An example of the calculation of the statistical power to detect population sub-division in North Pacific minke whales

Barbara L. Taylor and Susan J. Chivers
Southwest Fisheries Science Center, 8604 La Jolla Shores Drive, La Jolla, CA 92038 U.S.A.

Abstract

We estimated the statistical power to detect population subdivision for a plausible case for North Pacific minke whale population structure. We limited our investigation to the question of whether two stocks may exist to the east and north of Japan (within sub-areas 7, 8, 9, 11 & 12). We used historical numbers from one of the base-case implementation trials (N1-j1g0), which assumes that 30% of the animals in the Sea of Okhotsk (sub-area 12) are from the more coastal area (O-stock). We estimated power using simulations and the effect size of a dispersal rate of ½%/year between the stocks. Even for the most powerful statistic of population differentiation (χ^2) the power to detect population subdivision was 0.49 when $\alpha = 0.05$. Using $\alpha = 0.05$ as the criterion for significance would result in making Type 2 errors (under-protecting minke whales) over ten times more frequently than making Type 1 errors (over-protecting minke whales). When Type 1 and Type 2 errors were equalized the critical α-level was 0.23, i.e. any p-value less than 0.23 would result in rejecting the null hypothesis of no population structure. Even with this increased critical α-level the statistical power would only be 0.77. We suggest future directions for estimating statistical power for North Pacific minke whales.

Introduction

The statistical power to detect population subdivision in North Pacific minke whales remains an issue in interpreting analyses that use genetic data gathered during the course of scientific whaling (JARPN). There are two important questions: 1) if the western North Pacific (to the east and north of Japan) is subdivided, what is the probability that the JARPN sample would have detected such sub-division, and 2) how can observed p-values be interpreted. We provide an example of a statistical power calculation to shed light on these questions. We chose a plausible case for stock-structure (implementation trial base-case N1-j1g0). Note that the calculation here is for one of many plausible cases. The null hypothesis (H_0) is that there is no population subdivision in this area. We examined only a single alternate hypothesis (H_A): there are two stocks (defined in more detail below) with an annual dispersal rate of ½%/year (d = 0.005). We chose this dispersal rate because a rate this low could result in errors in management if quotas were based on the abundance of the entire area while harvest was taken primarily in the western area, which has a lower total abundance and was the historically harvested population. We also expect that genetic differentiation at this level of dispersal will be low and therefore difficult to detect.

The difference in magnitude between H_0 and H_A is called effect size. As explained in SC/F2K/J4 statistical power decreases as effect size decreases. In this case, the effect size is the amount of

genetic differentiation. For genetic differentiation (F_{ST}) the expected effect size depends on effective population size (N), dispersal (m) (Wright's formula (Wright 1931) modified for mitochondrial DNA [mtDNA] (Takahata & Palumbi 1985)) and the number of populations (p) ($\psi = [p/(p-1)]^2$) (Malécot 1975; Maruyama 1970, 1971)):

$$F_{ST} = \frac{1}{2\,N\,m\,\psi + 1} \qquad (1)$$

For an ideally panmictic population that is falsely subdivided into two strata $F_{ST} = 0$. Thus, the expected value for $H_0 = 0$. The effective population for mtDNA is roughly the number of breeding females. We obtained data to indicate historical population sizes for N1-j1g0 from Cherry Allison. Historical populations sizes are important because the genetic profile of the current population will reflect this historical abundance, not the current reduced abundance. The commercial exploitation of North Pacific minke whales occurred from the 1950s to the late 1980s, which is a short period from an evolutionary perspective. This short duration of exploitation combined with a maximum reduction that is relatively small, in an evolutionary sense, makes it unlikely that any genetic diversity was lost because of commercial whaling. The data file gave the initial abundances for each of 3 stocks: J (K1), O(K2) and W(K3). We are only interested in the latter two, which had abundances for breeding females of 4,814 (O stock) and 7,023 (W stock). This trial assumes that the Okhotsk Sea (sub-area 12) is comprised of 30% O stock. Equation 1 assumes that all populations are of equal size. If we use the mean abundance for N we get 5,919.

The term "m" refers to the migration rate per generation (T). Because this is just an example of power calculation that approximates the minke whale case, we made a number of assumptions about demographic rates to obtain the approximate generation time. Clearly, these assumptions could be improved at a later date, but they are likely to have a small effect (if any) on the results. Some parameters were taken from literature on minke whales, such as the age of first reproduction (AFR), while others were educated guesses based on the demography of similar whales, such as the relation between adult and juvenile survival in humpback whales. We solved the Lotka equation using the following parameters: survival from birth to first birthday (s_0) = 0.90, survival of all other ages (s_a) = 0.98, crude annual birth rate of female calves to females (b) = 0.2, AFR = 7, and oldest age (ω) = 36. These parameters result in a maximum growth rate (r_0) = 0.067. We effected density dependence entirely through the birth rate (b) in a linear fashion. At carrying capacity (K) the growth rate (r_K) is zero and b = 0.10. The generation time (T) is then 19.6 years. Since we set the annual dispersal (d) to 0.005/year the dispersal rate/generation (m) is 0.098.

Making these assumptions, the rough level of expected genetic differentiation (F_{ST}) using equation 1 is 0.00043. Thus, we are trying to detect a very small effect size and expect that statistical power will therefore be low. However, our experience using simulations rather than the analytical equations has shown that our ability to detect population subdivision is actually greater than

would be expected on the basis of the expected effect size from analytical equations. Further, the actual level of genetic differentiation can differ quite dramatically from the expected value when the assumptions made for analytical equations (such as all populations are of equal size) are violated (Taylor et al. In Press.). The actual genetic differentiation also varies through time, which affects the calculation of power and is a source of uncertainty not considered in the analytical equations (Taylor et al. In Press.). Therefore, we used the simulation technique of estimating statistical power (SC/F2K/J4) for this example based on western North Pacific minke whales.

Methods

We used the same techniques described in SC/F2K/J4 (Appendix 1). In this case, we considered only two stocks. The model considers only females where each "individual" can give birth, die, and/or disperse each year. Using the demographic parameters given above, the mean birth and death rate at K = 0.04. We use 40 variable base-pairs to approximate the expected variability in the hypervariable portion of the mtDNA d-loop. We used same the mutation rate (μ) as previous simulations (0.0001). This mutation rate was found iteratively to approximately yield the observed level of haplotypic diversity for most marine mammals (Taylor et al. In Press). We examined the most powerful of the commonly used statistics of population differentiation: χ^2 (Roff & Bentzen 1989), H_{ST}, F_{ST}, K^*_{ST} (Hudson et al. 1992), ϕ_{ST} (Excoffier et al. 1992). Each of these statistics uses a randomization process to create a null distribution. We performed a double-sample check procedure to ensure that both our sampling procedure and our randomization process were producing unbiased estimates of p-values for the statistics. At each time step when we sampled the populations, we also sampled the same population twice. In this case, the null hypothesis is true, i.e. the strata are artificial and only a single population exists. Thus, we expect that 5% of the time, we will get p-values less than 0.05. That is, when we set $\alpha = 0.05$ and expect to falsely reject H_0 5% of the time, that we will actually do so in our simulation procedure. P-values when H_0 is actually true should be uniformly distributed between zero and one.

We made the assumption that sampling was random within populations. We sampled n = 180 from each of the two stocks. This matches the case where sub-area 9 is a separate stock from sub-areas 7 and 8 together (see Taylor SC/F2K/J6). Note that we considered that the distribution of haplotypes would be the same for males and females within a stock, thus the number of actual males and females sampled is immaterial. It is, however, plausible that there is more dispersal of adult males than of adult females. Thus, even though mtDNA is not passed down from male to offspring, the frequency within a population might differ between males and females if adult males that had dispersed to the neighboring population were sampled there and considered to have originated from that population. Our assumption that the samples from within an area are representative of that area (i.e. that a preponderance of adult male samples will not bias our result) will result in higher power than if that assumption is false. Thus, the assumption we have made is optimistic as regards statistical power.

Results & Discussion

As expected the effect size varied with time (Fig. 1.). The effect size was greater for the haplotypic statistic F_{ST} (bold line) than it was for the K^*_{ST} (thin line). This matches with the different performance of different statistics as revealed by Type 1 versus Type 2 error trade-off curves (Fig. 2). The choice of statistic makes a large difference in statistical power (Table 1).

	β (with power in parentheses)				
	χ^2	F_{ST}	H_{ST}	K^*_{ST}	ϕ_{ST}
$\alpha = 0.05$	0.51 (0.49)	0.73 (0.27)	0.76 (0.24)	0.88 (0.12)	0.92 (0.08)
$\alpha = \beta$	0.23 (0.77)	0.30 (0.70)	0.30 (0.70)	0.39 (0.61)	0.44 (0.56)

Table 1. Statistical power and its complement β comparing different statistics for two cases: α = 0.05 and α = β.

Thus, if a dispersal rate of 0.005 is important in determining whether to manage the western North Pacific as one or two stocks, the model indicates that if decisions are based on $\alpha = 0.05$ there is a greater than 50% chance of incorrectly rejecting that stock structure exists and managers would be over 10 times more willing to commit an under- rather than an over-protection error (0.51/0.05). If the decision criterion was to equalize errors, then the critical α-level would be 0.23 for the most powerful statistic (χ^2). In this case if one obtained a p-value of 0.06, as occurred using χ^2 for western North Pacific minke whales (SC/F2K/J6), then the decision would be to reject the null hypothesis of no population structure.

The double-sample check resulted in the expected uniform distributions of p-values between zero and one. The average population size of O-stock was 4,773 and of W-stock was 6,889, which differed trivially from the input carrying capacities of 4,814 (O stock) and 7,023 (W stock). The average haplotypic diversity (H_T) was 0.97. Hudson et al. (1992) advise that sequence based statistics should prove more powerful if H_T exceeds 1-[1/min(n1,n2)], which in our case is 1-[1/180] = 0.994. Thus, our result that χ^2 is most powerful is consistent with this advice in this case. We note, however, that our results are also consistent with our previous comparative results for high dispersal (SC/F2K/J5), which showed that the sequence statistics performed much more poorly than the haplotypic statistics. Also consistent with past results (Taylor et al. In Press) the average $F_{ST} = 0.0012$ exceeded the expected analytical F_{ST} (0.00043), which is a factor of 2.78 times greater than the expected value.

The average number of haplotypes sampled (n = 360) during simulations was 92 (Fig. 3). The observed number of haplotypes sampled (excluding J-stock) was 58, which is possible, given the distribution from the simulations, but unlikely. The reduced observed level of haplotypes could

result from a number of assumptions that we made that could fail to represent minke whales. For example, the abundances in the simulations fluctuated around the different carrying capacities but no doubt under-represent the actual fluctuations of minke whale populations. Actual populations are affected by their environment and fluctuate according to habitat quality. In evolutionary time, this species has experienced dramatically different ice conditions. Because adult females seem to prefer ice-edge habitat, it is likely that ice ages had affects on abundance. The effective population size is the harmonic mean of abundances through time. Thus, it is possible that fluctuating populations contributed to a lower overall effective population size. The number of haplotypes that can be maintained decreases as abundance decreases. Thus, it is to be expected that this uncertainty will contribute to some lack of fit between the model results and observed levels of haplotypic diversity. The mutation rate is another variable that could result in differences between the simulated and observed levels of haplotypic diversity.

It is unclear, however, that this difference in haplotypic diversity will make much difference in our ability to detect population structure. The performance of χ^2 depends primarily on the haplotype frequency distribution. A comparison of the frequency distribution for the most common twenty haplotypes shows the simulation to be remarkably similar to the observed distribution (Fig. 4). Of course, the distribution of haplotype frequencies changes through time for the simulation but more or less maintains this form.[1]

Conclusions

Our results show that although statistical power is higher than one might expect from the analytical equations, it is still quite low from the viewpoint of making stock definition decisions. What steps could be taken to get better resolution of stock structure? Because simulations with over 10,000 individuals are very computer intensive, it would be most efficient to minimize the number of plausible scenarios. A logical first step would be to place initial boundaries so that they make some biological sense. It is likely that the boundaries between 7, 8 and 9 poorly represent

---

[1]As an interesting aside for the JARPN discussions, I found it interesting to compare the haplotypic diversity between J-stock (5 haplotypes, n = 28) and O-stock (48 haplotypes in 7 & 8, n = 180) in light of the estimated historical number from N1-j1g0 (5,983 and 4,814 respectively). The amount of haplotypic diversity a population can support depends on abundance. One would expect, therefore, J and O to have roughly the same number of haplotypes. The low number of haplotypes found in J-stock is indicative of a population that has on average been small over historical time. Although it is plausible that J-stock was not large in recent times, which fits with recent interpretations of CPUE data, it is also plausible that this population has been subject to severe reductions in the past (loss of habitat during ice ages?) or resulted from a relatively recent colonization (within hundreds of generations). An exercise could be carried out to estimate the plausible range of initial abundances given these scenarios and the observed level of haplotypic diversity.

biological boundaries. The power of the genetic data to detect subdivision is compromised by improper boundary location (SC/F2K/J3). Coming up with better initial strata can be done in several ways. Perhaps the best way is to modify the geographic clustering algorithm developed by Martien & Taylor to fit the case of minke whales. Although this modification task is not insurmountable, it will be a challenging and time consuming modeling effort. This technique has the advantage of using the data to rank the plausibility of boundaries. Our experience with harbor seals in Alaska has shown that strong boundaries were placed in areas that we never would have posited.

A less objective but easier route would be to closely scrutinize the distributional data obtained from surveys and use hiatuses in distribution as a guide to setting strata boundaries. This is likely to be a frustrating experience as surveys are always complicated by factors beyond our control, such as Beaufort state. Apparent hiatuses may result from rough conditions caused by the interface of currents or areas of constant high winds. The seasonal migration of minke whales together with the different timing of the surveys will also make interpreting densities difficult. Nevertheless, it should not be difficult to do better than the current boundaries that clearly cut through areas of high densities. This technique will always have the disadvantage of forcing our preconceived notions about minke whale behavior into the analysis of the data. Again, we emphasize that our experience with Alaskan harbor seals was very humbling as regards our ability to accurately assign initial strata. For example, three separate stocks are apparently found on Kodiak island. Support for these boundaries is very strong. Now that we have these boundaries delineated, scientists doing radio telemetry have noted that they match the actual movement of seals in this area very well. These boundaries are probably caused by fragmentation and colonization events dating back at least to the most recent ice ages. Harbor seals appear to be much more plastic in their behavior than do minke whales. It is likely quite beyond our ability to guess what motivates minke whale movements.

Once a set of ranked plausible boundaries have been established, there is the matter of what to do about areas where there are few to no data, in particular the Sea of Okhotsk where abundance is relatively high. One possibility is to use the demographic data from sub-areas 7 and 8 and 11 to deduce a range of plausible percentages of that population present in sub-area 12. For example, I considered the scenario where 7, 8 and 11 contained all O-stock juveniles and adult males and thus, 12 would contain all the females. Using the N1-g1j0 data I found that approximately 33% of the population were adult females. These assumptions resulted in 2,755 O-stock females in 12 (which is about 20% of the number estimated in that area). Roughly, the minke whale population is split into thirds comprised of juveniles, adult males and adult females. Even if we consider that 7, 8, and 11 contain only juveniles and the other 2/3 are found in 12, that results in 10,758 O-stock whales in 12. Because no juveniles are described as being in 12, it is therefore very likely that 12 is an area of mixed stocks because we cannot otherwise account for the approximately 14,000 animals in that area as being solely adults of the juveniles found in 7, 8 and 11. In the example in this paper, individuals were sampled and divided into the different strata without error. Thus, unless the stocks are perfectly segregated within 12 and we are able to place the boundary

correctly, it will be very difficult to use genetics to better partition that area because we know the effect size is already small.

The last, but not least, consideration for actual calculations of statistical power is defining the alternate hypothesis, i.e. what level of dispersal would be required such that a harvest could occur without undesirable effects if sub-areas 7,8, 9, 11 and 12 were managed a single stock. In this paper we arbitrarily chose to examine a dispersal rate of ½%/year (d = 0.005). Clearly, before doing further lengthy simulations, researchers would want to know fairly precisely what level of dispersal is of interest. Taylor (1997) described how to arrive at such a critical dispersal rate. In essence, the critical dispersal rate is the dispersal rate required such that even if there is population structure but the area is mistakenly managed as a single unit, dispersal is sufficient to meet management objectives. Consider a very simple example: N1 = 100 and N2 = 200. All harvest is coming from N1. If the population structure is known, then the sample objective would be to harvest half the productivity at a rate of 3%/year. If N1 and N2 are mistakenly pooled to set the quota, a harvest of 9 individuals would be allowed, which would be 9% of N1. Thus, we would need a net increase of 6 individuals from N2 to cover for our error in defining population structure. For this example the critical dispersal rate would need to be much greater than the ½%/year used as an effect size in this paper. This example shows that to estimate a critical dispersal rate one needs to posit how harvest will occur spatially. Clearly, if harvest was equal across space with respect to density there would be no need to define population structure. If harvest is unequal (as was the case for past exploitation of North Pacific minke whales) and densities are also unequal (as appears to be the case now) then the potential exists for serious mis-management if population structure is ignored.

References

Excoffier, L., P. E. Smouse, and J. M. Quattro. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131:479-491.

Hudson, R. R., D. D. Boos, and N. L. Kaplan. 1992. A statistical test for detecting geographic subdivision. Mol. Biol. Evol. 9:138-151.

Malécot, G. 1975. Heterozygosity and relationship in regularly subdivided populations. Theor. Popul. Biol. 8:212-841.

Maruyama, T. 1970. Analysis of population structure: I. One-dimensional stepping-stone models of finite length. Ann. Hum. Genet. 34:201-219.

Maruyama. T. 1971. Analysis of population structure: II. Two-dimensional stepping stone models of finite length and other geographically structured populations. Ann. Hum. Genet. 35:179-196.

Roff, D. A. and P. Bentzen. 1989. The statistical analysis of mitochondrial DNA polymorphisms: χ^2 and the problem of small samples. Mol. Biol. Evol. 6:539-545.

Taylor, B. L., S. J. Chivers, and A. E. Dizon. 1997. Using statistical power to interpret genetic data to define management units for marine mammals. In Molecular Genetics of Marine Mammals. eds. A. E. Dizon, S. J. Chivers, and W. F. Perrin. Special Publication 3:347-364 Allen Press, Inc., Lawrence, Kansas, U.S.A.

Taylor, B. L., S. J. Chivers, and A. E. Dizon. SC/F2K/J4. Estimating the statistical power to detect population subdivision using mitochondrial DNA.

Taylor, B. L., S. J. Chivers, S. Sexton and A. E. Dizon. In press. Using simulation models that incorporate uncertainty to estimate dispersal rates from mitochondrial DNA data. Conservation Biology.

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

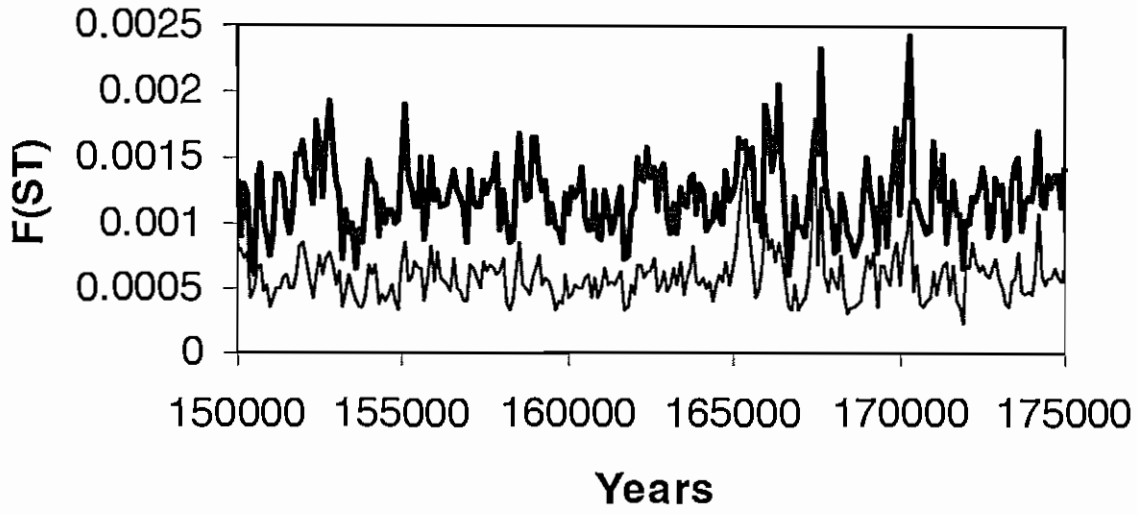

Figure 1. The actual level of genetic differentiation through time between the two posited populations of minke whales. Here we show two of the statistics F_{ST} (bold line) and K^*_{ST} (thin line) over a period of 25,000 years where statistics are calculated every 100 years, which is approximately every 5.1 generations.

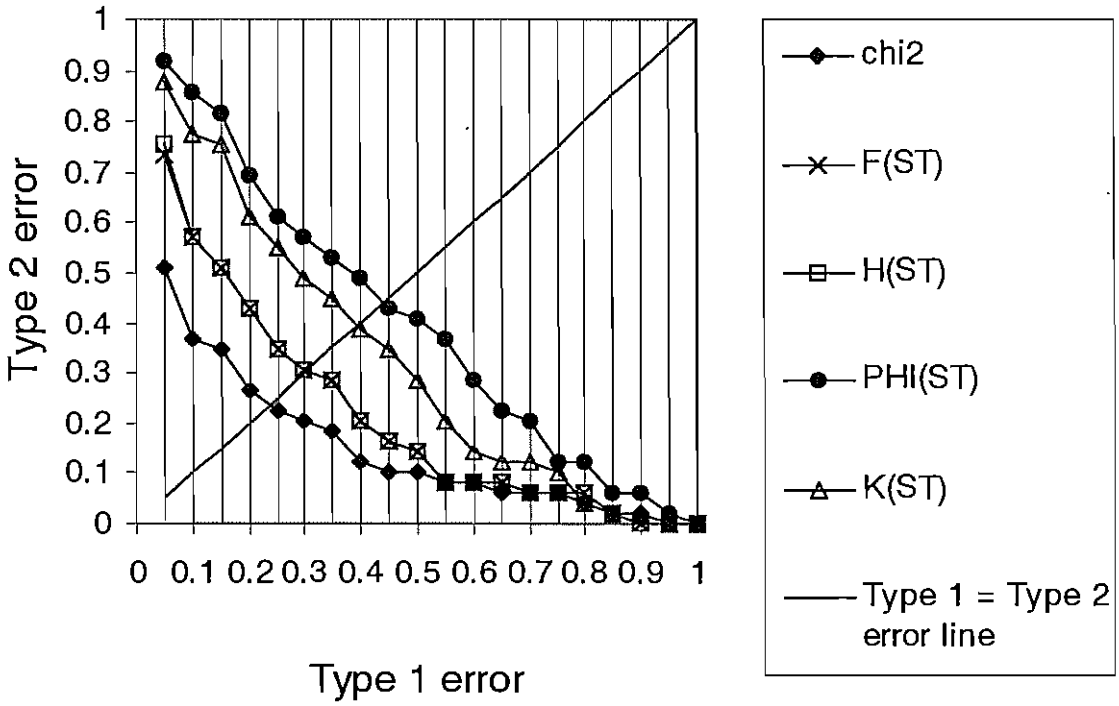

Figure 2. Type 1 versus Type 2 error curves for different statistics of population differentiation. Vertical grids at an interval of 0.05 are given so the reader can more easily read Type 2 error levels at typical α-levels. Power is (1 - β) where β is the Type 2 error. Thus, the lower the trade-off curve, the higher the power. The Type 1 = Type 2 error line is shown as the case when managers choose to equalize over and under-protection errors. Values for the different statistics can be seen in Table 1.

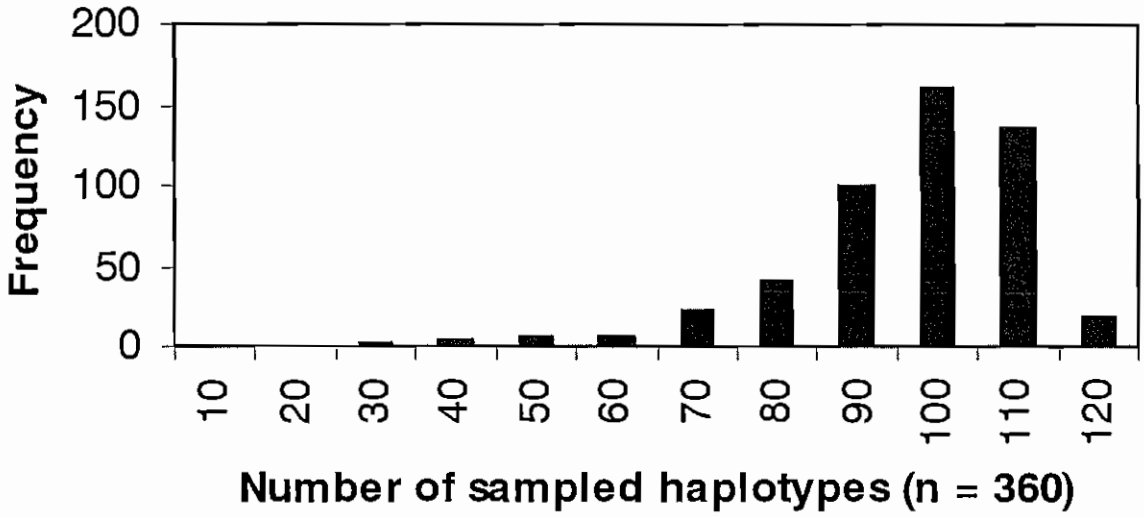

Figure 3. Frequency histogram of the number of haplotypes sampled when n = 360 (n1 = n2 = 180). The number of observed haplotypes samples from sub-areas 7, 8 and 9 was 58.

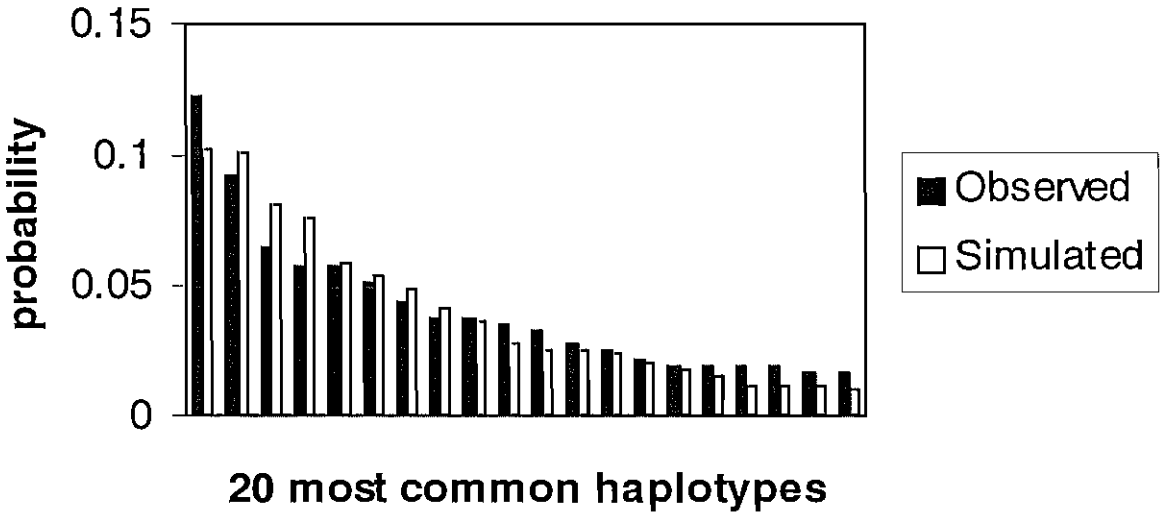

Figure 4. An example of the comparison of the haplotypic frequency distributions between the simulation (Simulated) and the haplotypes observed in sub-areas 7, 8, and 9 (Observed) for the year 200,000.