Average Number of Nucleotide Differences in a Sample From a Single Subpopulation: A Test for Population Subdivision

Curtis Strobeck

Department of Zoology, University of Alberta, Edmonton, Alberta, Canada T6G 2E9

Manuscript received November 24, 1986
Revised copy accepted May 20, 1987

ABSTRACT

Unbiased estimates of $\theta = 4N\mu$ in a random mating population can be based on either the number of alleles or the average number of nucleotide differences in a sample. However, if there is population structure and the sample is drawn from a single subpopulation, these two estimates of θ behave differently. The expected number of alleles in a sample is an increasing function of the migration rates, whereas the expected average number of nucleotide differences is shown to be independent of the migration rates and equal to $4N_T\mu$ for a general model of population structure which includes both the island model and the circular stepping-stone model. This contrast in the behavior of these two estimates of θ is used as the basis of a test for population subdivision. Using a Monte-Carlo simulation developed so that independent samples from a single subpopulation could be obtained quickly, this test is shown to be a useful method to determine if there is population subdivision.

If the DNA sequences are known for several copies of a gene sampled from a randomly mating population, there are two ways to estimate $\theta = 4N\mu$, where N is the population size and μ is the mutation rate. EWENS (1972) showed that the maximum likelihood estimate of θ assuming the infinite alleles model (KIMURA and CROW 1964) is a function of only the number of alleles in the sample and not the frequencies of the alleles. On the other hand, if the infinite sites model (KIMURA 1969) is assumed, then θ can be estimated from the number of segregating sites (WATTERSON 1975) or less efficiently from the average number of nucleotide differences between copies of the gene (TAJIMA 1983). If there is random mating and no intragenic recombination then both are unbiased estimates of θ. However, if either of these assumptions is not true, then these two methods may estimate different values. For example, intragenic recombination is expected to increase the number of alleles in a sample but not affect the number of segregating sites or the average number of nucleotide differences (WATTERSON 1975; STROBECK and MORGAN 1978; STROBECK and GOLDING 1983; HUDSON 1983b).

In this paper it is shown that for a general model of population structure, the expected number of nucleotide differences between two genes which are randomly chosen from a single subpopulation is independent of the migration rates between subpopulations and equal to $4N_T\mu$ where N_T is the total number of individuals in the population. This model includes the island model of WRIGHT (1931) and the circular stepping-stone model (MARUYAMA 1970). In contrast, the distribution of the number of alleles in a sample from a single subpopulation is approximately the distribution of alleles in a panmictic population with a θ' which is an increasing function of the migration rates (G. B. GOLDING, personal communication). For example, for the island model it has approximately the same sampling distribution as a panmictic population with a $\theta' = \theta + M(n-1)\theta/\{(n-1)\theta + M\}$ where n is the number of islands and $M = 4Nm$ (SLATKIN 1982; GOLDING and STROBECK 1983). Thus it may be possible to determine if a population is panmictic or structured by comparing the observed number of alleles with that expected from the estimate of θ based on the average number of nucleotide differences between the copies of the genes in the sample. Lastly, the power of this test to detect population subdivision is investigated using Monte-Carlo simulation to obtain independent samples from a population known to be subdivided.

EXPECTED NUMBER OF NUCLEOTIDE DIFFERENCES

In this section the expected number of nucleotide differences between two randomly chosen DNA sequences is derived for three models of population structure; the island model (WRIGHT 1931), the circular stepping-stone model (MARUYAMA 1970) and a general conservative migration model. Unless otherwise stated, it is assumed throughout this section that the population consists of n subpopulations each with N diploid individuals with $N \gg 1$. The mutation rate per DNA sequence per generation is denoted by μ and is assumed to be $O(1/N)$.

In the island model the migration rate from one subpopulation to any of the other $n - 1$ subpopulations is $m/n - 1(m \simeq O(1/N))$. The recurrence equations for the expected number of nucleotide differences between two randomly chosen DNA sequences from the same subpopulation, ξ_{ii}, and from two different subpopulations, ξ_{ij}, are

$$\xi'_{ii} = \left(1 - \frac{1}{2N} - 2\mu - 2m\right)\xi_{ii} + 2\mu(\xi_{ii} + 1) + 2m\xi_{ij} \tag{1a}$$

$$\xi'_{i=j} = \left(1 - 2\mu - 2\frac{m}{n-1}\right)\xi_{ij} + 2\mu(\xi_{ij} + 1) + 2\frac{m}{n-1}\xi_{ii} \tag{1b}$$

neglecting terms of $O(1/N^2)$ or less. The first term on the righthand side of (1a) is the probability that the two sequences came from two distinct sequences in the previous generation, no new mutants occurred, and neither were migrants times the expected number of nucleotide differences. The second term is the probability that there was a mutation in one of the two sequences times the expected number of nucleotide differences plus one (because of the new mutation). The third term is the probability that one of the two sequences is a migrant times the expected number of nucleotide differences when the two sequences come from different subpopulations. The terms containing the probabilities that there were two mutations, one sequence was a migrant and there was a mutation, etc. have been neglected. Equation 1b is derived in a similar fashion. The equations for the expected values of the stationary distributions for the number of nucleotide differences are therefore

$$(1 + 4\,Nm)\hat{\xi}_{ii} - 4Nm\hat{\xi}_{ij} = 4N\mu \tag{2a}$$

$$-\frac{4Nm}{n-1}\hat{\xi}_{ii} + \frac{4Nm}{n-1}\hat{\xi}_{ij} = 4N\mu \tag{2b}$$

and the solution to these equations is

$$\hat{\xi}_{ii} = n4N\mu \tag{3a}$$

$$\hat{\xi}_{ij} = n4N\mu + (n-1)\frac{\mu}{m}. \tag{3b}$$

These results were first obtained by LI (1976) using generating functions.

In the circular stepping-stone model, the subpopulations are arranged in a circle with migration only between adjacent subpopulations. The migration rate from a subpopulation to each of the two adjacent subpopulations is $m/2$. As the number of subpopulations goes to infinity, the circular stepping-stone model converges to the stepping-stone model of KIMURA and WEIS (1964). The recurrence equations for the expected number of nucleotide differences when the two sequences come from two subpopulations which are i steps apart, ξ_i, are

$$\xi'_0 = (1 - \frac{1}{2N} - 2\mu - 2m)\xi_0 + 2\mu(\xi_0 + 1) + 2m\xi_1 \tag{4a}$$

$$\xi'_i = (1 - 2\mu - 2m)\xi_i + 2\mu(\xi_i + 1) + m\xi_{i+1} + m\xi_{i-1} \quad \text{for } i = 1, \cdots, k-1 \tag{4b}$$

$$\begin{aligned} \xi'_k &= (1 - 2\mu - 2m)\xi_k + 2\mu(\xi_k + 1) + 2m\xi_{k-1} \quad \text{if } n = 2k \\ \xi'_k &= (1 - 2\mu - 2m)\xi_k + 2\mu(\xi_k + 1) + m\xi_k + m\xi_{k-1} \quad \text{if } n = 2k + 1. \end{aligned} \tag{4c}$$

Thus the equations for the expected values of the stationary distribution for the number of nucleotide differences are

$$(1 + 4Nm)\hat{\xi}_0 - 4Nm\hat{\xi}_1 = 4N\mu \tag{5a}$$

$$\hat{\xi}_{i+1} - 2\hat{\xi}_i + \hat{\xi}_{i-1} = -2\frac{\mu}{m} \quad \text{for } i = 1, \cdots, k-1 \tag{5b}$$

$$\begin{aligned} \hat{\xi}_k - \hat{\xi}_{k-1} &= \frac{\mu}{m} \quad \text{if } n = 2k \\ \hat{\xi}_k - \hat{\xi}_{k-1} &= 2\frac{\mu}{m} \quad \text{if } n = 2k + 1. \end{aligned} \tag{5c}$$

The general solution to the inhomogeneous Equation 5b is

$$\hat{\xi}_i = C_0 + C_1 i - \frac{\mu}{m} i^2 \tag{6}$$

The constants C_1 and C_0 are determined by substituting (6) into (5c) and then into (5a). The complete solution for the expected number of nucleotide differences is

$$\hat{\xi}_i = n4N\mu + \frac{\mu}{m} i(n - i) \quad \text{for } i = 0, \cdots, k. \tag{7}$$

For both the island model and the circular stepping-stone model, the expected number of nucleotide differences when both sequences are drawn randomly from the same subpopulation is independent of the migration rate and equal to $n4N\mu$. This is the expected number of nucleotide differences if the total population, nN, were panmictic. This result also holds for any isotropic conservative migration model. Both the island model and the circular stepping-stone model are isotropic conservative migration models.

Consider the general backward migration model in which m_{ij} is the proportion of the ith subpopulation that comes from the jth subpopulation each generation. (Thus $m_{ii} = 1 - \sum_{j \neq i} m_{ij}$.) Let N_i be the number of diploid individuals in the ith subpopulation. The migration will be considered conservative if for every subpopulation the number of individuals migrating into the subpopulation is equal to the number of individuals migrating out of the subpopulation, *i.e.*,

$$N_i \sum_{j \neq i} m_{ij} = N_i(1 - m_{ii}) = \sum_{j \neq i} m_{ji}N_j$$

or

$$0 = \sum_j m_{ji}N_j - N_i.$$

This implies that $\eta(M - I) = 0$ where $\eta = (N_1, N_2, \cdots, N_n)$ is the row vector of the population sizes, $M = (m_{ij})$ is the $n \times n$ migration matrix, and $I = \mathrm{diag}(1)$ is the $n \times n$ identity matrix. If all the N_i are of the same order of magnitude and the m_{ij} for $i \neq j$ are of the order of $1/N_i$ then the recurrence equations for ξ_{ii} and ξ_{ij} are

$$\begin{aligned} \xi'_{ii} = (1 - \frac{1}{2N_i} - 2\mu - 2 \sum_{j \neq i} m_{ij})\xi_{ii} \\ + 2\mu(\xi_{ii} + 1) + 2 \sum_{j \neq i} m_{ij}\xi_{ij} \qquad \text{(8a)} \\ \text{for } i = 1, \cdots, n \end{aligned}$$

$$\begin{aligned} \xi'_{ij} = (1 - 2\mu - \sum_{k \neq i} m_{ik} - \sum_{k \neq j} m_{jk})\xi_{ij} + 2\mu(\xi_{ij} + 1) \\ + \sum_{k \neq i} m_{ik}\xi_{jk} + \sum_{k \neq j} m_{jk}\xi_{ik} \qquad \text{(8b)} \\ \text{for } i, j = 1, \cdots, n; i \neq j \end{aligned}$$

neglecting terms of $O(1/N_i^2)$ or less. Therefore the equations for the expected values of the stationary distributions of the number of nucleotide differences are

$$\begin{aligned} \left(\frac{1}{2N_i} + 2 \sum_{j \neq i} m_{ij}\right) \hat{\xi}_{ii} \\ - 2 \sum_{j \neq i} m_{ij}\hat{\xi}_{ij} = 2\mu \qquad \text{for } i = 1, \cdots, n \end{aligned} \qquad \text{(9a)}$$

$$\begin{aligned} \left(\sum_{k \neq i} m_{ik} + \sum_{k \neq j} m_{jk}\right)\hat{\xi}_{ij} - \sum_{k \neq i} m_{ik}\hat{\xi}_{jk} \\ - \sum_{k \neq j} m_{jk}\hat{\xi}_{ik} = 2\mu \qquad \text{for } i, j = 1, \cdots, n. \end{aligned} \qquad \text{(9b)}$$

Since $\sum_{j \neq i} m_{ij} = 1 - m_{ii}$, these n^2 equations can be summarized as the matrix equation

$$D - (M - I)\xi - \xi(M^T - I) = 2\mu U \qquad (10)$$

where $D = \mathrm{diag}(\hat{\xi}_{ii}/2N_i)$, $\xi = (\hat{\xi}_{ij})$, and $U = (1)$ is the $n \times n$ matrix consisting of all ones.

Multiplying (10) by $2\eta = (2N_1, 2N_2, \cdots, 2N_n)$ on the left and $2\eta^T$ on the right, one obtains

$$\sum_i 2N_i\hat{\xi}_{ii} = 2\mu\left(\sum_i 2N_i\right)^2 = 2\mu(2N_T)^2$$

where $N_T = \sum_i N_i$ is the total population size. Thus the average of the $\hat{\xi}_{ii}$,

$$\bar{\xi}_{ii} = \sum_i 2N_i\hat{\xi}_{ii}/2N_T = 4N_T\mu$$

and is independent of the migration rates. If the migration model is isotropic, *i.e.*, the pattern of migration for each subpopulation is identical to the migration pattern of any other subpopulation and $N_i = N$ for all i, then

$$\hat{\xi}_{ii} = \bar{\xi}_{ii} = n4N\mu.$$

Formally, the migration model is isotopic if for any i and j there exists a permutation such that i goes to j and $m_{kl} = m_{k'l'}$ where k goes to k' and l goes to l'. Both the island model and the circular stepping-stone model satisfies this condition. During the preparation of this paper, it was learned that SLATKIN (1987) has obtained similar results for the symmetric migration model.

TEST FOR POPULATION SUBDIVISION

As shown in the previous section, the expected number of nucleotide differences between two DNA sequences from the same subpopulation is independent of the migrations rates and equal to $n4N\mu$ for a number of models of population structure, including the island model. However, for the island model it has been shown that the number of alleles in a sample from a single subpopulation has a EWEN's distribution with a modified value of θ, $\theta' = \theta + M\{(n - 1)\theta/[(n - 1)\theta + M]\}$ where $\theta = 4N\mu$ and $M = 4Nm$ (SLATKIN 1982; GOLDING and STROBECK 1983). As M goes from zero, *i.e.*, the subpopulations are completely isolated, to infinity, *i.e.*, the population is panmictic, θ' goes from θ to $n\theta$. Thus, if a population is effectively subdivided, the number of alleles in a sample from a single subpopulation should be much less than predicated using the θ estimated from the average number of nucleotide differences found in the sample.

This suggests the following test for population subdivision when the DNA sequences are known for s copies of a gene from a single population. First calculate the average number of nucleotide differences for the $s(s - 1)/2$ pairs of genes. This is an unbiased estimate of θ if the population were panmictic (TAJIMA 1983). Then use this estimate of θ to calculate the distribution for the number of alleles found in a sample of size s (EWENS 1972). If the probability of obtaining a sample with less than or equal to the number of alleles which were observed in the sample is less than a given value α, the null hypothesis that the population is panmictic is rejected at the $\alpha \times 100\%$ level. In Table 1 the values which the estimate of θ must exceed before the probability of observing

TABLE 1

Values which $\theta = 4N\mu$ must exceed before the probability of observing a given number of alleles in a sample of size 10 or 50 is less than 0.10, 0.05 or 0.01

Sample size	No. of alleles	Probability		
		0.10	0.05	0.01
10	2	2.18	2.85	4.65
	3	3.84	4.93	7.90
	4	6.30	8.08	13.00
	5	10.23	13.24	21.85
	6	17.19	22.71	39.59
50	2	1.04	1.30	1.91
	3	1.53	1.86	2.62
	4	2.06	2.46	3.36
	5	2.63	3.09	4.14
	6	3.24	3.78	4.97
	7	3.90	4.51	5.86
	8	4.60	5.29	6.81
	9	5.36	6.13	7.83
	10	6.17	7.03	8.92

a given number of alleles is less than α ($\alpha = 0.10$, 0.05 and 0.01) are listed for sample sizes of $s = 10$ and 50.

Two questions about this test come quickly to mind: "What is the power of this test to detect population subdivision if it exists?" and "Since the number of alleles and the average number of nucleotide differences are positively correlated, what is the true value of rejection of the null hypothesis?" Information about both of these questions can be obtained using Monte-Carlo simulation to obtain independent samples from populations with known structure, subpopulation size, migration rates, and mutation rate. A modification of the simulation method used by HUDSON (1983a) (see also TAJIMA 1983) provides an easy method to obtain independent sample for any given population structure.

To obtain independent samples from a panmictic population of size N, HUDSON first constructs the phylogenetic tree of the sample and determines the time associated with each branch of the tree and then determines the number of mutations that occur along each branch. The phylogenetic tree and the associated times are constructed recursively. The probability that k copies of a gene in one generation are derived from $k - 1$ copies of the gene in the previous generation is $k(k - 1)/4N$ plus terms of $O(1/N^2)$. The probability that the k copies come from k copies in the previous generation is $1 - k(k - 1)/4N$ plus terms of $O(1/N^2)$. Therefore, the time, t, measured in $4N$ generations for k copies of a gene to have come from $k - 1$ copies is exponentially distributed, *i.e.*,

$$f(t) = k(k - 1)e^{-k(k-1)t}.$$

The cumulative distribution is

$$F(t) = 1 - e^{-k(k-1)t}.$$

Therefore a value of t can be obtained by choosing x from a uniform distribution on (0, 1) and setting

$$t = \frac{-1}{k(k - 1)} \log(1 - x).$$

The two copies of the gene that came a single copy in the previous generation are chosen randomly. The number of mutations along a branch with an associated time T, measured in $4N$ generations is Poisson distributed with mean $\lambda = \theta T = 4N\mu T$.

To extend this method to the island model with islands, each island with a population size of N, and a migration rate m per gene per generation, it is only necessary to note that the probability that all k copies of a gene on an island one generation were all on the island the previous generation is $1 - mk$ and the probability that one copy was an immigrant is mk. Therefore the time, measured in $4N$ generations, till one of the copies was an immigrant is exponentially distributed

$$g(t_m) = kM\ e^{-kMt_m},$$

where $M = 4Nm$. Therefore a value of t_m can be obtained by choosing x from a uniform distribution on (0, 1) and setting

$$t_m = \frac{-1}{kM} \log(1 - x).$$

To construct the phylogenetic tree of a sample together with the associated time for each branch, it is necessary to obtain a value of t for each island that contains two or more copies of the gene (the distribution of t depends on the number of copies on each of the islands) and a value of t_m for each island that contains at least one copy of the gene. If the minimum value of all of these values is a t value, then on that island two genes are randomly selected and replaced by a single copy of the gene. If the minimum value is a t_m value, then on that island one copy of the gene is chosen to have been the immigrant and one island is chosen from which it had emigrated. A new set of t and t_m values are chosen and the process repeated until only one copy of the gene remains. At this time the number of mutations along each branch of the phylogenetic tree is determined as before.

For the purpose of answering the question on the power of the proposed test, the number of islands was assumed to be $n = 8$; the sample size, $s = 10$ or 50; $\theta = 4N\mu = 0.50$; and $M = 4Nm = 0.125$, 0.250, 0.500, 1.000 and 2.000. The results of 100 independent samples are shown in Table 2 for each of the above combinations. The test is moderately successful in detecting the presence of population subdivision especially with $s = 50$. Note that as either M becomes small or large the power of the test decreases. This is expected since as M becomes small the single subpop-

TABLE 2

Percentage of samples from a single subpopulation for which the null hypothesis of random mating was rejected when the island model was assumed

Sample size	$4Nm$	Level of significance (%)		
		10	5	1
10	0.125	15	13	11
	0.250	33	29	19
	0.500	33	25	13
	1.000	33	23	9
	2.000	22	10	4
50	0.125	35	31	24
	0.250	34	33	25
	0.500	49	45	35
	1.000	56	50	39
	2.000	49	44	21

A hundred independent samples were obtained for each set of parameters. It was assumed that the number of islands was $n = 8$; $\theta = 4N\mu = 0.5$; $4Nm = 0.125, 0.250, 0.500, 1.000$ or 2.00; and the sample size $s = 10$ or 50.

TABLE 3

Percentage of samples from a population for which the null hypothesis of random mating was rejected when random mating was assumed

Sample size	$4N\mu$	Level of significance (%)		
		10	5	1
10	0.5	0.6	0.1	0.0
	1.0	1.2	0.7	0.1
	2.0	1.4	0.5	0.0
50	0.5	1.7	0.6	0.2
	1.0	5.8	3.3	0.8
	2.0	7.1	5.2	2.2

A thousand independent samples were obtained for each set of parameters. It was assumed that $\theta = 4N\mu = 0.5$, 1.0 or 2.0 and the sample size $s = 10$ or 50.

ulation becomes completely isolated, whereas if M becomes large, the whole population behaves as if it were panmictic.

To answer the question of how conservative the proposed test is, 1000 independent samples were obtained for $\theta = 0.50$, 1.00 and 2.00 and $s = 10$ and 50. The results are shown in Table 3. The test is very conservative with $s = 10$, *i.e.*, only 0.6 to 1.4% are rejected at the 10% level. With $s = 50$ the test is much less conservative, *i.e.*, from 1.7 to 7.1% are rejected at the 10% level. It should be noted, however, that the test does not behave properly for $\theta = 2.0$ at the 1% level, *i.e.*, 2.2% is rejected. This is not just a stochastic error since another run of 1000 independent samples had approximately the same results. It seems to occur because although the probability of two, three, or four alleles in a sample is small with a high mutation rate, when such a sample does occur the average number of nucleotide differences in the sample is large and therefore more likely to be significant. For example, in the simulation with $\theta = 2.0$ and $s = 50$, there were 115 samples with two, three or four alleles of which 25, 21 and 9 were significant at the 10, 5 and 1% level, respectively. On the otherhand, there were 128 samples with ten or more alleles of which only one was significant (at the 5% level).

To summarize, the Monte-Carlo simulations show that population structure can be detected by comparing the observed number of alleles in a sample to that expected when $\theta = 4N\mu$ is estimated from the average number of nucleotide differences. However, it is a very conservative test for small sample sizes and it is not a proper statistical test if θ is large. It may be possible to overcome these problems by developing a test based on the joint distribution of the number of alleles and the average number of nucleotide differences in a sample.

LITERATURE CITED

EWENS, W. J., 1972 The sampling theory of selectively neutral alleles. Theor. Popul. Biol. **3:** 87–112.

GOLDING, G. B. and C. STROBECK, 1983 Variance and covariance of homozygosity in a structured population. Genetics **104:** 533–545.

HUDSON, R. R., 1983a Testing the constant-rate neutral allele model with protein sequence data. Evolution **37:** 203–217.

HUDSON, R. R., 1983b Properties of a neutral model with intragenic recombination. Theor. Popul. Biol. **23:** 183–201.

KIMURA, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics **61:** 893–903.

KIMURA, M. and J. F. CROW, 1964 The number of alleles that can be maintained in a finite population. Genetics **49:** 725–738.

KIMURA, M. and G. H. WEISS, 1964 The stepping stone model of population structure and the decrease of genetic correlation with distance. Genetics **49:** 561–576.

LI, W.-H., 1976 Distribution of nucleotide differences between two randomly chosen cistrons in a subdivided population: the finite island model. Theor. Popul. Biol. **10:** 303–308.

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

SLATKIN, M., 1982 Testing neutrality in a subdivided population. Genetics **100:** 533–545.

SLATKIN, M., 1987 The average number of sites separating DNA sequences drawn from a subdivided population. Theor. Popul. Biol. In press.

STROBECK, C. and G. B. GOLDING, 1983 The variance of linkage disequilibrium between three loci in a finite population. Can. J. Genet. Cytol. **25:** 139–145.

STROBECK, C. and K. MORGAN, 1978 The effect of intragenic recombination on the number of alleles in a finite population. Genetics **88:** 829–844.

TAJIMA, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105:** 437–460.

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.

WRIGHT, S., 1931 Evolution in Mendelian populations. Genetics **16:** 97–159.

Communicating editor: W.-H. LI