Genet. Res., Camb. (2002), **80**, *pp.* 15–25. *With 5 figures.* © 2002 Cambridge University Press
DOI: 10.1017/S0016672302005748 Printed in the United Kingdom

A statistical test for the difference in the amounts of DNA variation between two populations

HIDEKI INNAN† AND FUMIO TAJIMA*
Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Hongo 7-3-1, Tokyo 113-0033, Japan

(*Received 22 November 2001 and in revised form 15 March 2002*)

Summary

A statistical test for the difference in the amounts of DNA variation between two populations is developed. The test statistic involves the covariance of the amount of variation between two populations, which is given by a function of their divergence time, T_0. Accordingly, the power (rejection probability) of the test depends on T_0. In this article, T_0 is treated as unknown because it is very difficult to estimate. The test is most conservative when $T_0 = \infty$ is assumed because the covariance is zero. If $T_0 = 0$ is assumed, the largest value of the rejection probability is obtained. Thus, the test provides a range of rejection probability unless we have a reliable estimate of T_0. The test is applied to the *PgiC* region in three mustard species: *Leavenworthia stylosa*, *L. crassa* and *L. uniflora*. The results of our test show that the level of variation in *L. stylosa* is significantly higher than those in the other species.

1. Introduction

The amount of DNA variation varies among species, even among local populations in the same species (Kimura, 1983; Nei, 1987; Gillespie, 1991). In addition to random genetic drift, there are many possible reasons for such a difference. One is the difference in the population size and/or mutation rate, because the expectation of the amount of variation depends on the product of the population size and mutation rate under the neutral theory (Kimura, 1968, 1983). Natural selection also affects the level of DNA variation. For example, a recent selective sweep event reduces the amount of variation dramatically.

The purpose of this article is to develop a statistical test that examines whether or not the observed difference in the amounts of variation can be explained by random genetic drift alone. A test statistic is introduced, which is based on the observed numbers of segregating sites in two populations. As the reason for having to account for the positive correlation of the amounts of variation between two populations (Takahata & Nei, 1985; Wakeley, 1996; Wakeley & Hey, 1997), the test statistic involves their covariance. We develop a recurrence formula to calculate the covariance.

The test is applied to DNA polymorphism data of three mustard species. Liu *et al.* (1999) investigated DNA variation in the *PgiC* region of *Leavenworthia stylosa*, *L. crassa* and *L. uniflora*. The interaction between the amount of variation and breeding system was studied. These three species have different breeding systems. *L. stylosa* is an outcrossing species, while *L. uniflora* is a selfer. The rate of selfing in *L. crassa* is moderate. As expected from the difference in breeding system, the observed levels of nucleotide variations in these three species are different, but it is not clear whether the difference is statistically significant. The application of our test reveals that the amount of variation in *L. stylosa* is significantly larger than those in the other species.

2. Model and statistical test

To test whether the difference in the amounts of nucleotide variation between two populations can be explained by random genetic drift alone, a simple two-

* Corresponding author. Tel: +81 3 5841 4051. Fax: +81 3 3818 5399. e-mail: ftajima@biol.s.u-tokyo.ac.jp
† Present address: Department of Biological Science, University of Southern California, Los Angeles, CA 90089-1340, USA.

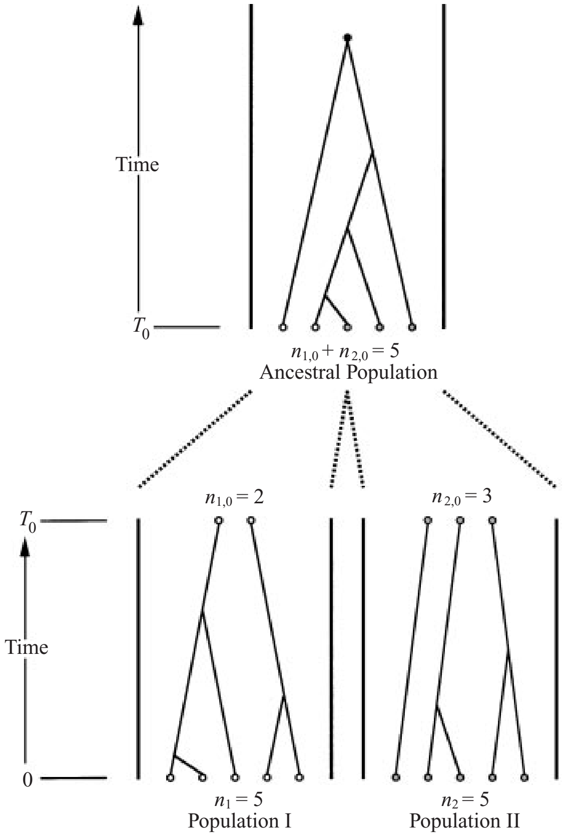

Fig. 1. Model. The open circles represent sampled sequences from population I, and the shaded circles those from population II. The filled circle represents the most recent common ancestor of the whole sample.

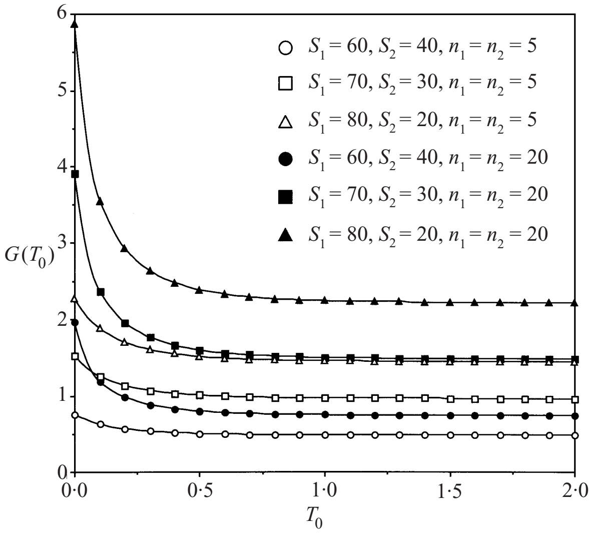

Fig. 2. Test statistic $G(T_0)$ for $n_1 = n_2 = 5$ and 20. $G(T_0)$ was calculated by (3).

population model is considered (Fig. 1). Each population consists of N diploid individuals. The two populations are derived from their ancestral population with the same size N, and the divergence time is given by T_0 (the unit is $4N$ generations). Random mating is assumed and random mutations occur at a constant rate, μ, per sequence per generation.

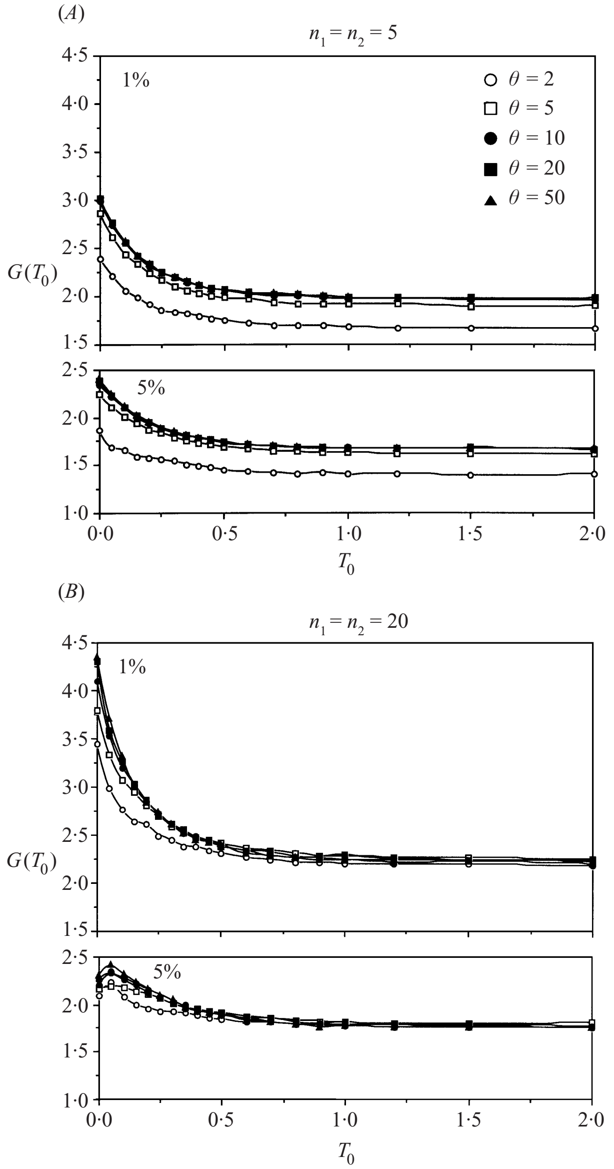

Fig. 3. Critical values of $G(T_0)$ obtained by computer simulations with 50000 replications. (*A*) Results for $n_1 = n_2 = 5$. (*B*) Results for $n_1 = n_2 = 20$.

Consider that n_1 and n_2 DNA sequences are sampled from the two populations, I and II, respectively. Under the neutral infinite site model (Kimura, 1969; Watterson, 1975), the amount of nucleotide variation, θ $(= 4N\mu)$, can be estimated from the number of segregating sites. Let S_1 and S_2 be the observed number of segregating sites in populations I and II, respectively. From S_1 and S_2, estimates of θ in populations I and II are

$$\hat{\theta}_1 = S_1/a_1 \text{ and } \hat{\theta}_2 = S_2/a_2, \tag{1}$$

where

$$a_1 = \sum_{k=1}^{n_1-1}\frac{1}{k} \text{ and } a_2 = \sum_{k=1}^{n_2-1}\frac{1}{k} \tag{2}$$

(Watterson, 1975).

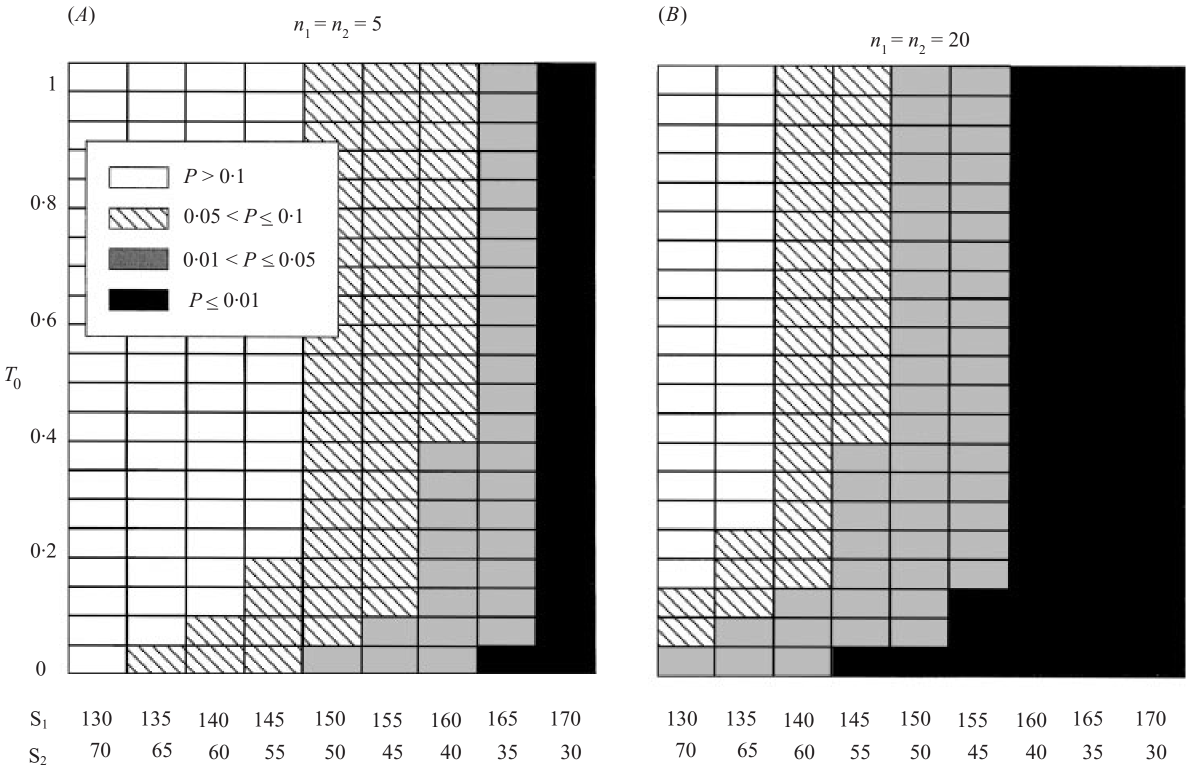

Fig. 4. Power of the test. The rejection probability was investigated when $S_1 + S_2 = 200$. (*A*) Results for $n_1 = n_2 = 5$. (*B*) Results for $n_1 = n_2 = 20$.

To examine whether the difference between $\hat{\theta}_1$ and $\hat{\theta}_2$ can be explained by random genetic drift alone, a test statistic is developed:

$$G(T_0) = \frac{|\hat{\theta}_1 - \hat{\theta}_2|}{\sqrt{\mathrm{Var}(\hat{\theta}_1 - \hat{\theta}_2)}}, \qquad (3)$$

where $\mathrm{Var}(\hat{\theta}_1 - \hat{\theta}_2)$ is the variance of $\hat{\theta}_1 - \hat{\theta}_2$. $G(T_0)$ is given by a function of T_0 because T_0 affects the variance. The derivations of $\mathrm{Var}(\hat{\theta}_1 - \hat{\theta}_2)$ are shown in Appendixes A and B. Fig. 2 shows numerical examples of $G(T_0)$ for $n_1 = n_2 = 5$ and 20. The test statistics decrease with increasing T_0, and their distributions are almost flat when $T_0 > 1$ because the covariance between $\hat{\theta}_1$ and $\hat{\theta}_2$ is nearly zero.

Computer simulations are carried out to investigate the distribution of $G(T_0)$, which may depend on θ, n_1, n_2 and T_0. The standard coalescent process without recombination is used in the simulations (Griffiths, 1980; Kingman, 1982; Hudson, 1983; Tajima, 1983). In each replication of the simulation, first, a random genealogy of n_1 sequences in population I was simulated backward in time from 0 to T_0, and we determine the number of sequences, $n_{1,0}$, at T_0 (Fig. 1*B*). In the same way, a random genealogy of n_2 sequences in population II is constructed, and the number of sequences, $n_{2,0}$, at T_0 is determined. Next a random genealogy of $n_{1,0} + n_{2,0}$ sequences in the ancestral population is constructed (Fig. 1*A*). After giving random mutations on the genealogy, S_1 and S_2 are obtained. Using (1)–(3), $G(T_0)$ is calculated given T_0. In the calculation, $(\hat{\theta}_1 + \hat{\theta}_2)/2$ is used as an estimate of $\theta(\hat{\theta})$.

Fig. 3 shows the results of computer simulations for $n_1 = n_2 = 5$ and 20. The critical values at the 5% and 1% levels are shown for $\theta = 2$, 5, 10, 20 and 50. In the case of $n_1 = n_2 = 5$ (Fig. 3*A*), the critical values decrease with increasing T_0, and the curves become almost flat when $T_0 > 1$. When $\theta = 2$ and 5, the critical values are smaller than those for $\theta = 10$, 20 and 50. The distributions of the critical values for $\theta =$ 10, 20 and 50 are similar. When $n_1 = n_2 = 20$ (Fig. 3*B*), the critical values at the 1% level decrease with increasing T_0, while the distributions of the 5% critical values have peaks at $T_0 \approx 0·1$. The distributions of critical values for $\theta = 2$ and 5 are not much smaller than those for the others. It is indicated that the effect of θ on the distribution of $G(T_0)$ is not large unless θ is very small.

The effect of T_0 on the power of the test is investigated when $n_1 = n_2 = 5$ and 20. Given S_1 and S_2, the probability that rejects the null hypothesis ($\theta_1 = \theta_2$) is obtained for $T_0 = 0$, 0·05, 0·1, 0·15, … , 1. $G(T_0)$ is calculated for each value of T_0, and the P value for each $G(T_0)$ is determined by a computer simulation. The results for $(S_1, S_2) = (130, 70)$, (135, 75), … , (165, 35), (170, 30) are shown in Fig. 4. The P value increases with increasing T_0 because the covariance between S_1 and S_2 decreases. For example,

Table 1. *Critical values of* G(T_0) *for* $T_0 = \infty$ *and 0*

		$T_0 = \infty$			$T_0 = 0$		
n_1	n_2	5%	2·5%	1%	5%	2·5%	1%
5	5	1·67	1·82	1·98	2·41	2·77	3·07
5	6	1·69	1·85	2·03	2·47	2·84	3·23
5	7	1·70	1·88	2·04	2·49	2·86	3·26
5	8	1·70	1·87	2·04	2·50	2·88	3·26
5	9	1·72	1·90	2·08	2·48	2·90	3·32
5	10	1·71	1·89	2·08	2·50	2·97	3·36
5	12	1·74	1·92	2·13	2·47	2·96	3·36
5	14	1·74	1·94	2·15	2·48	3·00	3·48
6	6	1·70	1·88	2·05	2·53	2·91	3·33
6	7	1·71	1·90	2·05	2·52	2·96	3·40
6	8	1·71	1·89	2·07	2·55	3·03	3·49
6	9	1·71	1·90	2·08	2·54	3·04	3·56
6	10	1·73	1·93	2·14	2·53	3·01	3·56
6	12	1·74	1·92	2·14	2·54	3·03	3·55
6	14	1·75	1·93	2·16	2·50	3·04	3·57
6	16	1·77	1·97	2·17	2·51	3·06	3·59
7	7	1·70	1·90	2·09	2·56	3·05	3·55
7	8	1·72	1·91	2·10	2·57	3·08	3·61
7	9	1·73	1·90	2·10	2·54	3·06	3·60
7	10	1·74	1·92	2·13	2·58	3·19	3·69
7	12	1·76	1·94	2·15	2·55	3·16	3·70
7	14	1·75	1·95	2·18	2·54	3·12	3·69
7	16	1·78	1·98	2·18	2·54	3·11	3·73
7	18	1·76	1·96	2·19	2·56	3·07	3·74
8	8	1·73	1·89	2·08	2·56	3·09	3·60
8	9	1·75	1·94	2·11	2·55	3·10	3·74
8	10	1·73	1·93	2·12	2·56	3·12	3·73
8	12	1·75	1·95	2·14	2·55	3·16	3·79
8	14	1·74	1·94	2·17	2·57	3·18	3·85
8	16	1·75	1·95	2·17	2·55	3·18	3·79
8	18	1·75	1·97	2·18	2·52	3·14	3·81
8	20	1·75	1·96	2·19	2·53	3·14	3·80
9	9	1·74	1·91	2·09	2·54	3·09	3·70
9	10	1·74	1·94	2·14	2·59	3·25	3·79
9	12	1·75	1·95	2·17	2·57	3·24	3·83
9	14	1·77	1·97	2·17	2·59	3·24	3·91
9	16	1·77	1·97	2·18	2·55	3·24	3·93
9	18	1·77	1·99	2·20	2·51	3·23	3·98
9	20	1·77	1·99	2·22	2·58	3·21	3·88
9	25	1·79	2·00	2·24	2·45	3·11	3·82
10	10	1·74	1·95	2·15	2·54	3·24	3·88
10	12	1·75	1·94	2·17	2·58	3·22	4·00
10	14	1·75	1·96	2·16	2·55	3·28	3·98
10	16	1·75	1·96	2·18	2·55	3·26	4·00
10	18	1·77	1·97	2·19	2·56	3·28	4·05
10	20	1·76	1·98	2·22	2·53	3·32	4·06
10	25	1·79	2·02	2·22	2·48	3·20	3·95
10	30	1·79	2·02	2·25	2·47	3·14	3·93
12	12	1·76	1·96	2·19	2·48	3·25	4·02
12	14	1·76	1·99	2·19	2·47	3·26	4·08
12	16	1·76	1·98	2·19	2·53	3·31	4·16
12	18	1·77	1·99	2·19	2·49	3·35	4·18
12	20	1·75	1·99	2·21	2·45	3·18	4·13
12	25	1·79	2·01	2·25	2·38	3·19	4·13
12	30	1·77	1·99	2·26	2·37	3·20	4·01
14	14	1·76	1·96	2·20	2·34	3·27	4·01
14	16	1·76	1·98	2·20	2·41	3·32	4·17
14	18	1·77	1·99	2·23	2·43	3·33	4·19
14	20	1·77	2·00	2·22	2·54	3·34	4·18

Table 1. (*cont.*)

		$T_0 = \infty$			$T_0 = 0$		
n_1	n_2	5%	2·5%	1%	5%	2·5%	1%
14	25	1·79	2·02	2·25	2·46	3·28	4·20
14	30	1·82	2·04	2·29	2·40	3·27	4·28
16	16	1·79	1·99	2·24	2·33	3·28	4·03
16	18	1·80	1·99	2·25	2·46	3·29	4·19
16	20	1·79	1·99	2·24	2·39	3·40	4·29
16	25	1·81	2·04	2·27	2·36	3·37	4·46
16	30	1·83	2·04	2·27	2·29	3·30	4·41
18	18	1·79	2·01	2·25	2·33	3·36	4·32
18	20	1·79	2·03	2·27	2·38	3·30	4·38
18	25	1·82	2·03	2·27	2·28	3·23	4·39
18	30	1·80	2·05	2·28	2·21	3·23	4·38
20	20	1·84	2·03	2·27	2·33	3·33	4·39
20	25	1·82	2·05	2·29	2·22	3·21	4·37
20	30	1·83	2·04	2·29	2·21	3·23	4·46
25	25	1·84	2·06	2·28	2·23	3·21	4·58
25	30	1·84	2·07	2·31	2·17	3·24	4·71
30	30	1·84	2·09	2·32	2·15	3·28	4·74

For a parameter set (n_1, n_2 and T_0), coalescent simulations with 10000 replications were conducted for $\theta = 2$, 3, 5, 10, 20, 30 and 50, and the most conserved values are shown. The simulation program is available on request from the authors.

when $n_1 = n_2 = 20$ and $(S_1, S_2) = (145, 55)$, P is smaller than 0·05 when $T_0 \leqslant 0{\cdot}35$, while P is larger than 0·05 when $T_0 \geqslant 0{\cdot}4$.

Here, we show how to evaluate the P value for a given DNA sequence data set when T_0 is unknown. Since the power of the test decreases with increasing T_0, we can conduct the most conservative test assuming $T_0 = \infty$. This test may be useful because it is usually very difficult to obtain a reliable estimate of T_0 from DNA sequence data. The covariance between S_1 and S_2 is zero when $T_0 = \infty$ so that calculation of $G(\infty)$ is easy. Table 1 shows the critical values of $G(\infty)$ at the 5%, 2·5% and 1% levels, which are determined by computer simulations. Following Fu & Li (1993), for a given parameter set of n_1, n_2 and T_0, the critical values are investigated for $\theta = 2$, 3, 5, 10, 20, 30 and 50, and the most conservative one is shown. Since the effect of θ on the distribution of the test statistic is not very large (see Fig. 3), the critical values in Table 1 are not very conservative. However, as pointed out by Fu (1996), the critical values obtained from a simulation with an estimate of θ from the data may be more powerful, especially when θ and sample size are small (see Fig. 3). If the result of this most conservative test is significant, the observed difference is significant with no condition. For example, in the cases of $(S_1, S_2) = (150, 50)$, $(155, 45)$, $(160, 40)$, $(165, 35)$ and $(170, 30)$ when $n_1 = n_2 = 20$, the test is significant at the 5% level for any value of T_0 (Fig. 4*B*).

On the other hand, the smallest P value is given when T_0 is assumed to be 0. Table 1 also shows the critical values of $G(0)$ at the 5%, 2·5% and 1% levels. When $G(0)$ is not significant, the null hypothesis cannot be rejected for any value of T_0. When $G(0)$ is significantly large (e.g. at the 5% level), it indicates that there is possibility that the null hypothesis is rejected even if $G(\infty)$ is not significant. For example, the cases of $(S_1, S_2) = (130, 70)$, $(135, 65)$, $(140, 60)$ and $(145, 55)$ when $n_1 = n_2 = 20$ are in this situation (Fig. 4*B*), and the rejection of the null hypothesis depends on T_0.

Thus, our test provides a range of P values for a given data set. Let $P(T_0)$ be the P value for $G(T_0)$. The maximum and minimum P values, $P_{\max}$ and $P_{\min}$, are given by $P_{\max} = P(\infty)$ and $P_{\min} = P(0)$, respectively. As an example, consider when $(S_1, S_2) = (130, 70)$ and $n_1 = n_2 = 20$, where $G(\infty) = 1{\cdot}15$ and $G(0) = 3{\cdot}08$. Computer simulations show that $P_{\max}$ and $P_{\min}$ are 0·215 and 0·030, respectively, indicating $0{\cdot}030 < P < 0{\cdot}215$. From additional simulations with $T_0 = 0{\cdot}01$, 0·02, 0·03, … , we obtain $P(0{\cdot}03) = 0{\cdot}049$ and $P(0{\cdot}04) = 0{\cdot}055$, indicating that T_0 that gives $P(T_0) = 0{\cdot}05$ is about 0·03. We refer to such T_0 as $T_{0,5\%}$. It is indicated that the difference is significantly large if the real divergence time is smaller than $T_{0,5\%} \approx 0{\cdot}03$. Although the real T_0 is not easy to know, $T_{0,5\%}$ is a useful summary of the observed difference in the amount of variation.

3. Application to *Leavenworthia* species

The test is applied to the DNA sequence data in the *PgiC* intron 12 region of three mustard species: *Leavenworthia stylosa*, *L. crassa* and *L. uniflora* (Liu *et*

Table 2. *Application of the test to* Levenworthia *species*

Species I	Species II	n_1	n_2	S_1	S_2	θ_1	θ_2	$G(\infty)$	$P(\infty)$
L. stylosa	*L. crassa*	26	45	33[a]	5	8·65	1·14	2·82	0·0009
L. stylosa	*L. uniflora*	26	11	29[a]	0	7·60	0·00	2·98	0·0001
L. crassa	*L. uniflora*	45	11	5	0	1·14	0·00	1·78	0·1285[b]

[a] The difference in S_1 of *L. stylosa* is due to the difference in the regions used in the analysis.

[b] The simulation method of this comparison is different from the other two because $S_1 + S_2$ is small. See text.

al., 1999). The largest amount of variation is observed in *L. stylosa*, an outcrossing species, while a selfer, *L. uniflora*, has no nucleotide variation. *L. crassa* is self-compatible with an intermediate level of outcrossing, and the amount of variation in this species is moderate.

First, we investigate the significance of the difference in the amounts of variation between *L. stylosa* and *L. crassa*. We refer to *L. stylosa* as species I and *L. crassa* as species II (Table 2). The observed numbers of segregating sites of the two species are 33 and 5, and estimates of θ are $\hat{\theta}_1 = 8·65$ and $\hat{\theta}_2 = 1·14$. The absolute value of the difference in $\hat{\theta}$ is 7·51 and $G(\infty)$ is 2·82. A simulation with $\theta = 4·90$ shows $P(\infty) =$ 0·0009, indicating that the difference is highly significant.

Next, *L. stylosa* and *L. uniflora* are compared. The observed numbers of segregating sites of the two species are 29 and 0. $G(\infty)$ is 2·98, and $P(\infty) < 0·0001$ is obtained by a simulation with $\theta = 3·80$, indicating that the difference is highly significant.

In the comparison between *L. crassa* and *L. uniflora*, $G(\infty) = 1·78$ is given. In this case, we conducted a simulation in which $S_1 + S_2 = 5$ is fixed, because we have the observation of $S_1 = S_2 = 0$ many times in a simulation with $\theta = 0·57$. It is demonstrated that the probability of $(S_1, S_2) = (5, 0)$ or $(0, 5)$ is 0·1285 when $T_0 = \infty$, indicating the difference is not significant. Because of a lack of knowledge of T_0, the test is not conducted for another smaller value of T_0.

4. Discussion

(i) *Theory and statistical test*

The amounts of variation of two populations have positive correlation unless their divergence time is very long. A recursion formula is developed to calculate the covariance between the number of segregating sites in the two populations. The theory assumes that the population sizes of two descendant populations and their ancestral population are the same, according to our purpose of developing a statistical test for the difference in the amount of variation.

Using this theoretical result, a statistical test for the difference in the amounts of nucleotide variation between two populations was developed. The test is based on the fact that we do not know T_0 because an estimate of T_0 usually has a huge variance. The value of the test statistic depends on the divergence time of the two populations, T_0, because the covariance of the amounts of variation in the two populations is involved in the test statistic. The power of the test increases with decreasing T_0. Therefore, the test provides a range of the rejection probability of the null hypothesis (i.e. $P_{\min} < P < P_{\max}$). For any value of T_0 the null hypothesis is rejected at the 5% level if $P_{\max} < 0·05$, while the test statistic is not significant when $P_{\min} > 0·05$.

Only when $P_{\min} < 0·05 < P_{\max}$ does the significance at the 5% level depends on T_0. In this case, $T_{0,5\%}$ is a useful summary for evaluating the difference. $T_{0,5\%}$ is defined such that $P(T_{0,5\%}) = 0·05$, indicating that the null hypothesis is rejected when the real T_0 is smaller than $T_{0,5\%}$.

The test assumes no migration after the divergence. Although the relationship between migration and coalescent is very difficult (Wakeley, 1996; Rosenberg & Feldman, 2002), it should be noted that this assumption makes the test conservative when the most conservative test with $T_0 = \infty$ is carried out.

We tried to develop another test that is independent of T_0. Using an estimate of T_0 from the data might make this possible. One of the successful examples is the HKA test (Hudson *et al.*, 1987). We investigated the possibility of developing such a test statistic. Unfortunately, simulations with a number of different parameter sets (n_1, n_2, T_0, θ) demonstrated that it does not work well as long as the estimate of T_0 has variance.

(ii) *On the estimation of* T_0

It is important to estimate the divergence time to evaluate the rejection probability of the test, especially when $P_{\min} < 0·05 < P_{\max}$. Several methods have been developed recently (e.g. Takahata *et al.*, 1995; Nielsen, 1998; Nielsen *et al.*, 1998; Edwards & Beerli, 2000). These methods are based on likelihood and take into account the effect of polymorphism in the ancestral population. It should be noted that an estimate of T_0 has a huge variance. The only way to reduce this variance is to use data from a number of independent (unlinked) loci. Although it may be very laborious to obtain DNA sequence data from many loci, we can obtain such data from RFLP (Nei & Li, 1979; Nei, 1987) or AFLP (Vos *et al.*, 1995; Innan *et al.*, 1999) analyses. RFLP and AFLP can reveal patterns of polymorphism and divergence of a random set of many DNA fragments, and we might be able to expect almost free recombination among fragments. With data from multiple loci, the above-mentioned methods give a likelihood function of T_0, $L(T_0)$. This function might be very useful for evaluating the rejection probability,

$$P = \int_0^\infty P(T_0)L(T_0)dT_0, \tag{4}$$

so that P can be obtained.

We thank Yun-Xin Fu for his helpful comments and suggestions. We also thank two anonymous reviewers for comments. This work was supported in part by a fellowship from Japan Society for the Promotion of Science to H.I. and by a grant-in-aid from the Ministry of Education, Science, Sports and Culture of Japan to F.T.

Appendix A

The variance of $\hat{\theta}_1 - \hat{\theta}_2$ is written as

$$\begin{aligned}\mathrm{Var}(\hat{\theta}_1 - \hat{\theta}_2) &= \mathrm{Var}(\hat{\theta}_1) + \mathrm{Var}(\hat{\theta}_2) - 2\mathrm{Cov}(\hat{\theta}_1, \hat{\theta}_2)\\ &= \frac{\mathrm{Var}(S_1)}{a_1^2} + \frac{\mathrm{Var}(S_2)}{a_2^2} - \frac{2\mathrm{Cov}(S_1, S_2)}{a_1 a_2},\end{aligned} \tag{A1}$$

where a_1 and a_2 are given by (2). The variances of S_1 and S_2 are given by

$$\mathrm{Var}(S_1) = a_1\theta + b_1\theta^2 \text{ and } \mathrm{Var}(S_2) = a_2\theta + b_2\theta^2, \tag{A2}$$

where

$$b_1 = \sum_{k=1}^{n_1-1} \frac{1}{k^2} \text{ and } b_2 = \sum_{k=1}^{n_2-1} \frac{1}{k^2} \tag{A3}$$

(Watterson, 1975).

Here, we consider the covariance between S_1 and S_2. S_1 is written as

$$S_1 = S_{1,0} + S_{1,1}, \tag{A4}$$

where $S_{1,0}$ is the number of mutations that occurred in the ancestral population before T_0 and $S_{1,1}$ is the number occurring in population I after divergence. In the same way, S_2 is given by

$$S_2 = S_{2,0} + S_{2,2}, \tag{A5}$$

where $S_{2,0}$ is the number of mutations in the ancestral population and $S_{2,2}$ is that in population II. Then, the covariance between S_1 and S_2 is given by

$$\begin{aligned}\mathrm{Cov}(S_{1,}, S_{2,}) &= \mathrm{Cov}(S_{1,0} + S_{1,1}, S_{2,0} + S_{2,2})\\ &= \mathrm{Cov}(S_{1,0}, S_{2,0}) + \mathrm{Cov}(S_{1,0}, S_{2,2}) + \mathrm{Cov}(S_{1,1}, S_{2,0}) + \mathrm{Cov}(S_{1,1}, S_{2,2})\\ &= \mathrm{Cov}(S_{1,0}, S_{2,0}),\end{aligned} \tag{A6}$$

because $\mathrm{Cov}(S_{1,0}, S_{2,2})$, $\mathrm{Cov}(S_{1,1}, S_{2,0})$ and $\mathrm{Cov}(S_{1,1}, S_{2,2})$ are 0.

The covariance of $S_{1,0}$ between $S_{2,0}$ is given by

$$\mathrm{Cov}(S_{1,0}, S_{2,0}) = E(S_{1,0}S_{2,0}) - E(S_{1,0})E(S_{2,0}), \tag{A7}$$

and the expectation of $S_{1,0}S_{2,0}$ is written as

$$E(S_{1,0}S_{2,0}) = \sum_{n_{1,0}=2}^{n_1} \sum_{n_{2,0}=2}^{n_2} P_{n_1,n_{1,0}} P_{n_2,n_{2,0}} E(S_{1,0}S_{2,0} \,|\, n_{1,0}, n_{2,0}). \tag{A8}$$

$P_{n_j,n_{j,0}}$ $(j = 1, 2)$ is given by

$$P_{n_j,n_{j,0}} = \sum_{k=n_{j,0}}^{n_j} \exp\{-k(k-1)T_0\} \frac{(2k-1)(-1)^{k-n_{j,0}} n_{j,0(k-1)} n_{j[k]}}{n_{j,0}!(k-n_{j,0})! n_{j(k)}}, \tag{A9}$$

where $n_{(k)} = n(n+1) \ldots (n+k-1)$ and $n_{[k]} = n(n-1) \ldots (n-k+1)$. (A9) is from equation (6.1) in Tavaré (1984) (also see Griffiths, 1979; Watterson, 1982; Takahata & Nei, 1985; Wakeley & Hey, 1997).

Since

$$\mathrm{Cov}(S_{1,0}, S_{2,0}\,|\,n_{1,0}, n_{2,0}) = E(S_{1,0}S_{2,0}\,|\,n_{1,0}, n_{2,0}) - E(S_{1,0}\,|\,n_{1,0})E(S_{2,0}\,|\,n_{2,0}), \tag{A10}$$

the expectation of $S_{1,0}S_{2,0}$ given $n_{1,0}$ and $n_{2,0}$ becomes

$$E(S_{1,0}S_{2,0}\,|\,n_{1,0}, n_{2,0}) = \mathrm{Cov}(S_{1,0}, S_{2,0}\,|\,n_{1,0}, n_{2,0}) - E(S_{1,0}\,|\,n_{1,0})E(S_{2,0}\,|\,n_{2,0}). \tag{A11}$$

Then, from (A7), (A8) and (A11), the covariance is given by

$$\mathrm{Cov}(S_{1,0}, S_{2,0}) = \sum_{n_{1,0}=2}^{n_1} \sum_{n_{2,0}=2}^{n_2} P_{n_1,n_{1,0}} P_{n_2,n_{2,0}} \mathrm{Cov}(S_{1,0}, S_{2,0}\,|\,n_{1,0}, n_{2,0}), \tag{A12}$$

because

$$\sum_{n_{1,0}=2}^{n_1} \sum_{n_{2,0}=2}^{n_2} P_{n_1,n_{1,0}} P_{n_2,n_{2,0}} E(S_{1,0}\,|\,n_{1,0})E(S_{2,0}\,|\,n_{2,0}) = E(S_{1,0})E(S_{2,0}), \tag{A13}$$

indicating that the unconditional covariance between $S_{1,0}$ and $S_{2,0}$ can be calculated given $\mathrm{Cov}(S_{1,0}, S_{2,0}\,|\,n_{1,0}, n_{2,0})$. The derivation for $\mathrm{Cov}(S_{1,0}, S_{2,0}\,|\,n_{1,0}, n_{2,0})$ is shown in Appendix B.

Appendix B

To obtain $\mathrm{Cov}(S_{1,0}, S_{2,0}\,|\,n_{1,0}, n_{2,0})$, it is helpful to consider the variance of $S_{1,0}+S_{2,0}$ given $n_{1,0}$ and $n_{2,0}$. The variance of $S_{1,0}+S_{2,0}$ can be obtained by using a recurrence formula according to the coalescence scheme as shown in Fig. B1. This approach is similar to that in Innan & Tajima (1997).

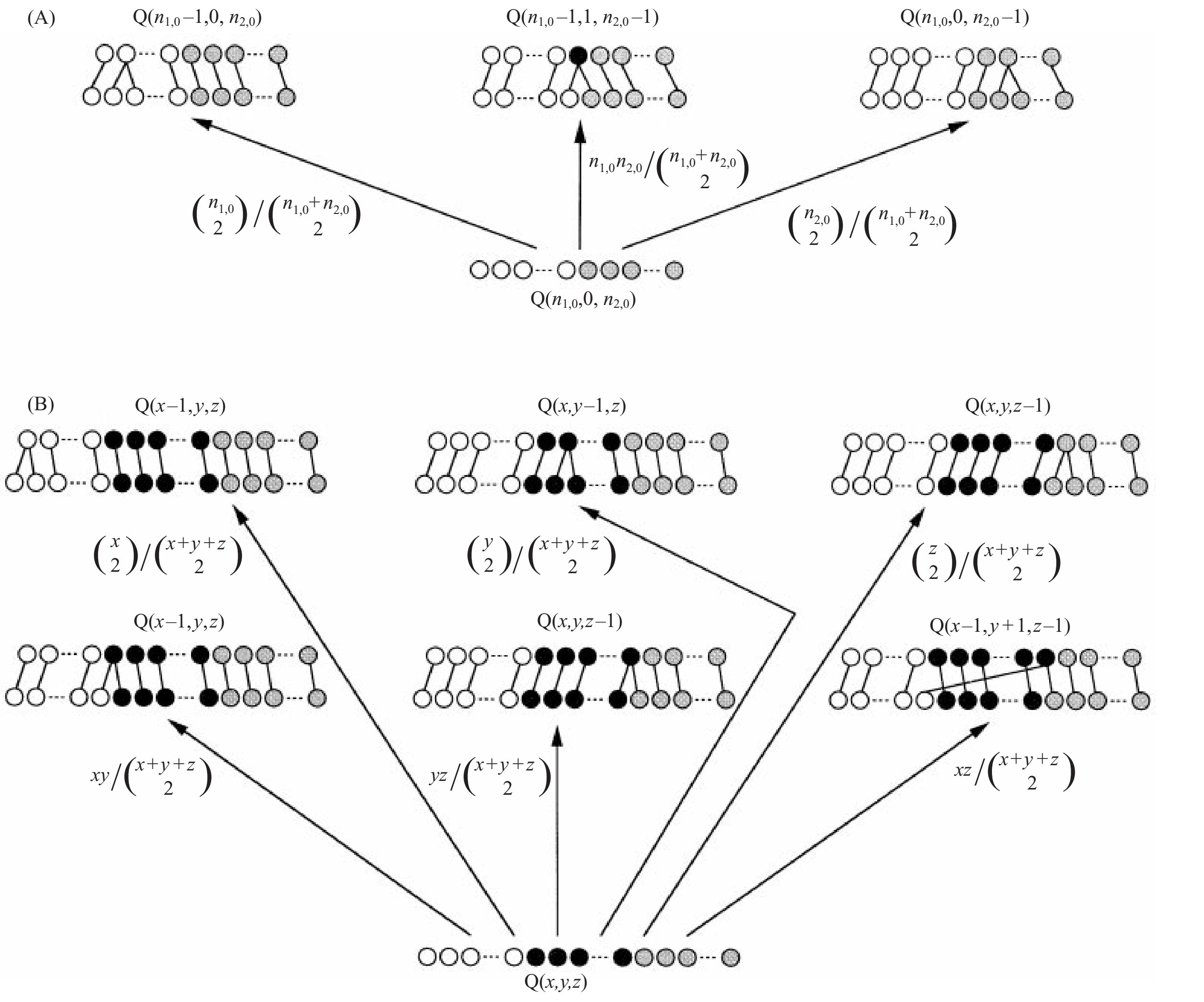
(A)
$Q(n_{1,0}-1, 0, n_{2,0})$
$Q(n_{1,0}-1, 1, n_{2,0}-1)$
$Q(n_{1,0}, 0, n_{2,0}-1)$
$n_{1,0}n_{2,0}/\binom{n_{1,0}+n_{2,0}}{2}$
$\binom{n_{1,0}}{2}/\binom{n_{1,0}+n_{2,0}}{2}$
$\binom{n_{2,0}}{2}/\binom{n_{1,0}+n_{2,0}}{2}$
$Q(n_{1,0}, 0, n_{2,0})$
(B)
$Q(x-1,y,z)$
$Q(x,y-1,z)$
$Q(x,y,z-1)$
$\binom{x}{2}/\binom{x+y+z}{2}$
$\binom{y}{2}/\binom{x+y+z}{2}$
$\binom{z}{2}/\binom{x+y+z}{2}$
$Q(x-1,y,z)$
$Q(x,y,z-1)$
$Q(x-1,y+1,z-1)$
$xy/\binom{x+y+z}{2}$
$yz/\binom{x+y+z}{2}$
$xz/\binom{x+y+z}{2}$
$Q(x,y,z)$

Fig. B1. Coalescent scheme used to calculate the variance of S_1+S_2.

Table B1. *Six patterns of coalescent events in* Q(x, y, z)

Coalescent event	Probability	Next state
Within X	$P_{XX} = \binom{x}{2} \Big/ \binom{x+y+z}{2}$	$Q(x-1, y, z)$
Within Y	$P_{YY} = \binom{y}{2} \Big/ \binom{x+y+z}{2}$	$Q(x, y-1, z)$
Within Z	$P_{ZZ} = \binom{z}{2} \Big/ \binom{x+y+z}{2}$	$Q(x, y, z-1)$
Between X and Y	$P_{XY} = xy \Big/ \binom{x+y+z}{2}$	$Q(x-1, y, z)$
Between Y and Z	$P_{YZ} = yz \Big/ \binom{x+y+z}{2}$	$Q(x, y, z-1)$
Between X and Z	$P_{XZ} = xz \Big/ \binom{x+y+z}{2}$	$Q(x-1, y+1, z-1)$

Assume that there are $n_{1,0}$ and $n_{2,0}$ sequences at time T_0, and consider the genealogical relationship among these $n_{1,0}+n_{2,0}$ sequences in the ancestral population. As shown in Fig. B1*A*, when $n_{1,0}+n_{2,0}$ sequences coalesce into $n_{1,0}+n_{2,0}-1$ sequences, there are three possible cases: (1) the coalescence occurs within $n_{1,0}$ sequences from population I; (2) the coalescence occurs within $n_{2,0}$ sequences from population II; (3) the coalescence occurs between a sequence from population I and a sequence from population II. The probabilities of these three cases are

$$\binom{n_{1,0}}{2} \Big/ \binom{n_{1,0}+n_{2,0}}{2}, \quad \binom{n_{2,0}}{2} \Big/ \binom{n_{1,0}+n_{2,0}}{2} \text{ and } n_{1,0}n_{2,0} \Big/ \binom{n_{1,0}+n_{2,0}}{2},$$

respectively. Denote the state before the coalescence by $Q(n_{1,0}, 0, n_{2,0})$, and let the three states after the coalescence be $Q(n_{1,0}-1, 0, n_{2,0})$, $Q(n_{1,0}, 0, n_{2,0}-1)$ and $Q(n_{1,0}-1, 1, n_{2,0}-1)$, respectively. The first number in parentheses represents the number of sequences from population I, and we call these sequences class X sequences, which are represented by open circles in Fig. B1*A*. The third number in the parentheses is the number of sequences from population II, which we call class Z sequences. In Fig. B1*A*, the sequences in this class are represented by shaded circles. The second number in parentheses is the number of sequences which are the ancestors of the samples from both populations. These sequences are produced by coalescent events between classes X and Z. We call these sequences class Y sequences, and they are represented by filled circles in Fig. B1*A*.

This coalescent process becomes more complicated after the first coalescent event (Fig. B1*B*). Consider a state $Q(x,y,z)$, where x, y and z represent the number of sequences belonging to classes X, Y and Z, respectively. When a coalescent event occurs in this state $Q(x, y, z)$ there are six possible patterns, which are summarized in Table B1. Following this coalescent process, we consider the expectation and variance of $S_{1,0}+S_{2,0}$ in $Q(x, y, z)$. Let $A(x, y, z)$ and $V(x, y, z)$ be the expectation and variance of $S_{1,0}+S_{2,0}$ in $Q(x, y, z)$, respectively. The expectation is easily obtained as

$$A(x, y, z) = \sum_{k=1}^{x+y-1} \frac{1}{k}\theta + \sum_{k=1}^{y+z-1} \frac{1}{k}\theta \qquad \text{(B1)}$$

(Watterson, 1975). Note that $x+y$ indicates the number of ancestral sequences of population I and $y+z$ indicates the number of ancestral sequences of population II. We develop a recurrence formula for $V(x, y, z)$ by using the relationship among the six states $Q(x, y, z)$, $Q(x-1, y, z)$, $Q(x, y-1, z)Q(x, y-1, z)$, $Q(x, y, z-1)$ and $Q(x-1, y+1, z-1)$ as shown in Fig. B1*B*. First, we consider the case where $x+y \geqslant 2$ and $y+z \geqslant 2$. $V(x, y, z)$ is written as

$$\begin{aligned} V(x, y, z) = {} & (P_{XX}+P_{XY})\left[\, V(x-1, y, z)+(A(x-1, y, z)-\bar{A})^2 \,\right] \\ & +P_{YY}\left[\, V(x, y-1, z)+(A(x, y-1, z)-\bar{A})^2 \,\right] \\ & +(P_{YZ}+P_{ZZ})\left[\, V(x, y, z-1)+(A(x, y, z-1)-\bar{A})^2 \,\right] \\ & +P_{XZ}\left[\, V(x-1, y+1, z-1)+(A(x-1, y+1, z-1)-\bar{A})^2 \,\right] \\ & +(x+4y+z)V_{x+y+z}+\left[\, (x+2y+z)^2-(x+4y+z) \,\right]\mathrm{Cov}_{x+y+z}, \end{aligned} \qquad \text{(B2)}$$

where

$$\bar{A} = (P_{XX}+P_{XY})A(x-1, y, z)+P_{YY}A(x, y-1, z) \\ +(P_{YZ}+P_{ZZ})A(x, y, z-1)+P_{XZ}A(x-1, y+1, z-1), \tag{B3}$$

$$V_{x+y+z} = \left[\frac{\theta}{(x+y+z)(x+y+z-1)}+\frac{\theta^2}{(x+y+z)^2(x+y+z-1)^2}\right] \tag{B4}$$

and

$$\mathrm{Cov}_{x+y+z} = \frac{\theta^2}{(x+y+z)^2(x+y+z-1)^2}. \tag{B5}$$

See Table B1 for P_{XX}, P_{YY}, P_{ZZ}, P_{XY}, P_{YZ} and P_{XZ}. Note that V_{x+y+z} is the variance of the number of mutations on a branch of the genealogy when $x+y+z$ sequences coalesce into $x+y+z-1$ sequences and Cov_{x+y+z} is the covariance of the number of mutations between a pair of branches. For details, see Tajima (1983) and appendix B in Innan & Tajima (1997).

Next we consider the case where $x+y = 1$ and $y+z \geqslant 2$. In this case, the sequences from population I have already coalesced into one sequence so that $V(x, y, z)$ is given by

$$V(x, y, z) = \sum_{k=1}^{y+z-1}\frac{1}{k}\theta+\sum_{k=1}^{y+z-1}\frac{1}{k^2}\theta^2. \tag{B6}$$

In the same way, when the sequences from population II have already coalesced into one sequence ($x+y \geqslant 2$ and $y+z = 1$), $V(x, y, z)$ is given by

$$V(x, y, z) = \sum_{k=1}^{x+y-1}\frac{1}{k}\theta+\sum_{k=1}^{x+y-1}\frac{1}{k^2}\theta^2. \tag{B7}$$

Finally, the case of $x+y+z = 2$ is considered, where there are six possible states: $Q(2, 0, 0)$, $Q(0, 2, 0)$, $Q(0, 0, 2)$, $Q(1, 1, 0)$, $Q(0, 1, 1)$ and $Q(1, 0, 1)$. The variances for these states are given by

$$V(2, 0, 0) = V(0, 0, 2) = \theta+\theta^2, \tag{B8a}$$

$$V(0, 2, 0) = 4\theta+4\theta^2, \tag{B8b}$$

$$V(1, 1, 0) = V(0, 1, 1) = \theta+\theta^2 \tag{B8c}$$

and

$$V(1, 0, 1) = 0. \tag{B8d}$$

From (B2)–(B8), $\mathrm{Var}(S_{1,0}+S_{2,0}\,|\,n_{1,0}, n_{2,0})$ can be calculated.

The covariance between $S_{1,0}$ and $S_{2,0}$ given $n_{1,0}$ and $n_{2,0}$, $\mathrm{Cov}(S_{1,0}, S_{2,0}\,|\,n_{1,0}, n_{2,0})$, is obtained from $\mathrm{Var}(S_{1,0}+S_{2,0}\,|\,n_{1,0}, n_{2,0})$. Since the relationship between the variance and covariance is given by

$$\mathrm{Var}(S_{1,0}+S_{2,0}\,|\,n_{1,0}, n_{2,0}) = \mathrm{Var}(S_{1,0}\,|\,n_{1,0})+\mathrm{Var}(S_{2,0}\,|\,n_{2,0})+2\mathrm{Cov}(S_{1,0}, S_{2,0}\,|\,n_{1,0}, n_{2,0}), \tag{B9}$$

the covariance becomes

$$\mathrm{Cov}(S_{1,0}, S_{2,0}\,|\,n_{1,0}, n_{2,0}) = [\mathrm{Var}(S_{1,0}+S_{2,0}\,|\,n_{1,0}, n_{2,0})-\mathrm{Var}(S_{1,0}\,|\,n_{1,0})-\mathrm{Var}(S_{2,0}\,|\,n_{2,0})]/2, \tag{B10}$$

where

$$\mathrm{Var}(S_{1,0}\,|\,n_{1,0}) = \sum_{k=1}^{n_{1,0}-1}\frac{1}{k}\theta+\sum_{k=1}^{n_{1,0}-1}\frac{1}{k^2}\theta^2 \text{ and } \mathrm{Var}(S_{2,0}\,|\,n_{2,0}) = \sum_{k=1}^{n_{2,0}-1}\frac{1}{k}\theta+\sum_{k=1}^{n_{2,0}-1}\frac{1}{k^2}\theta^2. \tag{B11}$$

References

Edwards, S. V. & Beerli, P. (2000). Gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. *Evolution* **54**, 1839–1854.

Fu, Y.-X. (1996). New statistical tests of neutrality for DNA samples from a population. *Genetics* **143**, 557–570.

Fu, Y.-X. & Li, W.-H. (1993). Statistical tests of neutrality of mutations. *Genetics* **133**, 693–709.

Gillespie, J. H. (1991). *The Cause of Molecular Evolution.* New York: Oxford University Press.

Griffiths, R. C. (1979). Exact sampling distributions from the infinite neutral alleles model. *Advances in Applied Probability* **11**, 326–354.

Griffiths, R. C. (1980). Lines of descent in the diffusion approximation of neutral Wright–Fisher models. *Theoretical Population Biology* **17**, 37–50.

Hudson, R. R. (1983). Testing the coalescent-rate neutral allele model with protein sequence data. *Evolution* **37**, 203–217.

Hudson, R. R., Kreitman, M. & Aguadé, M. (1987). A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**, 153–159.

Innan, H. & Tajima, F. (1997). The amounts of nucleotide variation within and between allelic classes and the reconstruction of common ancestral sequence. *Genetics* **147**, 1431–1444.

Innan, H., Terauchi, R., Kahl, G. & Tajima, F. (1999). A method for estimating nucleotide diversity from AFLP data. *Genetics* **151**, 1157–1164.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature* **217**, 624–626.

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. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press.

Kingman, J. F. C. (1982). On the genealogy of large populations. *Journal of Applied Probability* **19A**, 27–43.

Liu, F., Charlesworth, D. & Kreitman, M. (1999). The effect of mating system differences on nucleotide diversity at the phosphoglucose isomerase locus in the plant genus *Leavenworthia*. *Genetics* **151**, 343–357.

Nei, M. (1987). *Molecular Evolutionary Genetics*. New York: Columbia University Press.

Nei, M. & Li, W.-H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the USA* **76**, 5269–5273.

Nielsen, R. (1998). Maximum likelihood estimation of population divergence times and population phylogenies under the infinite sites model. *Theoretical Population Biology* **53**, 143–151.

Nielsen, R., Mountain, J. L., Huelsenbeck, J. P. & Slatkin, M. (1998). Maximum-likelihood estimation of population divergence times and population phylogeny in models without mutation. *Evolution* **52**, 669–677.

Rosenberg, N. A. & Feldman, M. W. (2002). The relationship between coalescence times and population divergence times. In *Modern Developments in Theoretical Population Genetics* (ed. M. Slatkin & M. Veuille), pp. 130–164. Oxford: Oxford University Press.

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

Takahata, N. & Nei, M. (1985). Gene genealogy and variance of interpopulational nucleotide differences. *Genetics* **110**, 325–344.

Takahata, N., Satta, Y. & Klein, J. (1995). Divergence time and population-size in the lineage leading to modern humans. *Theoretical Population Biology* **48**, 198–221.

Tavaré, S. (1984). Line-of-descendant and genealogical processes and their applications in population genetic models. *Theoretical Population Biology* **26**, 119–164.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., Frijters, A., Pot, J., Peleman, J., Kuiper, M. & Zabeau, M. (1995). AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research* **23**, 4407–4414.

Wakeley, J. (1996). The variance of pairwise nucleotide differences in two populations with migration. *Theoretical Population Biology* **49**, 39–57.

Wakeley, J. & Hey, J. (1997). Estimating ancestral population parameters. *Genetics* **145**, 847–855.

Watterson, W. A. (1975). On the number of segregating sites in genetic models without recombination. *Theoretical Population Biology* **7**, 256–276.

Watterson, W. A. (1982). Mutant substitutions at linked nucleotide sites. *Advances in Applied Probability* **14**, 206–224.