How to Use Fewer Markers in Admixture Studies?

Anamarija FRKONJA [1(✉)]
Birgit GREDLER [2]
Urs SCHNYDER [2]
Ino CURIK [3]
Johann SÖLKNER [1]

Summary

Swiss Fleckvieh has been established from 1970 as a composite of Simmental and Red Holstein Friesian cattle. Breed composition is currently reported based on pedigree information. Information on ancestry informative molecular markers potentially provides more accurate information.

For the analysis Illumina Bovine SNP50 Beadchip data for 495 bulls were used. Markers were selected based on difference in allele frequencies in the pure populations, using F_{ST} as an indicator. Performance of sets with decreasing number of markers was compared. The scope of the study was to see how much we can reduce the number of markers based on F_{ST} to get a reliability that is close to that with the full set of markers. On these sets of markers hidden Markov models (HMM) and methods used in genomic selection (BayesB, partial least squares regression, LASSO variable selection) were applied.

Correlations of admixture levels were estimated and compared with admixture levels based on pedigree information. F_{ST} chosen SNP gave very high correlations with pedigree based admixture. Only when using 96 and 48 SNP with the highest F_{ST}, correlations dropped to 0.92 and 0.90, respectively.

Key words

admixture, Swiss Fleckvieh, breed composition, AIM

[1] University of Natural Resources and Life Sciences Vienna, Department of Sustainable Agricultural Systems, Division of Livestock Sciences,
Gregor Mendel Str. 33, A-1180 Vienna, Austria
✉ e-mail: anamarija.frkonja@boku.ac.at
[2] Qualitas AG, Chamerstrasse 56, Ch-6300 Zug, Switzerland
[3] University of Zagreb, Faculty of Agriculture, Department of Livestock Sciences,
Svetosimunska 25, 10000 Zagreb, Croatia

Received: May 11, 2011 | Accepted: July 11, 2011

ACKNOWLEDGEMENTS
We gratefully acknowledge the generous assistance of Gabor Meszaros. We would like to thank the Swissherdbook cooperative Zollikofen for providing genotypes for analysis.

Introduction

Several analyses of individual admixture levels for cattle populations have been presented. Bray et al. (2009) investigated the ancestral components of Devon and Kerry cattle in the Dexter breed, using different ML approaches. Gorbach et al. (2010) analysed the genetic make-up of Kenyan dairy cattle with STRUCTURE software (Pritchard et al., 2001) employing HMM. Animal geneticists have developed a host of procedures for predicting genetic merit of animals for individual traits from large number of markers (Wu et al., 2010; Meuwissen, 2009). Information from a reference population of animals with accurate breeding values is used to predict the genetic merit of a test population of animals for which such accurate phenotypic information is not available. For two-breed crosses, these procedures can be used to predict proportion of genes of one breed in crossbred cattle when sets of genotypes from purebred animals are available as reference. We employed some of these methods for predicting levels of admixture of Swiss Fleckvieh cattle, a breed with Simmental (SI) and Red Holstein Friesian (RHF) being founder populations, using genotypes from the Illumina 50k SNP bovine bead chip (Illumina, 2009). Results of these methods were compared to that of the HMM approach implemented in STRUCTURE, taking pedigree breed composition as reference. In this study analyses were performed on small numbers of ancestry informative markers (Xu and Jin, 2008).

For extracting important markers for admixture analysis, various indicators have been employed. Michael et al. (2004) extracted SNP for admixture mapping based on differences in allele frequenies. Shannon information content served as an indicator for extracting important markers in the study of Alkes et al. (2007) while Paschou et al. (2010) used principal components analysis for the same purpose. Xu et al. (2008) extracted ancestry informative markers (AIM) according to $F_{ST,}$ a measure of genetic differentiation between the pure populations (Weir and Hill, 2002).

Data and methods

Swiss Fleckvieh was established in 1970 as a composite of Simmental and Red Holstein Friesian cattle, with the aim of substantially increasing milk production while keeping dual purpose characteristics of the Simmental breed. The formal definition of the Swiss Fleckvieh population has changed over time and currently includes animals with a pedigree based breed composition involving 1/8 to 7/8 Red Holstein Friesian (RHF) "blood". Animals <1/8 RHF are in the Simmental section of the herd book, animals >7/8 RHF are called Red Holstein Friesian. For analysis, 100 pure Red Holstein Friesian according to pedigree, 100 pure Simmental and 305 admixed animals were selected. We did not respect the range of breed proportions for the Swiss Fleckvieh breed but included animals along the range of pedigree composition of 0.02-0.99 RHF. The average admixture level was 0.716 (standard deviation 0.339). After pruning and quality control, applying Plink 1.07 (Purcell et al., 2007), 40492 SNP and 495 animals were used for further analysis.

Subsets of ancestry informative markers (AIM) were chosen according to F_{ST}. SAS/genetics 9.2 (proc ALLELE) was used to calculate F_{ST} for every SNP, based on variance in allele frequencies (Weir and Cockerham, 1984). Average F_{ST} from all markers was 0.11 (min -0.011, max 0.783). Subsets of SNP with F_{ST} higher than 0.25, 0.30, 0.35, …, 0.65, 0.70 and 96 and 48 SNP with the highest F_{ST} values were extracted. From the study Frkonja et al. (submitted) we compared our results with full set of markers. The methods employed to predict breed composition were HMM using STRUCTURE and ADMIXTURE software (Alexander et al., 2010), and three procedures frequently used in genomic selection: partial least squares regression (PLSR), a Bayesian approach called BayesB and the least absolute shrinkage and selection operator (LASSO).

STRUCTURE is using a model-based clustering algorithm to infer population structure using genotype data. We employed the admixture model using a burn-in period of 10000 and 10000 Markov Chain Monte Carlo (MCMC) repeats and considering frequencies of SNP correlated. To make comparison with pedigree possible, two genetic clusters were chosen. Checks with higher numbers of clusters using the approach of STRUCTURE confirmed that two clusters were the best choice indeed with the data at hand (Pritchard et al., 2010).

PLSR, originally developed by Wold (1966), is trying to minimize the sample response prediction error by seeking linear functions of the predictors that explain as much variation in each response as possible (proc PLS, SAS 2009). For applications in genomic selection, see Colombani et al. (2010). We employed SAS software (proc PLS, SAS 2009) and have used internal cross validation to improve predictive capacity. BayesB (Meuwissen et al., 2001) applies a Bayesian mixture model, which assumes that only part of the SNP involved provide information about the phenotype. Marker effects and resulting phenotype predictors were estimated using software bayesgg, kindly provided by T. Meuwissen. The user needs to provide information about the proportion of SNP with considerable effect, in our case for distinguishing breeds. Values given here were: 0.01, 0.1, 0.2, 0.3, 0.4, and 0.5. Results were similar; those giving the highest correlations for particular sets are presented here. LASSO is a very efficient variable selection method that adds and deletes parameters (regression coefficients) based on ordinary least squares. SAS 9.2 (proc GLMSELECT) was employed, choosing 96 and 48 SNP.

Predictions of individual breed composition based on all methods and data sets were compared by correlating them with values of pedigree admixture. Statistical testing of differences of correlation coefficients was done by Fishers Z-transformation, p-values <0.01 were considered significant; no correction for multiple testing was performed.

Results and discussion

Subsets of AIM selected based on F_{ST} values from the two samples of purebred animals (Table 1), correlations were similar for STRUCTURE, PLSR, and BayesB, and lower for LASSO except for the situation with F_{ST}>0.65, F_{ST}>0.70, 96 and 48 SNP, where actually no variable selection was performed any more with LASSO and LASSO results are those of multiple regression on those SNP.

Figure 1 is showing distribution of the highest 100 F_{ST} AIM on the chromosome 29. It indicates that such SNP are situated across the chromosome with some clustering. Please note that

Table 1. Pearson correlations among pedigree based admixture and different methods and subsets of SNP used (chosen according to F_{ST} value)

Number of SNP	STRUCTURE	PLSR	BayesB (1%)	LASSO 96
Full set of SNP 40492	0.972	0.976	0.974	0.934
5635 (F_{ST} > 0.25)	0.971	0.976	0.966	0.934
3904 (F_{ST} > 0.30)	0.971	0.974	0.965	0.934
2620 (F_{ST} > 0.35)	0.969	0.974	0.952	0.934
1677 (F_{ST} > 0.40)	0.968	0.973	0.949	0.934
1028 (F_{ST} > 0.45)	0.966	0.968	0.967	0.934
594 (F_{ST} > 0.50)	0.961	0.957	0.955	0.934
316 (F_{ST} > 0.55)	0.953	0.943	0.937	0.921
135 (F_{ST} > 0.60)	0.940	0.923	0.931	0.912
49 (F_{ST} > 0.65)	0.908	0.904	0.907	0.903
14 (F_{ST} > 0.70)	0.810	0.802	0.802	0.800
96 (F_{ST} > 0.623)	0.924	0.916	0.926	0.918
48 (F_{ST} > 0.651)	0.907	0.903	0.906	0.903

96 SNP are these with highest F_{ST}, 48 is the subset of 96 with the highest values of F_{ST}

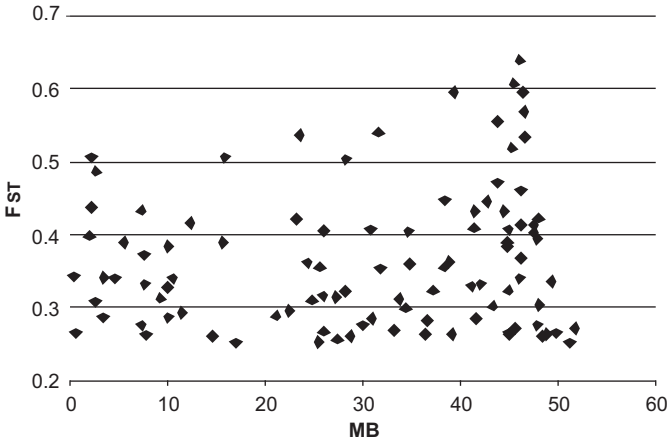

Figure 1.
The highest 100 F_{ST} SNP markers on the chromosome 29

we have not pruned SNP in extreme LD for our analyses. Figure 2 provides a graphical representation of predicted admixture levels for the case with 96 ancestry informative SNP.

Average genome wide F_{ST}, indicating the differentiation of Simmental and Red Holstein Friesian breeds was 0.11. This is larger than the 0.07 for Holstein Friesian and Angus (MacEachern et al., 2009) and close to the average of 0.12 among 10 taurine breeds reported by Chan et al. (2010). The pedigree of Swiss Fleckvieh traced the ancestry of animals until their pure Red Friesian or Simmental ancestors. The reliability of these pedigrees is very good; parentage tests have been obligatory for all male breeding animals since the inception of the breeding program in 1970. We have used pedigree admixture as a reference while we are aware that identity by descent calculations would potentially be more accurate. SNP data was not available for the founder individuals of the admixed population. We have selected ancestry informative markers based on $F_{ST.}$ Using 1028 SNP with F_{ST}>0.45 resulted in very similar correlations with pedigree admixture as using all SNP, using 594 SNP (F_{ST}>0.50) resulted in marginally (0.01-0.02 units) lower correlations. Estimating admixture based on 96 or 48 SNP with the highest F_{ST} resulted in substantially lower correlations (0.90-0.93). Our results indicate that there is considerable loss of information in predicting admixture when going below 1000 ancestry informative markers for our recently admixed cattle population. Rapid and cheap prediction of breed composition in cattle breeds will be worthwhile in case of incomplete pedigrees and the search for the best type of cross or composite

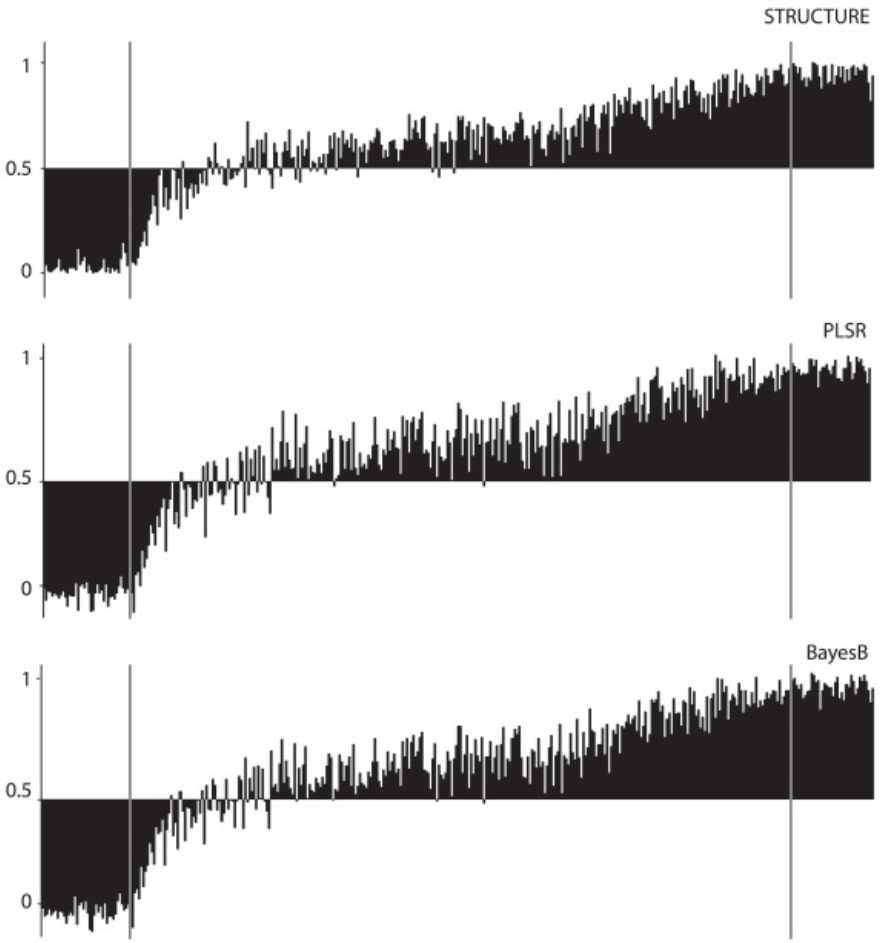

Figure 2.
Admixture predicted with STRUCTURE, PLSR and BayesB using 96 highest F_{ST} SNP-s. Animals are ordered according to pedigree admixture (proportion of RHF), the group before the first vertical bar represents pure Simmental, the group after the second vertical bar are pure Red Holstein Friesian.

of breeds. Given the low price for SNP chips featuring ~3000 markers (3k chips), trying to go for a solution with a very small number of SNP (e.g. 96) does not seem necessary while the information from such a chip gives similar results for admixture levels as a chip with much higher numbers of SNP (50k chip or high density chips featuring >500k SNP).

References

Bray T.C., Chikhi L., Sheppy A.J., Bruford M.W. (2009) The population genetic effects of ancestry and admixture in a subdivided cattle breed. Animal Genetics 40, 393-400.

Chan E.K.F., Nagaraj S.H. & Reverter A. (2010) The evolution of tropical adaptation: comparing taurine and zebu cattle. Animal Genetics 41, 467-77.

Colombani C., Legarra A., Croiseau P., Guillame F., Fritz S., Ducrocq V., Robert-Granie C. (2010) Application of PLS And Sparse PLS Regression In Genomic Selection. Proc. 9th WCGALP, 2010.

Frkonja A., Gredler B., Schnyder U., Curik I., Sölkner J. (2011) Prediction of breed composition in an admixed cattle population. "submitted"

Gorbach D.M., Makgahlela M.L., Reecy J.M., Kemp S.J., Baltenweck I., Ouma R. , Mwai O., Marshall K., Murdoch B., Moore S., Rothschild M.F. (2010) Use of SNP genotyping to determine pedigree and breed composition of dairy cattle in Kenya. Journal of Animal Breeding and Genetics 127, 348-351.

Illumina (2009) Bovine SNP50 Genotyping BeadChip Available at http://www.illumina.com/documents/products/datasheets/datasheet_bovine_snp5O.pdf (last accessed 19 December 2010).

MacEachern S., Hayes B., McEwan J. & Goddard M. (2009) An examination of positive selection and changing effective population size in Angus and Holstein cattle populations (Bos taurus) using a high density SNP genotyping platform and the contribution of ancient polymorphism to genomic diversity in Domestic cattle. BMC Genomics 10, 181.

Meuwissen T. H. E., Hayes B. J. , Goddard M. E. (2001) Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics 157, 1819–1829.

Meuwissen T. H. E. (2009) Accurancy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. Genetics Selection Evolution 41, 35

Nassir R., Kosoy R., Tian C., White P.A., Butler L.M., Silva G., Kittles R., Alarcon-Riquelme M.E., Gregersen P.K., Belmont J.W., De La Vega F.M. & Seldin M.F. (2009) An ancestry informative marker set for determining continental origin: Validation and extension using human genome diversity panels. BMC Genetics 10.

Pritchard J. K., Wen X., Falush D. (2001) Documentation for Structure software, Version 2.3 Available at http://pritch.bsd.uchicago.edu/structure_software/release_versions/v2.3.3/structure_doc.pdf (last accessed 14 January 2011).

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics, 81.

SAS Institute Inc. (2009) SAS/STAT® User's Guide, Version 9.2. Cary, NC

Weir, B.S., and Hill,W.G. (2002) Estimating F-statistics. Annual Review of Genetics 36, 721–750.

Weir B. S., Cockerham C. C. (1984) Estimating F-statistics for the analysis of population structure. Evolution 38, 1358-1370.

Wold H., (1966) Estimation of principal components and related models by iterative least squares. In Multivariate analysis, Academic Press. 391-420.

Wu X., Heringstand B., Gianola D. (2010) Bayesian structural equotation models for inferring relationships between phenotypes: a review of methodology, identifiability, and applications. Journal of Animal Breeding and Genetics 127, 3-15.

Xu S., Jin L. (2008) A Genome-wide Analysis of Admixture in Uyghurs and a High-Density Admixture Map for Disease-Gene Discovery. The American Journal of Human genetics 83, 322-336.

acs76_32