2004

Predicting population gene frequency from sample data

R. M. Lewis

Virginia Polytechnic Institute and State University, ron.lewis@unl.edu

B. Grundy

Virginia Polytechnic Institute and State University

L. A. Kuehn

Virginia Polytechnic Institute and State University, Larry.Kuehn@ars.usda.gov

Follow this and additional works at: <https://digitalcommons.unl.edu/animalscifacpub>

Part of the [Genetics and Genomics Commons](#), and the [Meat Science Commons](#)

Lewis, R. M.; Grundy, B.; and Kuehn, L. A., "Predicting population gene frequency from sample data" (2004). *Faculty Papers and Publications in Animal Science*. 841.

<https://digitalcommons.unl.edu/animalscifacpub/841>

This Article is brought to you for free and open access by the Animal Science Department at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Faculty Papers and Publications in Animal Science by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

Predicting population gene frequency from sample data

R. M. Lewis^{1, 2†}, B. Grundy² and L. A. Kuehn¹

¹*Department of Animal and Poultry Sciences (0306), Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061, USA*

²*Sustainable Livestock Systems Group, Scottish Agricultural College, West Mains Road, Edinburgh EH9 3JG, UK*

† E-mail : rmlewis@vt.edu

Abstract

With an increase in the number of candidate genes for important traits in livestock, effective strategies for incorporating such genes into selection programmes are increasingly important. Those strategies in part depend on the frequency of a favoured allele in a population. Since comprehensive genotyping of a population is seldom possible, we investigate the consequences of sampling strategies on the reliability of the gene frequency estimate for a bi-allelic locus. Even within a subpopulation or line, often only a proportion of individuals will be genotype tested. However, through segregation analysis, probable genotypes can be assigned to individuals that themselves were not tested, using known genotypes on relatives and a starting (presumed) gene frequency. The value of these probable genotypes in estimation of gene frequency was considered. A subpopulation or line was stochastically simulated and sampled at random, over a cluster of years or by favouring a particular genotype. Line was simulated (replicated) 1000 times. The reliability of gene frequency estimates depended on the sampling strategy used. With random sampling, even when a small proportion of a line was genotyped (0.10), the gene frequency of the population was well estimated from the across-line mean. When information on probable genotypes on untested individuals was combined with known genotypes, the between-line variance in gene frequency was estimated well; including probable genotypes overcame problems of statistical sampling. When the sampling strategy favoured a particular genotype, unsurprisingly the estimate of gene frequency was biased towards the allele favoured. In using probable genotypes the bias was lessened but the estimate of gene frequency still reflected the sampling strategy rather than the true population frequency. When sampling was confined to a few clustered years, the estimation of gene frequency was biased for those generations preceding the sampling event, particularly when the presumed starting gene frequency differed from the true population gene frequency. The potential risks of basing inferences about a population from a potentially biased sample are discussed.

Keywords: gene frequency, genetic drift, genotype probability, sampling.

Introduction

The increase in the number of genes being identified with important effects on performance traits including fitness (Goldmann *et al.*, 1990; Fujii *et al.*, 1991; Shuster *et al.*, 1992; Wilson *et al.*, 2001) has resulted in the wider incorporation of such quantitative trait loci (QTL) into selection programmes (Hunter *et al.*, 1993; Dawson *et al.*, 1998). Optimal methods to integrate a QTL into a breeding programme for a quantitative trait have been

developed which maximize genetic gain over multiple-generations (Dekkers and van Arendonk, 1998) while restricting the rate of inbreeding (Villanueva *et al.*, 2002; Woolliams *et al.*, 2002). Within such strategies, the gene frequency for the QTL at the start of selection affects its contribution to genetic response (Luo *et al.*, 1997; Dekkers and van Arendonk, 1998). For instance, a favourable allele with a very low starting frequency will contribute little in the early generations of mixed-inheritance

selection (Weller, 2001). Prior knowledge of gene frequency is thus useful when incorporating a QTL into practical breeding programmes.

In large panmictic populations, gene frequencies are inherently stable (Falconer, 1989) and thus can be reliably estimated. However, when a population is subdivided, gene frequencies within the smaller subpopulations or lines may fluctuate or drift erratically because of the random sampling of gametes (Wright, 1931). Since in practice genotype information may only be available on numerically small and perhaps selected lines, or even on subsets of individuals chosen from within lines, population gene frequencies inferred from such sample data may be misleading.

Scenarios exist where individuals of a given genotype, or born within a set of years, may disproportionately contribute to the sample data used to estimate gene frequency. One current example is DNA testing at the prion protein (*PrP*) locus in sheep, polymorphisms of which affect susceptibility to the disease scrapie. Techniques for *PrP* genotyping were established in the early 1990s (Goldmann *et al.*, 1990); *PrP* genotypes are thus restricted to individuals alive since that time, or on ancestors with stored tissues or semen. An experimental flock of Suffolk sheep was used in the development of the *PrP* genotype test (Hunter *et al.*, 1991). Animals affected with scrapie were preferentially genotyped, with confirmation of an association between natural scrapie and a particularly 'susceptible' *PrP* genotype (Hunter *et al.*, 1997). As a corollary, that sampling strategy favoured susceptible genotypes for DNA testing. The intent of the Hunter *et al.* (1991 and 1997) studies was not to infer the frequency of *PrP* alleles in Suffolk sheep in general. However, in commercial testing schemes (Dawson *et al.*, 1998; Department for Environment, Food and Rural Affairs, 2001) where animals with favourable genotypes often achieve a premium in the marketplace, non-random or biased sampling of individuals for genotyping is to be anticipated. Gene frequencies estimated from such data may therefore prove unreliable.

The reliability of an estimate of gene frequency may be improved simply by increasing the number of individuals with genotype information. Financial constraints, however, may limit the extent of DNA testing even within a line; in other words, only a selection of animals from a flock or herd may be tested. Kerr and Kinghorn (1996) suggested that by genotyping relatively few animals, segregation analysis could then be used to unambiguously assign some untested individuals to specific genotypes,

whilst assigning others to the alternative genotypes with varying degrees of confidence (Kinghorn, 1997). Several related approaches to obtain these probable genotypes have been proposed (van Arendonk *et al.*, 1989; Fernando *et al.*, 1993; Kerr and Kinghorn, 1996). Since probable genotypes are derived from known genotypes of relatives, the value of this extra information for estimating gene frequency in the population is unclear particularly if the individuals DNA tested are not chosen at random.

Our first objective in this study was to assess the consequence of sampling strategy for genotype testing on the estimate of gene frequency in a population. The strategies considered were to vary (i) the proportion of individuals genotype tested from within a subpopulation or line, and (ii) the way in which those individuals were chosen for DNA testing. The individuals were chosen at random or in biased fashion where a particular genotype or set of years was favoured. Our second objective was to assess the value of gratuitous genotype information, obtained from segregation analysis, on the quality of the estimate of population gene frequency. This was done for the different sampling strategies considered, and when including only genotypes predicted with a sufficient degree of confidence in the analysis.

Material and methods

Simulation

Stochastic simulation was used to model a closed line (representing a subdivided population such as a flock or herd drawn from a breed) with overlapping generations. A bi-allelic locus, with alleles *A* and *a*, was simulated where the initial frequency of *A* was either 0.25 or 0.5. The genotype of a base line of 50 males and 50 females was created by sampling allele *A* or *a* from a binomial distribution with probabilities *p* and *q* = 1 - *p*, respectively. By Hardy-Weinberg, the resulting genotypic frequencies were *p*², 2*pq*, and *q*² (*AA*, *Aa* and *aa*) among the base animals. The base males were 1 year of age whilst the base females were between 2 and 5 years of age.

In each subsequent generation, five sires and requisite replacement dams (dams were culled at 6 years of age) were chosen at random from the 50 animals of each sex available. These were mated hierarchically with each mating pair producing two individuals with their sex assigned at random. The genotype of the progeny was created by randomly sampling one allele from each parent. Each male was mated to 10 females. The same age structure for sires and dams was used as in the base. Ten generations were simulated (approx. 25 years), and replicated 1000 times. Each replicate was considered as a separate line drawn from the overall population.

An individual's true genotype was defined as the test result when an individual was DNA tested (i.e. a DNA extraction and analysis to establish whether an animal's genotype was *AA*, *Aa* or *aa*). We assumed that the true genotype was known without error for those animals DNA tested; that is, the true genotype was equal to the simulated genotype.

In the scenarios considered, it was assumed that only a proportion of animals were genotype tested. For the remainder, the probabilities that an individual had genotype *AA*, *Aa* or *aa* was obtained using the procedure of Kerr and Kinghorn (1996). Using rules of Mendelian inheritance, these probabilities are calculated from true genotypes, or current estimates of genotype probabilities, on (i) parents and full sibs, (ii) the individual itself, and (iii) its mates and progeny. The Kerr and Kinghorn (1996) approach extends that of van Arendonk *et al.* (1989) to better account for full-sib information in the prediction, following methods proposed by Fernando *et al.* (1993).

A computer algorithm developed at the University of New England, Armidale, Australia (R. J. Kerr and B. P. Kinghorn, personal communication) was used to obtain genotype probabilities, from which gene frequencies were estimated. Besides true genotypes, an initial gene frequency is used in the computations. Calculations are performed iteratively, with information from progressively more distant relatives of an individual used with each further iterate. This process is referred to as 'iterative peeling'. The procedure minimizes the difference between the prior gene frequency estimate and the estimate subsequently obtained from all the data, both known and predicted. This is achieved by updating the gene frequency at the start of iteration as the mean of the starting and ending values from the previous iteration. This procedure is repeated until the gene frequency no longer changes between iterations for a set convergence criterion. The algorithm did not account for inbreeding.

Depending on the scenario tested (described latterly) either actual (simulated) or 'presumed' gene frequencies were input as starting values. The stop criterion used was that the absolute difference between gene frequencies in the current and previous iteration was less than 0.01.

Within each replicate and generation, the gene and genotypic frequency of all animals in the line, and of all animals with true genotypes within the line, was estimated. As well, these statistics were obtained using the weighted average of genotype probabilities of all individuals, or of a defined subset of

individuals. For animals that were DNA tested, the probability of their true genotype (e.g. *AA*) was one, and that of the alternative genotypes (e.g. *Aa* and *aa*) was zero.

The level of certainty that an animal's true genotype was predicted correctly was also obtained within each replicate. This genotype probability index (GPI) is calculated from the dissimilarity between the genotype probabilities for an individual and those obtained by Hardy-Weinberg probabilities from gene frequencies for the base animals (Kinghorn, 1997). The gene frequency in the base is determined from segregation analysis. Where no genotype information was available on an animal (either because it was not DNA tested itself or it lacked any relatives that were), the GPI associated with its genotype probabilities was 0%. Where an animal had been DNA tested, the GPI of its true genotype was necessarily 100%. As an animal's genotype probability for a particular genotype approached one, the GPI of that prediction approached 100%. Further details of the properties of the GPI are provided in Kinghorn (1997). The GPI was used to form subsets of genotype probabilities from which gene frequencies were estimated.

Proportion sampled

Genotype probabilities were obtained for scenarios where the proportion of individuals within a line sampled for genotype testing was varied. As a proportion, from 0.10 to 0.90 of the line was sampled at 0.10 increments. Only a subset of these proportions was considered in some of the scenarios investigated. For instance, with a biased sampling procedure a particular genotype was favoured. In these cases, there were insufficient numbers of animals with the favoured genotype to choose when the proportion of the line sampled was large.

Method of sampling

The sampling of animals was undertaken under three regimes. In the first two, animals were sampled at random either across all years, or all individuals within a cluster of 3 years (one generation). The clustering regime represented a genotype test being introduced at a particular point in time, with concentrated testing immediately thereafter. The third strategy was to introduce a known bias favouring a particular genotype within the sample. This was achieved by sampling the *AA* genotype twice as frequently as would otherwise be the case with random sampling.

In a line of small finite size, gene frequency does not remain constant over time due to genetic drift. The alternative alleles are either fixed or lost as

generations proceed. In order to apply a consistent bias across years, it was therefore necessary to determine the expected gene and thus genotype frequency at any point in time. From this, the target number of individuals to sample from each genotype category within a year could be set. Since the genotypes of all individuals were known *a priori* in the simulated data, it was then possible to categorize animals by genotype and sample the prerequisite number accordingly.

Within individual lines (replicates), a deficiency of individuals within specific genotype categories prevented the bias from being explicitly applied. In these cases, individuals were sampled from the other two genotypes until the target proportion was sampled within a year. With genetic drift, over an extended time period, the *A* allele is fixed in p_0 percent of lines and the *a* allele is fixed in q_0 percent of lines, where p_0 and q_0 are the initial gene frequencies, respectively. The probability that a particular genotype is in surplus (or in deficit) is therefore balanced across lines. This means that taken across all replicates, the bias is applied properly.

The variance of gene frequency

The variance in gene frequency across lines for each proportion sampled, and for each method of sampling considered, was obtained as

$$Var(p) = \frac{\sum_{i=1}^n (p_i - \hat{p})^2}{n-1} \quad (1)$$

where p_i is the gene frequency from each line simulated, $\hat{p}$ is the mean gene frequency across lines, and n is the number of lines ($n = 1000$). The gene frequencies ($p_i; \hat{p}$) were specific to a generation of interest. This statistic is consistent with a mean squared error of prediction.

The variance of the gene frequency (σ_p^2) across lines at generation t could also be obtained deterministically (Falconer, 1989) from

$$\sigma_p^2 = p_0 q_0 \left(1 - \left(1 - \frac{1}{2N_e} \right)^t \right) \quad (2)$$

where N_e is the effective population size. N_e is calculated as $1/(2\Delta F)$, where ΔF is the incremental change in inbreeding defined as $(F_t - F_{t-1})/(1 - F_t)$ and F_t is the inbreeding coefficient of generation t . Note

that the variance includes lines that have become fixed for *A* or *a*.

With equation (2) the variance of gene frequencies amongst lines can be calculated, given that all the individuals within lines have genotype information. From this variance a confidence interval for the true gene frequency p can be set (Weir, 1996). Providing the sample size is reasonably large (see following remarks), there is a 95% chance that the interval

$$p \pm 1.96 \sqrt{\sigma_p^2} \quad (3)$$

includes the true population frequency. Equation (3) is intended for use when all individuals have true genotypes and thus σ_p^2 is known. However, since we obtained probable genotypes on all individuals, it seems reasonable to use the variance of these observations as a first approximation of σ_p^2 . The strength of this assumption is investigated latterly.

Results

Trends in genotype frequencies over generations

Due to genetic drift the frequency of an allele can change erratically over time, finally going to fixation or loss (Falconer, 1989). This process has an impact on the distribution of genotypic frequencies across generations. In Table 1 the across-line mean genotype frequencies, based on the actual (simulated) genotypes for all individuals, are shown by generation. Although the mean gene frequency across lines was consistent with starting values, genotype frequencies diverged from Hardy-Weinberg ($P < 0.05$ by χ^2 goodness-of-fit test).

Proportion sampled (at random)

The precision of estimating gene frequency when different proportions of individuals from the line were genotype tested was investigated. With random sampling, irrespective of the proportion genotype tested, the mean gene frequency in the sample was equal to that in the population ($P < 0.01$). When the genotype probabilities were included when calculating the mean, this result remained true.

As the proportion of individuals genotype tested was reduced, the variance of gene frequency among lines (replicates) increased (Table 2). In contrast, when probable and true genotypes were combined so that all individuals had observations the between-line variance remained relatively constant. This was irrespective of the proportion sampled, and the

Table 1 Mean genotype frequencies by generation for actual (simulated) genotypes on all individuals†

Generation	Initial gene frequency					
	f(A) = 0.5			f(A) = 0.25		
	AA	Aa	aa	AA	Aa	aa
1	0.249	0.501	0.250	0.063	0.375	0.562
2	0.251	0.501	0.248	0.058	0.384	0.559
3	0.256	0.493	0.251	0.067	0.357	0.575
4	0.266	0.474	0.260	0.069	0.346	0.584
5	0.269	0.465	0.266	0.077	0.334	0.589
6	0.280	0.446	0.274	0.078	0.339	0.583
7	0.285	0.434	0.281	0.081	0.326	0.593
8	0.291	0.424	0.285	0.084	0.311	0.605
9	0.298	0.412	0.290	0.090	0.301	0.610
10	0.304	0.400	0.296	0.094	0.292	0.614

† Standard errors range from 0.0009 to 0.0015.

variance obtained was consistent with that when all individuals within a line had true genotypes (i.e. all genotype tested).

Variance of gene frequency. When the effective population size and number of generations of random selection are accounted for, it is possible to compute deterministically (formula (2)) the variance in gene frequency between lines. These deterministic results were compared with those obtained for one scenario (Table 3) where, as a proportion, 0.10 of animals had been randomly sampled for genotyping and the remainder had probable genotype information; this scenario was chosen to represent an extreme case. When this was done, the predicted variance in gene frequency (from formula (2)) was very consistent with that obtained for the combined true and probable genotypes.

The variance in gene frequency in generation 10 for an initial gene frequency of 0.5 was predicted to be 0.051 (Table 3). Therefore, one would expect the gene frequency for a particular line to lie between 0.05 and 0.95, given a 95% confidence interval (formula (3)). This means that nearly the full range of gene frequency is possible for a single line. As evidence of this in practice, in the 1000 replicates of the simulation, 2% went to fixation before generation 10.

Method of sampling (non-random)

Sampling from a set of years. To simulate the scenario of a genotype test becoming available at a specific point in time, the sample of individuals chosen for genotype testing was restricted to a cluster born in generation 4. By sampling one generation, the same number of individuals was genotype tested as with, proportionally, 0.10 random sampling of the entire

line. Two general scenarios were considered. The first was where the initial gene frequency was equivalent to the actual gene frequency for the population (frequency of A of either 0.5 or 0.25). The second was where the presumed initial gene frequency was lower than the actual initial gene frequency ($f(A) = 0.1$). Probable genotypes were obtained for the remaining individuals.

The average gene frequency for allele A at generation 1 and 10 is presented in Table 4 for these scenarios. Since all individuals were sampled in generation 4 their gene frequency at that generation was known explicitly. It can be seen that the gene frequency in generation 10 was unaffected by the starting gene

Table 2 Observed variance of gene frequency between lines at generation 10 when gene frequency was obtained from genotypes for all individuals (combination of true and probable genotypes) or only from those individuals sampled (true genotypes only) within the line, when different proportions of individuals were sampled †

Proportion of line sampled	Initial gene frequency			
	f(A) = 0.5		f(A) = 0.25	
	All‡	Sample§	All‡	Sample§
0.10	0.0514	0.1056	0.0430	0.0829
0.30	0.0511	0.0993	0.0440	0.0653
0.50	0.0514	0.0710	0.0428	0.0534
0.70	0.0520	0.0631	0.0398	0.0501
0.90	0.0499	0.0513	0.0436	0.0444

† Variances obtained from equation (1) as described in text.

‡ Gene frequency within line obtained from true (individuals sampled and thus genotype tested) and probable genotypes.

§ Gene frequency within line obtained from true genotypes on the proportion sampled for genotype testing.

Table 3 Observed variance of gene frequency between lines and that obtained using deterministic methods, when a proportion of 0.1 was sampled at random (results from generation 6 to 10 are shown)

Generation	Initial gene frequency			
	f(A) = 0.5		f(A) = 0.25	
	Observed†	Deterministic‡	Observed†	Deterministic‡
6	0.0265	0.0233	0.0218	0.0214
7	0.0300	0.0312	0.0299	0.0285
8	0.0372	0.0376	0.0360	0.0350
9	0.0434	0.0439	0.0370	0.0376
10	0.0514	0.0511	0.0430	0.0430

† Variance obtained from equation (1) as described in text where gene frequencies for lines were estimated from genotypes on all individuals, both true and probable.

‡ Variance determined using equation (2) as described in text.

Table 4 Mean gene frequencies at generations 1 and 10 when the presumed initial gene frequency was equal to or lower than the actual initial gene frequency†

Initial gene frequency		Generation	Mean‡
Presumed	Actual		
0.5	0.5	1	0.498
		10	0.510
0.25	0.25	1	0.249
		10	0.249
0.1	0.5	1	0.430
		10	0.511
0.1	0.25	1	0.222
		10	0.240

† Generation 4 was sampled for genotype testing, which corresponds with, proportionally, 0.10 of the line. The gene frequency of a line was estimated from genotypes on all individuals, both true and probable.

‡ Standard errors range from 0.0310 to 0.0520.

frequency but that the ancestral generations were affected to an extent. When the presumed starting gene frequency was too low, the observed gene frequency at generation 1 was 0.86 and 0.89 times that of the actual gene frequency of 0.5 and 0.25, respectively.

Biased sampling. A bias was applied where twice the number of AA individuals was sampled as would otherwise have been the case with random sampling. This was done with, as a proportion, 0.10, 0.30 and 0.50 of the line sampled for genotype testing. With a biased sample, the gene frequencies based on true genotypes simply reflected that bias (Table 5). This was the case regardless of the proportion genotype tested.

Probable genotypes were grouped according to the reliability of their estimate. When information from untested individuals whose genotypes were

Table 5 Actual (simulated), sample and estimated means of gene frequencies at generation 10 when various proportions of individuals were sampled for genotype testing (a bias was applied so that genotype AA was sampled at twice the rate of that under random sampling)

Initial gene frequency	Proportion sampled	Mean gene frequency†			
		Actual	Sample	Estimated‡	
				GPI > 0%	GPI > 90%
0.25	0.10	0.250	0.294	0.276	0.290
	0.30	0.249	0.300	0.278	0.290
	0.50	0.250	0.295	0.278	0.290
0.5	0.10	0.506	0.596	0.568	0.587
	0.30	0.495	0.607	0.560	0.582
	0.50	0.498	0.593	0.567	0.586

† Standard errors range from 0.0301 to 0.0542.

‡ Combination of true and probable genotypes in the estimate of the mean, where the probable genotypes were included on all untested individuals [genotype probability index (GPI) > 0%] or only for those untested individuals whose probable genotypes were predicted reliably (GPI > 90%).

predicted with high reliability (GPI > 90%) was combined with that from individuals with true genotypes, the estimated gene frequency was similar to that obtained from the sample alone. This was because the untested individuals included were necessarily close relatives to those with true genotypes. When information from individuals with probable genotypes that were estimated less reliably was also considered, the estimated gene frequency more closely reflected the actual initial gene frequency. However, even when the probable genotypes on all individuals were included (GPI > 0%), the frequency of the favoured allele continued to be overestimated (by at least, as a proportion, 0.10 in the scenarios considered).

Discussion

For many reasons, it is useful to describe populations in terms of their gene frequencies. However, there are problems in doing so. Firstly, gene frequencies change over time. Secondly, a single line from within a population may be inadequate in describing the true gene frequency of the population. And thirdly, in most practical applications it is likely that only a subsample of individuals from within lines will have genotype information available from which gene frequencies are inferred. In this study we have focused on this third issue, investigating reasons why gene frequency estimates from a subsample may be misleading, and a possible way to improve the reliability of inferences about a population based on such a subsample.

Proportion sampled (at random)

When individuals genotype tested within lines were chosen at random, the mean gene frequency of the population was well estimated. This was true even when a small proportion of the line was genotype tested (0.10). However, when smaller proportions of individuals were genotype tested, the variance of gene frequency between lines increased. This increase in variance reflects a less precise estimate of the gene frequency within individual lines.

By using segregation analysis to obtain the genotype probabilities of untested individuals, as proposed by van Arendonk *et al.* (1989), Fernando *et al.* (1993), and Kerr and Kinghorn (1996), it was possible to get a closer approximation to the true variance of gene frequency between lines (Table 2). By effectively increasing the sample size, segregation analysis increases the precision of the gene frequency estimate within lines, thereby lowering the variance between lines. This result appeared to be independent of the proportion of individuals genotyped, at least within the boundary of the scenarios simulated in this study.

Genetic drift. As anticipated (Falconer, 1989), genetic drift induced a deviation from Hardy-Weinberg equilibrium (Table 1). Thus it is not appropriate to assume that a gene frequency available for a single line adequately describes the population. From a starting gene frequency of 0.5, nearly the entire spectrum of gene frequencies was observed within individual lines after 10 generations.

Method of sampling (non-random)

With random sampling of individuals within lines for genotype testing, the mean gene frequency of the population could be well estimated. However, this was not the case in two scenarios where the sampling was not random: sampling restricted to a limited number of generations (clustering) and sampling favouring a particular genotype (biased sampling). Of concern, is that both of these methods of non-random sampling are likely to occur in practice (albeit some genotyping following a cluster is likely) and both have an impact on our estimate of gene frequency.

Sampling from a set of years. When sampling was confined to a few clustered years, the estimate of the gene frequency was biased for those generations preceding the sampling event, particularly when the starting gene frequency differed from the true gene frequency (Table 4). With random sampling across generation, the prediction of genotypes and resulting gene frequencies were insensitive to the starting value assumed for gene frequency. The reason for this difference between random and clustered genotyping is the way the frequencies were estimated.

Two sources of information were used to predict genotype probabilities and thus determine gene frequencies: the presumed initial gene frequency and the genotype information on tested individuals (Kerr and Kinghorn, 1996). As noted previously, the algorithm used minimized the difference between the gene frequency in the current and prior iteration of the segregation analysis, with the prior updated each iteration. Consequently, it would seem little knowledge of the starting gene frequency would be necessary since genotyped individuals would dominate the updating of the prior. However, when the sampling was confined to a specific generation, it became apparent that the starting value was important (Table 4). Van Arendonk *et al.* (1989) reported a similar finding. Moreover, the impact on probable genotypes (from which the gene frequencies were calculated) was asymmetric with regard to time. Past generations were more influenced than future generations, and an example (results not shown) where sampling took place in

generation 10 revealed that the problem increased the greater the time gap.

The reason for this finding is that all individuals born in future years will be descendants of the clustered generation but not *vice versa*. The impact of this is threefold. Firstly, only through pedigree links with the cluster can predictions be made. Secondly, predictions for unrelated individuals are solely dependent on the prior information that is given to determine the gene frequency. As a consequence, thirdly, one cannot assess change in gene frequency over time. So for the same convergence criterion, it becomes more difficult to determine the gene frequency if the data are clustered. This is because it is difficult to move away from the prior if only distant genetic links to individuals with genotype information is present.

Biased sampling. If a sampling strategy favours a particular genotype, unsurprisingly the estimate of gene frequency is biased towards the allele favoured. In using probable genotypes the bias is lessened but the estimate of gene frequency still reflects the sampling strategy rather than the actual gene frequency.

By using probable genotypes, the effects of bias in a sample on the estimate of the population gene frequency can be reduced although not removed; the extent of the benefit is also likely to be dependent on the value of the prior gene frequency estimate. The practical difficulty is that the presence or extent of bias is likely unknown where genotyping is done commercially. If a biased sample has been taken, then it should not be used to estimate the gene frequencies for the population or breed concerned. Implementing a strategy to randomly sample a relatively small proportion of animals within breeds for genotype testing may be an important safeguard to minimize risks of selection policies being based on potentially unreliable estimates of gene frequencies.

Acknowledgements

The financial support of the Scottish Executive Environment and Rural Affairs Department (SEERAD) for this research is gratefully acknowledged. We are also thankful to Brian Kinghorn for providing software used in the investigations, and to him and Beatriz Villanueva for their useful comments on the manuscript.

References

Arendonk, J. A. M. van, Smith, C. and Kennedy, B. W. 1989. Method to estimate genotype probabilities at individual loci in farm livestock. *Theoretical and Applied Genetics* **78**: 735-740.

Dawson, M., Hoinville, L. J., Hosie, B. D. and Hunter, N. 1998. Guidance on the use of PrP genotyping as an aid to the control of clinical scrapie. *Veterinary Record* **142**: 623-625.

Dekkers, J. C. M. and Arendonk, J. A. M. van. 1998. Optimizing selection for quantitative traits with information on an identified locus in outbred populations *Genetical Research, Cambridge* **71**: 257-275.

Department for Environment, Food and Rural Affairs. 2001. *National scrapie plan for Great Britain*. Available at: <http://www.defra.gov.uk/animalh/bse/bse-science/scrapie/nsp/brochure.pdf>.

Falconer, D. S. 1989. *Introduction to quantitative genetics*. Longman Scientific and Technical, England.

Fernando, R. L., Stickler, C. and Elston, R. C. 1993. An efficient algorithm to compute the posterior genotypic distribution for every member of a pedigree without loops. *Theoretical and Applied Genetics* **87**: 89-93.

Fujii, J., Otsu, K., Zorzato, F., DeLeon, S., Khanna, V. K., Weiler, J. E., O'Brien, P. J. and McLennan, D. H. 1991. Identification of a mutation in porcine ryanodine receptor associated with malignant hyperthermia. *Science* **253**: 448-451.

Goldmann, W., Hunter, N., Foster, J. D., Salbaum, J. M., Beyreuther, K. and Hope, J. 1990. Two alleles of a neural protein gene linked to scrapie in sheep. *Proceedings of the National Academy of Sciences of the United States of America* **87**: 2476-2480.

Hunter, N., Foster, J. D., Goldmann, W., Stear, M. J., Hope, J. and Bostock, C. 1991. Restriction fragment length polymorphisms of the scrapie-associated fibril protein (PrP) gene and their association with susceptibility to natural scrapie in British sheep. *Journal of General Virology* **72**: 1287-1292.

Hunter, N., Goldmann, W., Benson, G., Foster, J. D. and Hope, J. 1993. Swaledale sheep affected by natural scrapie differ significantly in PrP genotype frequencies from healthy sheep and those selected for reduced incidence of scrapie. *Journal of General Virology* **74**: 1025-1031.

Hunter, N., Moore, L., Hosie, B. D., Dingwall, W. S. and Greig, A. 1997. Association between natural scrapie and PrP genotype in a flock of Suffolk sheep in Scotland. *Veterinary Record* **140**: 59-63.

Kerr, R. J. and Kinghorn, B. P. 1996. An efficient algorithm for segregation analysis in large populations. *Journal of Animal Breeding and Genetics* **113**: 457-469.

Kinghorn, B. P. 1997. An index of information content for genotype probabilities derived from segregation analysis. *Genetics* **145**: 479-483.

Luo, Z. W., Thompson, R. and Woolliams, J. A. 1997. A population genetics model of marker-assisted selection. *Genetics* **146**: 1173-1183.

Shuster, D. E., Kehrli Jr, M. E., Ackermann, M. R. and Gilbert, R. O. 1992. Identification and prevalence of a genetic defect that causes leukocyte adhesion deficiency in Holstein cattle. *Proceedings of the National Academy of Sciences of the United States of America* **89**: 9225-9229.

Villanueva, B., Pong-Wong, R. and Woolliams, J. A. 2002. Marker assisted selection with optimised contributions of

the candidates to selection. *Genetics, Selection, Evolution* **34**: 679-703.

Weir, B. S. 1996. *Genetic data analysis II*. Sinauer Associates Inc., Sunderland, MA.

Weller, J. I. 2001. *Quantitative trait loci analysis in animals*. CABI Publishing, Oxford.

Wilson, T., Wu, X. -Y., Juengel, J. L., Ross, I. K., Lumsden, J. M., Lord, E. A., Dodds, K. G., Walling, G. A., McEwan, J. C., O'Connell, A. R., McNatty, K. P. and Montgomery, G. W. 2001. Highly prolific Booroola sheep have a mutation in the intracellular kinase domain of bone morphogenetic protein IB receptor (ALK-6) that is expressed in both

oocytes and granulosa cells. *Biology of Reproduction* **64**: 1225-1235.

Woolliams, J. A., Pong-Wong, R. and Villanueva, B. 2002. Strategic optimisation of short- and long-term gain and inbreeding in MAS and non-MAS schemes. *Proceedings of the seventh world congress on genetics applied to livestock production, Montpellier*. CD-ROM communication no. 23-02.

Wright, S. 1931. Evolution in Mendelian populations. *Genetics* **16**: 97-159.

(Received 12 November 2002—Accepted 18 July 2003)