Scientific report for exchange grant
"Genome wide association study for functional traits in cattle"
Reference Number 4122

ESF activity:	LESC; Advances in Farm Animal Genomic Resources; Genomic Resources
Project title:	Genome wide association study for functional traits in cattle
Guest researcher:	Gábor Mészáros, post doc researcher at University of Natural Resources and Life Sciences, Vienna, Austria
Host researcher:	prof. Mattias Villani, Dr. Patrik Waldmann, Department of Computer and Information Science, Linköping University, Sweden
Time interval:	15.04.2013 – 03.05.2013 (3 weeks)

Contents

Purpose of the visit

The main objective of the stay was to carry out a genome-wide association studies (GWAS) for a range of functional traits in cattle. GWAS looks for associations between genotypes of the SNPs and some trait of interest. Most of these traits are continuous they are likely to be controlled by many loci of small effects, or a mixture of a few genes with large effects and many genes of small effects. It is challenging to perform analysis of GWAS data because the number of SNPs (p) is much larger than the sample size (n), commonly referred to as the "small n, large p" problem. A major difficulty in this problem is that the number and extent of spurious associations between predictors and response increase rapidly with increasing p.

There are several methodologies to perform genome wide associations, single SNP associations being the most common. In this case each SNP is considered as a single effect in a linear model, running as many times as many SNPs are considered in the analysis. Methodologies, such as lasso, ridge regression and elastic net with different penalty factors are simultaneously accounting for the effects of SNPs. Another possibility is to implement Bayesian variable selection, such as Bayesian lasso to find important regions for the traits of interest. Resulting effects for any of these methodologies might be caused by real association to the trait, or just be a product of "population structure" in the data set caused by e.g. differences in breeds, admixture levels or differences in ancestry for the analyzed data set. For this reason the implementation of a correction for population structure is needed to remove false positive results. Likewise, careful steps need to be taken not to delete associated markers, and so to avoid false negative results.

Description of the work carried out during the visit

The original plan was to use the 50k Illumina SNP chip genotypes for Pinzgau and Tyrol Grey cattle (~220 genotypes each) and the 50k genotype data for ~2.000 Austrian Fleckvieh bulls available for the "Genome wide association study for functional longevity and related traits in dairy cattle" project. After submission of the ESF application we got the permission to use 50k genotypes from the joint German-Austrian (DEA) genetic evaluation for ~6.000 Fleckvieh bulls. As the estimated breeding values, deregressed breeding values for all traits and all reliabilities were available for this big set, we decided to use the DEA set to perform genome wide analysis on a range of traits.

The PLINK (Purcell et al. 2007) software was used for quality control. Only SNPs that could be unambiguously mapped following the paper of Fadista and Bendixen (2012) were kept for the analysis. Also SNPs at sex chromosomes were removed. The R software (R Core Team, 2012) was used to perform the analyses and visualize the results. Due to the large number of genotypes involved the handling of the data and the actual analysis was challenging, especially when it involved correction for population structure. For this purpose the eigenvector decomposition with GEMtools R package was used. The computed eigenvectors were then fitted in the model together with the SNP effects. The computations were carried out using the Vienna Scientific Cluster. Unfortunately even this huge server allowed running only one or two analyses at the same time, depending on the methodology, supposedly due to memory limitations.

The paper "Evaluation of the lasso and elastic net in genome-wide association studies" previously submitted to Journal of Animal Breeding and Genetics was reviewed and we decided to improve it before resubmission. The GWAS on a real data set will be changed from longevity to fat content, to show the features of different methodologies. This however needs a re-analysis of each model type with and without population structure correction, which is currently under way.

For the "Genome wide association study for functional longevity and related traits in dairy cattle" project, deregressed breeding values for milk production, fat content, longevity, fertility, calving ease (maternal and direct), stillbirth rate (maternal and direct) and somatic cell count were analyzed using a single SNP association with and without population structure correction. The effect of using different numbers of eigenvectors for population structure correction was also explored.

Description of the main results obtained

The significance values from each model were extracted and transformed to a negative logarithm, so the higher values denote higher significance level. All –log(p) values were plotted, distinguishing the chromosomes using different colors.

The number of eigenvectors used for population structure correction was studied in subsequent runs, when deregressed breeding values for fat content were used as phenotypes. For this trait there is a huge signal on chromosome 14, presumably DGAT, with –log(p) values up to 150. To show the changes in smaller peaks we limited the y axis to 40.

Figures 1-3 show the –log(p) for no population structure correction, 15 and 117 eigenvectors used. The three scenarios intended to compare situations with no, low and high number of eigenvectors in the model. The 15 eigenvectors were chosen based on numbers found in

the literature (Hao et al. 2010). The 117 determined by the GEMtools package as the number of significant dimensions for the ~6.000 genotypes.

Some of the peaks visible with 15 eigenvectors in the model vanish when using the high number of eigenvalues for population structure correction. Notable examples are on chromosome 5 and 11, smaller previously significant SNPs on multiple other chromosomes. From these results it is not clear if the approach gets rid of false positives or some of the SNPs become false negative when using many eigenvectors. A more detailed look into this problem is needed.

Figure 1 Single SNP analysis without correction for population structure

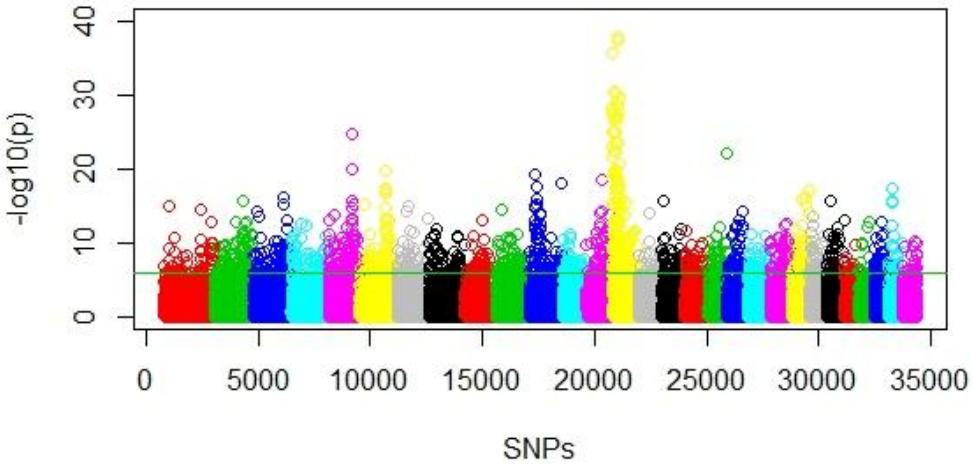


Figure 2 Single SNP analysis using 15 eigenvectors for population structure correction

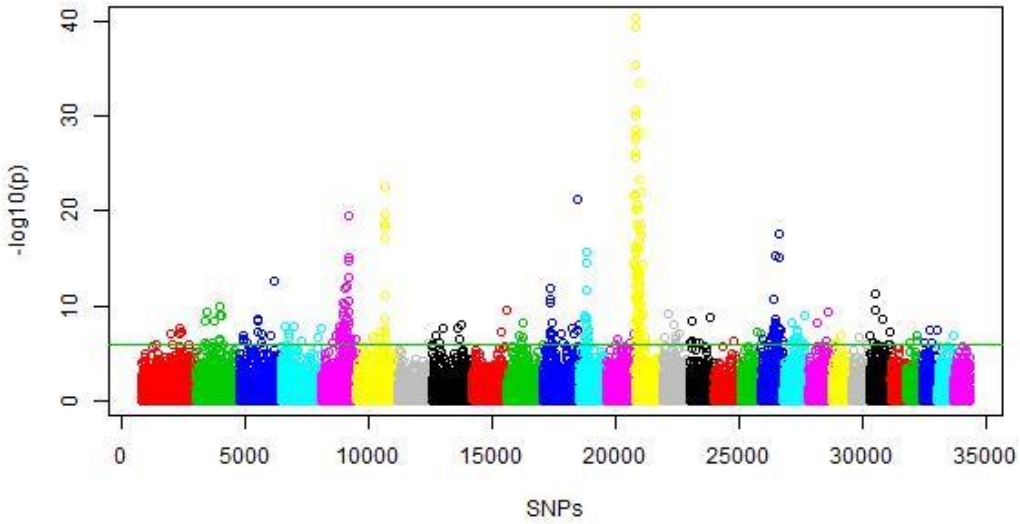
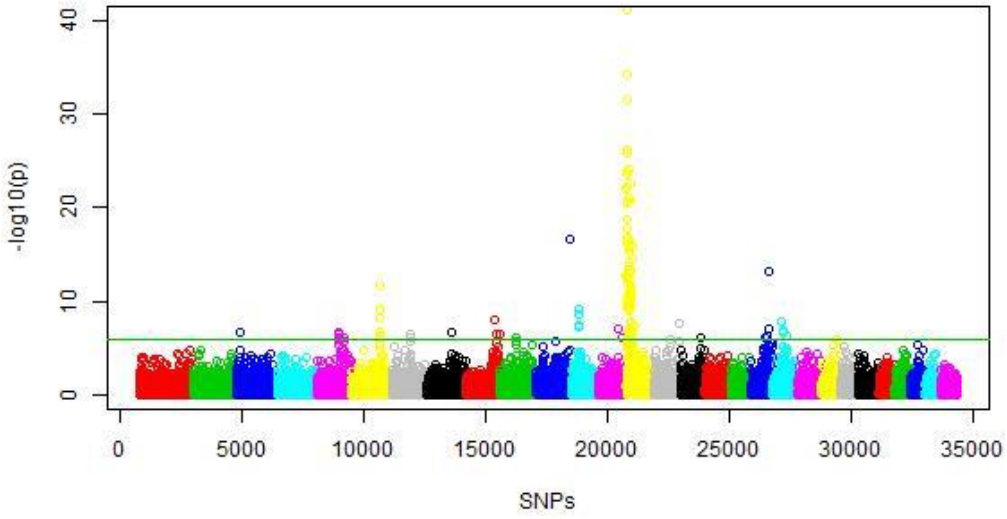


Figure 3 Single SNP analysis using 117 eigenvectors for population structure correction



In addition to the fat content presented above, single SNP analyses with and without population structure correction were conducted during the stay. Interestingly, for most of the traits there were SNPs above the Bonferroni line, which is a very conservative form of significance testing. Only some of the results are shown in this document, but based on the results obtained during the research stay a wide range of additional, more detailed analyses will be done.

The GWAS results for longevity are shown in figure 4. Previous analyses with a smaller data set of ~2.000 Fleckvieh bulls did not show any significant SNPs. With this data set we obtained several significant SNPs indicating signals on chromosomes 13, 14 and 19.

Results from single SNP analysis for fertility after population structure correction are shown in figure 5. Also here we see some interesting SNPs, but these are much less clear, not supported by many neighboring SNPs as in the previous cases.

For direct stillbirth rate there are two very clear signals on chromosomes 14 and 21, and several additional significant ones on chromosomes 5, 6, 10 and 17. The graph is very similar for direct calving ease (not shown), with the major peaks at the same chromosomes. The results confirm analyses published by Pausch et al. (2011), using a smaller set of Fleckvieh bulls.

In general, the obtained results provide exciting research opportunities for the future.

One of the first steps will be to check the false positive rate using the qvalue R function, which identifies the important SNPs and corrects for the false discovery rate. A trial run for one of the traits was conducted during the stay. Depending on the significance threshold, different numbers of significant SNPs and false positives can be expected.

Figure 4 Single SNP analysis for longevity with population structure correction

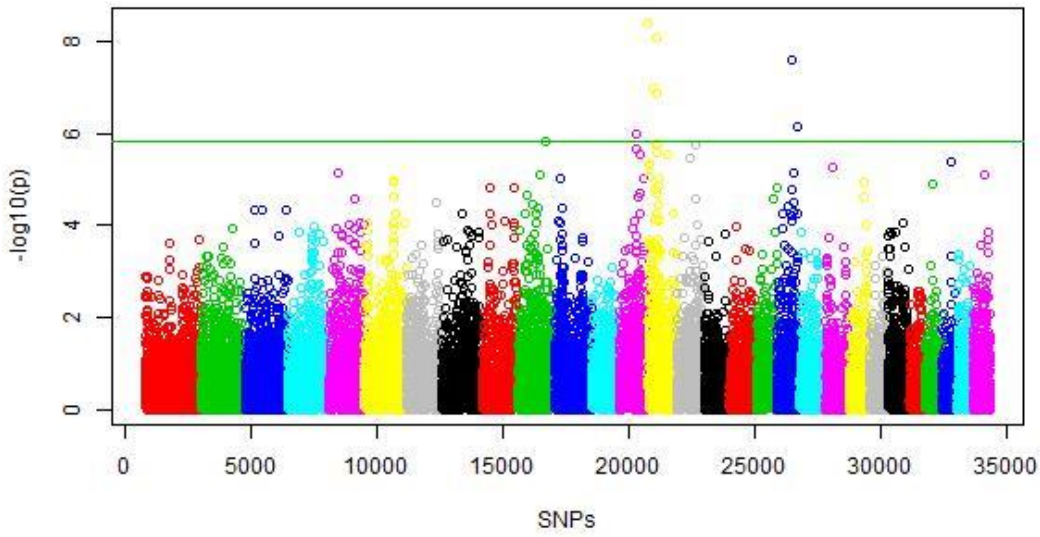


Figure 5 Single SNP analysis for fertility with population structure correction

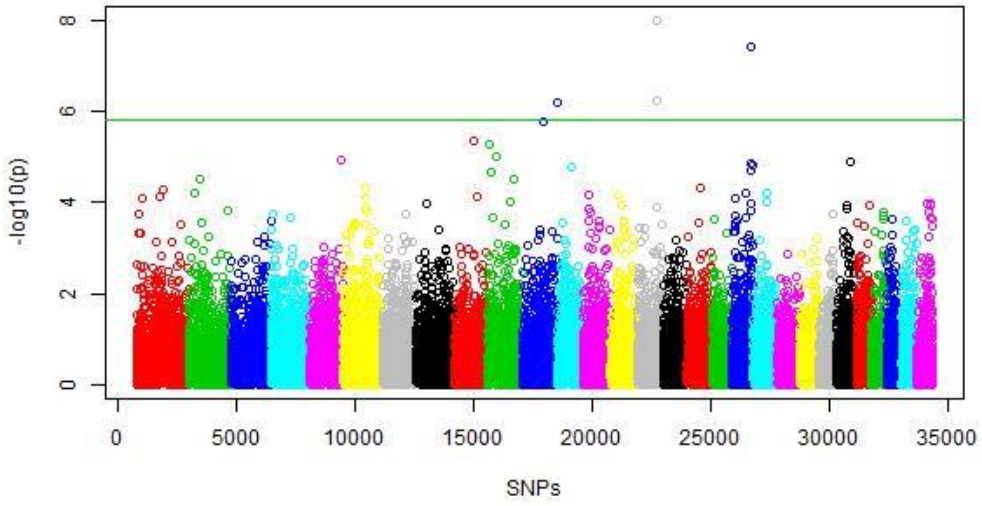
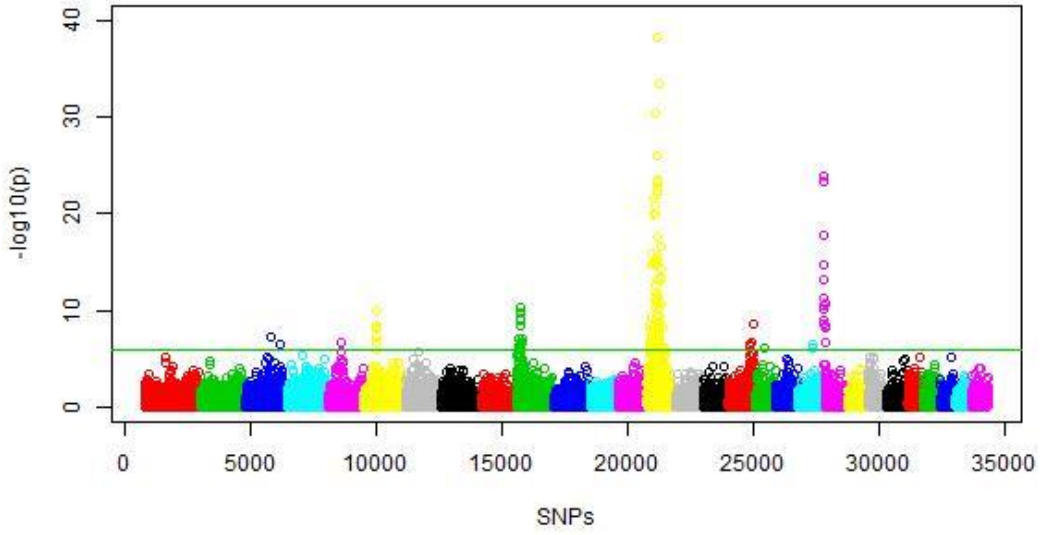


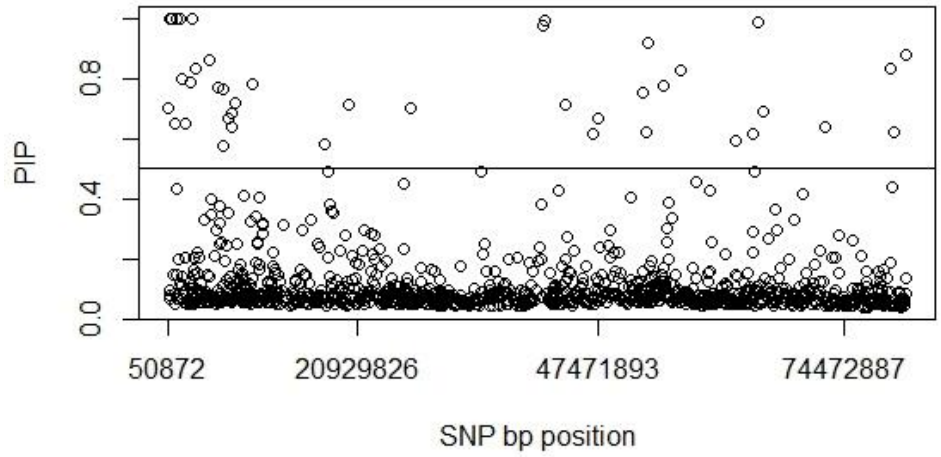
Figure 6 Single SNP analysis for longevity with population structure correction



The results from lasso, ridge regression and elastic net with diverse penalty factors are under way. In general the ridge regression is selecting the most SNPs, lasso the lowest number. The penalty factor in elastic net can be between 0 and 1 - the “proportion” of lasso in the analysis. If the penalty factor is 0 then the analysis is a ridge regression, if 1 the analysis is equal to a standard lasso. With increasing value of the penalty factor the number of selected SNPs is decreasing in our study, as expected.

We have also worked with the Bayesian lasso during the stay, using the bLASSO (Hans, 2010) R package. The run on the full data set was not possible due to computational constrains. Even with only 2.000 iterations and 1.000 burn-in steps the computation took about 3 days on the Vienna Scientific Cluster. After various convergence testing methodologies using the CODA R package we found that the number of iterations should be increased. Because of this we decided to go for a chromosome wise analysis using the bLASSO, which is a more feasible option even with many more iterations. The significance threshold for the results is the 0.5 posterior inclusion probability (PIP). The results using 20.000 iterations on chromosome 14 are shown in figure 7. Although the number of iterations was much higher in this case, the follow up tests showed that some of the SNPs still did not converge. Additional increase and fine tuning of the parameters is needed.

Figure 7 Posterior inclusion probability (PIP) for SNPs on chromosome 14 using fat content as phenotype



Future collaboration with host institution

The collaboration between the two groups at University of Natural Resources and Life Sciences, Vienna and Linköping University was non-existent before this project. This stay however opened new opportunities for the two groups to jointly work on statistical analysis of genotype data. A visit of Dr. Waldmann to BOKU is planned for summer 2013, with continued work on GWAS in cattle.

Projected publications/articles resulting or to result from your grant

An oral presentation of the results will be given (already accepted) at the 64th Annual Meeting of the European Federation of Animal Science in Nantes, dealing with diverse methodologies in GWAS. The paper “Evaluation of the lasso and elastic net in genome-wide association studies” will be resubmitted to a peer reviewed journal, currently we consider Frontiers in Genetics, upon completion of the results. One more paper is under preparation on genome wide association in functional traits in cattle.