FOUNDER ALLELE FREQUENCIES AND SEGREGATION ANALYSIS

B. Tier[1] and J.M. Henshall[2]

[1] Animal Genetics and Breeding Unit ,
University of New England and NSW Agriculture, Armidale NSW 2351, Australia
[2] CSIRO Livestock Industries, Ibis Avenue, North Rockhampton QLD 4601, Australia

INTRODUCTION

Segregation analysis generally requires knowledge of the allele frequencies in the founder population. These founder allele frequencies (FAF) are used to generate genotype probabilities for individuals in the population when no data are available, or are combined with small amounts of data. Different FAF can be used to examine the sensitivity of the probabilities to the choice of FAF. Alternatively, as shown by Henshall and Tier (2002), the impact of the FAF on the genotype probabilities can be separated from that of the data in a segregation analysis. FAF also have an important role in computing likelihoods. The use of FAF, with algorithms based on peeling (Elston and Stewart, 1971), has the potential to generate misleading confidence in the genotype probabilities that result from segregation analyses.

FAFs are frequently determined subjectively - often from the data under analysis. In animal breeding, the pedigree of a large population may be known for many generations, but only the genotypes of the most recently born animals are known. In these situations an estimate of the allele frequencies in the oldest genotyped generation is used as the FAF. The aim of this paper is to demonstrate how little we can know about the FAF when only the genotypes of the most recent animals are known and how different FAF affect genotype probabilities.

MATERIALS AND METHODS

A population was simulated with 5 founder males and 200 founder females. Females were placed in groups of 40. One male was allocated to each group of females. All females had a single offspring which had an equal chance of being a male or female. In each cycle, the oldest and worst females were replaced with the best female progeny and sires were replaced with their best male progeny with probability 0.5. 200 independent loci were simulated and a trait under the control of a polygenic effect with a heritability of 0.2. Each locus started with 4 alleles with frequencies 0.1, 0.2, 0.3 and 0.4 within founder sires and dams. Animals were chosen to replace parents on the basis of the their phenotype. 20 cycles were completed to give a total population size of 4005 animals. Gene frequencies were calculated for cohorts 5, 10, 15 and 20.

FAF were inferred for each of these observed cohorts by assigning observed alleles to founders, weighting the founders and using the weights. The method used to trace back contributions to the founders is similar to the method described by Boichard *et al.* (1997) to computing probabilities of gene origin. Each observed animal is assigned a value 1. Its parents are assigned a half, grand-parents one quarter each and so on back to the founder population. The weights applied to each founder were $1\text{-}0.5^r$, where r is the sum of these contributions from

observed descendants. Thus a founder with only one descendant would have a weight of a half (half its genes were observed) and a founder with three descendants would have a weight of 0.875. These weights reflect the chance of both their alleles being observed in a cohort. FAF were inferred from the whole cohort and also by using the alleles ascribed to the dams of the cohort - to avoid over-representation of the sires of the cohort (Genotypes of the sires of the last cohort were determined from their progeny's genotypes.). The number of alleles expected to be observed in cohorts 5, 10, 15 and 20 were also calculated by labeling the founder's alleles and randomly dropping them through the pedigree. An additional 10,000 loci were generated for the same population structure to examine the distributions of allele frequencies in early cohorts giving rise to small ranges of frequencies in cohort 20.

To illustrate the effect that the FAF has on the genotype probabilities, they were calculated for the small population described in Henshall and Tier (2002) using a number of different FAF.

RESULTS AND DISCUSSION

Table 1 presents summary statistics for FAF inferred from the dams' alleles. The means are similar to the initial frequency throughout but the variance increases with distance from the base population. By cohort 15 some alleles that started at low frequency have vanished and some which started at high frequency are at a much lower frequency than when they started. When all the alleles are used then these trends are more pronounced.

Table 1. Means, standard deviations, minimum and maximum of the dam's allele frequencies observed in different cohorts

	Founder frequency							
Cohort	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4
	Means				Standard Deviations			
5	0.100	0.196	0.303	0.396	0.030	0.038	0.040	0.042
10	0.097	0.200	0.301	0.401	0.054	0.074	0.090	0.089
15	0.090	0.201	0.309	0.401	0.067	0.100	0.120	0.116
20	0.089	0.193	0.316	0.400	0.088	0.121	0.150	0.149
	Minima				Maxima			
5	0.030	0.110	0.197	0.302	0.191	0.550	0.405	0.523
10	0.005	0.048	0.102	0.181	0.334	0.525	0.513	0.612
15	0.000	0.006	0.043	0.130	0.343	0.663	0.632	0.726
20	0.000	0.000	0.026	0.098	0.472	0.615	0.680	0.696

The method of weighting suggests that about 73, 48, 42 and 40 founder alleles remain in cohorts 5, 10, 15 and 20 generations respectively when FAF was inferred from all alleles. With the gene-drop method an average of 131, 64, 39 and 26 base alleles were estimated to remain in the population at cohorts 5, 10, 15 and 20 respectively. There is a disparity between the results of the two methods which arises from the gene-drop method constraining all individuals to have only two alleles, whereas the weighting method counts the expected number of expressions of each founder in a cohort, without limiting any individual to two alleles. The lower counts for the weighted method in early cohorts results from limiting the expressions of

the founders. Regardless of which method is used it is clear that by cohort 20 relatively few of the founders' alleles remain. For this population, the loci were independent of the trait under selection. However, this is a small population with a pedigree typical of sheep. The processes of selection and selective genotyping or errors in the pedigree could all contribute additional unreliability of such estimates of FAF.

Figure 1 illustrates the inferred FAF in cohorts 5, 10 and 15 for alleles that had a frequency between 0.07 and 0.13 in cohort 20. 2,778 of the 10,000 loci fell within this range of frequencies in cohort 20. The histograms demonstrate that the further back from the observed cohort the more variable the allele frequencies were. The distributions are also skewed with means increasing with distance from cohort 20. The means of these distributions were 0.17, 0.15 and 0.13 in cohorts 5, 10 and 15 respectively. The distribution of frequencies within this range on cohort 20 was flat.

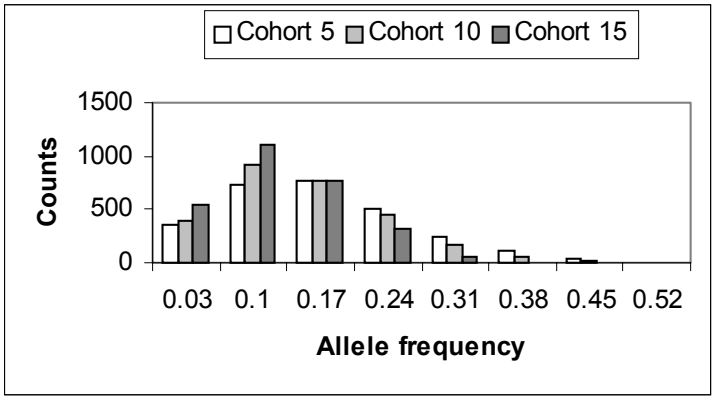

Figure 1. Distribution of allele frequencies in cohorts 5, 10 and 15 for alleles with a frequency between 0.07 and 0.13 in cohort 20

The variation in the allele frequencies found by looking down or up the pedigree, and the numbers of founder alleles remaining in the population suggest that caution is required when inferring allele frequencies in the founder population.

Table 2 illustrates the genotype probabilities for four individuals shown in table 1 of Henshall and Tier (2002) for different FAF. It is clear from these results that the FAF has a significant effect on individuals without supporting data. Consequently care must be taken when choosing a FAF, and when interpreting results when methods are employed that cannot differentiate between the influence of the FAF and the data.

Table 2. Genotype probabilities for four individuals shown in table 1 (Henshall and Tier, 2002) for founder allele frequencies P(a)=0.2 and P(a)=0.5

	Genotype					
	P(a)=0.2			P(a)=0.5		
Identity	aa	ab	bb	aa	ab	bb
Y1	10	50	40	25	50	25
Y2	10	50	40	16	50	34
Z1	0	18	82	4	37	59
Z2	0	19	81	0	23	77

Uncertainty in the FAF could be modeled by introducing a prior distribution for those parameters. A beta distribution, similar to that used for qualitative segregation analysis (Sorensen, 1997) would be suitable. It is readily adaptable to multiple alleles. Furthermore, it would be possible to treat different groups of founders with different prior distributions if desired.

CONCLUSION

Uncertainty with regard to estimating FAF among founders when only the pedigree and genotypes of the most recent cohort of individuals are known has been clearly demonstrated, as has the effect it can have on individuals with limited information. It is also clear that the further the observed cohort is from the base, the less reliable are the estimates. Incorporation of a prior distribution for the FAF needs to be examined.

REFERENCES

Boichard, D., Magniel, L. and Verrier, E. (1997) *Genet. Sel. Evol* **29** : 5-23.

Elston, R.C. and Stewart, J. (1971) *Hum. Hered.* **21** : 523-542.

Henshall, J.M. and Tier, B. (2002) *Proc. 7WCGALP*

Sorensen, D (1997) “Gibbs Sampling in Quantitative Genetics”, Internal Report No. **82**, Danish Institute of Animal Science, Foulum, Denmark.