Genomic prediction and genomic variance partitioning of daily and residual feed intake in pigs using Bayesian Power Lasso models

D.N Do[1], L.L.G.Janss[3], A.B Strathe[1, 2], J.Jensen[3] and H.N Kadarmideen[1]

[1]Faculty of Health and Medical Sciences, University of Copenhagen, [2]Pig Research Centre, [3]Aarhus University, Denmark

ABSTRACT: Improvement of feed efficiency is essential in pig breeding and selection for reduced residual feed intake (RFI) is an option. The study applied Bayesian Power LASSO (BPL) models with different power parameter to investigate genetic architecture, to predict genomic breeding values, and to partition genomic variance for RFI and daily feed intake (DFI). A total of 1272 Duroc pigs had both genotypic and phenotypic records for these traits. Significant SNPs were detected on chromosome 1 (SSC 1) and SSC 14 for RFI and on SSC 1 for DFI. BPL had similar accuracy and bias as GBLUP but power parameters had no effect on predictive ability. Genomic variance partitioning showed that SNP groups either by position (intron, exon, downstream, upstream and 5'UTR) or by function (missense and protein-altering) had similar average explained variance per SNP, except that 3'UTR had a higher value.

Keywords: Bayesian Power Lasso; genomic prediction; pigs

Introduction

Genomic selection using molecular markers covering the whole genome for predicting genomic breeding values (GEBVs) is widely used in both animal and plant species. In pigs, genomic selection is currently implemented and is attractive for traits which are expensive to measure or cannot be measured early in life. Feed efficiency is a very important trait in the breeding goal and is costly to record. Residual feed intake (RFI) is an alternative indicator for feed efficiency and selection for lower RFI may help to improve feed efficiency. Genomic selection reducing feed efficiency is interesting although choice of phenotypes is debated.

Various methods have been proposed for genomic selection based on different assumptions about genetic architecture underlying the trait. The accuracy of methods depends on how these assumptions are fulfilled for the traits of interest. Bayesian least absolute shrinkage and selection operator (Bayesian LASSO (BL), Park and Casella, 2008) and Bayesian Power LASSO (BPL, Gao et al. 2013) assume that SNP effects follow a double exponential distribution (with an extra power parameter for BPL) and are attractive models for genomic selection. This is due to simplicity, computational ease and little (or no) need to postulate prior information (Legarra et al. 2012). Moreover, the assumption about genetic architecture in Bayesian models are more flexible than Genomic best linear unbiased prediction method (GBLUP), therefore models can be applied to phenotypes with little understanding of trait biology. The main aims of this study were to investigate genetic architecture of RFI and DFI based on partitioning genomic variance, and to investigate predictive ability using the BPL models.

Materials and Methods

Data and quality control. Phenotypic records included the feed intake and feeding behaviors in a period from 2008 to 2012 for Danish Duroc pigs. Daily feed intake (DFI) was recorded from 30 to approximately 105 kg at the national pig test station. Residual feed intake was the residual in the regression of DFI on average daily gain and backfat with initial body weight was as a 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 was 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 SNPs were removed from the study. After quality control (QC), 30232 SNPs and 1272 pigs remained for analysis.

Genomic prediction methods. For reference purpose, GBLUP model used was: $\mathbf{y} = \mathbf{1}\mu + \mathbf{Xb} + \mathbf{Z}p + \mathbf{g} + e$ in which $\mathbf{y}$ was the vector of observed phenotypic values of the animals (RFI or DFI), $\mathbf{1}$ was a vector of ones, μ was the overall mean, $\mathbf{b}$ was vector of fixed effect (herd-year-section), $\mathbf{X}$ was a design matrix relating observations to the corresponding the fixed effect, p was a vector of random pen effect, $\mathbf{Z}$ was a design matrix relating observation to the corresponding the random pen effect, $\mathbf{e}$ was the vector of random error, and $\mathbf{g}$ was a vector of breeding value with $var(g) = \boldsymbol{G}\sigma_g^2$, in which σ_g^2 is genetic variance and $\mathbf{G}$ was the genomic relationship matrix. The GBLUP model was fitted using DMU package (Jensen and Madsen, 2010).

The Bayesian Power LASSO (BPL) is a sparse shrinkage model that uses an exponential power distribution for marker effects; details of the model are in Gao et al. 2013. BPL is an extension of BL by adding a power parameter that can modify the sparsity of the marker effects. The model was ($\mathbf{y} = \mathbf{1}\mu + \mathbf{X}b + \mathbf{M}\beta + e$); where SNPs effect ($\beta_i$) follow an exponential power distribution $p(\beta) = \prod_{j=1}^{m} \frac{\lambda}{2} e^{-\lambda|\beta|q}$, where λ was a rate parameter, m was the number of markers, and q was 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. The Bayesian analyses were performed using BayZ package (http://www.bayz.biz/). Each of the Bayesian analysis was run as a single chain with a length of 50000 samples, and the first 5000 cycles were regarded as the burn-in period.

Evaluation criterion. To investigate the accuracy of different genomic selection methods, we split the records

into a training dataset (968 pigs) and a validation dataset (304 pigs) by birth date at 1 January 2012. Moreover, we also corrected phenotypes for a fixed effect of herd-year-section and a random pen effect to avoid use of 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). The regression of y_c on GEBV was used to assess bias or inflation of prediction.

Partitioning genomic variance based SNPs groups. There were three different methods to group SNPs: (i) functional relevance of SNPs for RFI based on data mining, (ii) groups based on positions in the genome, and significance based on previous association study in the same data set. Data mining of candidate SNPs for RFI was performed based on 3 main databases (Pubmed at http://www.ncbi.nlm.nih.gov/pubmed, ISI web of knowledge at http://portal.isiknowledge.com/ and Scopus at http://www.scopus.com/). Positional and functional groups of SNPs were based on Sus Scrofa gene 10.2 at http://www.ensembl.org/biomart. The SNPs were classified upon these positions in a gene (intron, exon, coding sequence, 3'UTR, 5'UTR, upstream and downstream region) as well as based on these function if they are missense variants or protein-altering variants.

Results and Discussion

Genome wide mapping and predictive performance. Bayesian linear regression has been widely used for the GWA studies. We estimated SNP effects and SNP variance using the BPL models. The total estimated genomic variance varied with the sharpness (q value), however the top SNPs with highest variance (effects) were similar in models with different power parameters. We also found that SNPs on chromosome (SSC) 1 (30-31Mb) and 60-61Mb had the largest effects on variance of RFI (Figure 1) and DFI, respectively, which confirmed the results from linear analysis recorded in previous GWA study (Do et al. 2014). Moreover, we also detected some SNPs on SSC 14 for RFI, which were not recorded in our previous GWA. It is important to note that the standard BL (q = 1) estimated the highest value of additive genetic effects (Table 1). Reducing the q value allowed more intense shrinkage and higher sparsity of SNPs effects that explained the estimated SNP variance is higher at q = 0.3 than q = 1. Moreover, genetic architecture of traits influences on the estimation of SNP variance. Because RFI was defined as DFI corrected for growth and backfat, therefore numbers of QTL effect on RFI might be less than DFI. This was in agreement that the total genomic variance of RFI was smaller than that of DFI.

Accuracy of genomic prediction for RFI and DFI was similar (Table 1). For both RFI and DFI, all BPL methods showed similar accuracy of prediction as the GBLUP method. These results was in agreement with Ostersen et al. 2011, who reported that BL had same reliability with GBLUP method for food conversion ratio traits in pigs. There was no difference in prediction accuracy using different sharpness in the current study (Table 1). Gao et al. 2013 indicated that BPL model with power parameter of 0.3 had highest accuracy. In our data there were a closer relationship between training and testing populations than that found in Gao et al. 2013. GBLUP model had similar bias as BPL. Several studies found Bayesian methods improved the genomic selection accuracy but increased bias compared to GBLUP method (Gao et al. 2013 and Su et al. 2012).

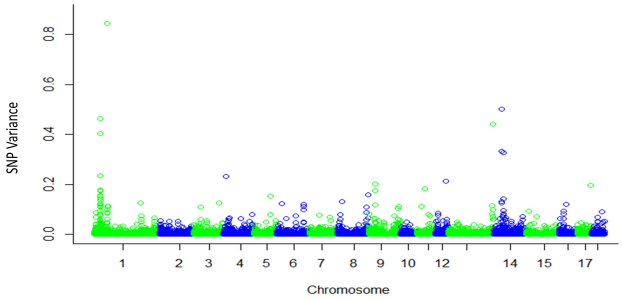

Figure 1: **Manhattan plot of estimated SNP variance for RFI with BPL0.3**

Table 1. **Estimation of total genomic variance (TGV), prediction accuracy of GEBVs for daily and residual feed intake**

Method	DFI			RFI		
	TGV	Cor	Reg	TGV	Cor	Reg
BEP0.3	16318.9	0.34	1.39	11705.6	0.31	1.23
BEP0.5	16262.1	0.34	1.40	12071.3	0.31	1.20
BEP0.7	15986.5	0.33	1.43	12295.4	0.31	1.20
BEP0.9	16472.9	0.34	1.41	12334.1	0.32	1.23
BL	16836.4	0.33	1.35	12583.1	0.31	1.19
GBLUP	19478	0.33	1.28	14380	0.31	1.23

Partitioning of genomic variance. Patterns of genomic variance from different groups of SNPs are shown in Table 2. To a large extent, the contribution of different groups was linearly associated with the number of SNPs in the groups. Among them, intron group contained biggest number of SNPs (5621 SNPs), therefore it contributed around 20% of total genomic variance for both RFI and DFI. SNPs in exons did not contribute more on average than the intron SNPs, but SNPs in 3' UTR explained on average three times more, and SNPs from literature mining explain on average two times more variance than intronic and exonic SNPs. Results also show that an average contribution to total genetic variance of RFI in the "data mining group" was lower than in the other groups. This reflects that: (i) many SNPs associated with RFI or feed traits are not on the 60 K SNPs chip or does not pass QC control, (ii) very few GWAS and QTL have been found for RFI in pigs and mostly done by using cross breeding

experiments or in other pig breeds, (iii) and the definition of RFI are currently inconsistent in different studies and associated SNPs may depend on the definition. Secondly, different groups by position using Ensembl database had similar average variance contribution per SNP, such as a SNP in exon or intron had almost the same values of variance contribution. With an exception that average contribution of SNPs in 3'UTR to genomic variance was more than one in other groups. This was probably due to the number of SNPs in the group was very small and the average contribution increased due to only several highly variant SNPs. However 3'UTR often contains regulatory regions that influence post-transcriptional gene expression, and therefore it could also influence specific traits.

Table 2. **Partitioning of genetic variance for different SNP groups**

		DFI		RFI	
Group[1]	No. SNPs	Ex.Var (%)	Average SNP.Var	Ex.Var (%)	Average SNP.Var
Exon	220	0.77	3.50E-05	0.74	3.39E-05
Intron	5621	18.99	3.38E-05	19.45	3.46E-05
Coding sequence	214	0.75	3.50E-05	0.73	3.41E-05
Missense variants	53	0.19	3.55E-05	0.19	3.54E-05
Proterin-alterning variant	54	0.19	3.58E-05	0.19	3.53E-05
3'UTR variants	26	0.38	1.47E-04	0.34	1.32E-04
5'UTR variants	110	0.09	7.76E-06	0.12	1.09E-05
Upstream variants	1051	3.51	3.34E-05	3.29	3.13E-05
Downstream variants	987	3.30	3.35E-05	3.16	3.20E-05
Significantly associated SNPs [2]	11	NA	NA	0.43	3.88E-04
Suggestively associated SNPs [3]	164	NA	NA	2.11	1.28E-04
Data mined SNPs/genes [4]	110	NA	NA	0.69	6.24E-05

[1]: a SNP can appear in some different groups

[2]and [3]: Significant and suggestive SNPs associated with RFI 2 in previous studies (Do et al., BMC Genetics, 2014)

[4]: SNPs have functionally involved in RFI from data mining, the summary for all breeds.

Conclusion

The study showed the BPL methods had similar prediction accuracy and bias as those reported with GBLUP method. The different shape parameters did not affect predictive performance of BPL methods. These results suggested that the choice of genomic selection method had less impact on predictive performance for RFI and DFI. However, the study helped better understanding of the genetic architecture of RFI and DFI as well as the contribution of different functional SNP groups to total genomic variance of these traits. Further work will investigate the predictive ability using different SNPs group for RFI as well as it component traits.

Acknowledgements

We thank the Department of Breeding and Genetics of the Danish Pig Research Centre for providing all data. DND is funded by Faculty of Health and Medical Sciences, University of Copenhagen. HNK thanks EU-FP7 Marie Curie Actions – Career Integration Grant (CIG-293511) granted to him for partially funding this study.

Literature Cited

Do, D.N., Strathe, A. B., Jensen, J., et al. (2013). J. Anim. Sci. 91, 4069-4079

Do, D.N., Ostersen, T., Strathe, A. B., et al .(2014) BMC Genetics. 15:27

Gao, H., Su, G., Janss, L., et al. (2013), J.Dairy Sci. 96(7): 4678-4687.

Legarra, A., Christèle, R., Pascal, C., et al. (2011). Gen Res. 93(01): p. 77-87.

Madsen, P., and Jensen, J. (2012). DMU. Version 6, release 5.2

Ostersen, T., Christensen, O.F., Henryon, M., et al.(2011). Gen.Sel.Evo. 43(1): 38.

Park, T. and G. Casella, (2008). JASA. 103(482): 681-686.

Su, G., Brondum, R.F., Ma, P., et al. (2012). J.Dairy Sci. 95:4 657-4665.