Analyzing follow-up studies to genome wide association studies

B. Guldbrandtsen[*], J.K. Höglund[†‡], M.S. Lund[*], and G. Sahana[*]

Introduction

The availability of genome wide marker sets has caused a flurry of activity in genome wide association studies, including in large livestock species. However, there is often a pronounced disagreement between studies in the results obtained.

There are multiple possible reasons for this disagreement. Studies are undertaken in different populations detecting different polymorphisms in different breeds, expression of the same polymorphism may differ when observed on different genetic backgrounds. Also, non-detection of QTL may be due to that the original detection represented a false positive. Particularly, the latter possibility leads to a need to confirm detected QTL before they either are included in selection decisions or large sums are invested into gene identification.

When new data are obtained there are two choices: either the analysis of new data are considered an independent replication or they are considered as new information used to fortify the conclusions through an analysis of a combined dataset. However, the latter approach has two problems: often data for a follow-up are obtained from a different population from the original one making interpretation of negative results problematic, and analyzing a combined dataset uses the information in the original dataset twice. The information information in the original data is used both in the original detection of QTL as well as in their subsequent confirmation. This makes interpretation as a confirmation study difficult.

A genome-wide association study for calving traits has been conducted for the Danish and Swedish populations of Holstein cattle (Sahana et al., *in prep*). In this study we want to compare the performance of two approaches to confirming QTL detected in a follow-up study. To this end we obtained substantial new data from the same study population as was analyzed in the previous study. We compare analyzing the combined dataset to only analyzing the new data alone.

Materials and Methods

Phenotypes Phenotypes were estimated breeding values (EBV) for two calving traits. Breeding values were estimated using the models described Danish Cattle Federation (2006) simultaneously incorporating direct and maternal additive genetic effects using a sire model and

[*]Dept. Genetics and Biotechnology, Faculty of Agricultural Sciences, PO. box 50, DK-8830 Tjele Denmark
[†]VikingGenetics, Ebeltoftvej 16, Assentoft, DK-8960 Randers SØ, Denmark
[‡]Sveriges Lantbruksuniversitet, PO Box 7070, 750 07 Uppsala, Sweden

ignoring correlations between traits. Breeding values were estimated based on recordings undertaken as part of the routine Danish recording system.

Traits presented here are EBV for direct genetic effects for calf size (SIZE) and calf survival (SURV) in first parity. Calf size is scored subjectively by the farmer on a categorical scale 1 to 4. Calf survival is scored as a binary variable where 1 represents a calf surviving the first 24 hours. Narrow-sense heritabilities for both traits are $h^2 = 0.04$. SURV and SIZE are strongly genetically correlated ($r = -0.55$). To remove double accounting of information, for this study genetic correlations between traits were set to zero during breeding value estiation.

Datasets. All animals were typed with the bovineSNP50 beadchip (Matukumalli et al., 2009). Results from two datasets were compared with the results of Sahana et al. (2010) (OLD). Only markers with positions in the UMD3 assembly of the taurine genome (Zimin et al., 2009) were included in the comparisons. Markers not shared between the studies were omitted. One was a set of new phenotypes and marker types not included in the previous study (NEW), another was a combined set containing all data (COMB). For OLD, NEW and COMB there were 1064, 1196 and 2265 genotyped animals with breeding values for SIZE, and 1341, 1917 and 3203 for SURV. The numbers for COMB do not equal the sum of the numbers for OLD plus NEW due to breeding values being retrieved at different times. Numbers of markers used were 36389, 38545 and 38545 in the datasets OLD, NEW and COMB.

Statistical analyses. The statistical models used was adapted from Yu et al. (2006). Analyses were done with the model

$$y_i = \mu + bx_i + s_i + e_i, \tag{1}$$

where y_i is the estimated breeding value of individual i for the trait, μ is a shared fixed effect, x_i is a count in individual i of one of the two alleles (with an arbitrary labeling), b is the fixed allele substitution effect, s_i is the fixed effect of the sire of individual i and e_i is a random residual of individual i assumed to a normal distribution with mean zero and unknown variance. Testing was done by a Wald test with a null hypothesis of $H_0 : b = 0$. Significance thresholds were determined using a Bonferroni correction; genome- or chromosome-wide significance thresholds were obtained by dividing the nominal significance threshold of 0.05 by the numbers of SNP included in the analysis. Thresholds of $1.37\,10^{-6}$, $1.30\,10^{-6}$ and $1.30\,10^{-6}$ were obtained for datasets OLD, NEW and COMB.

Results and discussion

The numbers of genome- and chromosome-wide significant SNP are shown in table 1. Clearly the larger COMB dataset allows the detection of far more significant effects than the two smaller datasets. For SIZE NEW confirmed 2 and COMB confirmed 12 out of 23 SNP chromosome-wide significant (CW) in OLD, for SURV NEW confirmed none and COMB confirmed 2 out of 4 SNP. Out of the 11 SNP from OLD not confirmed by COMB, 5 were isolated SNP. Isolated SNP are considered more likely to represent false positive results, suggesting that the failure to confirm reflects false positive detections in the OLD analysis.

Clearly the COMB analysis is far better suited to *identify* significant associations of EBV with SNP than either of the two other analyses are. The COMB analysis is much better at retrieving

Table 1: List of genome-wide significant (GW) and chromosome-wide (CW) markers for SIZE and SURV. The "Unique" column shows the number of SNP found to be associated with the trait in any of the analyses. In parentheses are shown the number of distinct groups ("QTL regions") of significant markers.

		Dataset							
Trait		**OLD**		**NEW**		**COMB**		**Unique**	
SIZE	**GW**	6	(1)	2	(1)	24	(3)	25	(3)
	CW	23		13		97		115	
SURV	**GW**	1	(1)	0	(0)	1	(1)	1	(1)
	CW	4		4		21		27	

the results obtained in the OLD analysis, in part reflecting the fact that the same data are being re-analyzed.

Even when restricting the analysis of the NEW dataset to analyzing only the markers CW significant in the OLD analysis, (only using the Bonferroni correction to correct for a number of tests that is the number of CW significant SNP in the OLD analysis) the NEW analysis only succeeds at retrieving a few of the SNP found in the OLD analysis. Specifically the analysis of the NEW dataset for SIZE failed to retrieve the powerful QTL on BTA18 (see e.g. Cole et al., 2009).

Here we have presented results for three choices of how to analyze a follow-up study. First, one can analyze the new data as an independent new study. Judging by our results clearly unless the new dataset is very large this method under-performs, in this case even failing to detect a well known QTL of large effect. In the present case this approach has the additional complication that the new data are not really independent of the first dataset. Data used for estimating breeding values are shared between the datasets, and animals in the two datasets are related. Second, the new dataset could be analyzed only for whether the new dataset confirms the effects already detected in the previous study. Third, one can analyze the whole dataset as one new large dataset. This approach has problems with interpretation due to double accounting of data. However, this problem seems to be more than offset by the increased power of the combined analysis.

Conclusion

Three approaches to analyzing follow-up data to GWA studies were compared. Analyses involving a follow-up data of the size presented here clearly fail as an approach to confirmation. The advantages of superior power of analyzing a combined dataset outweigh the disadvantages following from double accounting due to re-use of the first dataset in the follow-up analysis.

Acknowledgments. This work was part of "Genomic Selection — from function to efficient utilization in cattle breeding" supported by grant no. 3412-08-02253 from the Danish Directorate for Food, Fisheries and Agri Business, VikingGenetics, Nordic Genetic Evaluation, and Aarhus University, along with support from the Swedish Farmers' Foundation for Agricultural

Research.

References

Cole, J., VanRaden, P., O'Connel, J., Van Tassel, C., Sonstegard, T., Schnabel, R., Taylor, J., and Wiggans, G. (2009). Distribution and location of genetic effects in dairy traits. *J. Dairy Sci.*, 92:2931–2946.

Danish Cattle Federation (2006). Principles of danish cattle breeding. Technical Report Eigth edn., The Danish Agricultural Advisory Centre, Udkærsvej 15, Skejby, DK-8200 Aarhus N. http://www.landbrugsinfo.dk/KVAEG/AVL/Sider/principles.pdf.

Matukumalli, L., Lawley, C., et al. (2009). Development and characterization of a high density snp genotyping assay in cattle. *PLoS One*, 4(4):e5350.

Sahana, G., Guldbrandtsen, B., and Lund, M. ([2010]). Genome-wide association study for calving traits in danish and swedish holstein cattle. *in prep.*

Yu, J., Pressoir, G., Briggs, W., Bi, I., Yamasaki, M., Doebley, J., McMullen, M., Gaut, B., Nielsen, D., Holland, J., Kresovich, S., and Buckler, E. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38:203–208.

Zimin, A., Delcher, A., Florea, L., et al. (2009). A whole-genome assembly of the domestic cow, *bos taurus*. *Genome Biology*, 10:R42.