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, $T_w$. Accordingly, the power (rejection probability) of the test depends on $T_w$. In this article, $T_w$ is treated as unknown because it is very difficult to estimate. The test is most conservative when $T_w = \infty$ is assumed because the covariance is zero. If $T_w = 0$ 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 $T_w$.

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-
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).

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, \( \theta (=4N\mu) \), 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 \( \theta \) in populations I and II are

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

where

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

(Watterson, 1975).
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:

$$G(T_0) = \frac{|\hat{\theta}_1 - \hat{\theta}_2|}{\sqrt{\text{Var}(\hat{\theta}_1 - \hat{\theta}_2)}}$$  \hspace{1cm} (3)

where $\text{Var}(\hat{\theta}_1 - \hat{\theta}_2)$ 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 $\text{Var}(\hat{\theta}_1 - \hat{\theta}_2)$ 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 $\theta$, $n_1$, $n_2$, 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, $n_{1t}$, at $T_0$ (Fig. 1B). In the same way, a random genealogy of $n_2$ sequences in population II is constructed, and the number of sequences, $n_{2t}$, at $T_0$ is determined. Next a random genealogy of $n_{1t} + n_{2t}$ sequences in the ancestral population is constructed (Fig. 1A). 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 $\theta(T_0)$.

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. 3A), 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. 3B), 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 $\theta$ on the distribution of $G(T_0)$ is not large unless $\theta$ 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 $T_0 = 0, 0.05, 0.1, 0.15, \ldots, 1$. $G(T_0)$ 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 $(S_1, S_2) = (130, 70), (135, 75), \ldots, (165, 35), (170, 30)$ 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>
<thead>
<tr>
<th>$n_1$</th>
<th>$n_2$</th>
<th>$T_0 = \infty$</th>
<th>$T_0 = 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>5%</td>
<td>2.5%</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>1.67</td>
<td>1.82</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>1.69</td>
<td>1.85</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>1.70</td>
<td>1.88</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>1.70</td>
<td>1.87</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>1.72</td>
<td>1.90</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>1.71</td>
<td>1.89</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>1.74</td>
<td>1.92</td>
</tr>
<tr>
<td>5</td>
<td>14</td>
<td>1.74</td>
<td>1.94</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>1.71</td>
<td>1.90</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>1.71</td>
<td>1.89</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
<td>1.71</td>
<td>1.90</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
<td>1.73</td>
<td>1.93</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
<td>1.74</td>
<td>1.92</td>
</tr>
<tr>
<td>6</td>
<td>14</td>
<td>1.75</td>
<td>1.93</td>
</tr>
<tr>
<td>6</td>
<td>16</td>
<td>1.77</td>
<td>1.97</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>1.70</td>
<td>1.90</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>1.72</td>
<td>1.91</td>
</tr>
<tr>
<td>7</td>
<td>9</td>
<td>1.73</td>
<td>1.90</td>
</tr>
<tr>
<td>7</td>
<td>10</td>
<td>1.74</td>
<td>1.92</td>
</tr>
<tr>
<td>7</td>
<td>12</td>
<td>1.76</td>
<td>1.94</td>
</tr>
<tr>
<td>7</td>
<td>14</td>
<td>1.75</td>
<td>1.95</td>
</tr>
<tr>
<td>7</td>
<td>16</td>
<td>1.78</td>
<td>1.98</td>
</tr>
<tr>
<td>7</td>
<td>18</td>
<td>1.76</td>
<td>1.96</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>1.73</td>
<td>1.89</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>1.75</td>
<td>1.94</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td>1.73</td>
<td>1.93</td>
</tr>
<tr>
<td>8</td>
<td>12</td>
<td>1.75</td>
<td>1.95</td>
</tr>
<tr>
<td>8</td>
<td>14</td>
<td>1.74</td>
<td>1.94</td>
</tr>
<tr>
<td>8</td>
<td>16</td>
<td>1.75</td>
<td>1.95</td>
</tr>
<tr>
<td>8</td>
<td>18</td>
<td>1.75</td>
<td>1.97</td>
</tr>
<tr>
<td>8</td>
<td>20</td>
<td>1.75</td>
<td>1.96</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>1.74</td>
<td>1.91</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>1.74</td>
<td>1.94</td>
</tr>
<tr>
<td>9</td>
<td>12</td>
<td>1.75</td>
<td>1.95</td>
</tr>
<tr>
<td>9</td>
<td>14</td>
<td>1.77</td>
<td>1.97</td>
</tr>
<tr>
<td>9</td>
<td>16</td>
<td>1.77</td>
<td>1.97</td>
</tr>
<tr>
<td>9</td>
<td>18</td>
<td>1.77</td>
<td>1.99</td>
</tr>
<tr>
<td>9</td>
<td>20</td>
<td>1.77</td>
<td>1.99</td>
</tr>
<tr>
<td>9</td>
<td>25</td>
<td>1.79</td>
<td>2.00</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>1.74</td>
<td>1.95</td>
</tr>
<tr>
<td>10</td>
<td>12</td>
<td>1.75</td>
<td>1.94</td>
</tr>
<tr>
<td>10</td>
<td>14</td>
<td>1.75</td>
<td>1.96</td>
</tr>
<tr>
<td>10</td>
<td>16</td>
<td>1.75</td>
<td>1.96</td>
</tr>
<tr>
<td>10</td>
<td>18</td>
<td>1.77</td>
<td>1.97</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>1.76</td>
<td>1.98</td>
</tr>
<tr>
<td>10</td>
<td>25</td>
<td>1.79</td>
<td>2.02</td>
</tr>
<tr>
<td>10</td>
<td>30</td>
<td>1.79</td>
<td>2.02</td>
</tr>
<tr>
<td>12</td>
<td>12</td>
<td>1.76</td>
<td>1.96</td>
</tr>
<tr>
<td>12</td>
<td>14</td>
<td>1.76</td>
<td>1.99</td>
</tr>
<tr>
<td>12</td>
<td>16</td>
<td>1.76</td>
<td>1.98</td>
</tr>
<tr>
<td>12</td>
<td>18</td>
<td>1.77</td>
<td>1.99</td>
</tr>
<tr>
<td>12</td>
<td>20</td>
<td>1.75</td>
<td>1.99</td>
</tr>
<tr>
<td>12</td>
<td>25</td>
<td>1.79</td>
<td>2.01</td>
</tr>
<tr>
<td>14</td>
<td>14</td>
<td>1.76</td>
<td>1.96</td>
</tr>
<tr>
<td>14</td>
<td>16</td>
<td>1.76</td>
<td>1.98</td>
</tr>
<tr>
<td>14</td>
<td>18</td>
<td>1.77</td>
<td>1.99</td>
</tr>
<tr>
<td>14</td>
<td>20</td>
<td>1.77</td>
<td>2.00</td>
</tr>
</tbody>
</table>
Test for the difference in $G$

Table 1. (cont.)

<table>
<thead>
<tr>
<th>$T_0 = \infty$</th>
<th>$T_0 = 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_1$</td>
<td>$n_2$</td>
</tr>
<tr>
<td>14</td>
<td>25</td>
</tr>
<tr>
<td>14</td>
<td>30</td>
</tr>
<tr>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>16</td>
<td>25</td>
</tr>
<tr>
<td>16</td>
<td>30</td>
</tr>
<tr>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>18</td>
<td>20</td>
</tr>
<tr>
<td>18</td>
<td>25</td>
</tr>
<tr>
<td>18</td>
<td>30</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>20</td>
<td>30</td>
</tr>
<tr>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>25</td>
<td>30</td>
</tr>
<tr>
<td>30</td>
<td>30</td>
</tr>
</tbody>
</table>

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 $T_0 \leq 0.35,$ while $P$ is larger than 0.05 when $T_0 \geq 0.4.$

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 $\theta$ 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 $\theta$ from the data may be more powerful, especially when $\theta$ 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. 4B).

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. 4B), 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 $G(\infty) = 1.15$ and $G(0) = 3.08.$ Computer simulations show that $P_{\max}$ and $P_{\min}$ are 0.215 and 0.030, respectively, indicating $0.03 < P < 0.215.$ From additional simulations with $T_0 = 0.01, 0.02, 0.03, \ldots,$ we obtain $P(0.03) = 0.049$ and $P(0.04) = 0.055,$ indicating that $T_0$ that gives $P(T_0) = 0.05$ is about 0.03. We refer to such $T_0$ as $T_{0.05}.$ It is indicated that the difference is significantly large if the real divergence time is smaller than $T_{0.05} \approx 0.03.$ Although the real $T_0$ is not easy to know, $T_{0.05}$ 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
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 \( \theta \) are \( \hat{\theta}_1 = 8.65 \) and \( \hat{\theta}_2 = 1.14 \). The absolute value of the difference in \( \theta \) 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 \( T_o = \infty \), indicating the difference is not significant. Because of a lack of knowledge of \( T_o \), the test is not conducted for another smaller value of \( T_o \).

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

<table>
<thead>
<tr>
<th>Species I</th>
<th>Species II</th>
<th>( n_1 )</th>
<th>( n_2 )</th>
<th>( S_1 )</th>
<th>( S_2 )</th>
<th>( \theta_1 )</th>
<th>( \theta_2 )</th>
<th>( G(\infty) )</th>
<th>( P(\infty) )</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>L. stylosa</em></td>
<td><em>L. crassa</em></td>
<td>26</td>
<td>45</td>
<td>33</td>
<td>5</td>
<td>8.65</td>
<td>1.14</td>
<td>2.82</td>
<td>0.0009</td>
</tr>
<tr>
<td><em>L. stylosa</em></td>
<td><em>L. uniflora</em></td>
<td>26</td>
<td>11</td>
<td>29</td>
<td>0</td>
<td>7.60</td>
<td>0.00</td>
<td>2.98</td>
<td>0.0001</td>
</tr>
<tr>
<td><em>L. crassa</em></td>
<td><em>L. uniflora</em></td>
<td>45</td>
<td>11</td>
<td>5</td>
<td>0</td>
<td>1.14</td>
<td>0.00</td>
<td>1.78</td>
<td>0.1285*</td>
</tr>
</tbody>
</table>

\( * \) The difference in \( S_1 \) of *L. stylosa* is due to the difference in the regions used in the analysis.

\( * \) The simulation method of this comparison is different from the other two because \( S_1 + S_2 \) is small. See text.

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 \( T_o \) because an estimate of \( T_o \) usually has a huge variance. The value of the test statistic depends on the divergence time of the two populations, \( T_o \), 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 \( T_o \). 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 \( T_o \) 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 depend on \( T_o \). In this case, \( T_{o, 0.5\%} \) is a useful summary for evaluating the difference. \( T_{o, 0.5\%} \) is defined such that \( P(T_{0, 0.5\%}) = 0.05 \), indicating that the null hypothesis is rejected when the real \( T_o \) is smaller than \( T_{o, 0.5\%} \).

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 \( T_o = \infty \) is carried out.

We tried to develop another test that is independent of \( T_o \). Using an estimate of \( T_o \) 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 \( (n_1, n_2, T_o, \theta) \) demonstrated that it does not work well as long as the estimate of \( T_o \) has variance.

### 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.

(ii) **On the estimation of \( T_o \)**

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,
Appendix A

The variance of $\hat{\theta}_1 - \hat{\theta}_2$ is written as
\[
\text{Var}(\hat{\theta}_1 - \hat{\theta}_2) = \text{Var}(\hat{\theta}_1) + \text{Var}(\hat{\theta}_2) - 2\text{Cov}(\hat{\theta}_1, \hat{\theta}_2)
\]
\[
= \frac{\text{Var}(S_1)}{n_1} + \frac{\text{Var}(S_2)}{n_2} - \frac{2\text{Cov}(S_1, S_2)}{n_1n_2},
\]
where $a_1$ and $a_2$ are given by (2). The variances of $S_1$ and $S_2$ are given by
\[
\text{Var}(S_1) = a_1\theta + b_1\theta^2
\]
\[
\text{Var}(S_2) = a_2\theta + b_2\theta^2,
\]
where
\[
b_1 = \sum_{k=1}^{n_1} \frac{1}{k} \quad \text{and} \quad b_2 = \sum_{k=1}^{n_2} \frac{1}{k^2}.
\]
(Watterson, 1975).

Here, we consider the covariance between $S_1$ and $S_2$. $S_1$ is written as
\[
S_1 = S_{1,0} + S_{1,1},
\]
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 1 after divergence. In the same way, $S_2$ is given by
\[
S_2 = S_{2,0} + S_{2,2},
\]
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
\[
\text{Cov}(S_1, S_2) = \text{Cov}(S_{1,0} + S_{1,1}, S_{2,0} + S_{2,2})
\]
\[
= \text{Cov}(S_{1,0}, S_{2,0}) + \text{Cov}(S_{1,0}, S_{2,2}) + \text{Cov}(S_{1,1}, S_{2,0}) + \text{Cov}(S_{1,1}, S_{2,2})
\]
\[
= \text{Cov}(S_{1,0}, S_{2,0}),
\]
because $\text{Cov}(S_{1,0}, S_{2,2})$, $\text{Cov}(S_{1,1}, S_{2,0})$ and $\text{Cov}(S_{1,1}, S_{2,2})$ are 0.

The covariance of $S_{1,0}$ between $S_{2,0}$ is given by
\[
\text{Cov}(S_{1,0}, S_{2,0}) = E(S_{1,0}S_{2,0}) - E(S_{1,0})E(S_{2,0}),
\]
and the expectation of $S_{1,0}S_{2,0}$ is written as
\[
E(S_{1,0}S_{2,0}) = \sum_{n_{1,0}=0}^{n_1} \sum_{n_{2,0}=0}^{n_2} \frac{n_{1,0}(n_{1,0} - 1)!}{n_{1,0}!} \frac{n_{2,0}(n_{2,0} - 1)!}{n_{2,0}!} E(S_{1,0}S_{2,0} | n_{1,0}, n_{2,0}).
\]
\[
P_{n_{1,0}n_{2,0}} (j = 1, 2) \text{ is given by}
\]
\[
P_{n_{1,0}n_{2,0}} = \sum_{k=n_{1,0}}^{n_j} \exp\left(-k(k-1)T_0\right) \left(\frac{(2k-1)(-1)^{k-n_{1,0}}n_{1,0}!n_{2,0}!}{n_{1,0}!(k-n_{1,0})!n_{2,0}!(k-1)!}\right),
\]
where $n_{(k)} = n(n+1) \ldots (n+k-1)$ and $n_{(0)} = n(n-1) \ldots (n-k+1)$. (A9) is from equation (6.1) in Tavare (1984) (also see Griffiths, 1979; Watterson, 1982; Takahata & Nei, 1985; Wakeley & Hey, 1997).
Since
\[ \text{Cov}(S_{1,0}, S_{2,0} | n_{1,0}, n_{2,0}) = E(S_{1,0}S_{2,0} | n_{1,0}, n_{2,0}) - E(S_{1,0} | n_{1,0})E(S_{2,0} | n_{2,0}), \]
(A10)

the expectation of \( S_{1,0}S_{2,0} \) given \( n_{1,0} \) and \( n_{2,0} \) becomes
\[ E(S_{1,0}S_{2,0} | n_{1,0}, n_{2,0}) = \text{Cov}(S_{1,0}, S_{2,0} | n_{1,0}, n_{2,0}) - E(S_{1,0} | n_{1,0})E(S_{2,0} | n_{2,0}). \]
(A11)

Then, from (A7), (A8) and (A11), the covariance is given by
\[ \text{Cov}(S_{1,0}, S_{2,0}) = \sum_{n_{1,0}+n_{2,0}=2} P_{n_{1,0}, n_{2,0}} \text{Cov}(S_{1,0}, S_{2,0} | n_{1,0}, n_{2,0}), \]
(A12)

because
\[ \sum_{n_{1,0}+n_{2,0}=2} P_{n_{1,0}, n_{2,0}} E(S_{1,0} | n_{1,0})E(S_{2,0} | n_{2,0}) = E(S_{1,0})E(S_{2,0}), \]
(A13)

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

**Appendix B**

To obtain \( \text{Cov}(S_{1,0}, S_{2,0} | n_{1,0}, n_{2,0}) \), 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).

---

Fig. B1. Coalescent scheme used to calculate the variance of \( S_{1}+S_{z} \).
### Table B1. Six patterns of coalescent events in \( Q(x, y, z) \)

<table>
<thead>
<tr>
<th>Coalescent event</th>
<th>Probability</th>
<th>Next state</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within X</td>
<td>( P_{xx} = \frac{x}{2}, \frac{x+y+z}{2} )</td>
<td>( Q(x-1, y, z) )</td>
</tr>
<tr>
<td>Between X and Y</td>
<td>( P_{xy} = \frac{y}{2}, \frac{x+y+z}{2} )</td>
<td>( Q(x-1, y, z) )</td>
</tr>
<tr>
<td>Within Z</td>
<td>( P_{zz} = \frac{z}{2}, \frac{x+y+z}{2} )</td>
<td>( Q(x, y, z-1) )</td>
</tr>
<tr>
<td>Between X and Z</td>
<td>( P_{yz} = \frac{y}{2}, \frac{x+y+z}{2} )</td>
<td>( Q(x, y, z-1) )</td>
</tr>
<tr>
<td>Between Y and Z</td>
<td>( P_{yy} = \frac{x}{2}, \frac{x+y+z}{2} )</td>
<td>( Q(x-1, y+1, z-1) )</td>
</tr>
</tbody>
</table>

Assume that there are \( n_{1,o} \) and \( n_{2,o} \) sequences at time \( T_o \), and consider the genealogical relationship among these \( n_{1,o} + n_{2,o} \) sequences in the ancestral population. As shown in Fig. B1A, when \( n_{1,o} + n_{2,o} \) sequences coalesce into \( n_{1,o} + n_{2,o} - 1 \) sequences, there are three possible cases: (1) the coalescence occurs within \( n_{1,o} \) sequences from population I; (2) the coalescence occurs within \( n_{2,o} \) 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

\[
\begin{align*}
\frac{n_{1,o}}{2} & \bigg/ \left\{ \frac{n_{1,o}+n_{2,o}}{2} \right\}, \\
\frac{n_{2,o}}{2} & \bigg/ \left\{ \frac{n_{1,o}+n_{2,o}}{2} \right\} \text{ and } \frac{n_{1,o}n_{2,o}}{2} \bigg/ \left\{ \frac{n_{1,o}+n_{2,o}}{2} \right\},
\end{align*}
\]

respectively. Denote the state before the coalescence by \( Q(n_{1,o}, 0, n_{2,o}) \), and let the three states after the coalescence be \( Q(n_{1,o}-1, 0, n_{2,o}), Q(n_{1,o}, 0, n_{2,o}-1) \) and \( Q(n_{1,o}-1, 1, n_{2,o}-1) \), 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. B1A. The third number in the parentheses is the number of sequences from population II, which we call class Z sequences. In Fig. B1A, 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. B1A.

This coalescent process becomes more complicated after the first coalescent event (Fig. B1B). 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 \( S_{1,o} + S_{2,o} \) in \( Q(x, y, z) \). Let \( A(x, y, z) \) and \( V(x, y, z) \) be the expectation and variance of \( S_{1,o} + S_{2,o} \) in \( Q(x, y, z) \), respectively. The expectation is easily obtained as

\[
A(x, y, z) = \sum_{k=1}^{x+y} \frac{1}{k} \theta + \sum_{k=1}^{y+z} \frac{1}{k} \theta
\]

(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 \( Q(x, y, z), Q(x-1, y, z), Q(x, y-1, z), Q(x, y+1, z), Q(x-1, y+1, z-1) \) and \( Q(x-1, y+1, z-1) \) as shown in Fig. B1B. First, we consider the case where \( x+y \geq 2 \) and \( y+z \geq 2 \). \( V(x, y, z) \) is written as

\[
V(x, y, z) = (P_{xx} + P_{xy}) \left[ V(x-1, y, z) + (A(x-1, y, z) - \bar{A})^2 \right] \\
+ P_{yy} \left[ V(x, y-1, z) + (A(x, y-1, z) - \bar{A})^2 \right] \\
+ (P_{yz} + P_{zz}) \left[ V(x, y, z-1) + (A(x, y, z-1) - \bar{A})^2 \right] \\
+ P_{xy} \left[ V(x-1, y+1, z-1) + (A(x-1, y+1, z-1) - \bar{A})^2 \right] \\
+ (x+y+z)V_{x+y+z} + \left[ (x+2y+z)^2 - (x+4y+z) \right] \text{Cov}_{x+y+z},
\]
where

\[
A = (P_{XX} + P_{XY})A(x-1, y, z) + P_{YY}A(x, y-1, z) \\
+ (P_{XZ} + P_{ZZ})A(x, y, z-1) + P_{XY}A(x-1, y+1, z-1),
\]

(B3)

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

(B4)

and

\[
\text{Cov}_{x+y+z} = \frac{\theta^2}{(x+y+z)^2(x+y+z-1)^2}
\]

(B5)

See Table B1 for \(P_{XX}, P_{YY}, P_{ZZ}, P_{XY}, P_{XZ}\) and \(P_{XY}\). 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 \(\text{Cov}_{x+y+z}\) 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 \(y+z \geq 2\). In this case, the sequences from population I have already coalesced into one sequence so that \(V(x, y, z)\) is given by

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

(B6)

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

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

(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

\[
V(2, 0, 0) = V(0, 0, 2) = \theta + \theta^2,
\]

(B8a)

\[
V(0, 2, 0) = 4\theta + 4\theta^2,
\]

(B8b)

\[
V(1, 1, 0) = V(0, 1, 1) = \theta + \theta^2
\]

(B8c)

and

\[
V(1, 0, 1) = 0.
\]

(B8d)

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

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

\[
\text{Var}(S_{1,0} + S_{2,0} | n_{1,0}, n_{2,0}) = \text{Var}(S_{1,0} | n_{1,0}) + \text{Var}(S_{2,0} | n_{2,0}) + 2\text{Cov}(S_{1,0}, S_{2,0} | n_{1,0}, n_{2,0}),
\]

(B9)

the covariance becomes

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

(B10)

where

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

(B11)

References


