A Population Genetic Study of the Evolution of SINEs. I. Polymorphism With Regard to the Presence or Absence of an Element

Hidenori Tachida*1 and Masaru Iizuka†

*National Institute of Genetics, Mishima, Shizuoka-ken 411, Japan, and †General Education Course, Chikushi Jogakuen Junior College, Iizhaka 2-12-1, Dazaifu-shi, Fukuoka-ken 818-01, Japan

ABSTRACT

SINEs are short interspersed repeated DNA elements which are considered to spread throughout genomes via RNA intermediates. Polymorphisms with regard to the presence or absence of SINE are occasionally observed in a specific location of a genome. We modeled the evolution of SINEs with regard to this type of polymorphism. Because SINEs are rarely deleted, multiplication of elements is confined to a certain period, and a few master copies are considered to be responsible for their multiplication, the usual population genetic models of transposable elements assuming the equilibrium state are not applicable to describe the evolution of SINEs. Taking into account these properties and assuming selective neutrality, we computed conditional probabilities of finding a SINE at a specific site given that this site is first found because it is occupied by a SINE in an original sample. Using these probabilities, we investigated ways to estimate the multiplication period and infer relationships among populations. The latter inference procedures are shown to be strongly dependent on the multiplication period.

SHORT interspersed repetitive DNA elements (SINEs) are found in various eukaryotes and sometimes constitute a significant proportion of the total genome [for reviews, see Weiner, Deininger and Efstratiadis (1986), Deininger (1989) and Okada (1991)]. They are considered to transpose to other parts of the genome via RNA intermediates (Weiner, Deininger and Efstratiadis 1986). We still do not know the biological significance of these elements and how these elements evolved to the present status. To investigate these problems, there are at least two approaches. One is to examine the nature of these elements by observing transcription and transposition of them in vitro or in vivo (Deininger 1989; Wallace et al. 1991). Another approach is to survey the present pattern of variation in these elements and make inferences on what happened in the past (Jurka and Milosavljevic 1991; Shen, Batzer and Deininger 1991; Batzer et al. 1991). We take the second approach from the standpoint of population genetics and investigate the evolution of SINEs by first building and analyzing a model and then applying it to the data.

One type of datum available is on the presence or absence of a SINE at a specific site in the genome. In human populations, there are polymorphisms with regard to the presence or absence of an Alu element at specific sites (Matera et al. 1990; Batzer et al. 1991). Similar polymorphisms are found in salmon (N. Okada, personal communication) and rice (H. Hirano, personal communication). This kind of data is obtained by using neighboring short fragments of the SINE insertion sites as primers and amplifying DNA by polymerase chain reaction (PCR). After the amplification, presence of the element is detected by southern hybridization. Thus, this type of polymorphism is easier to be detected compared to DNA sequence polymorphisms. We are interested in what can be inferred from this kind of data.

Ohta (1986) and Kaplan and Hudson (1989) built population genetic models of SINEs and computed the probability of presence or absence of an element at a specific site. However, Ohta (1986) assumed that duplicative transpositions are possible for all elements like transposons in her model and Kaplan and Hudson (1989) analyzed the equilibrium state maintained by insertion and deletion. At present, SINEs are considered to expand by multiplication of one or a few master copies (Deininger and Slagel 1988). Also deletions of SINEs are not observed in a comparison of distantly related species (Koop et al. 1986) and we cannot consider the present status as the equilibrium state. Here, we build a population genetic model of SINE insertion into a site incorporating these features of SINEs and investigate the frequency pattern when samples are made from multiple populations separated sometime ago. Note that the site is found since a sampled genome has a SINE there. Thus, we investigate the pattern conditioned on the event that the first sample has a SINE at the site. The pattern de-

1 Present address: Department of Biology, Faculty of Science, Kyushu University 33, Fukuoka, 812 Japan

Accepted for publication December 4, 1992

Genetics 133: 1023–1030 (April, 1993)
depends on the population structure in the past and how the expansion of the elements occurred. Conversely, we can obtain information on the history by observing the present pattern.

**MODEL**

Consider three populations, A, B and C, whose population sizes are all \( N \). We made all population sizes to be the same for simplicity. When population sizes differ, the calculation goes similarly though the resulting expressions are complex. Assume that A and B separated \( t_1 \) generations ago and B and C separated \( t_2 \) generations ago (see Figure 1). We take a sample (designated as \( A_0 \)) from population A and assume that the sample has a SINE at a specific site. We call this site as a locus hereafter. For the SINE expansion, we assume the master copy model. In this model, we assume that all genomes in the population have one or a few master copies. Duplicates of the master copy are incorporated into the genome and these duplicates do not have a capability of further multiplication. Then, the rate of multiplication will not depend on the (mostly duplicates) copy number in the genome and we write the rate as a function, \( F(t) \) of time \( t \) measuring time backward in units of \( 2N \) generations.

The density, \( f(t) \), of the time when the SINE is incorporated into this locus is written as

\[
f(t) = \frac{F(t)}{\int_0^\infty F(s) \, ds}.
\]

In the following calculation, we use,

\[
f(t) = \frac{1}{t_2 - t_1} I(t, s_0)(t)
\]

for simplicity where \( I(t, s_0)(t) \) takes 1 in the interval \([t_2, t_1]\) and 0 otherwise. The density becomes this function if the expansion of the element occurred between \( t_2 \) and \( t_1 \) with a constant rate. This function is chosen to model one wave of SINE expansion. There seem to be several waves of SINE expansions in the lineage of rodents and primates (Quentin 1988, 1989). We assume that the insertion of the SINE into this locus does not affect fitness [the neutrality assumption (Kimura 1968)]. Furthermore, we assume that an insertion into a specific site is an unique event (no multiple insertion into a specific site) and that complete deletions of the element do not occur.

We represent the \( i \)th sample from population A, B or C by \( A, B, \) or \( C \), respectively. The indicator variable, \( A, B, \) or \( C \), takes a value of one or zero depending on whether the sample has the SINE at the locus or not, respectively (Cockerham 1969). We used the same symbols for samples and their indicator variables for an ease of representation. By definition \( A_0 = 1 \). We are interested in the probabilities that these samples have the SINE in the locus.

We first compute the probabilities when one gene is sampled. Define the following two conditional probabilities:

\[
Prob[A_1 = 1 | A_0 = 1] = \frac{\int_{t_2}^\infty F(s) \, ds}{\int_0^\infty F(s) \, ds}. (2)
\]

\[
Prob[B_1 = 1 | A_0 = 1]. (3)
\]

\( p_A(p_B) \) is the probability that another sample from population A (B) has the SINE at the same locus. Note that these probabilities are conditional on the event that the original sample \( A_0 \) has the SINE at the locus.

Let \( x, y \) and \( z \) denote the frequencies of genomes which have a SINE at this locus in population A, B and C, respectively. Note that \( x, y \) and \( z \) are random variables because of random genetic drift and a random insertion time of the element. Let \( l, m \) and \( n \) be numbers of gametes sampled from population A, B and C, respectively, in addition to the original sample \( A_0 \). Define \( S_A, S_B \) and \( S_C \) to be numbers of gametes which have a SINE at this locus sampled from population A, B and C, respectively, without counting the original sample \( A_0 \). Using the frequencies, \( x, y \) and \( z \), a general expression for these conditional probabilities is

\[
Prob[S_A = l, S_B = m1, S_C = n1 | A_0 = 1] = \binom{m}{l} \binom{n}{n1} \binom{N}{l1} \frac{E(x^{l+1}(1-x)^{l-1}y^{m1}(1-y)^{m-m1}z^{n1}(1-z)^{n-n1})}{E[x]}. (4)
\]

if \( l, m, n \ll N \) where \( E \) denotes an expectation. Since a SINE insertion into a specific locus occurs only in one genome in the population at first and because of the neutrality assumption and no deletion,

\[
E[x] = 1/2N. (5)
\]

To compute \( p_B \), let \( x_{i1} \) be the frequency of genomes which have a SINE at this locus in population A \( t_1 \) generations ago when the populations A and B are split. Then the numerator of (4) for this probability is

\[
E[x_{i1}^2] = E[x_{i1}^2]. (6)
\]

![Figure 1](https://academic.oup.com/genetics/article/133/4/1023/6009231) Description of the model. Samples are taken from three populations, A, B and C. The case where population A diverged first from the ancestral population of population B and C \( (t_1 \gg t_2) \) is shown. The genealogical relationships of samples are also shown.
because gene frequencies evolve independently after the separation of the two populations. Let \( t \) be the time when the element was inserted into this locus. Then, because of the neutrality assumption and since the initial frequency is \( 1/2N \),

\[
E[x^2] = \begin{cases} 
\frac{1}{2N} \left(1 - \exp[-(t - t_1)]\right) & \text{if } t \geq t_1 \\
0 & \text{otherwise}
\end{cases}
\]  

(7)

using Equation 7.4.25 of Crow and Kimura (1970) neglecting higher order terms. We can compute \( p_B \) by integrating the function with regard to \( t \) using the density \( f(t) \) of the insertion time,

\[
p_B = \begin{cases} 
\frac{\left((t_0 - t) - e^{-\theta}(e^{\theta t_0} - e^{\theta t})\right)}{(t_0 - t_0)} & \text{if } t_1 \leq t_0 \\
\frac{\left((t_0 - t_1 - 1) + e^{-\theta}(e^{\theta t_0} - e^{\theta t_1})\right)}{(t_0 - t_1)} & \text{if } t_1 \leq t_0 \\
0 & \text{if } t_1 \leq t_0.
\end{cases}
\]  

(8)

\( p_A \) is obtained by setting \( t_0 = 0 \) in this equation,

\[
p_A = \frac{\left((t_0 - t) - (e^{-\theta t_0} - e^{-\theta t})\right)}{(t_0 - t_0)}.
\]  

(9)

Numerical values of \( p_A \) and \( p_B \) as functions of \( t \) are shown for various value of \( t_0 \) in Figure 2. We set \( t_0 = 5.0 \) in the example. These probabilities are monotone increasing function of \( t \). The change of \( p_B \) depends much on the value of \( t_1 \). Large changes occur when \( t_1 \) is close to \( t_0 \). Note the following relationship when \( t_1 \leq t_0 \):

\[
1 - p_B = e^{\theta t_0}(1 - p_A).
\]  

(10)

This relationship will be used to estimate \( t_1 \) in the next section.

Next we compute the probabilities when two genes are sampled. There are four possible ways of sampling two genes if we pay attention only to the relationship of them to the original sample \( A_0 \). Define the following four probabilities:

\[
p_{AA} = \text{Prob}[A_1 = 1, A_2 = 1 | A_0 = 1]
\]  

(11)

\[
p_{AB} = \text{Prob}[A_1 = 1, B_1 = 1 | A_0 = 1]
\]  

(12)

\[
p_{BB} = \text{Prob}[B_1 = 1, B_2 = 1 | A_0 = 1]
\]  

(13)

\[
p_{BC} = \text{Prob}[B_1 = 1, C_1 = 1 | A_0 = 1].
\]  

(14)

The other probabilities, \( p_{AC} \) and \( p_{CC} \), can be computed by interchanging \( t_1 \) with \( t_2 \) in the expressions for \( p_{AB} \) and \( p_{BB} \), respectively. The first three probabilities are computed as special cases of \( p_{BC} \). We first consider the case of \( t_1 \geq t_2 \) [case (b) of Figure 3]. Note that

\[
p_{BC} = \frac{E[x_{xy}]}{E[x]}.
\]  

(15)

Using (7.4.25) and (7.4.27) of Crow and Kimura (1970), we can compute the numerator as

\[
E[x_{xy}] = E[x y_{rb}^2]
\]  

(16)

\[
= E[x_1x_2(1 - (1 - x_1)e^{-t_1-rb})]
\]  

\[
= \frac{1}{2N} E[1 - \frac{1}{2}(2e^{t_1} + e^{t_2})e^{-t} + \frac{1}{2}e^{2(t_1 + t_2)}e^{-2t}]
\]

where \( t \) is the time of insertion of the element into the locus as before and \( x_1 \) and \( y_2 \) are frequencies of genomes which have the element in population A at \( t_1 \) and population B at \( t_2 \), respectively. Again we neglected higher order terms in this computation. Combining (15) and (16), we obtain \( p_{BC} \).

\[
p_{BC} = \begin{cases} 
\frac{\left((t_0 - t) - \frac{1}{2}(2e^{t_1} + e^{t_2})(e^{-\theta t_0} - e^{-\theta t}) + \frac{1}{2}e^{2(t_1 + t_2)}e^{-3t_0}\right)}{(t_0 - t_0)} & \text{if } t_1 \leq t_0 \\
\frac{\left((t_0 - t_1 - 1) + \frac{1}{2}(2e^{t_1} + e^{t_2})e^{-\theta t} - \frac{1}{2}e^{2(t_1 + t_2)}e^{-3t_0}\right)}{(t_0 - t_1)} & \text{if } t_1 \leq t_0 \\
0 & \text{if } t_1 \leq t_0.
\end{cases}
\]  

(17)

The expression for the case where population C first splits from the ancestral population of A and B and then A and B split \( t_1 < t_2 \), case (a) of Figure 3] can be obtained by exchanging \( t_1 \) and \( t_2 \) in the above equation.

By setting \( t_2 = 0 \), we obtain \( p_{BB} \),

\[
p_{BB} = \begin{cases} 
\frac{\left((t_0 - t) - \frac{1}{2}(2e^{t_1} + 1)(e^{-\theta t_0} - e^{-\theta t}) + \frac{1}{2}e^{2(t_1 + 1)}e^{-3t_0}\right)}{(t_0 - t_0)} & \text{if } t_1 \leq t_0 \\
\frac{\left((t_0 - t_1 - 1) - \frac{1}{2}e^{-t_1} + \frac{1}{2}e^{2(t_1 + 1)}e^{-t} - \frac{1}{2}e^{2t_1}e^{-3t_0}\right)}{(t_0 - t_1)} & \text{if } t_1 \leq t_0 \\
0 & \text{if } t_1 \leq t_0.
\end{cases}
\]  

(18)
Since \( p_{AB} = \frac{E[x^2]}{E[x]} = \frac{E[x y^2]}{E[x]}, \) \( p_{AB} = p_{BB}. \) \( p_{AA} \) can be obtained by setting \( t_1 = 0 \) in (18),

\[
p_{AA} = \frac{\left[(t_b - t_a) - \frac{1}{2}(e^{-t_a} - e^{-t_b}) + \frac{1}{2}(e^{-3t_a} - e^{-3t_b})\right]}{t_b - t_a}
\]

(19)

APPLICATIONS

Here, we explore some applications of the probabilities computed in the previous section for inference.

**Estimation of \( p_A \) and \( p_B \):** We can estimate \( p_A \) and \( p_B \) if we sample \( l \) and \( m \) genomes from population A and B, respectively.

\[
\hat{p}_A = \frac{\sum_{i=1}^{l} A_i}{l} \quad \text{(20)}
\]

\[
\hat{p}_B = \frac{\sum_{i=1}^{m} B_i}{m}. \quad \text{(21)}
\]

We indicate an estimator by attaching a hat. These are observed gene frequencies of SINE at this locus in the two populations.

From (21), the variance of \( \hat{p}_B \) is

\[
\text{Var}(\hat{p}_B) = \frac{E[B_i^2 | A_0 = 1]}{m} + \frac{(m - 1)E[B_i B_j | A_0 = 1]}{m} - \hat{p}_B^2
\]

\[
= \frac{p_B + (m - 1)p_{BB} - \hat{p}_B^2}{m} \quad \text{(22)}
\]

\[
= p_B(1 - \hat{p}_B) \left(1 - \frac{m - 1}{m} k_B \right) \quad \text{(23)}
\]

where \( k_B = \frac{(p_B - p_{BB})/p_{BB}(1 - p_B)}{[p_B(1 - p_B)]} \) and \( i \neq j. \) Similarly, the covariance of \( \hat{p}_B \) and \( \hat{p}_A \) and the variance of \( \hat{p}_A \) are computed to be

\[
\text{Cov}(\hat{p}_B, \hat{p}_A) = p_{AB} - p_{AB} \hat{p}_B \quad \text{(24)}
\]

\[
\text{Var}(\hat{p}_A) = p_A(1 - p_A) \left(1 - \frac{m - 1}{m} k_A \right) \quad \text{(25)}
\]

where \( k_A = \frac{(p_A - p_{AA})/p_{AA}(1 - p_A)}{[p_A(1 - p_A)]}. \) As shown in (25) and (23), \( k_A \) and \( k_B \) are maximum reductions of the variances of the estimators when sample sizes are infinite. Some numerical examples when \( t_4 = 5.0 \) are shown in Figure 4. They are usually not large. This means that increasing the sample size using only one locus is not fruitful for the estimation of \( p_A \) and \( p_B. \) This is especially true when the separation time, \( t_1, \) of the two populations are large.

If \( t_1 \leq t_4, \) we can estimate \( t_1 \) from \( \hat{p}_A \) and \( \hat{p}_B \) using the relationship (10).

\[
\hat{t}_1 = \log \frac{1 - \hat{p}_B}{1 - \hat{p}_A}. \quad \text{(26)}
\]

For usual loci, \( t_1 \) can be estimated by Wright's \( F_{ST} \) which is a function of second moments of gene frequencies. Here, the estimation uses the first moments of gene frequencies because of the conditioning. Note that this method of estimation can be used only when we are sure that \( t_1 \leq t_4, \) since the relationship (10) does not hold if \( t_1 \geq t_4. \)

**Estimation of \( t_a \) and \( t_b: \)** As shown in (9) and (18), \( p_A \) and \( p_{BB} \) are functions of \( t_1, t_a, t_b. \) Thus, if we have estimates of \( p_A \) and \( p_{BB} \) and also have an estimate of \( t_1 \) by other means such as a measurement of DNA sequence divergence, we can estimate \( t_b \) and \( t_a \) by solving the two equations. Unfortunately, we could not find explicit solutions for the two equations. So we need to solve them numerically. We investigated the behavior of these estimators for \( t_a \) and \( t_b \) by simulation. First gene frequency data are generated by building gene trees for two populations which were separated \( t_1 \) generations ago and then randomly putting mutations in the lineage of the first sample from population A according to the density \( f(t) \) defined in (1) (HUDSON 1983). The probabilities \( p_A \) and \( p_{BB} \) are estimated from the frequency data. To obtain meaningful estimates, we sampled from multiple independent loci. We computed the averages of \( \hat{p}_A \) and \( \hat{p}_{BB} \) and used them as estimators. Then, \( t_a \) and \( t_b \) are estimated by solving (9) and (18) for \( t_a \) and \( t_b \) replacing the left-hand sides with the estimates of \( p_A \) and \( p_{BB} \) using Newton's method.
SINE Polymorphism

Estimates of \( t_A \) and \( t_B \) using data generated by simulation

| \( t_A \) | \( t_B \) | \( t_{AB} \) | \( i_A \) (std) | \( i_B \) (std) | Success
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>2.0</td>
<td>1.0</td>
<td>10</td>
<td>-0.09 (0.41)</td>
<td>2.13 (0.66)</td>
</tr>
<tr>
<td>10</td>
<td>-0.04 (0.32)</td>
<td>2.08 (0.42)</td>
<td>0.86</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>-0.08 (0.22)</td>
<td>2.04 (0.24)</td>
<td>0.94</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>-0.03 (0.15)</td>
<td>2.02 (0.18)</td>
<td>0.98</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>2.0</td>
<td>1.0</td>
<td>50</td>
<td>0.45 (0.29)</td>
<td>2.04 (0.30)</td>
</tr>
<tr>
<td>50</td>
<td>0.29 (0.29)</td>
<td>2.28 (0.45)</td>
<td>0.65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>2.0</td>
<td>1.0</td>
<td>50</td>
<td>0.59 (0.33)</td>
<td>2.44 (0.44)</td>
</tr>
<tr>
<td>0.0</td>
<td>5.0</td>
<td>1.0</td>
<td>50</td>
<td>0.02 (0.39)</td>
<td>5.06 (1.05)</td>
</tr>
</tbody>
</table>

For each parameter set, 10,000/1 simulations are carried out. The values shown in the columns \( t_A \) and \( t_B \) are averages of these estimates.

\(^a\) Number of loci used for the estimation.

\(^b\) Proportions of cases where \( t_A \) and \( t_B \) are estimated (see text).

Table 1. We sampled eleven and ten genomes from population A and B, respectively, for each locus in the simulation. Sometimes no solution for \( t_A \) and \( t_B \) is found. This is because the range of \( (p_A, p_{AB}) \) as a function of \( (t_A, t_B) \) does not cover the rectangle, [0,1] × [0,1], and \( p_A \) and \( p_{AB} \) may take any values in this rectangle due to stochastic factors. If we can find a solution, we call the estimation successful. Proportions of successful estimations are shown in the right-most column. First, we note that many loci are required to obtain reliable estimates. If the number of loci sampled is less than 10, the estimates are very approximate. Secondly, \( t_A \) is fairly well estimated but \( t_B \) is difficult to estimate. Especially when \( t_B \) is near or greater than \( t_{AB} \), standard errors are large and there is a downward bias. Also the proportion of successful estimations decreases in these cases. As \( t_B \) becomes larger, the estimation becomes more difficult since coalescence of the two genes in population B likely occurs (data not shown).

Net's distance: Net's distance (NEI 1972) is used in many field studies to estimate the time of the separation of two populations [see NEI (1987) chapter 9]. Let \( x_i \) and \( y_j \) be the frequencies of the i-th allele in the two populations. The distance \( D \) is defined as

\[ D = -\log \frac{j_{XY}}{(j_x j_y)^{1/2}} \]  

(27)

where \( j_x = \sum_{i=1}^{k} x_i^2 \), \( j_y = \sum_{i=1}^{k} y_i^2 \), \( j_{XY} = \sum_{i=1}^{k} x_i y_i \). The number of alleles is designated by \( k \). If we assume the infinite allele model of Kimura and Crow (1964), the equilibrium state and the equal sizes of the two populations, the expectation of \( D \) is approximately the mean number of net codon substitutions. Here, we modify the definition (27) of \( D \) to

\[ D = -\log \frac{P_{AB}}{(P_{AA} P_{BB})^{1/2}} \]  

(28)

where \( P_{AB} \), \( P_{AA} \) and \( P_{BB} \) denote the following probabilities:

\[ P_{AB} = \text{Prob}[\text{Two genes each from pop. A and B are the same allele}] \]  

(29)

\[ P_{AA} = \text{Prob}[\text{Two genes from pop. A are the same allele}] \]  

(30)

\[ P_{BB} = \text{Prob}[\text{Two genes from pop. B are the same allele}] \]  

(31)

With this definition, the expected number of net codon substitutions becomes \( D \) under the three assumptions mentioned above. We can estimate \( D \) using appropriate estimators for \( P_{AB} \), \( P_{AA} \) and \( P_{BB} \). Consider the case where \( D \) is estimated using data on the absence or presence of a SINE in a specific locus considered in the previous section. Considering the absence and the presence as two allelic states, we can estimate \( D \) by

\[ -\log \frac{\sum_{i=1}^{l} \sum_{j=1}^{m} [A_i B_j + (1 - A_i)(1 - B_j)]}{(\sum_{i=1}^{l} \sum_{j=1}^{m} A_i + (1 - A_i)(1 - A_j))^{1/2} (\sum_{i=1}^{l} \sum_{j=1}^{m} B_j + (1 - B_i)(1 - B_j))^{1/2}} \]  

(32)

using the previous notation. \( l \) and \( m \) are numbers of samples taken from population A and B, respectively. Since this locus is not evolving as the infinite allele model specifies and is not in the equilibrium state, we cannot expect \( D \) to behave as originally envisaged by NEI (1972). Our question is how \( D \) changes as a function of the separation time, \( t_{AB} \), of the two populations. Using \( p_{AA} \), \( p_{AB} \) and \( p_{BB} \), \( D \) is written as

\[ D = -\log \frac{1 - p_A - p_B + 2p_{AB}}{(1 - 2p_A + 2p_{AA})(1 - 2p_B + 2p_{BB})^{1/2}} \]  

(33)

By inserting (9), (8), (19) and (18) and noting \( p_{AB} = p_{BB} \), we can compute \( D \) as a function of \( t_{AB} \). We do not show the resulting complex expression but show it as a function of \( t_{AB} \) for several values of \( t_B \) when \( t_B = 5.0 \) in Figure 5. When \( t_B \) is zero, \( D \) increases linearly as in the infinite allele model. However, when \( t_B \) is large, the shape of the curve differs much from linear. Thus, in order to estimate the separation time of the two populations using SINEs as markers, we need information on \( t_A \) and \( t_B \).

Phylogenetic incongruence: Okada (1991b) and Kido et al. (1991) proposed to utilize SINE’S as markers and infer the phylogeny of species. Since SINEs are supposed not to be deleted, if one finds a SINE in a specific locus in species (population) A and B but not in C, one deduces that first C diverged from the ancestral species of A and B and then the divergence between A and B occurred. However, because of polymorphisms in the ancestral species, this inference is not always true (Takahata 1989; Wu 1991). We investigate this incongruence when an insertion of a SINE in a specific locus considered in the previous section. Considering the absence and the presence as two allelic states, we can estimate \( D \) by

\[ -\log \frac{\sum_{i=1}^{l} \sum_{j=1}^{m} [A_i B_j + (1 - A_i)(1 - B_j)]}{(\sum_{i=1}^{l} \sum_{j=1}^{m} A_i + (1 - A_i)(1 - A_j))^{1/2} (\sum_{i=1}^{l} \sum_{j=1}^{m} B_j + (1 - B_i)(1 - B_j))^{1/2}} \]  

(32)
FIGURE 5.—Nei’s distance as a function of the separation time of the two populations. \( t_s = 5.0 \).

(A0 = 1). If we observe \( B_1 = 1 \), \( C_1 = 0 \) or \( B_1 = 0 \), \( C_1 = 1 \), we obtain a correct or wrong phylogeny, respectively. Otherwise we cannot make an inference. We define the incongruence probability (Wu 1991), \( Q \), as

\[
Q = \frac{\text{Prob}[B_1 = 0, C_1 = 1 | A_0 = 1]}{\text{Prob}[B_1 = 1, C_1 = 0 | A_0 = 1] + \text{Prob}[B_1 = 0, C_1 = 1 | A_0 = 1]}
\]  (34)

The probabilities appearing on the right-hand side of the equation are calculated to be

\[
\text{Prob}[B_1 = 1, C_1 = 0 | A_0 = 1] = E(B_1(1 - C_1)) = p_B - p_{BC}
\]  (35)

\[
\text{Prob}[B_1 = 0, C_1 = 1 | A_0 = 1] = E(1 - B_1)C_1) = p_C - p_{BC}
\]  (36)

Using (8) and (17) with \( t_s \) exchanged, we obtain \( Q \) as

\[
Q = \frac{1}{2} \left[ \frac{1 - e^{-(t_1 - t_2)}}{1 - A} \right] \quad (t_2 \leq t_1)
\]

\[
= \frac{2 + [(t_2 - t_2 + 1) - e^{(t_1 - t_2)} - e^{(t_2 - t_1)}]B}{1 - e^{t_1 - t_2}B/2} \quad (t_1 \leq t_1 \leq t_2 \leq t_0)
\]

\[
= 0 \quad (t_1 = t_1 \leq t_2 \leq t_0)
\]

non-informative \( (t_b \leq t_{11}) \)  (37)

where

\[
A = \frac{e^{-2t_2} + e^{-(t_1 + t_2)} + e^{-2t_1}e^{(t_1 + t_2)}}{3}
\]

\[
B = \frac{6}{2e^{t_1 - t_2} + e^{2t_2}e^{t_1 - 3t_2}}
\]

Note that if multiplication of the SINE occurs only in the interval \([ t_{11}, t_{12} ]\), the incongruence probability is zero. Thus, an informative locus in which one can choose a phylogeny gives a correct inference on the phylogeny. However, if the multiplication is not confined in this period, incongruence may occur. Some numerical examples of the incongruence probability when this happens are shown in Table 2. As the proportion of the multiplication period which overlaps the interval \([ t_{11}, t_{21} ]\) becomes smaller, the incongruence probability increases. Thus, again, knowledge of the multiplication period \([ t_s, t_b ]\) is important.

DISCUSSION

Members of the HS subfamily (Batzer and Deininger 1991) in the Alu family are known to show polymorphisms with regard to their presence or absence at specific sites in the genome in human populations. Since data from different species is scant, we cannot estimate \( t_s \) and \( t_b \) using the method described in the previous section. Nevertheless, if we know either parameter, \( t_s \) or \( t_b \), we can estimate the other only from data within population. Here, as a worked example, we estimate \( t_b \) from the human data assuming \( t_s \) to be zero. \( t_s \) is assumed to be zero because an insertion event of the subfamily member is recently observed (Wallace et al. 1991). The average frequency of Alu-carrying genomes in human populations in 13 sites is 0.93 (Batzer et al. 1991; Materia et al. 1990; Lim et al. 1986). By equating \( p_A \) [see (9)] to 0.93 and solving the resulting equation for \( t_b \) with \( t_s = 0 \), we obtain \( t_b = 14.3 \). Note that time is measured in units of 2\( N \) generations. If we assume that generation time and effective population size are 15 years and 10,000, respectively (Nei 1987), this amounts to 43.3 million years. In fact, gorilla seems to have a small number of the HS (also called PV) family Alus (Materia, Hellmann and Schmid 1990; Leeflang et al. 1992) and human and gorilla are supposed to be separated about 5 million years ago (Hasegawa and Kishino 1991). This estimation is very approximate because of the small number of loci sampled and the assumption of \( t_s = 0 \). Still, our estimation shows that the data is compatible with the hypothesis that the expansion of the subfamily started a little before the separation of human and gorilla and has continued.
with a more or less constant rate since then considering a rather large variance of the estimator.

In our model, we assumed that deletion of an element never occurs. This assumption is based on the observation that many SINEs are found to occupy the same sites in distantly related species (Koop et al. 1986). For transposable elements which are known to be excised, elements are rarely observed to occupy the same site even within a species (Charlesworth and Langley 1989). Although the deletion rate is thus considered to be very low, we briefly consider the effect of deletion. Let \( u \) be the deletion rate of a SINE per element per generation. The probabilities \( p_A \) and \( p_B \) are computed similarly as in the case without deletion computing the moments of gene frequencies,

\[
p_B = \begin{cases} 
C\left[e^{-2Nu} - C\left(e^{-4Nu} - e^{-2Nu}\right)\right] & (t_A \leq t_B) \\
C\left[e^{-4Nu} - e^{-2Nu}\right] & (t_A \leq t_B, t_A \leq t_B) \\
0 & (t_B \leq t_A) 
\end{cases}
\]

and

\[
p_A = C\left[1 - D\left(e^{-(4Nu+1)} - e^{-2Nu}\right)\right]
\]

where

\[
C = \frac{1}{2Nu + 1}, \quad D = \frac{2Nu}{4Nu + 1}.
\]

From these expressions, we can see that \( Nu \) is an important parameter when deletion occurs. If \( Nu \) is much smaller than 1, (38) and (39) reduce to (8) and (9), respectively, unless \( t_B \) is much larger than one. There are many sites which are occupied, for example, both in human and orangutan (Koop et al. 1986). Since these two species have been separated for more than ten million years, \( Nu \) is considered to be much smaller than one and our assumption of no deletion is justified. If we let \( t_A = 0 \) and \( t_B \to \infty \), we obtain \( p_A \) in the equilibrium state when the duplication rate is constant throughout time,

\[
p_A = \frac{1}{4Nu + 1}.
\]

This expression is the same as the equilibrium allelism of Ohta (1985) and Equation 13.10 of Kaplan and Hudson (1989).

We used a rectangular shape function for the duplication rate. At present we do not have much information on the shape of this function except that there seemed to be bursts of duplication in the history of SINEs which resulted in the subfamily structures of the SINE sequences (Jurka and Smith 1988; Britten et al. 1988; Quentin 1989). This subfamily structure is considered to be results of successive waves of SINE expansions (Quentin 1988). Our model is for one (most recent) episode of duplication bursts and if the shape of \( f(t) \) is unimodal, the consequences would be not much different. So we think that our approach is at least justified as a first step to obtain more information on the history of SINE expansions. If the shape is more complex, we can still compute those probabilities just by changing the function with which the integrations are computed.

Our model of SINE differs from the usual one-locus two-allele model in two respects. Firstly, a locus is found because it is occupied by an element in the first sample. This situation is incorporated in our model by considering conditional probabilities. Secondly, insertions are assumed to occur only in a limited period. Because of these properties, we need to be careful when we analyze data. For example, Wright's \( F \) statistics have different expectation because the population in which the locus was originally found is special [compare (18) with (19)]. Only when \( t_A \geq t_B \) and when we use data from populations B and C [case (b) in Figure 3], the expected value of \( F_{ST} \) becomes 1 - \( \exp(t_B) \). Also the expectation of Nei's distance and the incongruence probability [compare with those in Wu (1991)] take different expressions in the present model. We must utilize a model which incorporates features of the loci in question in order to obtain appropriate inferences from data.

Here, we explored methods of inference which utilize data on the presence or absence of a SINE element at a specific site. Types of inferences obtained using this kind of data are limited. For example, only certain ranges of \( t_A \) and \( t_B \) can be fairly well estimated by such data (see Table 1). Also our inferences are confined to (sub)families of SINES which were recently expanding so that there are polymorphisms with regard to the presence or absence of the element. In order to get more information, we need to analyze sequence data which are now abundant (Jurka and Milosavljevic 1991; Shen, Batzer and Deininger 1991). Model analyses of sequence divergence incorporating the non-equilibrium assumption are the next step of population genetical assumption to the evolution of SINES.

We thank N. Okada for introducing us to this problem and N. Okada and his group for stimulating discussion. We thank B. Walsh and an anonymous reviewer for comments on the manuscript. This research was partially supported by NIG Cooperative Research Program (92-29) and a grant-in-aid from the Ministry of Education, Science and Culture of Japan. This is contribution no. 1937 from the National Institute of Genetics, Mishima, Japan.
LITERATURE CITED


