The statistical power of the hypothesis testing for the elucidation of genetic population structure in the North Pacific minke whales using allele frequency data

Hiroshi Okamura[1] and Mutsuo Goto[2]

[1] *National Research Institute of Far Seas Fisheries, 5-7-1 Orido , Shimizu, Shizuoka 424-8633, Japan.*

[2] *The Institute of Cetacean Research, 4-18 Toyomi-cho, Chuo-ku, Tokyo 104-0055, Japan.*

ABSTRACT

Genetic structure between sub-areas of the North Pacific minke whale was examined using allozyme allele frequency data from Adh-1 loci. Using the mixed model of J-stock and O-stock as a hypothesized model for W-stock, enabled us to calculate the estimates of the statistical power of hypothesis testing for determining whether or not there is significant difference between samples from sub-area 7 and sub-area 9. Furthermore, the conversion of mixing rate to dispersal rate promoted a better understanding of the result calculated and gave us a reliable guide for the relation between gene flow and the statistical power. Our results indicated the possibility that the genetically different stock exists in sub-area 9 was little.

INTRODUCTION

When the result of hypothesis testing did not reject the null hypothesis, we can consider two situations. One is that the null hypothesis is true and the other is that the null hypothesis is not to be rejected because of low statistical power although the alternative hypothesis is true. Therefore when the null hypothesis cannot be rejected, we need to estimate the statistical power to say whether or not we gave correct decision when the null hypothesis was true. The explicit representation of the alternative hypothesis is needed to estimate the statistical power.

The result of many analyses using data collected from JARPN surveys indicated no evidence of the existence of W-stock in the North Pacific. However documents that estimated the statistical power is rare because the alternative hypothesis could be not specified. We try to estimate the statistical power using allozyme allele frequency data and the simulation

method in this paper. Using allele frequency data makes analyses simpler than using mtDNA data.

Taylor *et al.* (1996b) proposed use of dispersal rate for the explicit representation of the alternative hypothesis. When we estimated the statistical power by simulation, we made an attempt to consider the result through dispersal rate.

MATERIALS AND METHODS

The genotype frequency data employed in this study are shown in Table 1. These data were collected from JARPN surveys between 6 years of 1994 to 1999 (Wada, this meeting). We pay attention to the particular allele, *d* and *h*, in allele frequency data from Adh-1 loci because they are main alleles in the North Pacific minke whales and such limitation surely gives us simple calculations and clear results.

Almost all methods below are introduced in Nei (1987) unless we especially mention references.

The relative frequencies of genotypes X_{dd}, X_{dh}, X_{hh} in each sub-area (6, 7 and 9) are calculated from the data in Table 1. Using their Xs, the relative gene frequency of *d* is given by

$$x_d = X_{dd} + X_{dh}/2 \tag{1}$$

while the gene frequency of *h* is $x_h = 1 - x_d$.

The fixation index (F) is used here to know whether or not the observed genotype frequencies in each sub-area follow Hardy-Weinberg equilibrium. When n individuals are examined, and n_{11} *dd*, n_{12} *dh*, and n_{22} *hh* individuals are observed, the maximum likelihood estimate of F is given by

$$F = \frac{4n_{11}n_{22} - n_{12}^2}{(2n_{11} + n_{12})(2n_{22} + n_{12})}. \tag{2}$$

The deviation of F from 0 can be tested by $\chi^2(1) = n\hat{F}^2$ with one degree of freedom.

Consider a population divided into two subpopulations. We disregard the effect of population size. When the harmonic mean of two subpopulations is more than 50, the gene frequency difference between two sub-areas is calculated by

$$\hat{F}_{ST} = 1 - \frac{\bar{x}_d - \overline{x_d^2}}{\bar{x}_d - \bar{x}^2} \tag{3}$$

where $\bar{x}_d = \sum_{i=1}^{2} x_{id} / 2$ is the mean gene frequency, x_{id} is the frequency of d in the i th subpopulation and $\overline{x_d^2} = \sum_{i=1}^{2} x_{id}^2 / 2$.

The dispersal rate (the rate per generation at which each of subpopulations exchanges genes) is estimated by the Wright's island model (Nei 1987, Taylor 1996b, 1999) :

$$F_{ST} = \frac{1}{1+4N_e d} \tag{4}$$

where N_e is the effective population size.

If there exists W-stock in sub-area 9, which is other than O-stock, we need the explicit model of W-stock to estimate the statistical power of analyses. Here it is assumed to be the mixed model of J-stock and O-stock as a model of W-stock. The mixed model with the mixing rate of π is given by

$$\pi \times \mathrm{J} + (1-\pi) \times \mathrm{O} \tag{5}$$

where the representative of J-stock is the sample from sub-area 6 and O-stock is from sub-area 7. Some samples are drawn from this model and the sample size of sub-area 7 is left alone (n = 138). The difference between samples from the mixed model given above and samples from sub-area 7 is investigated by χ^2-test. To repeat this operation 100 times and count the number of the possession of statistical difference gives us one estimate of the statistical power for χ^2-test of O-stock and the hypothesized W-stock when n samples are drawn from W-stock. Furthermore, when we repeat 100 repetitions 100 times, we gets the estimates of the expectation of the statistical power and the standard error of that. The differentiation between O-stock and the hypothesized W-stock (the mixed model of J and O) is given by equation (3) in the form of F_{ST}. The mixing rate is translated into the dispersal rate using the F_{ST} and equation (4). The simulation method used here to estimate the statistical power is similar to that of Pastene *et al.* (1996).

RESULTS

The deviation from Hardy-Weinberg equilibrium of sub-area 6, 7 and 9 was calculated from equation (2) using data in Table 1. The estimates of F in each sub-area were 0.0768, 0.2857 and –0.0556. The deviations of F from 0 were all not statistically significant (nF^2 = 0.8142, 3.6735 and 0.5500 ($\chi^2(1)$=3.8415 at the significance level of 0.05)).

F_{ST} between sub-area 7 and 9 calculated using equation (3) was 0.00025. It is needed to estimate effective population size for the estimate of dispersal rate. We used the mean of the

two areas as the abundance and assume approximately 60% are breeding males and females in imitation of Taylor *et al.* (1996b). Then we got the estimated N_e is almost equal to 3000 ((2202 + 8264)/2×0.6=3139.8). The estimate of dispersal rate obtained using the estimated F_{ST}, N_e and equation (4) was 0.33062. The estimates of F_{ST} and dispersal rate between sub-area 6 and 7 were 0.40723 and 0.00012 in the same manner (N_e = 3000).

The result that we estimated the statistical power for χ^2 test at the significance level of 0.05 using the simulation model of equation (5) is shown in Table 2. Some appropriate values between 0.1 and 0.3 were used as the mixing rates. The estimates of F_{ST} and dispersal rate between sub-area 7 and each mixed model with different mixing rate π were obtained using equation (3) and (4) etc. The statistical powers for these dispersal rates and sample sizes are shown in Table 3.

DISCUSSION

Taylor (1999) categorized the degree of gene flow into a stocks into low, medium and high as follows: low is less than one disperser per generation, medium is more than one disperser but less than 1%/generation, and high is any level greater than 1%/generation. Because the present sample size of sub-area 9 is 178, it is probably considered that the statistical power is large if the gene flow is medium or low when we look through Table 3. When the gene flow is high, we will need enormous sample size of sub-area 9 to get enough statistical power if sample size of sub-area 7 does not increase from now on.

Some uncertainties were not included in these analyses. The result may be sensitive to the violation of some assumptions. Taylor *et al.* (1996a) examined the influence on the violation of assumption of equal abundance and the effect on the stochastic model using simulation model. The result from our analyses also needs to be strengthened by approach like that we are interested in the development of model that took satisfactorily into account the relation between geographical distance and dispersal rate, too.

The difference between sub-area 7 and 9 is very small. The dispersal rate of two sub-areas is estimated to be 0.33062, whose value is very high. Because enormous sample size will be needed to detect this subtle difference and we may have no meaning of distinguishing such slight difference for management objectives, we can conclude the genetically identical stock in sub-area 7 and sub-area 9 even though imperfection of our analyses is to some extent. The expectation of the result from the analyses including the present available information indicates at least that the present statistical power is large and many analyses other than that of allele frequency data shows no statistically significant difference. We

think the possibility of existence of two genetically different stocks in the North Pacific, W-stock, is little in view of the estimated dispersal rate from the present data, large statistical power from simulation model and information from other analyses.

ACKNOWLEDGEMENTS

We greatly appreciate that S. Wada (NRIFS) kindly gave us leave to analyze his data. We also appreciate K. Hiramatsu (NRIFSF), S. Kawahara (NRIFSF), H. Hatanaka (NRIFS), L.A. Pastene (ICR), L. Walloe (University of Oslo) and D.S. Butterworth (University of Cape Town) for helpful discussions and comments. We are deeply indebted to the scientists and crew involved in the JARPN research cruise for their great efforts.

REFERENCES

Nei,M. 1987. Molecular Evolutionary Genetics, Columbia University Press, New York.

Pastene,L.,A., Kishino,H. and Goto,M. 1996. Preliminary RFLP analysis of mitochondrial DNA in the Antarctic minke whale from Areas III and VI. Paper SC/48/SH13 presented to the IWC Scientific Committee (unpublished), 19 pp.

Taylor,B.L., Chivers,S.J., Sexton,S. and Dizon,A.E. 1996a. Simulations to translate between dispersal rates and measures of genetic differentiation. Paper SC/48O13 presented to the IWC Scientific Committee (unpublished), 8pp.

Taylor,B.L., Brownell,B., Perrin,B., DeMaster,D. and Lankester,K. 1996b. Consideration of the hypotheses concerning the O and W stocks of North Pacific minke whales. Working paper SC/48/WP7 presented to the IWC Scientific Committee (unpublished), ?pp.

Taylor,B.L. 1999. Refining the definition of stocks: a discussion paper. Paper SC/51/RMP23 presented to the IWC Scientific Committee (unpublished), 8 pp.

Wada,S. The stock structure of the western North Pacific minke whale based on analyses of allozymes. SC/F2K/J12 in this meeting.

Table 1. Allele frequency data from Adh-1 loci.

Area	sub-area 6	sub-area 7	sub-area 9
allele			
hh	1	67	85
dh	4	55	79
dd	40	16	14
Total	45	138	178

Table 2. The statistical power in relation to the mixing rate π and the sample size n
(the standard error in pairenthesis)

π	0.1	0.11	0.12	0.13	0.14	0.15		0.2		0.3
n										
10	0.112	0.113	0.111	0.111	0.110	0.249		0.2526		0.518
	(0.035)	(0.030)	(0.031)	(0.036)	(0.030)	(0.047)		(0.038)		(0.058)
50	0.211	0.313	0.309	0.314	0.435	0.555		0.7818		0.995
	(0.044)	(0.045)	(0.052)	(0.042)	(0.044)	(0.052)		(0.040)		(0.008)
100	0.307	0.382	0.477	0.562	0.655	0.736		0.9734		1.000
	(0.045)	(0.044)	(0.051)	(0.047)	(0.052)	(0.045)		(0.018)		(0.000)
150	0.335	0.398	0.561	0.703	0.773	0.825	···	0.9955	···	1.000
	(0.045)	(0.045)	(0051)	(0.046)	(0.041)	(0.036)		(0.007)		(0.000)
200	0.390	0.524	0.642	0.761	0.856	0.922		0.9992		1.000
	(0.044)	(0.046)	(0.048)	(0.042)	(0.036)	(0.027)		(0.003)		(0.000)
1000	0.633	0.866	0.969	–	–	–		–		–
	(0.052)	(0.036)	(0.016)	–	–	–		–		–
5000	0.891	–	–	–	–	–		–		–
	(0.031)	–	–	–	–	–		–		–

– : not calculated

Table 3. The statistical power in relation to the dispersal rate d and the sample size n
(the standard error in pairenthesis)

d	0.020	0.016	0.014	0.012	0.010	0.009		0.005		0.002
n										
10	0.112	0.113	0.111	0.111	0.110	0.249		0.2526		0.518
	(0.035)	(0.030)	(0.031)	(0.036)	(0.030)	(0.047)		(0.038)		(0.058)
50	0.211	0.313	0.309	0.314	0.435	0.555		0.7818		0.995
	(0.044)	(0.045)	(0.052)	(0.042)	(0.044)	(0.052)		(0.040)		(0.008)
100	0.307	0.382	0.477	0.562	0.655	0.736		0.9734		1.000
	(0.045)	(0.044)	(0.051)	(0.047)	(0.052)	(0.045)		(0.018)		(0.000)
150	0.335	0.398	0.561	0.703	0.773	0.825	···	0.9955	···	1.000
	(0.045)	(0.045)	(0051)	(0.046)	(0.041)	(0.036)		(0.007)		(0.000)
200	0.390	0.524	0.642	0.761	0.856	0.922		0.9992		1.000
	(0.044)	(0.046)	(0.048)	(0.042)	(0.036)	(0.027)		(0.003)		(0.000)
1000	0.633	0.866	0.969	–	–	–		–		–
	(0.052)	(0.036)	(0.016)	–	–	–		–		–
5000	0.891	–	–	–	–	–		–		–
	(0.031)	–	–	–	–	–		–		–

– : not calculated