Ranking Gene Sets and Biological Pathways Underlying Production Traits

M.M. Shariati[1], *L. Janss*[1], *A. Skarman*[1] *and P. Sørensen*[1]

Introduction

Genomic selection is used to predict breeding values of selection candidates using dense markers (Meuwissen et al. (2001)). In this approach one estimates an effect that each marker or haplotype has on the trait under study. Despite great success that genomic selection has had in animal breeding practice, the biological mechanisms underlying the trait remains largely unknown. Extension of the genomic models to include more biological information, such as gene pathways, can be interesting to increase the understanding of the selection mechanisms exploited, and to improve robustness, cross-population predictions, and also accuracy of genomic predictions. In this study we describe the inclusion of gene sets based on biological pathways in genomic models. We used gene sets that are associated to biological pathways from KEGG database (Kanehisa et al. (2008)) and analysed protein yield EBVs in Holstein bulls.

Material and methods

Data. Consisted of 1293 Holstein sires with highly reliable (>0.95) EBV for protein production with 50k genotyped data. After editing and quality control, 39264 segregating markers were used to estimate marker effects.

Genomic prediction of marker effects. Two Bayesian models were used; a Bayesian stochastic search variable selection (VS) model (George and McCulloch, (1993)); and a Bayesian model where allele effects have a long-tailed common prior (CP) distribution. These models are similar to the Bayes B and Bayes A models from (Meuwissen et al. (2001)), except that the VS model switches effects from big to (negligibly) small, instead of zero as in Bayes B, and where we implemented extensions to estimate the parameters in the distributions of allele effects from the data. Bayesian analysis was performed using software package iBay (Janss (2009)). Single markers were used in analyses. After preliminary analyses, two prior probabilities for "switched on" markers were set to either 20 (VS model) or 100 percent (CP model) resulting in, respectively, 80 or zero percent of markers were considered to be "switched off", *a priori*. In the former (20 percent prior), Bayesian variable selection model assigns each marker to one of these two groups, *a posteriori*, in each round of Gibbs sampler. In the CP model, there is no assignment to groups but, an effect is estimated for all markers. See Villumsen et al. (2009) for full description of the model. For VS model analysis, the average posterior probability for each marker to belong to "switched on" group was used as input for gene set analysis. When all markers were in the model (CP model), marker effects were used as input for gene set analysis. Samples from posterior

[1] Faculty of Agricultural Sciences, Department of Genetics and Biotechnology, Aarhus University, Blichers Alle 20, 8830 Tjele, Denmark

distribution were generated using a Gibbs sampler with total chain length of 40000 where first 10000 draws discarded as burn in.

Gene set. KEGG database (Kanehisa et al. (2008)) and annotated *Bos Taurus* genome (Carlson et al. (2009)) were used to associate each SNP marker to different pathways. First, based on the starting and ending positions of each gene in database, SNPs that were in that region were considered to be a part of that gene. Then using KEGG database, these SNPs were eventually connected to the pathways. At the end, 2559 SNP markers were covered by 210 KEGG pathways.

Gene set effects. Using SNP solutions from genomic analysis as input, gene set effects were assessed based on a random-set scoring method (Newton et al. (2007)). This method use predefined sets of genes (e.g. SNP) and for each gene set a z-score is calculated. The score describes the degree of association between the SNP solutions of the genes in the gene set and the phenotype of interest.

Results and discussion

Figure 1 shows the estimates of different gene sets in terms of z-score for the two Bayesian genomic models, CP model and VS model, respectively with 100 percent and 20 percent "switched on" markers. Model with all markers failed to rank gene sets in a clear way. In fact it could only rank unimportant gene sets with very low z-scores. The rest of gene sets got kind of null effects. When all markers are forced to get an effect and the number of markers are very high, the effects is distributed among markers even though there are possibly few SNPs with large effects. In statistical sense, there is a strong shrinkage on marker effects. This is not the case in variable selection model where only a fraction of markers are allowed to enter the model. Therefore, markers with larger effects are more likely to be in the model and markers with no effect or very small effects are not effective (George and McCulloch (1993)). As a result, this model performs better in ranking gene sets (Figure 1).

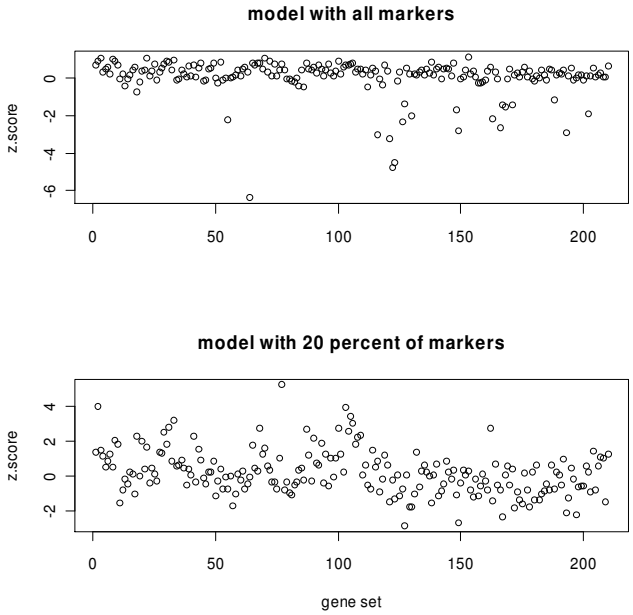

Figure 1: Estimated gene set effects (z.score) for two types of genomic analyses. Top: In genomic analysis all markers were in the model, and button: Only 20 percent of markers were switched on

Table 1 contains the list of top ranked gene sets estimated from Variable Selection analysis. Most of these gene sets are involved in metabolic pathways and it is logical to have an effect on protein production. But it should be noticed that coverage of gene sets for SNP markers we had was not high because we could not associate most of SNPs to annotated *Bos Taurus* genes. It will not be the case in near future with availability of more complete annotated genes and SNP chips.

Gene set effects estimated using genomic estimation of marker effects are peculiar to the population under study. This is because pattern of linkage disequilibrium is different across populations and this linkage disequilibrium is taken into account in genomic prediction. Consequently, marker effects and gene set effects will be population specific. Further, it was shown that gene set effects vary a lot depending on the method for estimating marker effect. Further studies are needed to investigate which methods are better for gene set studies.

Table 1: The first 20 top ranked gene sets and their associated pathways in protein production of Holstein[α]

Gene set	Pathway	Number of SNPs	z.score
720	Reductive carboxylate cycle (CO2 fixation)	8	5.21630

20	Citrate cycle (TCA cycle)	14	3.97178
1061	Biosynthesis of phenylpropanoids	21	3.91508
1063	Biosynthesis of alkaloids; from shikimate pathway	37	3.38581
380	Tryptophan metabolism	21	3.19646
1064	Biosynthesis of alkaloids; from ornithine, lysine	29	3.03213
360	Phenylalanine metabolism	10	2.78147
982	Drug metabolism - cytochrome P450	27	2.72298
4640	Hematopoietic cell lineage	29	2.71397
630	Glyoxylate and dicarboxylate metabolism	14	2.70494
830	Retinol metabolism	25	2.67334
1062	Biosynthesis of terpenoids and steroids	32	2.53018
340	Histidine metabolism	13	2.50153
1070	Biosynthesis of plant hormones	31	2.34414
230	Purine metabolism	128	2.28958
471	D-Glutamine and D-glutamate metabolism	2	2.24434
1066	Biosynthesis of alkaloids; from terpenoid ..	36	2.20693
903	Limonene and pinene degradation	7	2.17549
62	Fatty acid elongation in mitochondria	9	2.04713
240	Pyrimidine metabolism	69	1.96870

[α]KEGG names were used for gene sets and pathways.

Conclusion

High throughput genotyping technologies can be exploited to uncover biological mechanisms underlying production and functional traits. In future more comprehensive and more complete biochemical and annotation databases will be available. Merging all these information provides an exciting opportunity to study influential factors on production and survival potential of farm animals and enables making precise decisions for health problems.

References

Carlson, M., Falcon, S., Pages, H., et al. (2009). http://www.bioconductor.org/packages/devel/data/annotation/html/org.Bt.eg.db.html

George, E.I and McCulloch, R.E. (1993). *J Amer. Stat. Assoc*., 88: 881-889.

Janss, L, (2009) Janss Biostatistics. Leiden. The Netherlands.

Kanehisa, M., Araki, M., Goto, S., *et al*. (2008). *Nucleic Acids Rec.*, 36(Database issue):D480-4.

Meuwissen, T.H., Hayes, B.J., and Goddard, M.E. (2001). *Genetics,* 157: 1819-1829.

Newton, M.A., Quintana, F.A., den Boon, J.A, et al. (2007). *The Annals Appl. Stat*., 1: 85-106.

Villumsen, T.L., Janss, L., and Lund, M.S., (2009). *J. Anim. Breed. Genet*., 126: 3-13.