Sampling Method For Estimating Neutral Allele Frequency In A Pedigreed Population

T.Honda[*], S. Sasazaki[†], K. Oyama[*], T. Nomura[‡], and F. Mukai[§]

Introduction

Estimates of allele frequencies at neutral DNA marker loci are fundamental information for characterizing a livestock breed and establishing a conservation program for a set of breeds (e.g. Bennewitz *et al.* (2006)). FAO (1998) presented a guideline that allele frequency in a breed should be estimated from at least 25 randomly sampled animals. For a pedigreed population, several authors have proposed methods for sampling individuals with no common ancestors within several generations in the past to remove possible bias in the estimation of allele frequencies due to inclusion of related individuals (e.g. Konovalov and Heg (2008)). However, it is unknown how the pedigree information can be used to obtain the most representative samples for allele frequencies in the population. In this study, supposing a population with *a priori* known pedigree, we propose a sampling method, which is expected to minimize the estimation error of neutral allele frequency.

Material and methods

Theory. Consider a population of N individuals with known pedigree. Let q be the frequency of an arbitrary allele in the population, and $\hat{q}$ be the allele frequency estimated from n individuals sampled from the population. We consider a problem of finding a set of n individuals that minimizes $E[(\hat{q}-q)^2]$. From a theoretical standpoint underlying genetic management of pedigreed animals, the allele frequency q is regarded as a product of genetic drift accumulated from the founder generation (Caballero and Toro (2000)). On the basis of this view, the problem can be formulated as an optimization problem of minimizing the objective function

$$\bar{f}_N + (1/n^2)X - (2/nN)Y \qquad (1)$$

under a constraint of $\sum_{r=1}^{N} z_r = n$, where z_r is a decision variable taking the value 1 if individual r is sampled or the value 0 if not sampled, $\bar{f}_N$ is average coancestry among N individuals relative to the founder generation, and X and Y are functions of the coancestries and the decision variables.

[*] Food Resources Education and Research Center, Graduate School of Agricultural Science, Kobe University, Hyogo, Japan
[†] Graduate School of Agricultural Science, Kobe University, Hyogo, Japan
[‡] Faculty of Life Sciences, Kyoto Sangyo University, Kyoto, Japan
[§] Wagyu Registry Association, Kyoto, Japan

Furthermore, with the frequencies of hypothetical founder alleles, it can be theoretically shown that the objective function (1) is rewritten as

$$((N-n)/N)^2 E\left[\sum_{i=1}^{N_f}\sum_{j=1}^{2}(q_{ij\cdot s}-q_{ij\cdot ns})^2\right],$$

where N_f is the number of founders, $q_{ij\cdot s}$ and $q_{ij\cdot ns}$ are the frequencies of the hypothetical allele j (= 1 or 2) of founder i (= 1 to N_f) in the sampled and the non-sampled individuals, respectively. Thus, our proposed method is equivalent to minimizing the expectation of squared Euclidian distance of founder allele frequencies between the sampled and the non-sampled individuals. In this context, we refer to this method as minimum distance (MD) method.

Simulation. In order to assess the performance of the proposed method, we carried out computer simulation. Twenty five or fifty (L=25 or 50) unlinked neutral marker loci each with n_{af} (=5 or 10) alleles were simulated. Genotype of founders at each locus was generated by assigning alleles randomly sampled from an infinite (conceptual) gene pool, in which the frequency of each allele is $1/n_{af}$. The population, for which the allele frequency is estimated, was obtained by random mating among N_m=5 males and N_f=100 or 500 females over 10 generations. Through the progress of generations, pedigree of the animals was recorded to compute the coancestry. From the population in the final generation, n (=10, 25, or 50) animals were sampled by the proposed method. For comparison, two additional sampling methods were also examined: random sampling (RND) and the minimum coancestry sampling (MC), which samples n animals with the least relationships via minimizing an objective function

$$\sum_{r=1}^{N}\sum_{s=1}^{N} z_r z_s f_{rs} / n^2$$

under a constraint of $\sum_{r=1}^{N} z_r = n$, where f_{rs} is coancestry coefficient between individuals r and s. Optimizations in MC and MD were conducted with the simulated annealing algorithm (Press *et al*. (1992)).

Performance of the three sampling methods were evaluated by an indicator

$$RASE(q)=\sqrt{\sum_{l=1}^{L}\sum_{a=1}^{n_{af}}(\hat{q}_{l,a}-q_{l,a})^2 / \sum_{l=1}^{L} n_{all,l}},$$

where $n_{all,l}$ is the number of alleles segregating in the final generation at marker locus l (=1 to L), and $q_{l,a}$ and $\hat{q}_{l,a}$ are frequencies of the allele a (=1 to n_{af}) at locus l in the population and the sampled animals, respectively. The average expected heterozygosity and the average number of alleles per locus were also compared between the population and the samples. Simulation for each combination of the parameters was run with 100 replicates, and the results were evaluated with the averages over all the replicates. Only the results from the parameters of N_m=5, N_f=500, L=25, and n_{af}=5 are presented, since essentially the same tendencies were observed in all combinations of variables.

Numerical analysis with actual data. Microsatellite data and pedigree records of 251 dams in a closed herd of the Japanese Black cattle were used for numerical analysis. The dams were genotyped at 20 microsatellite marker loci (Sasazaki *et al*. (2004)). From 251 dams,

n=10, 25 or 50 dams were sampled by the three sampling methods, and the performances of those methods were also evaluated in an analogous way.

Results and discussion

Table 1 shows the average expected heterozygosity and the average number of alleles per locus in the simulated population and the actual cattle population, and the corresponding values in the samples obtained by the three methods. In both populations, MC consistently gave higher expected heterozygosity and larger number of alleles than RND and MD, which is a favorable property for establishing a conserved population.

Table 1: Average expected heterozygosity (H_e) and average number of alleles per locus (N_a) in the simulated population and the actual cattle population, and the corresponding estimates ($\hat{H}_e$ and $\hat{N}_a$) in the n individuals sampled by the random (RND), minimum coancestry (MC), and the minimum distance (MD) methods. n=number of sampled individuals.

	RND			MC			MD		
	n=10	25	50	n=10	25	50	n=10	25	50
	Simulated population: H_e =0.621, N_a =4.66								
$\hat{H}_e$	0.588	0.611	0.617	0.611	0.623	0.628	0.599	0.613	0.618
$\hat{N}_a$	3.69	4.05	4.22	3.79	4.09	4.25	3.74	4.07	4.25
	Actual cattle population: H_e =0.594, N_a =4.75								
$\hat{H}_e$	0.567	0.582	0.590	0.611	0.628	0.619	0.562	0.574	0.580
$\hat{N}_a$	3.46	3.90	4.17	4.00	4.40	4.40	3.50	3.80	4.10

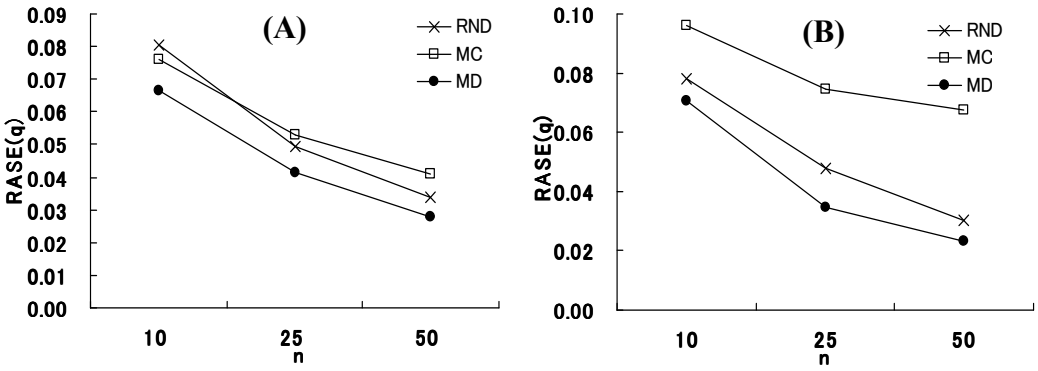

Figure 1: *RASE*(q) under the three sampling methods (RND, MC and MD) in the simulated population (A), and the actual cattle population (B). n=number of sampled individuals.

In Figure 1 (A), *RASE*(*q*) in the simulation is compared among the three sampling methods. Similar comparison is also made for the actual cattle data in Figure 1 (B). Although MC would be preferred for the purpose of establishment of a conserved population (Table 1), the larger values of *RASE*(*q*) in the both figures indicate that this method is not appropriate for sampling individuals to estimate allele frequencies in the current population. This can be more clearly seen in Figure 1 (B), in which MC gave a much higher *RASE*(*q*) than RND and MD. The low performance of MC would be due to the pedigree structure in the actual data. Despite the high average coancestry (0.182) among the 251 dams, MC gave samples with the average coancestries of 0.148, 0.137, and 0.138 when n=10, 25, and 50, respectively. By this reduction of the average conacestries, the amount of genetic drift accumulated in the sampled individuals was minimized (cf. Caballero and Toro (2000); Saura *et al*. (2008)). Thus, the sample obtained by MC more or less resembles the founder population in the allele frequencies, and could not be the most representative sample of the current population.

Contrary, *RASE*(*q*) under MD were consistently smaller than those under RND and MC in both the simulated and the actual populations, indicating that individuals sampled by MD can describe the genetic composition of the current population more accurately than the other two methods. Correlation coefficients between the objective functions (computed by equation (1)) and *RASE*(*q*) for the simulated populations ranged from 0.701 to 0.960, suggesting that the proposed objective function can be a good indicator for *RASE*(*q*).

Conclusion

Results from the simulation and numerical analysis with the actual cattle data showed that individuals sampled by MD could represent the genetic composition (allele frequencies) at neutral loci in the population of interest more accurately than RND and MC.

References

Bennewitz, J., Kantanen, J., Tapio, I. *et al*. (2006). *Genet. Sel. Evol*., 38: 201-220.

Caballero, A. and Toro, M.A. (2000). *Genet. Res. Camb*., 75: 331-343.

FAO (1998). Secondary Guidelines for Development of National Farm Animal Genetic Resources Management Plans.

Konovalov, D.A. and Heg, D. (2008). *Proc. 6th APBC*, 321-332.

Press, W.H., Teukolsky, S.A., Vetterling, W.T. *et al*. (1992). Numerical Recipes in Fortran.

Sasazaki, S., Honda, T., Fukushima, M. *et al*. (2004). *Asian-Aust. J. Anim. Sci.*, 17: 1355-1359.

Saura, M., Pérez-Figueroa, A., Fernández, J. *et al*. (2008). *Conserv. Biol*., 22: 1277-1287.