Separating signal from noise
Estimating SNP-effects and Decomposing Genetic Variation to the level of QTLs in Pure Breed Duroc Pigs

P. Sarup*, S.M. Edwards*, Just Jensen*, Tage Ostersen[†], Mark Henryon[†], P. Sørensen*.
* Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University;
[†]Danish Pig Research Centre

ABSTRACT: Genetic variance for complex traits in animal breeding are often estimated using linear mixed-models that incorporate information from SNP-markers using a realized genomic-relationship matrices. In these models, individual genetic markers are weighted equally and the variation in the genome is treated as a "black box". While this approach has proved useful in selecting animals with high genetic potential, it does not generate insight into the biological mechanisms underlying trait variation. We propose to build a linear mixed model approach to evaluate the collective effects of sets of SNPs in genomic features and open the "black box". Using data on ADG and BG from 6,112 entire Duroc boars and a high-density SNP chip, we show here, that the QTL categories with highest relative importance of the SNP set were indeed biological meaningful.

Keywords: genomic feature models; average daily gain; back fat depth; growth

Introduction

Genetic variance for complex traits such as body growth rate can be estimated by fitting linear mixed-models that that accounts for genetic relationships. Genetic relationships can be calculated from genetic markers and can be used to construct realized genomic relationship matrices. In this approach, the individual genetic markers are weighted equally and the variation in the genome is generally treated as a "black box".

A disadvantage with this "black box" modelling approach is that it does not provide any insight into the biological mechanisms underlying the trait variation. Evidence collected across genome-wide association studies shows that, while many genetic variants with small or moderate effects contribute to genetic variation, it appears that many independently associated variants are located in the same genes and many of these genes are connected via biological pathways (Lango Allen et al. (2010)).

In this paper, we propose to lift the lid on the "black box" and take a peak inside. We present a linear mixed-model approach that evaluates the collective action of sets of SNPs on the trait phenotypes using genomic features (e.g., QTL regions from previous studies or biological pathways). Novel insights into the biological mechanisms causing variation in the traits was generated by identifying genomic features that are causally related to trait variation.

We applied our approach to three growth traits in pure-bred Duroc boars (*Sus scrofa*). The genetic variation was decomposed according to genomic features defined by the QTL categories listed on the Pig QTLdb database (Rothschild et al. (2007)).

Materials and Methods

Animals: We used growth rate phenotypes from 6112 pure-bred Duroc boars. The boars were part of the Danish pig-breeding system (Pig Research Centre, Danish Agriculture and Food Council, Denmark).

Data: The phenotypic records were deregressed proofs for Average Daily Gain from 30kg-50kg body weight (ADG3050), Average Daily Gain from 30kg-100kg (ADG30100) and Back Fat depth (BF) (Ostersen et al. (2011)). The genotypic records were obtained from all phenotyped animals using the Illumina Porcine SNP60 BeadChip (Illumina). The criteria for SNP editing was the same as in Ostersen (2011), with the exception that the minimum minor allele frequency was set to 0.01. This resulted in 33,756 validated SNPs available for our analysis. The SNPs were grouped according to the genomic location of QTLs for 167 trait categories downloaded from the Pig QTLdb database (Rothschild et al. (2007)). The maximum genomic region spanned by any QTL was limited to 2 Mb.

Statistical analyses: An iterative REML approach was used to estimate the genetic variance (Wang et al. (2012)). The approach builds on the following linear mixed-model:

$$\mathbf{y} = \mu + \mathbf{g} + \mathbf{e}$$

where $\mathbf{y}$ the vector of observations, μ is the overall mean, the random genetic effects $\mathbf{g} \sim \mathrm{N}(0, \mathbf{G}\sigma_g^2)$, and residuals $\mathbf{e} \sim \mathrm{N}(0, \mathbf{I}\sigma_e^2)$. The genomic relationship matrix, $\mathbf{G}$, is constructed using all SNP markers as:

$$\mathbf{G} = \mathbf{WDW}'/\mathrm{N},$$

where $\mathbf{W}$ is the centered and scaled genotype matrix, $\mathbf{D}$ is a diagonal matrix containing the weight for each SNP, and N is the sum of the diagonal elements $\mathbf{D}$ The SNP weights were initially set to unity. The genetic values, **g,** and estimates of the variance components, σ_g^2 and σ_e^2, were obtained using the software package, DMU (Madsen & Jensen (2012)). In subsequent iterations, each SNP was weighted according to its variance contribution equal to the squared SNP effect. The individual SNP effects were obtained from:

$$\hat{\mathbf{b}} = \mathbf{DW}'(\mathbf{WDW}')^{-}\hat{\mathbf{g}},$$

where $\hat{\mathbf{b}}$ is the vector of estimated SNP effects. In each iteration the log-likelihood for the fitted model was determined and this iterative procedure was repeated until we observed a decrease in model fit as determined by a decrease in the log-likelihood. During this process the values of $\hat{\mathbf{b}}$ become more extreme and should result in SNPs that are causative, or highly correlated to the causative genetic variant, having a high weight in the model disregarding whether the effect on the trait is positive or negative. We determined a genetic value for each SNP set defined by the genomic feature using:

$$\hat{\mathbf{g}} = \hat{\mathbf{g}}_{\iota} + \hat{\mathbf{g}}_{-i} = \mathbf{W}_{i}\hat{\mathbf{b}} + \mathbf{W}_{-i}\hat{\mathbf{b}}$$

where $\hat{\mathbf{g}}_{i}$ is the genetic value associated to the i'th SNP set and $\hat{\mathbf{g}}_{-i}$ denotes the genetic values associated to the remaining SNPs. From these partitioned genetic values we decomposed the genomic variance using:

$$\mathbf{Var}(\hat{\mathbf{g}}) = \begin{bmatrix} \mathbf{Var}(\hat{\mathbf{g}}_{\mathbf{i}}) & \mathbf{Cov}(\hat{\mathbf{g}}_{\mathbf{i}}, \hat{\mathbf{g}}_{-\mathbf{i}}) \\ \mathbf{Cov}(\hat{\mathbf{g}}_{-\mathbf{i}}, \hat{\mathbf{g}}_{\mathbf{i}}) & \mathbf{Var}(\hat{\mathbf{g}}_{-\mathbf{i}}) \end{bmatrix}$$

We calculated the relative importance of the SNP set as the proportion of genomic variance that could be attributed to either the QTL or SNP set itself or the covariance between $\hat{g}_i$ and $\hat{g}_{-i}$:

$$\boldsymbol{\gamma}(\hat{\mathbf{g}}_{i}) = (\mathbf{Var}(\hat{\mathbf{g}}_{\mathbf{i}}) + \mathbf{Cov}(\hat{\mathbf{g}}_{\mathbf{i}}, \hat{\mathbf{g}}_{-\mathbf{i}}))\mathbf{Var}(\hat{\mathbf{g}})^{-\mathbf{1}}$$

This approach gives us a framework where we can easily decompose the variance contributed by different types of genomic feature classification schemes.

Results and Discussion

For all of the three investigated traits the log-likelihood ratio peaked at 4 iterations. As expected the estimated SNP effects became more extreme with each iteration resulting in a relatively low proportion of SNPs explaining almost all of the genetic variation (Fig.1).

Although the optimal number of iterations was the same for all investigated traits, back fat seemed to have fewer important genes with larger effect than ADG (Fig.2). Even though the genetic correlation between early and late growth has been reported to be low (Hermesch et al. (2000)), the lack of difference between ADG30-50kg and ADG30-100kg in the current study is not surprising as the later encompasses the data of the first.

Table 1 lists the top ten QTL categories for which the SNP set defined by the category had the highest relative importance of the SNP set for ADG30-100kg. Most of these categories are directly linked to growth, the proportion of adipose tissue or the water holding capacity. The balance between lean meat and fat in the carcass composition is one of the major determinants of ADG in pigs (Gjerlaug-Enger et al. (2011)). Although, we see a clear correlation between the set size of QTL markers and the proportion of total genomic variance it explains (Fig 3), the traits highlighted by our method were indeed biological relevant to ADG. However, as the 10 QTL marker sets with highest Υ_{QTL} were also the SNP sets that contained the largest number of SNPs (Fig 3), we were not able to see if the QTL SNP sets explained more than expected from their size in this data set.

We found leukocyte quantity among the top ten QTL categories (Table 1). While it seems logical that immune function is important to growth, as growth is hampered by disease, the relationship between leukocyte quantity and growth is not straight forward. A high leukocyte quantity might reflect either a high resistance towards diseases or a high infection rate. In addition, in a highly controlled production environment with low infection risk diversion of large proportions of the available

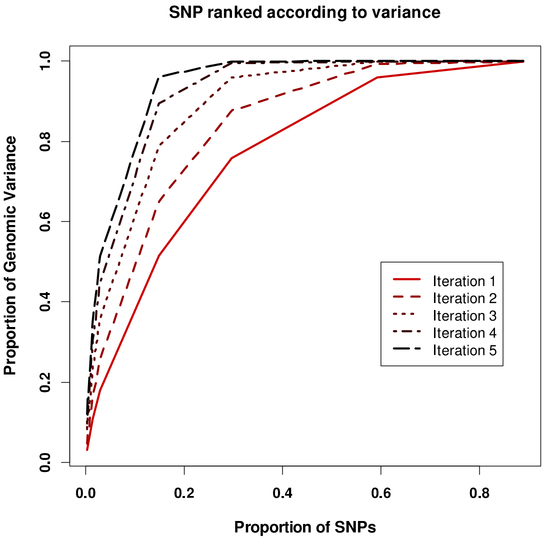

Figure 1. Cumulative proportion of genomic variance in ADG 30-100 kg explained by individual SNPs for each of 5 iterations. In the last iteration 20% of the SNPs explained as much of the genomic variance as 80% of the SNPs in iteration 1.

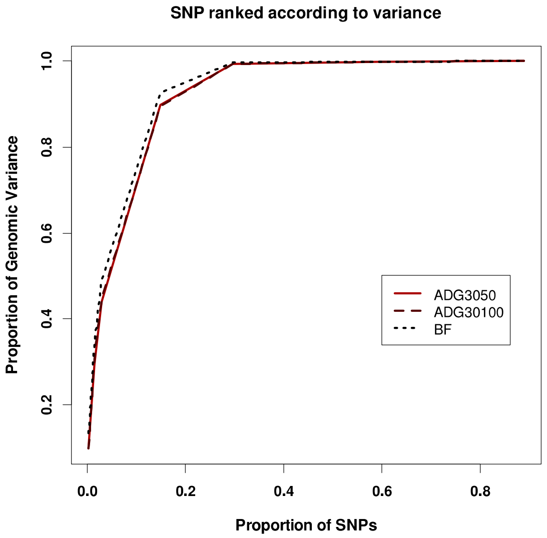

Figure 2. Proportion of genomic variance explained of ADG30-50kg, ADG30-100kg and BF at iteration 4. The curve for ADG30-50kg and ADG30-100kg is practical identical while there is a small tendency for fewer genes with larger effect contributing to the genetic variance of BF.

energy into immune function might not be advantageous.

Regardless of these complications immune functions remains an interesting avenue for research on factors affecting ADG.

Conclusion

We used a linear mixed model approach that evaluated the collective effect of sets of SNPs in genomic features on average daily gain and back fat depth in pigs. The QTL categories with highest relative importance of the SNP set were biological meaningful and directly linked to growth. However as they also were the QTL marker sets with the largest number of SNPs, we cannot conclude that they explained more of the variance than expected from their size.

Acknowledgement

The presented work was done as part of the ECO-FCE project. ECO-FCE is funded by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 311794. In addition, the work was in part funded by Quantomics a collaborative project under the 7th Framework Programme (contract no. KBBE-2A-222664) and the Danish Strategic Research Council (GenSAP: Centre for Genomic Selection in Animals and Plants, contract no. 12-132452).

Literature Cited

Lango Allen, H., Estrada, K., Lettre, G., et al. (2010). *Nature*. 467: 832–838.

Gjerlaug-Enger, E., Kongsro, J. , Ødegård, J., et al. (2011). *Animal*. 6:9–18.

Hermesch, S., Luxford, B. G. and Graser, H.-U. (2000). 65: *Nat.* 249–259.

Madsen, P., and Jensen, J. (2012). DMU. Version 6, release 5.1.

Ostersen, T. , Christensen, O. F., Henryon, M., et al. (2011). *Genet. Sel. Evol.* 43: 38.

Rothschild, M. F., Hu, Z.-L. and Jiang, Z. (2007). *Int. J. Biol. Sci.* 3: 192–197.

Wang, H., Misztal, I., Aguilar, I. et al. (2012). *Genet. Res.* 94: 73–83.

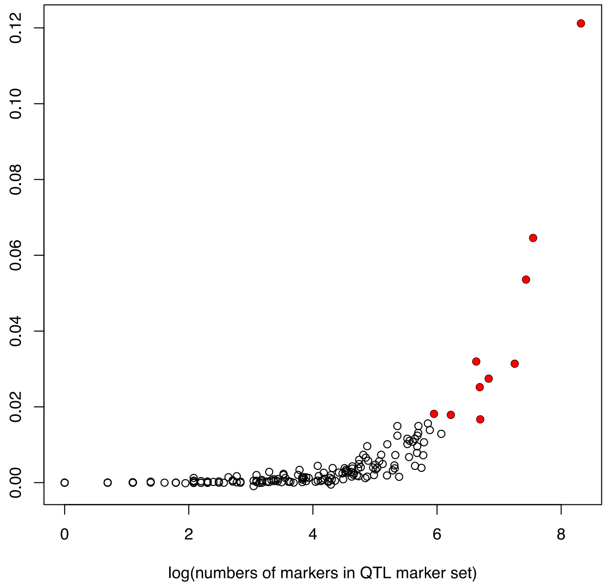

Figure 3. The correlation between the logarithm of the size of the QTL SNP set and the proportion of genomic variation it explained. The ten QTL categories from Table 1 are marked by closed red symbols.

Table 1. The top ten QTL categories for ADG30-100kg, ordered by relative importance of the SNP set (Υ_{QTL}).

QTL trait[1]	Var_{QTL}[2]	Var_{rest}[3]	$Cov_{QTL,rest}$[4]	Υ_{QTL}[5]
subcutaneous adipose thickness	48.40	1209	137.29	0.121
postnatal growth	16.15	1350	82.79	0.065
body mass	13.24	1381	68.86	0.054
hind limb mass	7.24	1441	41.76	0.032
longissimus dorsi muscle area	9.58	1446	38.53	0.031
intramuscular adipose mass	5.39	1453	36.66	0.027
skeletal muscle conductivity	5.17	1460	33.45	0.025
white adipocyte size	4.71	1481	23.13	0.018
leukocyte quantity	3.27	1481	24.14	0.018
nipple quantity	2.84	1484	22.76	0.017

[1] Trait associated with the QTLs used to define the g_{QTL} marker set.

[2] Genetic variation attributed to QTL marker set.

[3] Genetic variation attributed to rest marker set.

[4] Covariance between QTL and rest marker sets.

[5] $\Upsilon_{QTL} = (Var_{QTL} + Cov_{QTL,rest})/(Var_{QTL} + Var_{rest} + 2\ Cov_{QTL,rest})$.