CALCULATING EXACT PROBABILITIES OF ALLELE FREQUENCY DIFFERENCES IN DIVERGENT SELECTION LINES

K.G. Dodds and J.C. McEwan

AgResearch, Invermay Agricultural Centre, Private Bag 50034, Mosgiel, New Zealand

SUMMARY

One approach to searching for genes which influence quantitative traits involves detecting an association between the trait of interest and the genotype of a "candidate gene". An application of this approach is to test for allele frequency differences in divergent selection lines. However a simple binomial test for allele frequency differences ignores genetic drift within the lines, and therefore is too liberal. A simulation method, using the actual selection line pedigrees, which accounts for genetic drift is presented and compared with several approximate methods.

Keywords: Divergent selection lines, QTL, statistical significance, effective population size

INTRODUCTION

There is currently a large international research effort directed at finding genes or regions of the genome which influence quantitative traits (Archibald *et al.* 1994). Such genes are known as quantitative trait loci (QTL). Experiments designed to detect such loci usually involve scanning the genome with evenly spaced markers in an F_2 or backcross between divergent lines. A disadvantage of this approach is the additional resources required to generate the QTL mapping population. The delay in time and quantity of resources involved can be considerable in the case of large domestic species.

Another approach is to look for allele frequency differences in closed divergent selection lines. The rationale is that if alleles of a polymorphic locus influence a trait under divergent selection, the allele frequencies will also diverge. A significant difference in allele frequencies is inferred to be due to selection for the trait, either due to the allele itself, or by the linkage disequilibrium generated by selection on a close locus. Many large animal populations of this structure are available as a result of quantitative trait selection experiments, and so there is no delay in DNA sampling. Such a study would usually be conducted using "candidate genes" (polymorphic loci chosen because they may influence the trait of interest) or loci closely linked to candidate genes. Alternatively, if a putative QTL has already been detected, and its position localised, this technique allows verification in independent selected populations. Also, the process of several generations of selection may result in decreased genotyping for a specified power. Thus, under a particular model, Keightley and Bulfield (1993) found that 192 selection line individuals were equivalent to 1250 F_2 individuals.

This technique has been used by Fotouhi *et al.* (1993) to associate polymorphisms in the growth hormone locus with abdominal fat content using high and low chicken selection lines. Høj *et al.*

(1993) used it to suggest an association between a growth hormone gene polymorphism and milkfat production in cattle.

However use of the binomial test for allele frequency differences, as in the above studies, ignores genetic drift (due to founder effects and inbreeding) within the lines, and therefore is too liberal. We present a simulation method which accounts for these effects, using actual selection line pedigrees, and therefore provides the correct level of significance. We also discuss several alternative approaches, some of which are applicable when pedigree information is unavailable.

METHOD

The effect of genetic drift on allele frequency differences can be taken into account where pedigree information on the selection lines is available. It is assumed that the locus of interest is autosomal and follows the Mendelian laws of inheritance. The procedure simulates genetic sampling under the null hypothesis of no selection pressure on the locus, as follows. Firstly it requires the allele frequencies in the population from which the selection lines were derived. These may be estimated as the mean of the lines, using the individuals genotyped. The selection line founders are then assigned genotypes at the locus of interest by randomly sampling alleles with these frequencies. Subsequent generations are assigned genotypes by randomly choosing one of each parent's alleles. After all individuals in the lines have been assigned genotypes, those that were actually genotyped in the experiment are used to calculate the simulated allele frequencies for the lines. A test statistic which is suitable for comparing the allele frequencies between lines is then calculated. The simulation procedure is then repeated many times (say 1000) to generate the null distribution for the test statistic. The test statistic can then be calculated on the actual data, and its significance assessed by comparison with the simulated distribution.

The technique was examined using data from divergent selection lines and a simulated population with similar structure, but with non-overlapping generations and no selection. The selection flock consisted of lines selected for high (Fat+) or low (Fat-) subcutaneous backfat depth (Fennessy *et al.* 1987). Each line was established with 60 females (born 1979) and 8 males (born 1980 and 1981) and is maintained at about 60 ewes mated annually with 4 rams. The rams are used once only, either as lambs or two tooths and the average age of dams was about 3.5 years. All records to the 1992 birth year were used and constituted approximately 5 generations of selection (12 years / 2.4 years generation interval) after which 60 individuals (120 alleles) per line over 2 birth years were (hypothetically) genotyped with equal numbers per sire. The simulated population consisted of 2 lines of 5 non-overlapping generations from the founders, with 8 sires and 40 dams per line contributing to each successive generation. The appropriate numbers of males and females to be parents of the next generation were generated, and their sire and dam randomly assigned, along with an allele from each. In the final generation 60 offspring per line were generated and assigned genotypes.

The simulation method was then compared with 3 other methods. The first, which ignores genetic drift, is to calculate the binomial standard error of the difference (SED) in line allele frequencies. Suppose n_i individuals from line i are genotyped, and let p_0 be the observed proportion (line-

averaged) of a particular allele in the genotyped animals. The second method, accounts for genetic drift by approximating the effective population size (EPS) of the lines (Falconer 1983; Caballero 1994; see below) and assuming it is the same for each line. Let N_{eg} be the EPS for the gth generation starting from the founders (g=0) and continuing until the tth generation, where it is the number of individuals genotyped. The respective SED's for these two methods are:

$$SED_b=\sqrt{p_0(1-p_0)\left(\frac{1}{2n_1}+\frac{1}{2n_2}\right)} \quad \text{and} \quad SED_{EPS}=\sqrt{2p_0(1-p_0)\left(1-\prod_{g=0}^{t}(1-\frac{1}{2N_{eg}})\right)}.$$

The third method (SED_F) avoids EPS calculations by replacing the middle t-1 terms of the product in SED_{EPS} by 1-F, the average inbreeding calculated on the pedigrees of the sampled individuals.

RESULTS

For the simulated population N_{eg}=27, g=0,...,4 and N_{e5}=60 (the number of individuals genotyped) while p_0=0.488 (mean of genotyped animals) and F=0.107. Results are presented in Table 1 and Figure 1. For the selection flock, we assume p_0=0.5 in the base population. Calculation of EPS (generations 1 to 4) was approximated by ignoring complicating factors such as overlapping generations, restrictions on family size and the effects of selection for a trait of interest (see e.g. Nomura 1995) and by assuming the number of breeding males and females in each generation was 8 and 36 respectively. This gives N_{e0}=28, N_{eg}=26, g=1,...,4 and N_{e5}=60. For the sampled individuals, F=0.073. Results are presented in Table 1.

Table 1: Estimated SEDs, calculated by 4 methods, for line allele frequencies in both actual and simulated populations

Method	Population	
	Simulated	Fat- and Fat+ lines
Binomial	0.065	0.065
EPS	0.221	0.222
Inbreeding	0.256	0.221
Simulation (4000 replicates)	0.233	0.228

DISCUSSION

Comparison of methods. As illustrated above, use of the binomial can grossly underestimate the true SED. Much better results are obtained by using SED_{EPS}. However, this method requires an assumption of normality and departures from an idealised population require either approximate or complex calculations. The latter of these can be circumvented by using SED_F. However both difficulties are overcome by the simulation method, which correctly accounts for genetic drift and bias in the choice of individuals mated and genotyped. In addition, significance points from the simulated distribution are used, so normality assumptions, which are more likely to be violated as p_0 moves away from 0.5, are not required.

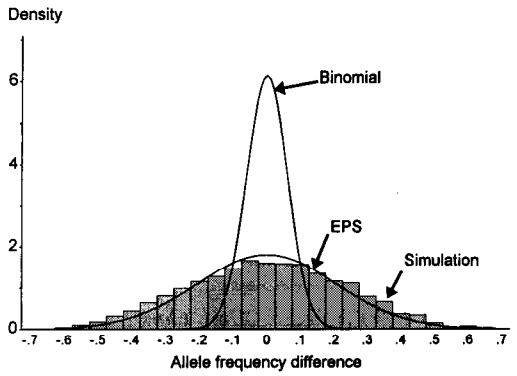

Figure 1. Comparison of allele frequency difference distributions for simulated data.

Missing pedigree information. If there is missing pedigree information because the selection lines are open, and new animals have entered the lines, genotypes are generated by sampling from the base population. If information is missing because of incomplete pedigree recording, genotypes could be generated using the average allele frequencies for that cohort. If no pedigree information is available then alternative techniques are required. Knowledge of the mating structure and time since divergence, would allow the use of SED_{EPS} by calculating the EPS with the appropriate formula (Caballero 1994). Alternatively it may be possible to approximate the effect of genetic drift by using neutral genetic markers to estimate the inbreeding (F) and then using $\sqrt{2p_0(1-p_0)(F+(1-F)/2N_{et})}$ for the SED. This equation ignores the contribution arising from founder effects (sampling the foundation animals). Keightley and Bulfield (1993) essentially apply the EPS method to their data, but add a selection effect parameter by simulating populations with a particular selection effect and EPS, to allow estimation of the selection effect.

Test statistic. The simulation method results in this paper use the line difference in frequency of a particular allele as the test statistic. This enabled comparison with other methods. However, any statistic suitable for comparing allele frequencies between lines could be used. For 2 selection lines and 2 alleles, we could choose the difference in frequency of a particular allele, a two-sample t statistic for this difference, or any 2×2 contingency table test statistic. If there are more than 2 selection lines or alleles, a contingency table X^2 statistic, with rare alleles possibly collapsed into a single class, would normally be used. Any statistics which are monotonically related will yield the same result, since the simulated critical values will be equivalent.

CONCLUSION

We have illustrated the importance of accounting for genetic drift when testing for line differences in allele frequency. This can be achieved by following the simulation procedure described which is available from the authors as a SAS/IML macro PEDDRIFT. Alternative approximate methods are available, but their accuracy may be uncertain. An additional benefit of the simulation procedure is

that it the automatically takes account of the sample size and table structure, so we do not need to be concerned with the asymptotic approximation to the finite sample test statistic.

REFERENCES

Archibald, A.L., Burt, D.W. and Williams, J.L. (1994) *Proc. 5th World Cong. Genet. Appl. Livest. Prod.* **21**:5.

Caballero, A. (1994) *Heredity* **73**:657.

Falconer, D.S. (1983) "Introduction to Quantitative Genetics" 3rd ed. Longman, London.

Fennessy, P.F., Greer, G.J. and Bain, W.E. (1987) *Proc. 4th Asian-Australasian Assoc. Anim. Prod.* **4**:382.

Fotouhi, N., Karatzas, C.N., Kuhnlein, U. and Zadworny, D. (1993) *Theor. Appl. Genet.* **85**:931.

Høj, S., Fredholm, M., Larsen, N.J. and Nielsen, V.H. (1993) *Anim. Genet.* **24**:91.

Keightley, P.D. and Bulfield, G. (1993) *Genet. Res.* **62**:195.

Nomura, T. (1996) *J. Anim. Breed. Genet.* **113**:1.