeds in the training population to reduce their average relatedness gave more accurate genomic predictions than using the same single breed for training, especially in simulated datasets [15, 25, 64, 76, 77]. In contrast, greater relationships between training and prediction animals can improve the prediction accuracy.

Macleod et al. [19] demonstrated that the accuracy of genomic prediction increased for all traits with increasing relatedness between training and prediction sets. Other studies also reported that a closer relationship between training and prediction increases the accuracy of genomic prediction [11, 52, 78, 79].

A small effective population size, which contributes to high LD [15], is another factor that influences the prediction accuracy. The effective size of pig breeding populations has been estimated to be relatively small (55 to 113 [80, 81]), so a small number of SNPs (80K) may capture most genetic variance, especially for ADFI and FAT, which may be determined by a small number of QTL with relatively large effects (at least in this population). Therefore, the potential increase in the accuracy of genomic prediction from using whole-genome sequence data is expected to be limited. A similar situation was also observed in dairy cattle [19, 24] and sheep [23] populations with small effective sizes.

Conclusions

In conclusion, although the reference population used was small, the genotype imputation accuracies were as high as 92.1% from 80K to 650K, and 85.6% from 650K to whole sequence. Increasing marker density, however, had no or little advantage for genomic prediction for FAT and ADFI, such that the available 80K SNP panel is sufficient for these traits. BayesB resulted in higher prediction accuracy than the other methods tested for these two traits. For LMD and ADG, GBLUP gave higher genomic prediction accuracies than BayesB, and BayesRC in SEQ data gave the best prediction accuracies. However, pedigree-based BLUP outperformed all genomic methods and produced the highest prediction accuracies for ADG and LMD, likely because the SNPs captured less genetic variance for these traits than pedigree data. In the future, with decreasing costs for whole-genome sequencing, a better understanding of the functional annotation of the genome and variants [63], and larger reference population sizes, BayesRC is anticipated to be a superior method for genomic prediction and application in genetic improvement.

Additional files

Additional file 1: Fig. S1. Histograms of MAF distribution for the variants from final 80K, 650K and SEQ data.

Additional file 2: Table S1. Variance component and heritability estimates using different information. The data provided presented the genetic variance, total phenotypic variance and estimated heritability for the traits using different information and methods.

Authors' contributions

GP conceived and designed the experiments; PS and KK performed the whole-sequence data management and functional annotation prediction; CZ performed the statistical analysis; GP, RK, ZW, NB and JD provided valuable advice to refine the statistical analyses; RK and NB contributed materials and application advice; CZ, GP and JD were major contributors in writing the manuscript. All authors read and approved the final manuscript.

Author details

¹ Department of Agricultural, Food and Nutritional Science, University of Alberta, Edmonton, AB T6G 2R3, Canada. ² Genesus Inc., Oakville, MB R0H 0Y0, Canada. ³ Department of Animal Science, Iowa State University, Ames, IA 50011, USA.

Acknowledgements

Thanks are given to all the staff and technicians who collected the samples and data. The authors are grateful to Dr. Iona MacLeod's team (University of Melbourne) who provided the BayesRC program, Dr. Feng Zhang (University of Alberta and Jiangxi Agricultural University) who provided help on the data analysis and Dr. Mario Calus who provided useful suggestions for analysis including sequence data.

Ethics approval and consent to participate

The proposed work was approved by the University of Alberta Animal Care and Use Committee. No other specific permissions were required.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Consent for publication

Not applicable.

Funding

This work was financially supported by grants from Genome Alberta, Alberta Livestock and Meat Agency, MITACS Elevate postdoctoral fellowship and the industry partner Genesus Inc.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 1 July 2017 Accepted: 27 March 2018

Published online: 06 April 2018

References

1. Patience JF, Rossoni-Serao MC, Gutierrez NA. A review of feed efficiency in swine: biology and application. *J Anim Sci Biotechnol.* 2015;6:33.
2. Rothschild MF, Ruvinsky A. The genetics of the pig. Wallingford: CAB International; 2011.
3. Daetwyler HD, Villanueva B, Bijma P, Woolliams JA. Inbreeding in genome-wide selection. *J Anim Breed Genet.* 2007;124:369–76.
4. Sonesson AK, Meuwissen TH. Testing strategies for genomic selection in aquaculture breeding programs. *Genet Sel Evol.* 2009;41:37.
5. VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, et al. Invited review: reliability of genomic predictions for North American Holstein bulls. *J Dairy Sci.* 2009;92:16–24.
6. Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME. Invited review: genomic selection in dairy cattle: progress and challenges. *J Dairy Sci.* 2009;92:433–43.
7. Hayes BJ, Lewin HA, Goddard ME. The future of livestock breeding: genomic selection for efficiency, reduced emissions intensity, and adaptation. *Trends Genet.* 2013;29:206–14.

8. Schaeffer LR. Strategy for applying genome-wide selection in dairy cattle. *J Anim Breed Genet.* 2006;123:218–23.
9. Badke YM, Bates RO, Ernst CW, Fix J, Steibel JP. Accuracy of estimation of genomic breeding values in pigs using low-density genotypes and imputation. G3 (Bethesda). 2014;4:623–31.
10. Guo X, Christensen OF, Ostersen T, Wang Y, Lund MS, Su G. Genomic prediction using models with dominance and imprinting effects for backfat thickness and average daily gain in Danish Duroc pigs. *Genet Sel Evol.* 2016;48:67.
11. Jiao S, Maltecca C, Gray KA, Cassady JP. Feed intake, average daily gain, feed efficiency, and real-time ultrasound traits in Duroc pigs: I. Genetic parameter estimation and accuracy of genomic prediction. *J Anim Sci.* 2014;92:2377–86.
12. Ostersen T, Christensen OF, Henryon M, Nielsen B, Su G, Madsen P. Deregressed EBV as the response variable yield more reliable genomic predictions than traditional EBV in pure-bred pigs. *Genet Sel Evol.* 2011;43:38.
13. 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.
14. 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 (Edinb).* 2014;112:39–47.
15. 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.
16. Meuwissen T, Goddard M. Accurate prediction of genetic values for complex traits by whole-genome resequencing. *Genetics.* 2010;185:623–31.
17. Calus MP, 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.
18. Heidaritabar M, Calus MP, Megens HJ, Vereijken A, Groenen MA, Bastiaansen JW. Accuracy of genomic prediction using imputed whole-genome sequence data in white layers. *J Anim Breed Genet.* 2016;133:167–79.
19. 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.
20. Ober U, Ayroles JF, Stone EA, Richards S, Zhu D, Gibbs RA, et al. Using whole-genome sequence data to predict quantitative trait phenotypes in *Drosophila melanogaster*. *PLoS Genet.* 2012;8:e1002685.
21. van Binsbergen R, Calus MP, Bink MC, 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.
22. Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA. The impact of genetic architecture on genome-wide evaluation methods. *Genetics.* 2010;185:1021–31.
23. Daetwyler HD, Kemper KH, Van der Werf JH, Hayes BJ. Components of the accuracy of genomic prediction in a multi-breed sheep population. *J Anim Sci.* 2012;90:3375–84.
24. 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.
25. Iheshiolor OO, Woolliams JA, Yu X, Wellmann R, Meuwissen TH. Within- and across-breed genomic prediction using whole-genome sequence and single nucleotide polymorphism panels. *Genet Sel Evol.* 2016;48:15.
26. VanRaden PM. Efficient methods to compute genomic predictions. *J Dairy Sci.* 2008;91:4414–23.
27. Fernando RL, Garrick D. Bayesian methods applied to GWAS. In: Gondro C, van der Werf J, Hayes B, editors. *Genome-wide association studies and genomic prediction. Methods in molecular biology (Methods and Protocols)*, vol. 1019. Totowa: Humana Press; 2013. p. 237–74.
28. Garrick DJ, Fernando RL. Implementing a QTL detection study (GWAS) using genomic prediction methodology. In: Gondro C, van der Werf J, Hayes B (editors). *Genome-wide association studies and genomic prediction. Methods in molecular biology (Methods and Protocols)*, vol. 1019. Totowa: Humana Press; 2013. p. 275–98.
29. MacNeil MD, Kemp RA. Genetic parameter estimation and evaluation of Duroc boars for feed efficiency and component traits. *Can J Anim Sci.* 2015;95:155–9.
30. Casey DS, Stern HS, Dekkers JC. Identification of errors and factors associated with errors in data from electronic swine feeders. *J Anim Sci.* 2005;83:969–82.
31. Boichard D. PEDIG: a Fortran package for pedigree analysis suited for large populations. In: *Proceedings of the 7th world congress on genetics applied to livestock production: 19–23 August 2002; Montpellier; 2002.*
32. 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.
33. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protoc Bioinform.* 2013;43:11.10.1–33.
34. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics.* 2010;26:589–95.
35. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20:1297–303.
36. Sargolzaei M, Chesnais JP, Schenkel FS. A new approach for efficient genotype imputation using information from relatives. *BMC Genomics.* 2014;15:478.
37. Gilmour AR, Gogel B, Cullis B, Thompson R, Butler D. *ASReml user guide release 3.0.* Hemel Hempstead: VSN International Ltd; 2009.
38. Grant JR, Arantes AS, Liao X, Stothard P. In-depth annotation of SNPs arising from resequencing projects using NGS-SNP. *Bioinformatics.* 2011;27:2300–1.
39. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007;81:559–75.
40. Cleveland MA, Hickey JM. Practical implementation of cost-effective genomic selection in commercial pig breeding using imputation. *J Anim Sci.* 2013;91:3583–92.
41. Gualdron Duarte JL, Bates RO, Ernst CW, Raney NE, Cantet RJ, Steibel JP. Genotype imputation accuracy in a F2 pig population using high density and low density SNP panels. *BMC Genet.* 2013;14:38.
42. Huang Y, Hickey JM, Cleveland MA, Maltecca C. Assessment of alternative genotyping strategies to maximize imputation accuracy at minimal cost. *Genet Sel Evol.* 2012;44:25.
43. Berry DP, Kearney JF. Imputation of genotypes from low- to high-density genotyping platforms and implications for genomic selection. *Animal.* 2011;5:1162–9.
44. Pausch H, Aigner B, Emmerling R, Edel C, Götz K-U, Fries R. Imputation of high-density genotypes in the Fleckvieh cattle population. *Genet Sel Evol.* 2013;45:3.
45. Druet T, Schrooten C, de Roos A. Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. *J Dairy Sci.* 2010;93:5443–54.
46. Bolormaa S, Pryce JE, Kemper K, Savin K, Hayes BJ, Barendse W, et al. Accuracy of prediction of genomic breeding values for residual feed intake and carcass and meat quality traits in, and composite beef cattle. *J Anim Sci.* 2013;91:3088–104.
47. Van Binsbergen R, Bink MC, Calus MP, Van Eeuwijk FA, Hayes BJ, Hulsege I, et al. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. *Genet Sel Evol.* 2014;46:41.
48. O'Connell JR, Tooker ME, Bickhart DM, VanRaden PM. Selection of sequence variants to improve genomic predictions. *Interbull Bull.* 2016;50:58–66.
49. Heidaritabar M, Calus MP, Vereijken A, Groenen MA, Bastiaansen JW. Accuracy of imputation using the most common sires as reference population in layer chickens. *BMC Genet.* 2015;16:101.
50. Bouwman AC, Veerkamp RF. Consequences of splitting whole-genome sequencing effort over multiple breeds on imputation accuracy. *BMC Genet.* 2014;15:105.
51. Silva RM, Fragomeni BO, Lourenco DA, Magalhães AF, Irano N, Carvalheiro R, et al. Accuracies of genomic prediction of feed efficiency traits using different prediction and validation methods in an experimental Nelore cattle population. *J Anim Sci.* 2016;94:3613–23.

52. Weng Z, Wolc A, Shen X, Fernando RL, Dekkers JC, Arango J, et al. Effects of number of training generations on genomic prediction for various traits in a layer chicken population. *Genet Sel Evol*. 2016;48:22.
53. Christensen OF, Madsen P, Nielsen B, Ostensen T, Su G. Single-step methods for genomic evaluation in pigs. *Animal*. 2012;6:1565–71.
54. Goddard M. Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica*. 2009;136:245–57.
55. Yang J, Bakshi A, Zhu Z, Hemani G, Vinkhuyzen AA, Lee SH, et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat Genet*. 2015;47:1114–20.
56. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet*. 2010;42:565–9.
57. Makowsky R, Pajewski NM, Klimentidis YC, Vazquez AI, Duarte CW, Allison DB, et al. Beyond missing heritability: prediction of complex traits. *PLoS Genet*. 2011;7:e1002051.
58. Meuwissen T, Hayes B, Goddard M. Accelerating improvement of livestock with genomic selection. *Annu Rev Anim Biosci*. 2013;1:221–37.
59. Cole JB, VanRaden PM, O'Connell JR, Van Tassell CP, Sonstegard TS, Schnabel RD, et al. Distribution and location of genetic effects for dairy traits. *J Dairy Sci*. 2009;92:2931–46.
60. Lu D, Akanno EC, Crowley J, Schenkel F, Li H, De Pauw M, et al. Accuracy of genomic predictions for feed efficiency traits of beef cattle using 50K and imputed HD genotypes. *J Anim Sci*. 2016;94:1342–53.
61. Kemper KE, Reich CM, Bowman PJ, Vander Jagt CJ, Chamberlain AJ, Mason BA, et al. Improved precision of QTL mapping using a nonlinear Bayesian method in a multi-breed population leads to greater accuracy of across-breed genomic predictions. *Genet Sel Evol*. 2015;47:29.
62. Pérez-Enciso M, Forneris N, de los Campos G, Legarra A. Evaluating sequence-based genomic prediction with an efficient new simulator. *Genetics*. 2017;205:939–53.
63. Andersson L, Archibald AL, Bottema CD, Brauning R, Burgess SC, Burt DW, et al. Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biol*. 2015;16:57.
64. Hayes BJ, MacLeod IM, Daetwyler HD, Bowman JP, Chamberlain AJ, Vander Jagt CJ, et al. Genomic prediction from whole genome sequence in livestock: the 1000 bull genomes project. In: Proceedings of the 10th world congress on genetics applied to livestock production, Vancouver, 19–23 August 2014.
65. Gunia M, Saintilan R, Venot E, Hozé C, Fouilloux MN, Phocas F. Genomic prediction in French Charolais beef cattle using high-density single nucleotide polymorphism markers. *J Anim Sci*. 2014;92:3258–69.
66. Solberg TR, Heringstad B, Svendsen M, Grove H, Meuwissen TH. Genomic predictions for production and functional traits in Norwegian red from BLUP analyses of imputed 54K and 777K SNP data. *Interbull Bulletin*. 2011;44:240–3.
67. Su G, Brondum RF, Ma P, Gulbrandsen B, Aamand GP, Lund MS. Comparison of genomic predictions using medium-density (approximately 54,000) and high-density (approximately 777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy cattle populations. *J Dairy Sci*. 2012;95:4657–65.
68. Perez-Enciso M, Rincon JC, Legarra A. Sequence- vs. chip-assisted genomic selection: accurate biological information is advised. *Genet Sel Evol*. 2015;47:43.
69. 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.
70. Brondum RF, Su G, Janss L, Sahana G, Gulbrandsen 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.
71. Ortega MS, Denicol AC, Cole JB, Null DJ, Hansen PJ. Use of single nucleotide polymorphisms in candidate genes associated with daughter pregnancy rate for prediction of genetic merit for reproduction in Holstein cows. *Anim Genet*. 2016;47:288–97.
72. VanRaden PM, Null DJ, Sargolzaei M, Wiggans GR, Tooker ME, Cole JB, et al. Genomic imputation and evaluation using high-density Holstein genotypes. *J Dairy Sci*. 2013;96:668–78.
73. Veerkamp RF, Bouwman AC, Schrooten C, Calus MP. Genomic prediction using preselected DNA variants from a GWAS with whole-genome sequence data in Holstein-Friesian cattle. *Genet Sel Evol*. 2016;48:95.
74. de Los Campos G, Sorensen D, Gianola D. Genomic heritability: what is it? *PLoS Genet*. 2015;11:e1005048.
75. Meuwissen TH. Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. *Genet Sel Evol*. 2009;41:35.
76. Lund MS, van den Berg I, Ma P, Brondum RF, Su G. Review: how to improve genomic predictions in small dairy cattle populations. *Animal*. 2016;10:1042–9.
77. van den Berg S, Calus MP, Meuwissen TH, Wientjes YC. Across population genomic prediction scenarios in which Bayesian variable selection outperforms GBLUP. *BMC Genet*. 2015;16:146.
78. Habier D, Fernando RL, Dekkers JC. The impact of genetic relationship information on genome-assisted breeding values. *Genetics*. 2007;177:2389–97.
79. Wolc A, Arango J, Settar J, Fulton P, O'Sullivan N, Preisinger R, et al. Application of a weighted genomic relationship matrix to breeding value prediction for egg production in laying hens. In: Plant and animal genome XXI conference, 12–16 January 2013; San Diego; 2013.
80. Welsh CS, Stewart TS, Schwab C, Blackburn HD. Pedigree analysis of 5 swine breeds in the United States and the implications for genetic conservation. *J Anim Sci*. 2010;88:1610–8.
81. Zhang C, Plastow G. Genomic diversity in pig (*Sus scrofa*) and its comparison with human and other livestock. *Curr Genomics*. 2011;12:138–46.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

