J Mol Evol (1993) 37:240–244

Journal of Molecular Evolution

Effective Population Size, Genetic Diversity, and Coalescence Time in Subdivided Populations

Masatoshi Nei,[1] Naoyuki Takahata[2]

[1] Department of Biology and Institute of Molecular Evolutionary Genetics, 328 Mueller Laboratory, Pennsylvania State University, University Park, PA 16802, USA
[2] National Institute of Genetics, Mishima, Japan

Received: 30 November 1992/Revised: 16 February 1993

Abstract. A formula for the effective population size for the finite island model of subdivided populations is derived. The formula indicates that the effective size can be substantially greater than the actual number of individuals in the entire population when the migration rate among subpopulations is small. It is shown that the mean nucleotide diversity, coalescence time, and heterozygosity for genes sampled from the entire population can be predicted fairly well from the theory for randomly mating populations if the effective population size for the finite island model is used.

Key words: Nucleotide diversity — Coalescence time — Average heterozygosity — Subdivided populations — Effective population size

Introduction

Nucleotide diversity and coalescence time (see Nei 1987 for the definitions of these quantities) play an important role in the study of molecular population genetics. Recently, Tajima (1989) and Takahata (1991) developed gene genealogy theories for the finite island model of subdivided populations (Maruyama 1970) and studied the expected values of nucleotide diversity and coalescence time of randomly chosen genes. The mathematical theories developed by these authors (see also Takahata 1988; Notohara 1990; Takahata and Slatkin 1990) are quite complicated, and it is not always easy to compute nucleotide diversity and coalescence time for different sets of population parameters. However, there is another approach for computing these quantities for subdivided populations. That is, if we can derive an equation for the effective population size (N_e) of a subdivided population, the expectations of nucleotide diversity (π) and coalescence time (T) may be given by

$$\pi = 4N_e\mu \tag{1}$$

$$T = 4N_e\left(1 - \frac{1}{n}\right) \tag{2}$$

where μ is the mutation rate per nucleotide site per generation, and n is the number of genes sampled from the entire population. Equations (1) and (2) are known to hold in randomly mating populations (Kimura 1969; Watterson 1975; Kingman 1982; Tajima 1983). Furthermore, if we know N_e, the expected heterozygosity (H) may also be given by

$$H = 4N_e v/(1 + 4N_e v) \tag{3}$$

where v is the mutation rate per locus.

The purpose of this paper is to derive an equation for N_e for the finite island model and to examine the accuracy of π, T, and H values obtained by this approach.

Correspondence to: M. Nei

Effective Population Size

In the following we consider the finite island model in which a population is subdivided into s subpopulations of effective size N and migration occurs from one subpopulation to another with a probability of $m/(s-1)$ per generation. (See Takahata and Nei 1984.) If this model is used, there are two different ways of deriving the formula for the effective size (N_e) for the entire population. One is to consider the drift variance of the mean gene frequency of the entire population, and the other is to use the theory of gene genealogy. Wright (1943, p. 133) used the former approach to derive an equation for N_e, although he did not clearly define the model of population structure he used. Wright's formula is given by

$$N_e = \frac{sN}{1 - F_i} \tag{4}$$

where F_i is the fixation index later redefined as F_{ST} (Wright 1951). Essentially the same formula can be derived from equation (4) of Maruyama (1972) if we note $1 - F_T = (1 - F_s)(1 - F_{ST})$, where F_T and F_S are the fixation indices for the entire population and the subpopulations, respectively.

In Maruyama's (1970) finite island model the infinite-allele model of neutral mutation (Kimura and Crow 1964) is used to study the extent of genetic polymorphism. In this model F_i in equation (4) should be replaced by G_{ST}, which is an extension of F_{ST} to the case of multiple alleles (Nei 1975). Takahata (1983) and Takahata and Nei (1984) derived a steady-state equation for G_{ST} for the finite island model. Their formula includes a term for the contribution due to mutation, but this term should be eliminated in the present case, because the effective size is related to the variance of gene frequency change per generation and has nothing to do with mutation. It then becomes

$$G_{ST} = \frac{1}{1 + 4Nm\left(\frac{s}{s-1}\right)^2} \tag{5}$$

The same equation has been derived by Crow and Aoki (1984). Therefore, if we replace F_i by this G_{ST}, equation (4) becomes

$$N_e = sN\left[1 + \frac{(s-1)^2}{4Nms^2}\right] \tag{6}$$

Note that G_{ST} reaches the steady-state value even if there is no mutation and s is finite (Nei et al. 1977).

To derive a formula for N_e by using the gene genealogy theory, we consider the case where two genes ($n = 2$) are randomly chosen from s subpopulations of effective size N. In this case the two genes may be drawn either from two different subpopulations or from the same subpopulation. The probability of occurrence of the former event is $1 - 1/s$, whereas the probability of occurrence of the latter event is $1/s$. Let T_b and T_w be the mean coalescence time of two randomly chosen genes when they are drawn from different subpopulations and that when they are drawn from the same subpopulation, respectively. (In the gene-genealogy theory the genealogical relationships of genes are examined retrospectively, and the coalescence time is the time at which two or more genes sampled from the present population trace back to a common ancestral gene in the past when time is measured in generations.) The mean coalescence of two genes randomly chosen from the entire population is then given by $T = T_w/s + (s-1)T_b/s$. Slatkin (1991) (see also Hey 1991) worked out the formula for T for the case of $n = 2$. It is given by

$$T = 2sN\left[1 + \frac{(s-1)^2}{4Nms^2}\right] \tag{7}$$

If we compare this equation with equation (2) for the case of $n = 2$, we obtain an equation for N_e, which is identical with equation (6).

Equation (6) indicates that when Nm is large, N_e is practically sN, as expected, but that N_e can be much greater than the total population size (sN) if $4Nm$ is smaller than 1.

Genetic Diversity and Coalescence Time

Nucleotide Diversity

Nucleotide diversity (π) is defined as the average number of nucleotide differences per site for all pairwise comparisons of the genes sampled from the same population. Tajima (1989) examined the expected value (H_1) of the number of nucleotide differences per DNA sequence for the case of two subpopulations ($s = 2$). In our approach, H_1 is given by $4N_e v$, where N_e is defined by (6) and v is the mutation rate per sequence. Obviously, $v = m_T\mu$, where m_T is the total number of nucleotides examined per sequence. Tajima considered the case of $4Nv = 1$ with various values of $4Nm$. According to his computation, the H_1 values [$S(2)$ in his notation] were 52.000, 7.000, 2.500, 2.050, and 2.005 for $4Nm = 0.001, 0.1, 1, 10$, and 100, respectively. (See Table 1 in Tajima's paper.) In our approach, the

Table 1. The mean and standard deviation (±) of the simulated coalescence time for the island model with s subpopulations[a]

$2Nm$	0.001	0.01	0.1	1.0	5	10
			$n = 2$			
$s = 2$	62.7 ± 106	6.96 ± 11.4	1.13 ± 1.46	0.56 ± 0.58	0.51 ± 0.51	0.51 ± 0.50
	(63.0 ± 63.0)	(6.75 ± 6.75)	(1.13 ± 1.13)	(0.56 ± 0.56)	(0.51 ± 0.51)	(0.51 ± 0.51)
$s = 4$	141 ± 182	15.1 ± 19.4	1.93 ± 2.29	0.64 ± 0.65	0.53 ± 0.53	0.50 ± 0.50
	(141 ± 141)	(14.6 ± 14.6)	(1.19 ± 1.19)	(0.64 ± 0.64)	(0.53 ± 0.53)	(0.51 ± 0.51)
$s = 8$	194 ± 216	19.4 ± 22.2	2.37 ± 2.61	0.69 ± 0.71	0.54 ± 0.54	0.52 ± 0.50
	(192 ± 192)	(19.6 ± 19.6)	(2.41 ± 2.41)	(0.69 ± 0.69)	(0.54 ± 0.54)	(0.52 ± 0.52)
$s = 16$	220 ± 236	22.8 ± 24.0	2.69 ± 2.82	0.72 ± 0.72	0.55 ± 0.55	0.53 ± 0.53
	(220 ± 220)	(22.5 ± 22.5)	(2.70 ± 2.70)	(0.72 ± 0.72)	(0.54 ± 0.54)	(0.52 ± 0.52)
$s = 32$	233 ± 240	24.1 ± 24.4	2.90 ± 2.93	0.74 ± 0.75	0.54 ± 0.54	0.52 ± 0.54
	(235 ± 235)	(24.0 ± 24.0)	(2.85 ± 2.85)	(0.73 ± 0.73)	(0.55 ± 0.55)	(0.52 ± 0.52)
			$n = 3$			
$s = 2$	92.7 ± 121	10.1 ± 12.4	1.60 ± 1.55	0.75 ± 0.60	0.68 ± 0.54	0.67 ± 0.53
	(84.0 ± 66.4)	(9.00 ± 7.12)	(1.50 ± 1.19)	(0.75 ± 0.59)	(0.68 ± 0.54)	(0.68 ± 0.53)
$s = 4$	203 ± 196	20.3 ± 19.5	2.66 ± 2.36	0.86 ± 0.69	0.69 ± 0.54	0.69 ± 0.54
	(188 ± 149)	(19.4 ± 15.4)	(2.54 ± 2.01)	(0.85 ± 0.68)	(0.70 ± 0.56)	(0.69 ± 0.54)
$s = 8$	264 ± 233	26.9 ± 23.7	3.22 ± 2.71	0.93 ± 0.73	0.73 ± 0.56	0.70 ± 0.55
	(256 ± 202)	(26.2 ± 20.7)	(3.22 ± 2.54)	(0.92 ± 0.73)	(0.72 ± 0.57)	(0.69 ± 0.55)
$s = 16$	290 ± 239	30.5 ± 25.6	3.60 ± 2.94	0.96 ± 0.75	0.73 ± 0.60	0.70 ± 0.56
	(294 ± 232)	(30.0 ± 23.7)	(3.60 ± 2.84)	(0.96 ± 0.76)	(0.73 ± 0.57)	(0.70 ± 0.55)
$s = 32$	316 ± 257	32.1 ± 25.9	3.81 ± 3.07	0.98 ± 0.77	0.73 ± 0.58	0.70 ± 0.54
	(314 ± 248)	(32.0 ± 25.3)	(3.80 ± 3.00)	(0.98 ± 0.77)	(0.73 ± 0.58)	(0.70 ± 0.55)
			$n = 10$			
$s = 2$	127 ± 129	13.4 ± 12.7	2.15 ± 1.62	1.01 ± 0.60	0.92 ± 0.54	0.91 ± 0.55
	(113 ± 67.8)	(12.2 ± 7.26)	(2.03 ± 1.21)	(1.01 ± 0.61)	(0.92 ± 0.55)	(0.91 ± 0.54)
$s = 4$	254 ± 196	26.6 ± 20.9	3.41 ± 2.46	1.13 ± 0.70	0.94 ± 0.57	0.91 ± 0.55
	(254 ± 152)	(26.2 ± 15.7)	(3.43 ± 2.05)	(1.15 ± 0.67)	(0.95 ± 0.57)	(0.93 ± 0.55)
$s = 8$	350 ± 234	35.5 ± 23.6	4.31 ± 2.86	1.23 ± 0.76	0.95 ± 0.58	0.94 ± 0.56
	(345 ± 207)	(35.4 ± 21.1)	(4.34 ± 2.60)	(1.25 ± 0.74)	(0.97 ± 0.57)	(0.93 ± 0.56)
$s = 16$	400 ± 251	40.4 ± 24.7	4.91 ± 3.02	1.28 ± 0.78	0.96 ± 0.58	0.94 ± 0.57
	(396 ± 237)	(40.5 ± 24.2)	(4.86 ± 2.90)	(1.30 ± 0.77)	(0.98 ± 0.59)	(0.94 ± 0.56)
$s = 32$	435 ± 263	43.0 ± 26.0	5.22 ± 3.14	1.30 ± 0.78	0.97 ± 0.59	0.93 ± 0.56
	(423 ± 253)	(43.1 ± 25.8)	(5.12 ± 3.06)	(1.32 ± 0.79)	(0.98 ± 0.59)	(0.94 ± 0.56)

[a] The coalescence time is measured in units of $4Ns$ generations. n is the number of genes sampled randomly from s subpopulations. The number of replications for each set of parameter values was 5,000. m: migration rate per generation. N: number of breeding individuals in each subpopulation. The values in the parentheses are the mean and standard deviation of coalescence time that were obtained from equation (8) and equation (9), respectively.

effective population size for the entire population ($s = 2$) is $N_e = 2N[1 + 1/(16Nm)]$, and $H_1 = 8Nv[1 + 1/(16Nm)] = 2 + 1/(8Nm)$, since $4Nv = 1$. Thus, Tajima's results are in complete agreement with the H_1 values obtained by the effective size approach.

It should be noted, however, that the effective size approach does not necessarily give the correct values for the expected number of segregating sites [$S(n)$] except for the case of sample size (n) equal to 2, for which $S(n) = H_1$. This approach gives the correct results for $n > 2$ only when $4Nm \geq 10$. It should also be noted that for this approach to work for H_1 the genes must be randomly chosen from the entire population.

Coalescence Time

Takahata (1991) derived an approximate formula for the mean coalescence time. The formula requires specification of the number (r) of subpopulations in which at least one gene is sampled. In the present approach genes are sampled at random from the entire population, and the mean coalescence time is given by

$$T = 4sN\left[1 + \frac{(s-1)^2}{4Nms^2}\right]\left(1 - \frac{1}{n}\right) \qquad (8)$$

Takahata's formula becomes essentially the same as the above when his r is equal to s and n is large (say $n > 20$).

To check the accuracy of equation (8), we conducted a computer simulation. The method of computer simulation was similar to that of Takahata (1991) except that all genes were sampled from the entire population. The results of the simulation, which are presented in Table 1, show that the above formula is very accurate when $n = 2$. For $n = 3$ or

10, however, it gives a slight underestimate of the mean coalescence time when s is small and $2Nm \leq 0.1$. When s is equal to 8 or larger and $2Nm \geq 0.01$, the agreement between the simulation results and the predictions from equation (8) is excellent for any value of n.

Table 1 also gives the standard deviations (s_T) of coalescence time obtained from the computer simulation and those obtained from Tajima's (1983) formula for the case of a randomly mating population. The latter formula is given by

$$s_T = 4N_e\left[\sum_{i=2}^{n}\left\{\frac{1}{i\,(i-1)}\right\}^2\right]^{1/2} \qquad (9)$$

Table 1 indicates that in a randomly mating population the standard deviation is identical with the mean when $n = 2$, as expected from comparison of equations (2) and (9), but is slightly smaller than the latter for $n \geq 3$. The results from our computer simulation indicate that equation (9) is applicable even to subdivided populations as long as $2Nm$ is 1.0 or larger. When $2Nm$ is less than 0.1, however, it gives an underestimate of the standard deviation for subdivided populations, particularly when s is small. Nevertheless, equation (9) is useful for obtaining an approximate value of the standard deviation of coalescence time when s is large.

Heterozygosity

Nei (1975) and Takahata (1983) studied the expected heterozygosity (H_T) for the total population when the population is subdivided following the finite island model. According to Takahata (1983), the expected homozygosity (J_T), which is defined as $1 - H_T$, is given by

$$J_T = \left[1 + 4Nsv + \frac{(s-1)^2 v}{sm + (s-1)v}\right]^{-1} \qquad (10)$$

If we assume $m >> v$, the above equation reduces to

$$J_T = \left[1 + 4Ns\left\{1 + \frac{(s-1)^2}{4Nms^2}\right\}v\right]^{-1} \qquad (11)$$

approximately. Therefore, if we use N_e defined in equation (6), J_T is given by $(1 + 4N_e v)^{-1}$, and $H_T \equiv 1 - J_T$ is given by equation (3). This indicates that equation (3) applies even to a subdivided population. However, when the condition $m >> v$ does not apply, H_T in (3) tends to give an overestimate of the true value.

Discussion

The concept of effective population size, which was first developed by Wright (1931), has been very useful for simplifying mathematical treatments in population genetics. In the present paper we have shown that this is exactly the case with expected nucleotide diversity, coalescence time, and heterozygosity for the finite island model. There is no need to use complicated formulas previously proposed as long as those quantities are concerned. Maruyama's (1972) complicated formulation of the distribution of gene frequencies in a geographically structured population can also be simplified tremendously if we use the N_e approach. The finite island model has been used for studying various properties of genetic polymorphism in subdivided populations. Yet, Wright's formula [equation (4)] has never been used as far as we know. This is probably because it was not clear what value of F_i should be used in his formula.

As mentioned earlier, equation (6) shows that the effective population size of a subdivided population can be much larger than the total population size when the migration rate among subpopulations is very small. This is intuitively obvious because the genetic variability among subpopulations is expected to increase continuously with time if there is no migration. However, it should be emphasized that our formulation of N_e depends on the assumption that the population structure remains the same for a long evolutionary time. In many natural populations this assumption may not hold. In some species such as *Escherichia coli* and *Drosophila,* subpopulations of a species are often subject to extinction and rapid multiplication. In this case the effective population size can be much smaller than the actual size (sN) (Nei 1976; Maruyama and Kimura 1980).

Nevertheless, there are cases in which structured populations remain more or less the same for a long period of time. Good examples are macaque species, which are subdivided into many matrilineally isolated troops. In these species intertroop nucleotide diversity for mitochondrial DNA (mtDNA) is much higher than intratroop diversity (Hayasaka *et al.* 1988; D.J. Melnik, personal communication). The coalescence time of mtDNA types is also much longer than that for other mammalian species (e.g., humans). Our mathematical formulas developed in this paper should be useful for analyzing data obtained from these species. They will also be useful

for resolving the current controversy over the evolution of *Homo sapiens* from *H. erectus*. (See Takahata 1991, 1993.)

Acknowledgments. This research was supported by NIH grant GM20293 and NSF grant DEB-9119802.

References

Crow JF, Aoki K (1984) Group selection for a polygenic behavioral trait: estimating the degree of population subdivision. Proc Natl Acad Sci USA 81:6073–6077

Hayasaka K, Horai S, Gojobori T, Shotake T, Nozawa K, Matsunaga E (1988) Phylogenetic relationships among Japanese, Rhesus, Formosan, and crab-eating monkeys, inferred from restriction-enzyme analysis of mitochondrial DNAs. Mol Biol Evol 5:270–281

Hey J (1991) A multi-dimensional coalescence process applied to multi-allelic selection models and migration models. Theor Popul Biol 39:30–48

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, Crow JF (1964) The number of alleles that can be maintained in a finite population. Genetics 49:725–738

Kingman JFC (1982) On the genealogy of large populations. J Appl Probab 19A:27–43

Maruyama T (1970) Effective number of alleles in a subdivided population. Theor Popul Biol 1:273–306

Maruyama T (1972) Distribution of gene frequencies in a geographically structured population. I. Distribution of neutral genes and of genes with small effect. Ann Hum Genet 35: 411–423

Maruyama T, Kimura M (1980) Genetic variability and effective population size when local extinction and recolonization of subpopulations are frequent. Proc Natl Acad Sci USA 77: 6710–6714

Nei M (1975) Molecular population genetics and evolution. North Holland and American Elsevier, Amsterdam, New York

Nei M (1976) The cost of natural selection and the extent of enzyme polymorphism. Trends Biochem Sci 1:247–248

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York

Nei M, Chakravarti A, Tateno Y (1977) Mean and variance of F_{ST} in a finite number of incompletely isolated populations. Theor Popul Biol 11:291–306

Notohara M (1990) The coalescent and the genealogical process in geographically structured population. J Math Biol 29:59–75

Slatkin M (1991) Inbreeding coefficient and coalescence times. Genet Res 58:167–175

Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437–460

Tajima F (1989) DNA polymorphism in a subdivided population: the expected number of segregating sites in the two-subpopulation model. Genetics 123:229–240

Takahata N (1983) Gene identity and genetic differentiation of populations in the finite island model. Genetics 104:497–512

Takahata N (1988) The coalescent in two partially isolated diffusion populations. Genet Res 52:213–222

Takahata N (1991) Genealogy of neutral genes and spreading of selected mutations in a geographically structured population. Genetics 129:585–595

Takahata N (1993) Allelic genealogy and human evolution. Mol Biol Evol 10:2–22

Takahata N, Nei M (1984) F_{ST} and G_{ST} statistics in the finite island model. Genetics 107:501–504

Takahata N, Slatkin M (1990) Genealogy of neutral genes in two partially isolated populations. Theor Popul Biol 38:331–350

Watterson GA (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

Wright S (1943) Isolation by distance. Genetics 28:114–138

Wright S (1951) The genetical structure of populations. Ann Eugenics 15:323–354