ON BREED COMPOSITION ESTIMATION OF CROSS-BRED ANIMALS USING NON-LINEAR OPTIMISATION

Vinzent Boerner

Animal Genetics and Breeding Unit, University of New England Armidale, 2351, NSW, Australia

SUMMARY

Genetically admixed animals are common in most quantitative genetic analysis, and usually are a result of intended crosses between two or more pure breed populations to enhance productivity. Disregarding the genetic heterogeneous architecture of admixed individuals may lead to poor or even wrong inference about the quality, quantity and genome location of genetic factors affecting phenotypes, and it could reduce the accuracy of estimates of genetic merit. In this article a non-linear optimisation approach (constrained genomic regression, CGR) is presented to describe the marker genotype of a focus animal as a linear function of marker allele frequencies of possible populations of origin. The algorithm was tested on a beef cattle data set consisting of 11639 animals from 11 different breeds with marker genotypes of 4022 single nucleotide polymorphisms, which were used to generate 5000 artificially cross-bred animals. For comparison the data set was also analysed with the ADMIXTURE software (ADM). CGR outperformed ADM with a maximum difference between the true and estimated breed proportion of 0.25 and 0.28 for the 5 and 25 cross-over data set respectively. For ADM this parameter was 0.83 and 0.66. The mean squared estimation error was 15 and 5 times larger for ADM compared to CGR for the 5 and 25 cross-over data set respectively. In addition, CGR always outperformed ADM in terms of speed by factor 20.

INTRODUCTION

The quantification of pure breed proportions of cross-bred animals' genomes is of relevance for genome wide association studies, estimation of population parameters, breeding value estimation and cross-breeding program optimisation. The most widely used methodology for marker based breed proportion estimation is likelihood formulation of the animals' genotype probability conditional on the pure breed population allele frequencies, where the latter are estimated in turn from the animals' genotypes and the assigned breed proportion (Pritchard et al. 2000). The whole system is evaluated using Gibbs Sampling (Pritchard et al. 2000; Raj et al. 2014), expectation maximisation (Tang et al. 2005), or, as a sped-up version, a block relaxation algorithm (Alexander et al. 2009). Since often the allele frequencies of pure breed populations can be estimated from animals of known pure breed origin, Alexander et al. (2009) shortcut their method to facilitate quicker breed proportion estimation for cross bred animals. However, the likelihood based method has two major shortcomings: a) the likelihood formulation assumes the absence of linkage disequilibrium between markers and orthogonality of pure breed population allele frequency vectors, and b) processing time becomes an obstacle if there are many marker genotypes (e.g. 700k or full genome sequences). This article describes a non-linear optimisation method (constrained genomic regression, CGR) for the estimation of pure breed proportions of cross-bred animals' maker genotypes, which overcomes both the limitations of the likelihood based method and allows a meaningful interpretation of the results even if the number of possible pure breeds is huge (see Chiang et al. 2010; Kuehn et al. 2011, for an unconstrained version of this approach). The algorithm was applied to 4k single nucleotide polymorphisms (SNP) genotypes of 5000 cross-bred animals artificially generated from real genotypes of 11639 animals from 11 different breeds. Result were compared to results from ADMIXTURE (ADM) (Alexander et al. 2009).

METHODS

Model. The problem to solve can be written as $argmin_b f(b) = y'y - 2y'Xb + b'X'Xb(1)$ subject to $b_i \geqslant 0\{i = 1, .., N\}(2)$ and $\sum b_i = 1(3)$where y is the marker genotype vector of the cross-bred animal, X is a column matrix of pure breed population allele frequency vectors, and N is the number of pure breeds. Note that equations (2) and (3) comprise constraints to the solutions of equation (1). Values in vector b are regression coefficients regressing y on the columns of X. Minimising equation 1 with respect to equation 2 and 3 will yield a vector b of which values will not only explain the genotype in y as a linear function of population allele frequencies in X, coefficients also have the straight forward interpretation of what proportion of Xb is explained by each column in X.

Data. The cattle data set consisted of 11639 animals from 11 different cattle breeds (Brahman (1492), Angus (1473), Murray Grey (316), Limousin (1395), Charolais (899), Hereford (1500), Simmental (337), Shorthorn (1126), Wagyu (1497), Santa Gertrudis (1474) and Drought Master (130)). Since genotypes of these animals were from various SNP panels, the 4022 SNP were selected which all panels had in common. The SNP genotypes were randomly phased to obtain haplotypes. Cross-bred animals were generated over five rounds. In round one the sex was randomly assigned to the 11639 pure-bred animals and 1000 males and 1000 females were randomly chosen (with replacement) to serve as parents. From each pair of parents one offspring was generated by joining their gametes generated from their haplotypes assuming 25 or 5 randomly located cross-overs. In the subsequent four rounds the 2000 parents were selected among previous 1000 offspring implying more than one offspring per parent. Thus, the total number of artificial admixed offspring was 5000. Table 1 summarises the number of cross-bred animals with 1 to 11 pure breed proportions in their genome.

Table 1: Summary of number of cross-bred animals with genome proportions of 1 to 11 pure breeds.

Number of cross-overs	Number of pure breeds contributing to a cross-bred genome										
	1	2	3	4	5	6	7	8	9	10	11
5	121	970	465	594	424	529	584	618	477	205	13
25	120	968	453	610	394	478	465	559	576	312	65

Note that table rows sum up to 5000, which is the number of cross-bred animals.

Result evaluation. Let bT be the row matrix of true breed proportions, and bE its estimated equivalent, with row dimension equal to the number of cross-bred animals and column dimension equal to the number of possible pure breeds. Results were evaluated by a parameter M calculated as the maximum of $|bT - bE|$, and by a parameter S calculated as the mean of $(bT - bE)^2$.

Software. CGR was implemented in a FORTRAN program which called the NLopt library (Johnson 2011). The optimisation solver used the augmented Lagrangian algorithm as global solver and the method of moving asymptotes as a local solver. All computations were carried out on a desktop computer with an Intel(R) Core(TM) i7-3770 processor and 32GB of memory.

RESULTS

Table 2 summarises the results for the cross-bred animals when the number of cross-overs during gametogenesis was 5 and 25 respectively. Invariably of the number of cross-overs CGR always performed better than ADM. The greatest absolute difference between the true and estimated breed proportion estimated by CGR was 0.24 and 0.28 for the 5 and 25 cross-over data set respectively, whereas for ADM that parameter was 0.85 and 0.67. The parameter S for the ADM results was 15 times larger than that for CGR results when the 5 cross-over data set was used. This difference to shrunk to 5 times larger when the 25 cross-over data set was used.

Table 2: Statistics of the breed proportion estimation error subject to the number of cross-overs when generating cross-bred animals and the used algorithm, where M is the maximum absolute error across all cross-bred animals and all possible breeds, and S is the mean of the squared estimation error calculated across all animals and possible breeds.

Number of cross-overs	CGR		ADM	
	M	S	M	S
5	0.24691	0.00103	0.85393	0.01578
25	0.28217	0.00107	0.67077	0.00566

CGR needed about 16 real time seconds for estimating the pure breed proportions of all 5000 cross-bred animals, whereas ADM needed 292 and 336 real time seconds for the 5 and 25 cross-over data set, respectively, which is an increase in processing time by a factor of 20. Note that the processing time was obtained without exploiting the parallel processing capabilities of both algorithms.

DISCUSSION

Results show that when pure breed population allele frequencies are known, the less elaborate modelling approach of CGR performs better than the ADM approach. Both algorithms do not account for linkage disequilibrium between marker. However, in addition to not assuming any LD between markers, the likelihood formulation of the ADM algorithm assumes also orthogonality between pure breed population allele frequency vectors. While this might be the case between very distant breeds having diverged many generations ago, it is unlikely to be the case for commercial beef cattle breeds. While CGR in its current formulation is not accounting for LD explicitly, it accounts for non-orthogonality between pure breed allele frequency vectors which might be one reason for the better performance. However, CGR could also account for LD by reformulating formula 1 to a generalised least square problem with the co-variance matrix of vector y reflecting the LD between markers, although this approach is limited by the number of markers. Beside better performance CGR generated more accurate results in a processing real time of only 5 % of that of ADM. This will becoming even more relevant when the number of marker used increases to 50k or more.

CONCLUSION

The results show that the simple modelling approach implemented in CGR provides accurate estimations of breed proportions in cross-bred animals. Moreover, CGR proved to be robust against LD, accounts for non-orthogonality of allele frequency vectors of founder breeds and is fast enough to deliver results for tens of thousands of animals in a reasonable time.

REFERENCES

Alexander D. H., Novembre J. and Lange K. (2009) *Genome Res.* **19**(9):1655.

Chiang C., Gajdos Z., Korn J. M., Kuruvilla F. G., Butler J. L., Hackett R., Guiducci C., Nguyen T. T., Wilks R., Forrester T. et al. (2010) *PLoS Genet.* **6**(3):e1000866.

Johnson S. G. (2011) The NLopt nonlinear-optimisation package.

Kuehn L., Keele J., Bennett G., McDaneld T., Smith T., Snelling W., Sonstegard T. and Thallman R. (2011) *J. Anim. Sci.* **89**(6):1742.

Pritchard J. K., Stephens M. and Donnelly P. (2000) *Genetics* **155**(2):945.

Raj A., Stephens M. and Pritchard J. K. (2014) *Genetics* **197**(2):573.

Tang H., Peng J., Wang P. and Risch N. J. (2005) *Genet. Epidemiol.* **28**(4):289.