Association Mapping Across Sheep Breeds Using A (Generalised) Linear Mixed Model Approach

P.C. Thomson[*], M.S. Khatkar[*], and H.W. Raadsma[*] on behalf of the International Sheep Genomics Consortium

Introduction

Association mapping across strains of mice has been used as a model for studying human disease (Cervino *et al*. 2007, Kang *et al*. 2008). It has the advantage of being able to control environmental causes of variation, through the replication of genetically identical animals within each strain. They rely on the availability of dense single nucleotide polymorphism (SNP) arrays. With the availability of such arrays in livestock species, these methods also have potential use in this area, with mapping across breeds. While breed averages for phenotypes may not always be currently available, often gross traits (major breed characteristics) are available for specific breeds, such as horns (presence / absence) and wool type (fine wool / coarse wool / hair), coat colour, litter size, and tail length.

Material and methods

Phenotype Database: We used the data base made available through the International Sheep Genomics Consortium Ovine Hapmap Project (http://www.sheephapmap.org/). In total, 70 sheep breeds were available for this study, and breeds were classified for the following traits: horned (Y / N), footrot resistant (Y / N), fat tail (Y / N), long tail (Y / N), and wool (hair / coarse wool / fine wool). Note that not all the phenotype information was available for each of the breeds. If means of continuous traits were available across breeds (as is the case for mouse strains, see for example the Mouse Phenome Database, http://www.jax.org/), this type of data could also be used in a breed association mapping,

Genotype Database: Across the 70 breeds, 2,819 sheep were genotyped with the Ovine SNP50 array, resulting in genotypes on 49,034 SNPs. Quality control on these SNPs included filtering for non-Mendelian inheritance, departure from Hardy-Weinberg equilibrium, and low minor allele frequency (< 0.05). In the present study, only OAR10 and OAR25 were evaluated, using 1681 SNPs on OAR10 and 906 SNPs on OAR25 for association mapping.

Since the unit record is the breed, genotypes were expressed as the proportion of animals having genotypes AA, AG, GA, or GG. Since these necessarily sum to unity, only the last three have been used as predictors in association mapping.

Breed Similarity: In association analyses, it has been shown that spurious associations will be detected if the population structure is not taken into consideration (Marchini *et al*. 2004,

[*] ReproGen – Animal Bioscience Group, Faculty of Veterinary Science, The University of Sydney, Private Bag 4003, Narellan NSW 2567, Australia

Cervino *et al*. 2007, Kang *et al*. 2008). To account for this, a breed similarity matrix of all 70 breeds was constructed. To compute this, the data were phased (Dr Paul Sheet) before using GERMLINE (Gusev *et al*. 2009) to derive IBD shared segments. The genome-wide sum for these shared segments was computed for each pair of animals and this was further reduced to a 70 × 70 matrix by taking the mean of haplotype sharing between animals from breed A with animals from breed B. Similarity was expressed on a scale from 0 (no similarity) to 1 (complete similarity). The resulting similarity matrix (**G**) was therefore similar in structure to a correlation matrix.

Statistical Model: At each locus, a separate linear mixed model for testing the association was fitted. This takes the form of a linear mixed model for continuous (i.e. normally distributed) data, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \varepsilon$, where $\mathbf{y}$ is the $n \times 1$ vector of phenotypes, $\mathbf{X}$ is the $n \times 3$ matrix of SNP genotype frequencies, $\boldsymbol{\beta}$ is the 3×1 vector of the SNP effects, $\mathbf{Z}$ is an incidence matrix that links records to breeds (here $\mathbf{Z} = \mathbf{I}_n$, since there is one observation per breed), $\mathbf{u}$ is an $n \times 1$ vector of random breed effects, assumed $\mathbf{u} \sim N(\mathbf{0}, \sigma_u^2\mathbf{G})$, and $\boldsymbol{\varepsilon}$ is the $n \times 1$ vector of independent random errors, assumed $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma_\varepsilon^2\mathbf{I}_n)$. Note that the non-independence of the data as a result of the breed similarity is addressed through inclusion of the correlated breed effects (**u**).

For data with a categorical classification (presence / absence), a linear mixed model is replaced by a logistic generalised linear mixed model, $\text{logit}(\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$, where $\text{logit}(\pi) = \log_e[\pi/(1 - \pi)]$ is the link function, and where $\boldsymbol{\pi}$ is the $n \times 1$ vector of "success" probabilities.

Computational Aspects of Model Fitting: Given the repeated fitting of models over the length of the chromosome, the statistical programming language R provides the best medium for analysis. However, neither the `lme()` function within the `nlme` library in R, nor the `asreml()` function as part of ASReml-R allow easy specification of arbitrary structures of the correlation structure matrix, **G**. However, Ball (2007) has indicated how to achieve this using a Choleski decomposition of **G**, allowing standard (generalised) linear mixed model procedures to be used. Using the `chol()` function in R, **G** can be written as $\mathbf{G} = \mathbf{T}\mathbf{T}'$, where **T** is the lower-triangular matrix from the Choleski decomposition. Then calculating $\mathbf{Z}^* = \mathbf{Z}\mathbf{T}$, the models can be re-written as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}^*\mathbf{u}^* + \varepsilon$, or $\text{logit}(\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}^*\mathbf{u}^*$, where now the resultant $\mathbf{u}^*$ are independent, i.e. $\mathbf{u}^* \sim N(\mathbf{0}, \sigma_u^2\mathbf{I}_n)$. The elements of $\mathbf{Z}^*$ will no longer be 0/1 values but take non-integer values. However, these models can be readily fitted in `lme()` or `asreml()`, but need to be constrained so that the variances associated with $u^*_1, \ldots, u^*_n$ are all equal (to estimate a common σ_u^2). This can be achieved using the `pdIdent` structure in `lme()` or the variance `constraint` option in ASReml-R.

Note that whilst ASReml-R includes a method for specifying logistic models (using the `family = asreml.binomial()` option), no such facility exists within `lme()`. However, it may be possible to achieve this using the `lmer()` function (in the `lme4` library), or the `glmmPQL()` function (in the `MASS` library).

Results and discussion

Results are presented here for two of the traits considered, horned/polled and hair/wool, and both being binary (presence / absence) traits. Breed association maps are shown for OAR10 and OAR25, being two chromosomes with known major genes for polled (OAR10) and fibre diameter (OAR25).

For the horned trait on OAR10 (Figure 1), there was a highly significant association at SNP "OAR10_29538398.1" (–$\log_{10}P$ = 8.4) and another at "s27070.1" (–$\log_{10}P$ = 6.1). The most significant SNP maps to a region known to contain the poll locus in sheep (Montgomery *et al.*, 1996).

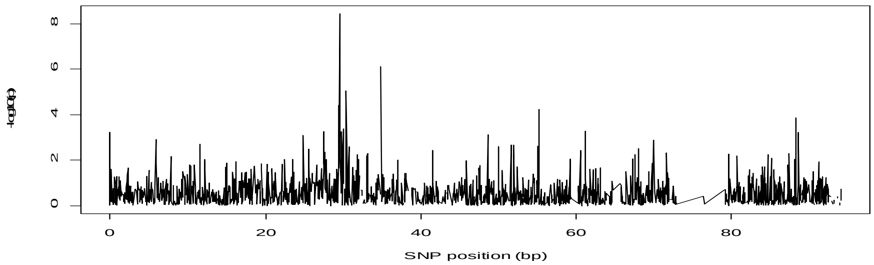

Figure 1: Association with SNP and horn (presence / absence) at positions along sheep chromosome 10. Association is expressed as –$\log_{10}$*P*, where *P* is the *P*-value for the test of association at the SNP.

Associations of SNPs on OAR25 with the hair / wool trait is shown in Figure 2. A region of SNPs with significant associations was detected, the most significant association (–$\log_{10}P$ = 9.3) being with the SNP "s25195.1" at 7,203,123 bp. This region is known to contain a major gene/QTL for wool traits (Ponz *et al.*, 2001). No further distinction was made between coarse wool and fine wool breeds, or between hair and fine wool breeds but obviously this could be done.

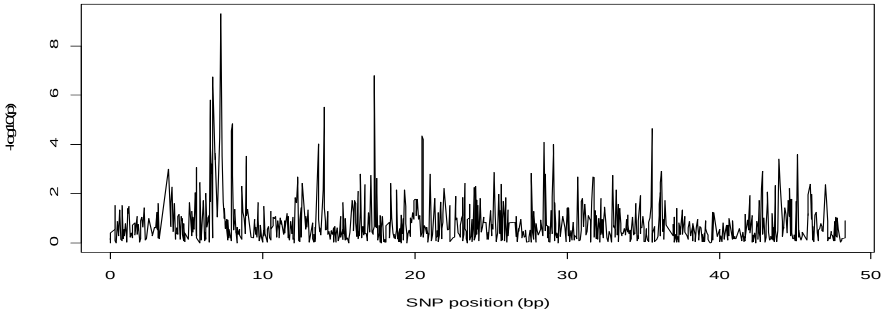

Figure 2: Association with SNP and hair (presence / absence) at positions along sheep chromosome 25. Association is expressed as –$\log_{10}$*P*, where *P* is the *P*-value for the test of association at the SNP.

The localisation of both the polled locus and the major QTL for fibre diameter shows the utility of combining high density SNP genotype data across breeds. Livestock breeds differ enormously from the inbred mouse models which are typically used for strain mapping approaches. Although many breeds are located in geographically separated environments, major breed characteristics often remain intact. It is also likely that breeds may indeed be fixed for major genes which contribute to such traits, and by utilizing between-breed comparisons, the location of such genes can be pinpointed in the genome with a high degree of accuracy. Relationships between breeds are taken care of by using the same molecular data by calculating an identity relationship matrix. Indeed the use of multiple breeds leads to inclusion of historical recombination. The inclusion of within-breed variation on individual animals may add an even greater degree of precision of mapping since individual phenotype can be matched against genotype. The method could thus be extended to traits for which the major loci within breed have not been fixed, and still allowing across-breed utilisation of historical recombination. Use of reference strains/breeds across environments may also allow for a greater precision of mapping in both continuous and categorical traits, making the method suitable for a wide range of applications in livestock where there is often intense interest in describing performance/phenotype characteristics across breeds, strains, and bloodlines within breed.

Conclusion

Breed association mapping has the potential to reveal highly localised regions of the genome showing associations with traits of interest. As use of SNP arrays becomes more widely available and data sets move into the public domain, its use could further increase. However, what will underpin its use will be the availability of reliable tables of breed-specific phenotypes.

Acknowledgements

We are grateful to Dr Paul Sheet of The University of Texas for providing the phased data from the PHASE program.

References

Ball, R. D. (2007). *Statistical Analysis and Experimental Design*. In *Association Mapping in Plants*, N. C. Oraguzie, E. H. A. Rikkerink, S. E. Gardiner, and H. N. De Silva (eds). New York: Springer.

Cervino, A. C. L., Darvasi, A., Fallahi, M., *et al*. (2007). *Genetics* 175:321-333.

Gusev, A., Lowe, J. K., Stoffel, M., *et al*. (2009). *Genome Res.* 19:318-326.

Kang, H. M., Zaitlen, N. A., Wade, C. M. *et al*. (2008). *Genetics* 178:1709-1723.

Marchini, J., Cardon, L. R., Phillips, M. S., and Donnelly, P. (2004). *Nature Genet*. 36:412-517.

Montgomery, G. W., Henry, H. M., Dodds, K. G., *et al*. (1996). *J. Hered.* 87:358-363.

Ponz, R., Moreno, C., and Allain, D., *et al*. (2001). *Mamm. Genome* 12:569-572.