*Proc. Assoc. Advmt. Anim. Breed. Genet. 18:251-254*

# A PECULIARITY OF GENE FREQUENCY ESTIMATION

**J.W. James[1], Vicki A. Whan[2] and Belinda J. Norris[2]**

[1]Faculty of Veterinary Science, University of Sydney, NSW 2006
[2] CSIRO Livestock Industries, St Lucia, Queensland 4067

## SUMMARY

Black wool is due to a recessive gene, the dominant white trait being caused by a duplication, with one or more repeats. The total number of repeats in the diploid genotype can be detected, but a 1/1 and a 2/0 repeat cannot be distinguished. This complicates gene frequency estimation, and in the absence of black sheep, it is shown that very similar data on repeat numbers can give drastically different gene frequency estimates. The (possible) presence of black sheep can resolve the problem.

## INTRODUCTION

The presence of black sheep in a flock is highly undesirable, and if such animals appear they are usually culled. Black wool has long been known to be due to a recessive gene, with heterozygotes for the recessive allele indistinguishable from the dominant white homozygotes. This has made it very difficult to eliminate the recessive allele, as only the production of black progeny indicates that a ram is a carrier. However, the molecular basis of the trait has recently been shown by Norris and Whan (2008) to be a duplication at the ovine agouti locus, and a test has been developed. Unfortunately, there is more than one dominant allele, with some having more than one repeat. The test counts the number of repeats in the diploid genotype and thus cannot distinguish between a homozygous dominant 1/1 and a heterozygous 2/0. The first genotype would produce all white progeny, but the second could produce black offspring, and half of all progeny would be carriers of the recessive allele. The practical value of the test clearly depends on the frequencies of the various alleles in the population to which it is applied. Samples of white sheep were therefore taken and tested. In the course of analysis of these data a peculiarity of gene frequency estimation was found, and this is reported in this paper.

## DATA and ANALYSIS

For the present purpose only a part of the data is relevant, so other data collected are not mentioned. From the Falkiner Field Station research flock, 60 White Suffolk and 76 Poll Dorset ewes were sampled and tested for number of repeats. The results are shown in Table 1.

**Table 1. Distributions of number of repeats in White Suffolk and Poll Dorset ewes.**

| Breed | Number of ewes | Number of repeats | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| WS | 60 | 2 | 23 | 18 | 11 | 5 | 1 |
| PD | 76 | 2 | 31 | 20 | 11 | 9 | 3 |

The two distributions look quite similar, and a chi-squared test gave a value of 1.56 on 5 degrees of freedom with a probability of about 0.9, confirming the apparent similarity.

If we assume that there are four alleles present with 0, 1, 2 and 3 repeats, with gene frequencies $p_0$, $p_1$, $p_2$ and $p_3$, and that the parents mated at random, the expected proportions of the six phenotypes are:

$f_1 = 2p_0p_1/(1 - p_0^2)$, $f_2 = (p_1^2 + 2p_0p_2)/(1 - p_0^2)$, $f_3 = (2p_1p_2 + 2p_0p_3)/(1 - p_0^2)$, $f_4 = (p_2^2 + 2p_1p_3)/(1 - p_0^2)$, $f_5 = 2p_2p_3/(1 - p_0^2)$, $f_6 = p_3^2/(1 - p_0^2)$. (1)

These frequencies have been calculated assuming a Hardy-Weinberg distribution at birth, with any black lambs being culled.

Letting $n_j$ denote the number of ewes with j repeats, the logarithm of the likelihood function is

$L = n_1 \log f_1 + n_2 \log f_2 + n_3 \log f_3 + n_4 \log f_4 + n_5 \log f_5 + n_6 \log f_6$. (2)

Maximising the likelihood for general values of $f_j$ gives estimated frequencies as $n_j/N$, where N is the total number of animals. Assuming the values of $f_j$ satisfy the relations given in (1), finding the maximum likelihood estimates and comparing the resulting maximum of the likelihood with the general maximum gives a test for the adequacy of the assumed model.

An analytical approach to finding the maximum with the assumed model is not rewarding, so a Monte Carlo method was used. The method was to sample 4 values from a uniform distribution on (0, 1), then to divide each by their sum to give 4 gene frequencies adding to unity. These were then used in the equations (1) to find the $f_j$ which were inserted in (2) to find the corresponding likelihood. This was replicated $10^8$ times, and the largest value of L and its associated values of $f_j$ and gene frequencies were taken as the maximum likelihood and the ML estimates. This is not necessarily computationally efficient, but is quick to program, and each run takes a couple of minutes. A chi-squared was computed from the difference in likelihoods for the general and assumed models, with 2 degrees of freedom as 3 independent parameters were estimated for the assumed model. In addition, the expected numbers were calculated from the estimated $f_j$ and tested against the observed numbers by chi-squared. The two methods of computing chi-squared agreed very well. As an additional test of the model, an additional allele with 4 repeats was added, and the analysis repeated, but the outcome was that the estimated frequency of the extra allele was negligible, and the likelihood was essentially unchanged, so the results are not given here. To compute standard errors, a simplex was constructed starting from the ML estimates, and the method of Nelder and Mead (1965) was used.

**RESULTS**

The maximum likelihood estimates of gene frequencies in the two samples are shown in Table 2.

**Table 2. Maximun likelihood estimates (standard errors) of gene frequencies in White Suffolk and Poll Dorset ewes, and chi-squared values for testing the assumed model.**

| Breed | $p_0$ | $p_1$ | $p_2$ | $p_3$ | Chi-squared |
|---|---|---|---|---|---|
| WS | .0276 (.0336) | .5970 (.0798) | .2487 (.0815) | .1267 (.0501) | 0.67 |
| PD | .5125 (.1007) | .0201 (.0271) | .2998 (.0713) | .1676 (.0437) | 0.37 |

Clearly there is no need to reject the assumed model. However, the great discrepancy in estimates of $p_0$ and $p_1$ is very surprising, since the distributions do not differ significantly. The Poll Dorset result seems to be discrepant, since it implies that about one quarter of Poll Dorset lambs would be born black, which is not the case. It appears that in the absence of black lambs, the frequencies of the recessive and single repeat alleles can compensate for each other. This can be illustrated by the fact that if the estimates for White Suffolk are used to compute expected numbers for Poll Dorset and vice versa, the chi-squared values for comparing the observed and expected values are 8.77 and 4.05 respectively on 5 degrees of freedom, with probabilities of

approximately 0.15 and 0.5. Thus, although the ML estimates fit the data most closely, the ML estimates from the other breed do not fit significantly worse. In this case it was not hard to see that the Poll Dorset ML estimate was misleading, but this will not always be the case.

**DISCUSSION**

The data used here were obtained from flocks of white ewes, but if the true frequency of the recessive allele is about 0.02 it is very likely that these sheep came from a population in which no black lambs were actually culled. If we therefore assume that in fact there was an observed value of zero for ewes with no repeats, we have an expected proportion of $f_0 = p_0^2$, while the $f_j$ values in (1) are all multiplied by $(1 - p_0^2)$. With these modified data we have estimated gene frequencies using the method described above with the obvious changes. The results are shown in Table 3.

**Table 3. Maximum likelihood estimates of gene frequencies (standard errors) for White Suffolk and Poll Dorset ewes with data augmented with zero homozygous recessives.**

| Breed | $p_0$ | $p_1$ | $p_2$ | $p_3$ |
|---|---|---|---|---|
| White Suffolk | .0275 (.0324) | .5974 (.0790) | .2476 (.0810) | .1275 (.0503) |
| Poll Dorset | .0221 (.0269) | .5954 (.0660) | .2230 (.0602) | .1594 (.0446) |

The chi-squared tests gave values of 6.30 and 0.74 respectively on 3 degrees of freedom, so that the augmented data are satisfactorily fitted. It is striking that by forcing the estimates in the Poll Dorset to give low expected numbers of homozygous recessives the analysis has resulted in a switch between $p_0$ and $p_1$. On the other hand, the White Suffolk estimates from the augmented data are nearly identical to those from the actual data.

As an approach to obtaining reasonable estimates for the Poll Dorset without augmenting the data we have modified the estimation program to find the maximum of the likelihood over the other gene frequencies for fixed values of $p_0$ and the resulting values of chi-squared comparing this likelihood with the unrestricted one are shown in figure 1. The likelihood curve plotted against fixed values of $p_0$ has two maxima for both the Poll Dorset and the White Suffolk, one near 0.02 and one near 0.5. As can be seen, the two cases show similar curves, but the lower of the two minimum chi-squareds is different in the two breeds. Obviously, the variation in the samples has had the effect of shifting the curve for the Poll Dorset so that the "wrong" likelihood is globally maximum.

Clearly in this case the information in the data does not allow the gene frequencies to be found with great precision, nor does it allow the two models, with or without an "observed" zero for the number of recessives, to be definitively distinguished. However, the fact that both breeds show two maxima for the likelihood, and that the global maximum may be incorrect in some cases, even when the sample distributions do not differ significantly, is a warning to be careful when analysing data of this kind.

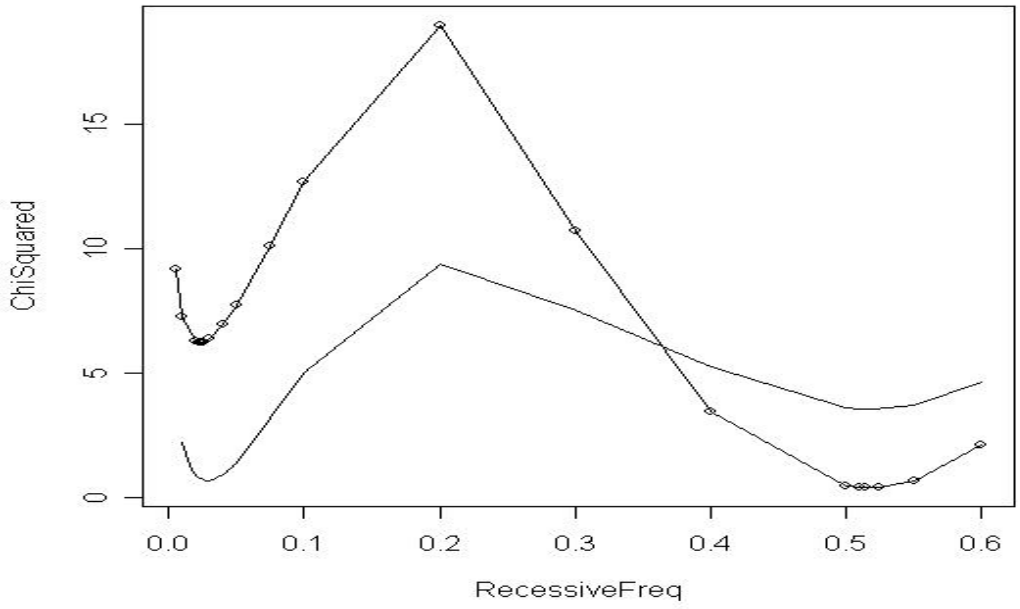


**Figure 1. Chi-squared values from unrestricted maximum likelihood versus maximum likelihood for different set values of $p_0$ in Poll Dorset (with dots) and White Suffolk (no dots).**

The main lesson to be learnt from this is that even estimates obtained by highly regarded methods such as maximum likelihood can be seriously wrong, and need to be regarded with scepticism if they do not make sense. Another lesson is that the absence of a class from the data can have important consequences for the analysis. An obvious possible situation is with a recessive lethal. A third lesson is that gene frequency estimation is always harder when the genes present in the sample cannot be counted.

**ACKNOWLEDGMENTS**

Professor Frank Nicholas was responsible for JWJ's involvement in this work. We thank a referee for suggesting the calculation of standard errors.

**REFERENCES**

Nelder, J.A. and Mead, R. (1965) A simplex method for function minimization. *The Computer Journal* **7**:308.

Norris, B.J. and Whan, V.A. (2008) A gene duplication affecting expression of the ovine ASIP gene is responsible for white and black sheep. *Genome Research* **18**:1282.