Annotation-Based Whole Genomic Prediction and Selection

Kadarmideen, Haja; Do, Duy Ngoc; Janss, Luc; Jensen, Just

Publication date:
2015

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

Citation for published version (APA):
Kadarmideen, H., Do, D. N., Janss, L., & Jensen, J. (2015). Annotation-Based Whole Genomic Prediction and Selection: An Application to Feed Efficiency and Growth Traits in Pigs. Abstract from The Plant and Animal Genome Conference XXIII, San Diego, United States.

Published May 15, 2015

SNP annotation-based whole genomic prediction and selection: An application to feed efficiency and its component traits in pigs[1]

D. N. Do,* L. L. G. Janss,† J. Jensen,† and H. N. Kadarmideen*[2]

*Department of Veterinary Clinical and Animal Sciences, Faculty of Health
and Medical Sciences, University of Copenhagen, 1870 Frederiksberg C, Denmark;
†Aarhus University, Department of Molecular Biology and Genetics, 8830 Tjele, Denmark

ABSTRACT: The study investigated genetic architecture and predictive ability using genomic annotation of residual feed intake (RFI) and its component traits (daily feed intake [DFI], ADG, and back fat [BF]). A total of 1,272 Duroc pigs had both genotypic and phenotypic records, and the records were split into a training (968 pigs) and a validation dataset (304 pigs) by assigning records as before and after January 1, 2012, respectively. SNP were annotated by 14 different classes using Ensembl variant effect prediction. Predictive accuracy and prediction bias were calculated using Bayesian Power LASSO, Bayesian A, B, and Cπ, and genomic BLUP (GBLUP) methods. Predictive accuracy ranged from 0.508 to 0.531, 0.506 to 0.532, 0.276 to 0.357, and 0.308 to 0.362 for DFI, RFI, ADG, and BF, respectively. BayesCπ100.1 increased accuracy slightly compared to the GBLUP model and other methods. The contribution per SNP to total genomic variance was similar among annotated classes across different traits. Predictive performance of SNP classes did not significantly differ from randomized SNP groups. Genomic prediction has accuracy comparable to observed phenotype, and use of genomic prediction can be cost effective by replacing feed intake measurement. Genomic annotation had less impact on predictive accuracy traits considered here but may be different for other traits. It is the first study to provide useful insights into biological classes of SNP driving the whole genomic prediction for complex traits in pigs.

Key words: Bayesian models, feed intake, genome annotation, genomic prediction, pigs

J. Anim. Sci. 2015.93:2056–2063
doi:10.2527/jas2014-8640

INTRODUCTION

Genomic selection using dense molecular markers for predicting GEBV is widely used in both animal and plant species. In pigs, genomic selection is currently implemented and is especially attractive for traits that are expensive to measure or cannot be measured early in life. Feed efficiency is a very important trait in breeding due to its moderate to high heritability, but it is costly to record (Kadarmideen et al., 2004; Do et al., 2013). Residual feed intake (**RFI**) is a measure of feed efficiency that is independent of production traits, and selection for lower RFI may help to improve feed efficiency. Various biometrical methods (such as genomic BLUP [**GBLUP**] [VanRaden, 2008], Bayesian least absolute shrinkage and selection operator Bayesian LASSO [**BL**] [Legarra et al.], Bayesian Power LASSO [**BPL**] [Gao et al., 2013], or Bayesian Alphabet [Habier et al., 2011; Meuwissen et al., 2001]) have been proposed for genomic prediction. However, genomic prediction is also performed as black box prediction due to lack of information about the genetic architecture of traits. Genetic architectures of feed efficiency have been investigated in our recent studies in Duroc and Yorkshire pigs (Do et al., 2014a,b). Moreover, it is possible to investigate the role of different genomic regions in prediction of GEBV using

[1] We thank the Department of Breeding and Genetics of the Danish Pig Research Centre for providing all data. D. N. Do is funded by the Faculty of Health and Medical Sciences, University of Copenhagen. H. N. Kadarmideen thanks the EU-FP7 Marie Curie Actions–Career Integration Grant (CIG-293511) presented to him to partially fund this study. We would like to thank Anders Bjerring Strathe for useful discussions on the modeling.

[2] Corresponding author: hajak@sund.ku.dk

Received October 21, 2014.
Accepted February 10, 2015.

the SNP annotation information. For instance, Morota et al. (2014) indicated that genomic regions could affect the predictive accuracy for quantitative traits in chickens. In addition, the identification of which genomic regions enriched for traits will improve a priori set up for association analysis, and will, consequently, increase the power of detection of variants (Koufariotis et al., 2014). This study aimed to investigate the performance of genomic prediction methods for RFI and its component traits (daily feed intake [**DFI**], ADG, and back fat [**BF**]) and to examine predictive ability of different annotated genomic classes for these traits.

MATERIALS AND METHODS

Data, Genotyping, and Quality Control

Phenotypic records for DFI and BF were made during a period from 2008 to 2012 for Danish Duroc pigs. A detailed description of data collection and a definition of feed efficiency phenotypes and their calculations are given in Do et al. (2013). In brief, the pigs were moved from the nursery barns to the test barns (each barn contained several pens) when they were approximately 30 kg. Every pen contained around 11 pigs and had an automatic dry feeding station ACEMA64 (ACEMO, Pontivy, France). Pigs were fed ad libitum with the same feed composition. Feed was given through a single feeder, and time, duration, and feed consumption was recorded for each pig for each individual visit. Average daily feed intake was computed by the total amount of recorded feed intake divided into the number of corresponding days at the feeder. Average daily gain was calculated as linear regressions of body weight from 30 to 100 kg on test days. Residual feed intake was the residual in the regression of DFI on ADG and BF with initial BW as an extra covariate in the model (Do et al., 2013). Pigs were genotyped using the PorcineSNP60 BeadChip (Illumina, San Diego, CA). The criteria for screening the genomic data were a call rate per animal of 0.95, call rate per SNP marker of 0.95, Hardy Weinberg equilibrium test with $P < 0.0001$, and minor allele frequency > 0.05. Unmapped SNP were removed from the study. After quality control, 30,234 SNP and 1,272 pigs remained for genomic prediction.

Genomic Prediction Methods

Genomic BLUP. For reference purposes, a GBLUP model was used where: $\mathbf{y} = \mathbf{1}\mathrm{m} + \mathbf{Xb} + \mathbf{Zp} + \mathbf{g} + \mathbf{e}$, in which **y** is the vector of observed phenotypic values of the animals, **1** is a vector of ones, m is the overall mean, **b** is the vector of fixed effect (herd-year-barn), **X** is a design matrix relating observations to the corresponding fixed effect, p is the vector of random effect (pen) and **Z** is a design matrix relating observations to the corresponding random pen effect, **e** is the vector of random error, and **g** is a vector of breeding value with var(**g**) = $\mathbf{G}\boldsymbol{\sigma}_g^2$, in which σ_g^2 is genetic variance and **G** is the genomic relationship matrix. The GBLUP method was similar to Ostersen et al. (2011) and was fitted using the DMU package (Jensen and Madsen, 1994).

The Bayesian Power LASSO Models. The BPL is a sparse shrinkage model that uses an exponential power distribution for marker effects. Details of the model appeared in Gao et al. (2013). In brief, BPL is an extension of BL by adding a power parameter that can modify the sparsity of the marker effects. The model is ($\mathbf{y} = \mathbf{1}\mathrm{m} + \mathbf{Xb} + \mathbf{M}\boldsymbol{\beta} + \mathrm{e}$) where SNP effect ($\mathbf{b}_j$) follow an exponential power distribution $p(\beta) = \Pi_{j=1}^{m} \frac{\lambda}{2} e^{-\lambda|\beta|^q}$, where λ is a rate parameter, *m* is the number of markers, and *q* is the power parameter controlling the sparsity. The rate parameter was estimated from the data using a uniform prior. The power parameter was set to 0.3, 0.5, 0.7, 0.9, or 1.0 (q = 1 corresponds to the ordinary BL). These models were denoted as BPL0.3, BPL0.5, BPL0.7, BPL0.9, and BL. It is important to note that **b** is the vector of environment effects (pen and herd-year-barn effects); therefore, the BL and BPL models are equivalent to above GBLUP model.

BayesA, B, and Cπ Models Three different Bayesian methods including BayesA, BayesB, and BayesCπ were used to estimate GEBV using raw phenotypes as response variables and fitted with the same environmental effects as the above BL model. These Bayesian methods have different assumptions for the prior distribution of SNP effects. BayesA assumes that all SNP have an effect, but each has a different variance (Meuwissen et al., 2001). BayesB and BayesCπ assume that each SNP has either an effect of zero or nonzero with probabilities π and 1 − π (Habier et al., 2011) and a group of nonzero effects has a different variance or a common variance, respectively. Moreover, the BayesB treated π as a known parameter (set π = 0.95 for BayesB in this study), while BayesCπ treated it as an unknown parameter with a uniform (0, 1) prior distribution (Bayes Cp1.1). We also used a slightly informative prior distribution ~*Beta*(10,1; Bayes Cp10.1) and *Beta*(100,1; Bayes Cp100.1) to predict breeding values. All the Bayesian analyses were performed using the BayZ package (http://www.bayz.biz/). Each of the Bayesian analyses was run as a single chain with a length of 50,000 to 200,000 samples, and the first 5,000 to 20,000 cycles were regarded as the burn-in period (depending on the convergence of these models). Convergences (Markov Chain error and effective sample size) were checked using the R Coda package (Plummer et al., 2006), and the length of sample and burn-in was optimized when the effective number of samples >300 for model effects and the Markov Chain error was smaller than 1%. A Bayes factor was calculated for every SNP using the prior probability (π and $1 - \pi$) and the posterior probability (p) as Bayes factor = $\dfrac{p/(1-p)}{\pi/(1-\pi)}$.

SNP having Bayes factors with values above 10 and 3 were considered genome wide significantly and suggestively associated with trait of interest, respectively (Kass and Raftery, 1995).

Evaluation Criterion

To investigate the accuracy of different genomic prediction methods, we split the records into a training dataset (968 pigs) and a validation dataset (304 pigs) before and after January 1, 2012, respectively. Moreover, we also corrected phenotypes for a fixed effect of section and a random pen effect to avoid using overlapping information between the reference and validation datasets. The adjusted phenotypes (y_c) were computed based on the full data, and the adjusted phenotypes were the sum of EBV and the estimated residual errors ($y_c = \hat{g} + \hat{e}$) (Ostersen et al., 2011). Estimated breeding values were estimated based on a pedigree tracked back 30 generations (approximately 8,000 pigs) as described in (Do et al., 2013). The predictive accuracy was computed as the correlation between y_c and GEBV divided by the square root of heritability. The linear regression of y_c on GEBV was used to assess bias inflation of prediction (how far the regression slope differs from 1). The Hotelling-Williams *t* test (Dunn and Clark, 1971) was applied to each trait with a significance level of 5% (Ostersen et al., 2011) to test the equality of prediction performance of these Bayesian methods with that of GBLUP.

Partitioning Genomic Variance Based on Genome Annotation

SNP were classified based on Ensembl variant effect predictor annotated pig 60K SNP chip data (ftp://ftp.ensembl.org/pub/release-75/variation/VEP/arrays/), and a detailed summary of the annotated pig 60K SNP chip is in Table S1. The SNP classes (annotated by at least 20 SNP after quality control) were used for further analysis. As a result, 14 classes were used including intergenic (variants that occur in-between genes), gene (variants found within genes), upstream (variants found 5 kb upstream of a transcription start site), downstream (variants found 5 kb downstream of a gene), gene ± 5 kb, intron, exon, intron_non-coding (**NC**) transcript, synonymous variants, missense variants (nonsynonymous variants), both 5¢ and 3¢ untranslated regions (**UTRs**), splice region, and intron_nonsense-mediated decay transcript (**NMD**) variant. The details of how variants were classified are given in sequence ontology terms at http://www.sequenceontology.org/index.html. The total genomic variance of each group was computed based on the *gbayz* function using the BayZ package. Briefly, the *gbayz* function uses allelic effects from every round of the Markov chain Monte Carlo chain and combines them to compute GEBV; the genomic variance is then computed as the variance of GEBV. Then the SNP group-specific total genomic variance is computed based on selected SNP for a specific group. The *gbayz* function also takes into account allele frequencies as well as linkage disequilibrium (**LD**) between markers to compute the total genomic variance. To investigate if the predictive ability of SNP in each group significantly differs from randomized groups, we compared this value with empirical group obtained from randomized samples with the same number of SNP in annotated groups. First, randomized SNP groups were generated by randomly sampling with replacement 1,000 times, and then the breeding values of animals and the accuracy of these groups were calculated based on posterior estimates of SNP effects for each of the 1,000 samples. The predictive ability of an annotated SNP class was significantly improved from a randomized one if accuracy of its prediction was greater than at least 950 of 1,000 values (95% accuracy threshold) obtained from 1,000 random samples.

RESULTS

Predictive Performance Using Different Methods and SNP Patterns

The accuracy of genomic prediction and regression coefficient of y_c on GEBV are shown in Table 1. Accuracy ranged from 0.508 to 0.531, 0.506 to 0.532, 0.276 to 0.357, and 0.308 to 0.362 for DFI, RFI, ADG, and BF, respectively. All methods showed bias because the regression coefficient differed from 1. There were no significant differences between GBLUP and other methods for DFI and ADG. The BPL0.3, BPL0.5, Bayes A, and Bayes Cπ100.1 methods differed significantly from GBLUP for RFI, and the Bayes A and Bayes Cπ100.1 methods were also significantly higher than the GBLUP method ($p < 0.05$). Overall, Bayes Cπ100.1 was among the highest accuracy methods for all traits.

Prediction accuracy also varied according to the size of the SNP panels. The changes in accuracy as a function of the number of SNP (results based on the average accuracy obtained by using 5 replicates of randomly sampled SNP) are shown in Fig. 1. However, the accuracy remained stable when the number of SNP reached was 2,000 for ADG and BF and 5,000 SNP for RFI and DFI. Approximately 1,000 SNP were able to predict more than 80 to 90% of the accuracy estimated by all SNP for ADG.

Table 1. Predictive accuracy (Acc) and bias (regression coefficient [Reg]) of GEBV for residual feed intake (RFI) and its component traits

	DFI[2]		RFI		ADG		BF	
Method[1]	Acc	Reg	Acc	Reg	Acc	Reg	Acc	Reg
GBLUP	0.517	1.281	0.517	1.237	0.283	0.725	0.317	0.665
BPL0.3	0.523	1.391	0.506*	1.233	0.329	1.054	0.351*	0.787
BPL0.5	0.528	1.399	0.508	1.200	0.318	0.992	0.310	0.710
BPL0.7	0.518	1.432	0.508	1.200	0.313	0.909	0.316	0.754
BPL0.9	0.531	1.413	0.519	1.232	0.320	0.931	0.313	0.746
BL	0.515	1.349	0.509	1.190	0.318	0.949	0.308	0.704
BayesA	0.528	1.287	0.535*	1.088	0.357	0.811	0.362*	0.410
BayesB	0.508	1.252	0.519	1.237	0.295	0.754	0.331	0.653
Bayes Cp1.1	0.515	1.271	0.521	1.240	0.286	0.729	0.336	0.672
Bayes Cp10.1	0.525	1.282	0.516	1.233	0.276	0.702	0.350*	0.703
Bayes Cp100.1	0.531	1.350	0.532*	1.185	0.302	0.775	0.344*	0.682

[1] GBLUP, genomic BLUP; BPL, Bayesian Power LASSO; BL, Bayesian LASSO.

[2] DFI, daily feed intake; RFI, residual feed intake; BF, back fat.

*Significantly different from the GBLUP model ($P < 0.05$).

Partitioning of Genomic Variances and Predictive Ability Among SNP Classes

The markers ALGA0082564 and ASGA0077969 were significantly associated with BF whereas the marker ASGA0077969 and ASGA0077977 were significantly associated with ADG (with a Bayes factor >10; Table S2). We also detected 12, 400, 30, and 82 SNP suggestively associated (Bayes factor >3) with DFI, RFI, ADG, and BF traits, respectively (Table S1). The genomic variance by annotated SNP groups using BayesCp100.1 is shown in Table 2. The biggest numbers of SNP were annotated to intergenic regions (62.75%). Among the genic classes, intron contained the biggest numbers of SNP (24.3%). The intergenic and intronic region also contained major numbers of suggestive SNP across the traits (Table S1). The intergenic class contributed the largest variance (61.4 to 65.1%) of total genomic variance. There are slightly different contributions to total genomic variance of different classes across these traits. Among genic classes, intron also contributed most to total genomic variance (23.87 to 25.15%). Because 30,234 SNP were used to estimate genomic variance, the expected average genomic variance by an SNP was 1/30,234 = 3.31E-05 of the total. Across these traits, the contribution to total genomic variance per SNP among these classes is also similar to the average value of 3.31E-05, although the contribution of the mRNA NMD transcript is higher than the average value across all the traits.

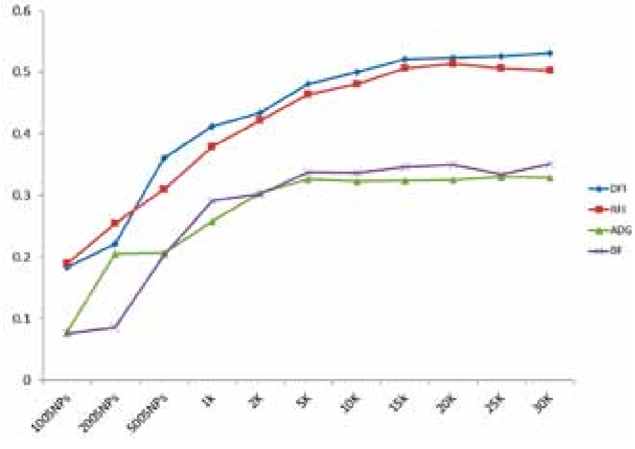

Figure 1. Predictive accuracy using different SNP panels. The y axis shows the prediction accuracy as a correlation between corrected phenotypes and GEBV divided by the square root of heritability. DFI, daily feed intake; RFI, residual feed intake; BF, back fat.

Prediction accuracy using intergenic class slightly lowered the average value generated from 1,000 randomized groups across these traits (Table 3). In contrast, using 3´UTR, 5´UTR, intron, gene, gene ± 5 kb, and missense variant classes slightly improved predictive accuracy (higher than an average value from 1,000 randomized groups). Moreover, the accuracy of prediction using missense was significant for DFI. The lower predictive ability (than average values from randomized groups) was also found in some classes such as intron_noncoding transcript (for DFI, RFI, and ADG traits) and intergenic class (for all traits).

DISCUSSION

Genomic Prediction Using Different Methods

The comparison of accuracy of genomic prediction using different methods and response variables has been done in many studies (reviewed by Meuwissen et al. [2013]). Many factors influence prediction accuracy including the extent of LD between SNP markers and QTL, the number of phenotype records and genotyped animals, the heritability, and the distribution of QTL effects for the trait (Hayes et al., 2009). Several studies also reported that

Table 2. Genomic partitioning-based annotated classes as genomic variance explained per SNP for residual feed intake (RFI) and component traits

		DFI[2]		RFI		ADG		BF	
Class	SNP[1]	Var·exp[3] (%)	Var·exp/ SNP[4]	Var·exp (%)	Var·exp/ SNP	Var·exp (%)	Var·exp/ SNP	Var·exp (%)	Var·exp/ SNP
Gene ± 5 kb	10,405	35.19	3.38E-05	35.88	3.45E-05	31.83	3.06E-05	34.53	3.32E-05
Downstream	1,110	3.82	3.45E-05	3.68	3.31E-05	3.93	3.54E-05	3.75	3.38E-05
Upstream	1,211	4.09	3.38E-05	3.89	3.21E-05	4.37	3.61E-05	4.06	3.36E-05
Gene	8,084	27.28	3.37E-05	28.31	3.50E-05	23.53	2.91E-05	26.72	3.31E-05
3´UTR[5]	154	0.49	3.15E-05	0.51	3.28E-05	0.53	3.45E-05	0.52	3.35E-05
5´UTR	36	0.11	3.08E-05	0.11	3.04E-05	0.12	3.29E-05	0.11	3.00E-05
Intron	7,347	24.67	3.36E-05	23.87	3.25E-05	25.15	3.42E-05	24.43	3.33E-05
Intron_NMD _transcript[6]	28	0.10	3.61E-05	0.1	3.52E-05	0.13	4.79E-05	0.09	3.37E-05
Intron_NC_ transcript[7]	51	0.19	3.72E-05	0.17	3.39E-05	0.16	3.23E-05	0.17	3.25E-05
Splice_region	32	0.10	3.27E-05	0.11	3.30E-05	0.11	3.54E-05	0.11	3.32E-05
Exon	426	1.42	3.25E-05	1.4	3.22E-05	1.54	3.54E-05	1.47	3.38E-05
Missense	109	0.34	3.14E-05	0.36	3.34E-05	0.38	3.45E-05	0.38	3.51E-05
Synonymous	305	1.00	3.29E-05	0.96	3.16E-05	1.09	3.58E-05	1.02	3.34E-05
Intergenic	18,974	61.99	3.27E-05	61.4	3.24E-05	65.1	3.43E-05	62.64	3.30E-05

[1] Number of SNP in each class after quality control.

[2] DFI, daily feed intake; BF, back fat.

[3] Percentage of total genomic variance explained by each SNP class.

[4] Average genomic variances explained by SNP in each class.

[5] UTR, untranslated region.

[6] Intron_nonsense-mediated mRNA decay transcript.

[7] Intron_non coding transcript.

Bayesian or their extensions perform better than GBLUP (Legarra et al., 2011; Gao et al., 2013). In the current population, accuracy was slightly increased by using some of the Bayesian methods (such as BayesCp100.1) compared with standard GBLUP (Table 2). In general, our results agreed with Ostersen et al. (2011) who reported that BL had the same reliability as the GBLUP method for feed conversion ratio and ADG traits in pigs. The Bayesian methods only benefit when these traits are influenced by a few large QTL or when the relationship between training and testing populations was low. Residual feed intake and its component traits are highly polygenic and high LD has been reported in pig populations: therefore, it is probable that these conditions are not conducive to the application of Bayesian methods. Sharpness in BPL did not affect accuracy in the current study (Table 1). In contrast, Gao et al. (2013) indicated that the BPL model with a power parameter of 0.3 had the highest accuracy. This is probably due to the close relationship of the training and testing populations in the current study than that found in Gao et al. (2013). The accuracy of genomic prediction for RFI (approximately 0.5) is higher compared to values reported for feed conversion ratio (**FCR**) in the same Duroc population (approximately 0.16; Ostersen et al., 2011). Jiao et al. (2014) also reported lower predictive accuracy (0.094) for RFI in American Duroc boars using a similar population size (1,047 boars genotyped with the pig 60K SNP chip). It must be noted that our prediction accuracy is calculated using the square root of heritability as a scale parameter, which differed from the accuracy calculation by Jiao et al. (2014). Moreover, there are several possible reasons for the low accuracies, such as reference population size and numbers of markers. All methods had high bias, with both inflation and deflation of genomic variances. Christensen et al. (2012) also reported high bias prediction using GBLUP for ADG (0.8) and FCR (0.57) in the same populations (except they had more records) but less when using the single step method. These results suggested the current data might have some problem with (pre) selection, and there is a need to further investigate the sources of bias. The accuracy of prediction can be reached by a reduced set of SNP. Our results based on cross validation were in agreement with previous reports (Moser et al., 2010). For instance, Moser et al. (2010) showed that subsets containing 3,000 SNP provided more than 90% of the accuracy that could be achieved with a high-density assay for cows. The accuracy of prediction using a small subset of selected SNP was also high in pigs. However, it is important to note that the accuracy achieved by a small set of SNP varies by methods of selection, traits, and reference populations. Moreover, the bias using a small set of SNP is potentially higher than that reported by the full SNP array. Nevertheless, breeding programs (in Denmark) currently impute low-density (7K) to medium size SNP chips to capture the whole genomic variances of traits.

Table 3. Predictive accuracy of annotated classes versus mean accuracy of 1,000 randomly generated groups for residual feed intake (RFI) and component traits

	DFI[1]			RFI			ADG			BF		
Class	Acc	Mean×Acc random[2]	95% Acc threshold[3]	Acc	Mean×Acc random	95% Acc threshold	Acc	Mean×Acc random	95% Acc threshold	Acc	Mean×Acc random	95% Acc threshold
Gene ± 5 kb	0.507	0.471	0.536	0.492	0.489	0.533	0.331	0.288	0.334	0.333	0.322	0.358
Downstream	0.231	0.378	0.490	0.290	0.384	0.492	0.318	0.208	0.348	0.243	0.284	0.418
Upstream	0.455	0.385	0.492	0.425	0.391	0.496	0.222	0.212	0.353	0.386	0.291	0.420
Gene	0.511	0.458	0.527	0.493	0.483	0.535	0.393	0.282	0.583	0.331	0.316	0.356
3´UTR[4]	0.253	0.190	0.351	0.243	0.200	0.353	0.175	0.098	0.279	0.182	0.134	0.299
5´UTR	0.177	0.097	0.269	0.160	0.103	0.273	0.01	0.052	0.251	0.091	0.065	0.248
Intron	0.500	0.483	0.536	0.482	0.481	0.535	0.391	0.280	0.350	0.420	0.396	0.461
Intron_NMD _transcript[5]	0.055	0.115	0.288	0.129	0.120	0.292	0.091	0.063	0.265	0.01	0.081	0.257
Intron_NC_ transcript[6]	0.015	0.087	0.256	0.063	0.090	0.265	0.01	0.046	0.240	0.060	0.057	0.234
Splice_region	0.154	0.091	0.262	0.027	0.097	0.270	0.186	0.049	0.245	0.090	0.061	0.246
Exon	0.284	0.337	0.393	0.283	0.293	0.426	0.199	0.141	0.278	0.109	0.245	0.264
Missense	0.331*	0.162	0.319	0.179	0.173	0.329	0.209	0.084	0.266	0.190	0.114	0.293
Synonymous	0.221	0.251	0.397	0.233	0.261	0.410	0.137	0.133	0.312	0.053	0.180	0.346
Intergenic	0.471	0.500	0.533	0.477	0.498	0.530	0.208	0.296	0.342	0.389	0.417	0.460

[1] DFI, daily feed intake; BF, back fat.

[2] The average value of prediction accuracy (Acc) of 1,000 randomly generated selected groups with the same number of SNP as in each class.

[3] The 95% of prediction accuracy of 1,000 randomly generated selected groups with the same number of SNP as in each class. The prediction accuracy of annotated class is significant if it is higher than this value.

[4] UTR, untranslated region.

[5] Intron_nonsense-mediated mRNA decay transcript.

[6] Intron_non coding transcript.

*Significant at $P < 0.05$.

Genomic Partitioning and Predictive Ability of Annotated SNP Classes Using 60K SNP Chip

Using the BayesCp100.1 method, we also detected 2 significant SNP (ASGA0077977 in the intron of the CBLN4 gene and ASGA0077969 in the intergenic region) for ADG on SSC 17, which also have been reported in a single SNP linear mixed model analysis in a previous study (Do et al., 2014a). Moreover, 2 novel SNP (ALGA0082564 and ASGA0027339 in the intron regions of *TACC2* [SSC 14] and *TBC1D22A* gene [SSC 5], respectively) were also detected for BF. Recently, a mutation in *TBC1D22A* was reported to be associated with fat traits (waist circumference) in humans (Liu et al., 2014). It is important to note that several SNP detected by Bayesian methods were not reported in the previous linear mixed model mapping. According to Sahana et al. (2010), Bayesian variable selection regression (BayesCπ) mapping resulted in higher power and more precise QTL locations than single-marker in a simulation study.

We annotated intergenic variants as the most common variants (56.3 and 62.8% of SNP before and after quality control). This result agreed with the annotation results from other species. For instance, Koufariotis et al. (2014) annotated 67.8 and 67.3% intergenic SNP for dairy and beef, respectively. Moreover, we also observed that all top 10 associated SNP were located in either intronic or intergenic regions across the traits, except that marker ALGA0011482 associated with DFI and marker DIAS0004797 associated with BF were in the downstream and synonymous regions, respectively (Table S2). The results implied the possible existing LD between the top SNP and the causal SNP to regulate phenotypes; however, the results might be limited because of the very few top SNP investigated here. To a greater extent, the contribution of different annotated classes is linearly associated with the number of SNP in these classes. The intergenic class consists of SNP outside genic regions and contributes to around 61 to 65% of total genomic variance, depending on the trait. Among genic groups, intron contained the highest number of SNP (7,347 SNP—24.3% of total SNP); therefore, it contributed around 23 to 25% of total additive genomic variance for all traits. Therefore, the average genetic contribution was similar between different groups and among different traits. In humans, Hindorff et al. (2009) reported 43% of traits/diseases associated with SNP from Genome wide association studies (**GWAS**) were intergenic and 45% were intronic. Yang et al. (2011) showed that SNP in or near genes explain more variation than SNP between genes for complex traits such as height and body mass index. Kindt et al. (2013) suggested SNP further away from the transcription start site were less likely to be significantly associated with trait. These results implied diversity in regulation of complex traits, which can be by either genic regions or nearby gene region, and even by intragenic region as well as by interaction among regulators in these regions. Nevertheless, it is also important to note that approximately 90% of SNP on the pig 60K SNP chip were annotated (Table S1); therefore, enrichment analysis can be affected by this limitation of annotation. Moreover, the pig 60K chip was designed based on common variants, although rare variants might explain significant variance in phenotypes. The annotation was based on the position of the SNP without considering their LD in this study. Consequently, partitioning of the variance might be affected if SNP in an annotated class are in high LD with causative SNP in other classes. Overall, there is a need for comprehensive bioinformatics or systems genetics tools that capture and group SNP based on all possible criteria including LD, genomic positions, and biological functions. One such tool developed specifically for livestock species and for humans is FunctSNP (Goodswen et al., 2010), which captures not only LD between SNP and QTL or genes but also the SNP classes.

In general, the predictive performance of different classes followed a pattern similar to genomic variance partitioning. Predictive performances for different annotated groups did not significantly differ from randomized groups, with the only exception that the predictive performance of missense was significantly higher than the randomized groups for DFI. The missense variant is defined as a point mutation leading to the substitution of one AA in protein for another, and this substitution can lead to a change in phenotype. For instance, a missense variant (Asp124Asn) of the porcine melanocortin-4 receptor was found to be associated with variation in fatness, growth, and feed intake (Kim et al., 2000). A missense mutation in the peroxisome proliferator-activated receptor delta gene caused a major QTL effect on ear size in pigs (Ren et al., 2011). Recently, Morota et al. (2014) showed coding sequences, genes, gene ± 1 kb, and exon parts of the chicken genome provided better predictive power than a randomized set, depending on the traits of interests. For instance, the authors reported that predictive performance for ultrasound of breast meat from genic regions was consistently better than that of SNP in intergenic regions. However, these differences were small; therefore, these authors suggested using all markers to predict GEBV of complex traits. The prediction accuracy and additive genomic variance explained by markers also depends on the number of markers on causative sites, LD between them with causative genes, and finally LD among markers and genes at the family level (Jensen et al., 2012). In fact, the understanding of genetic architecture of complex traits is limited in pigs; numbers of known causative variants are still very few for most traits. The dissection of annotated SNP may help researchers to understand which genomic regions provide higher predictive performance if better annotation data/methods become available. Genomic prediction is still in a black box, and a true understanding of the biology remains a challenge (van der Steen et al., 2005; van der Werf, 2007; Kadarmideen, 2014). The genomic prediction methods including biological knowledge have been performed and showed interesting results. For instance, Zhang et al. (2014) showed that the inclusion of known QTL can improve accuracy and reduce the bias of prediction of production traits in cattle. Snelling et al. (2013) indicated that networks and pathways can help to improve genomic selection. Finally, Kadarmideen (2014) provided a framework in which various types of information on SNP and other genetic variants can be included in a formal systems GBLUP method. The future perspectives of genomic predictions may be targeted at revealing different biological classes of SNP that are highly predictive for different complex traits and appropriately including them in genomic selection decisions.

Conclusion

This study calculated the accuracy of GEBV for RFI and its component traits, which can be of interest to pig breeders. Genomic prediction based on a relatively small training data set reached accuracies comparable to observed phenotypes. These results suggest that the choice of genomic prediction method has less impact on predictive performance for RFI and its component traits in pigs. Classification of SNP by genomic annotation had little impact on the accuracy of prediction for traits investigated here but could be different for other traits depending on the genetic architecture. Better annotation and classification methods, such as using the functional biological relevance of SNP in addition to the structural relevance of SNP, are needed when considering future genomic prediction strategies and when attempting to understand functional genomic architecture of complex traits. However, this is one of the very first studies to provide useful insights into the contribution of different biological classes of SNP in explaining genetic variation using markers and in driving the whole genomic prediction for complex traits in pigs.

LITERATURE CITED

Christensen, O. F., P. Madsen, B. Nielsen, T. Ostersen, and G. Su. 2012. Single-step methods for genomic evaluation in pigs. Animal 6:1565–1571.

Do, D. N., T. Ostersen, A. B. Strathe, J. Jensen, T. Mark, and H. N. Kadarmideen. 2014a. Genome-wide association and systems genetic analyses of residual feed intake, daily feed consumption, backfat and weight gain in pigs. BMC Genet. 15:27. doi:10.1186/1471-2156-15-27.

Do, D. N., A. B. Strathe, J. Jensen, T. Mark, and H. N. Kadarmideen. 2013. Genetic parameters for different measures of feed efficiency and related traits in boars of three pig breeds. J. Anim. Sci. 91:4069–4079. doi:10.2527/jas.2012-6197.

Do, D. N., A. B. Strathe, T. Ostersen, S. D. Pant, and H. N. Kadarmideen. 2014b. Genome-wide association and pathway analysis of feed efficiency in pigs reveal candidate genes and pathways for residual feed intake. Front. Genet. 5:307.

Dunn, O. J., and V. Clark. 1971. Comparison of tests of the equality of dependent correlation coefficients. J. Am. Stat. Assoc. 66:904–908. doi:10.1080/01621459.1971.10482369.

Gao, H., G. Su, L. Janss, Y. Zhang, and M. S. Lund. 2013. Model comparison on genomic predictions using high-density markers for different groups of bulls in the Nordic Holstein population. J. Dairy Sci. 96:4678–4687. doi:10.3168/jds.2012-6406.

Goodswen, S. J., C. Gondro, N. S. Watson-Haigh, and H. N. Kadarmideen. 2010. FunctSNP: An R package to link SNPs to functional knowledge and dbAutoMaker: A suite of Perl scripts to build SNP databases. BMC Bioinformat. 11:311. doi:10.1186/1471-2105-11-311.

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick. 2011. Extension of the Bayesian alphabet for genomic selection. BMC Bioinformat. 12:186. doi:10.1186/1471-2105-12-186.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009. Invited review: Genomic selection in dairy cattle: Progress and challenges. J. Dairy Sci. 92:433–443. doi:10.3168/jds.2008-1646.

Hindorff, L. A., P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. U S A 106:9362–9367. doi:10.1073/pnas.0903103106

Jensen, J., and P. Madsen. 1994. DMU: A package for the analysis of multivariate mixed models. Proceedings of the 5th World Congress on Genetics Applied to Livestock Production, August 7-12, 1994, Guelph, Canada, Vol 22:45-46.

Jensen, J., G. Su, and P. Madsen. 2012. Partitioning additive genetic variance into genomic and remaining polygenic components for complex traits in dairy cattle. BMC Genet. 13:44. doi:10.1186/1471-2156-13-44.

Jiao, S., C. Maltecca, K. A. Gray, and J. P. Cassady. 2014. 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. 92:2377–2386. doi:10.2527/jas.2013-7338.

Kadarmideen, H. N. 2014. Genomics to systems biology in animal and veterinary sciences: Progress, lessons and opportunities. Livest. Sci. 166:232–248. doi:10.1016/j.livsci.2014.04.028.

Kadarmideen, H. N., D. Schwörer, H. Ilahi, M. Malek, and A. Hofer. 2004. Genetics of osteochondral disease and its relationship with meat quality and quantity, growth, and feed conversion traits in pigs. J. Anim. Sci. 82:3118–3127.

Kass, R. E., and A. E. Raftery. 1995. Bayes factors. J. Am. Stat. Assoc. 90:773–795. doi:10.1080/01621459.1995.10476572.

Kim, K. S., N. Larsen, T. Short, G. Plastow, and M. F. Rothschild. 2000. A missense variant of the porcine melanocortin-4 receptor (MC4R) gene is associated with fatness, growth, and feed intake traits. Mamm. Genome 11:131–135. doi:10.1007/s003350010025.

Kindt, A. S. D., P. Navarro, C. A. M. Semple, and C. Haley. 2013. The genomic signature of trait-associated variants. BMC Genom. 14:108. doi:10.1186/1471-2164-14-108.

Koufariotis, L., Y.-P. P. Chen, S. Bolormaa, and B. J. Hayes. 2014. Regulatory and coding genome regions are enriched for trait associated variants in dairy and beef cattle. BMC Genom. 15:436. doi:10.1186/1471-2164-15-436.

Legarra, A., C. Robert-Granié, P. Croiseau, F. Guillaume, and S. Fritz. 2011. Improved Lasso for genomic selection. Genet. Res. (Camb.) 93:77–87. doi:10.1017/S0016672310000534.

Liu, A. Y., D. Gu, J. E. Hixson, D. C. Rao, L. C. Shimmin, C. E. Jaquish, D.-P. Liu, J. He, and T. N. Kelly. 2014. Genome-wide linkage and regional association study of obesity-related phenotypes: The GenSalt study. Obesity (Silver Spring) 22:545–556. doi:10.1002/oby.20469.

Meuwissen, T., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829.

Meuwissen, T., B. Hayes, and M. Goddard. 2013. Accelerating improvement of livestock with genomic selection. Annu. Rev. Anim. Biosci. 1:221–237. doi:10.1146/annurev-animal-031412-103705.

Morota, G., R. Abdollahi-Arpanahi, A. Kranis, and D. Gianola. 2014. Genome-enabled prediction of quantitative traits in chickens using genomic annotation. BMC Genom. 15:109. doi:10.1186/1471-2164-15-109.

Moser, G., M. S. Khatkar, B. J. Hayes, and H. W. Raadsma. 2010. Accuracy of direct genomic values in Holstein bulls and cows using subsets of SNP markers. Genet. Sel. Evol. 42:37. doi:10.1186/1297-9686-42-37.

Ostersen, T., O. F. Christensen, M. Henryon, B. Nielsen, G. Su, and P. Madsen. 2011. Deregressed EBV as the response variable yield more reliable genomic predictions than traditional EBV in purebred pigs. Genet. Sel. Evol. 43:38. doi:10.1186/1297-9686-43-38.

Plummer, M., N. Best, K. Cowles, and K. Vines. 2006. CODA: Convergence diagnosis and output analysis for MCMC. R News 6:7–11.

Ren, J., Y. Duan, R. Qiao, F. Yao, Z. Zhang, B. Yang, Y. Guo, S. Xiao, R. Wei, Z. Ouyang, N. Ding, H. Ai, and L. Huang. 2011. A missense mutation in PPARD causes a major QTL effect on ear size in pigs. PLoS Genet. 7:e1002043. doi:10.1371/journal.pgen.1002043.

Sahana, G., B. Guldbrandtsen, L. Janss, and M. S. Lund. 2010. Comparison of association mapping methods in a complex pedigreed population. Genet. Epidemiol. 34:455–462. doi:10.1002/gepi.20499.

Snelling, W. M., R. A. Cushman, J. W. Keele, C. Maltecca, M. G. Thomas, M. R. S. Fortes, and A. Reverter. 2013. Breeding and Genetics Symposium: Networks and pathways to guide genomic selection. J. Anim. Sci. 91:537–552. doi:10.2527/jas.2012-5784.

van der Steen, H. A. M., G. F. W. Prall, and G. S. Plastow. 2005. Application of genomics to the pork industry. J. Anim. Sci. 83:E1–E8.

van der Werf, J. 2007. Animal Breeding and the black box of biology. J. Anim. Breed. Genet. 124:101. doi:10.1111/j.1439-0388.2007.00657.x.

VanRaden, P. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423. doi:10.3168/jds.2007-0980.

Yang, J., T. A. Manolio, L. R. Pasquale, E. Boerwinkle, N. Caporaso, J. M. Cunningham, M. De Andrade, B. Feenstra, E. Feingold, M. G. Hayes, W. G. Hill, M. T. Landi, A. Alonso, G. Lettre, P. Lin, H. Ling, W. Lowe, R. A. Mathias, M. Melbye, E. Pugh, M. C. Cornelis, B. S. Weir, M. E. Goddard, and P. M. Visscher. 2011. Genome partitioning of genetic variation for complex traits using common SNPs. Nat. Genet. 43:519–525. doi:10.1038/ng.823.

Zhang, Z., U. Ober, M. Erbe, H. Zhang, N. Gao, J. He, J. Li, and H. Simianer. 2014. Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. PLoS ONE 9:e93017. doi:10.1371/journal.pone.0093017.