A Population Genetic Study of the Evolution of SINEs. II. Sequence Evolution Under the Master Copy Model

Hidenori Tachida

Department of Biology, Faculty of Science, Kyushu University 33, Fukuoka 812, Japan

Manuscript received January 5, 1996
Accepted for publication February 29, 1996

ABSTRACT

A transient population genetic model of SINE (short interspersed repetitive element) evolution assuming the master copy model is theoretically investigated. Means and variances of consensus frequency of nucleotides, nucleotide homozygosity, and the number of shared differences that are considered to have caused by mutations occurring in the master copy lineages are computed. All quantities investigated are shown to be monotone functions of the duration of the expansion period. Thus, they can be used to estimate the expansion period although their sampling variances are generally large. Using the theoretical results, the Sb subfamily of human Alu sequences is analyzed. First, the expansion period is estimated from the observed mean and variance of homozygosity. The expansion period is shown to be short compared to the time since the end of the expansion of the subfamily. However, the observed number of the shared differences is more than twice that expected under the master copy model with the estimated expansion period. Alternative models including that with multiple master copy loci to explain this observation are discussed.

SHORT interspersed repetitive elements (SINES) are found in many eukaryotes and considered to transpose to other locations of the genome through RNA intermediates (for review, see Deininger 1989; Okada 1992). One of the most studied SINE families is the primate Alu family whose copy number in human genome is considered to be around a million. There are basically two types of models concerning the expansion of SINES. One extreme is the master copy model in which one or a few elements in the genome have capacity to duplicate and all other elements are pseudogenes (Shen et al. 1991; Deininger et al. 1992). The other extreme is the transposon model in which all elements have capacity to duplicate. Intermediate models between these two extremes are also proposed (Matera et al. 1990; Brookfield 1993). Although difference in the mode of duplication results in difference of the shape of SINE genealogies and thus would be reflected in sequence divergence among SINE elements, discrimination among models is difficult because (1) SINE sequences are short and thus it is difficult to infer genealogies due to stochastic errors and (2) population polymorphisms blur the distinctions among the models. Thus, to judge which model is appropriate for the evolution of a SINE family taking into account these factors, we need to develop population genetic models in which modes of duplication are specified.

A pioneering work in this direction was done by Brookfield (1986) assuming the equilibrium transposon model. He applied his theoretical results to the analysis of the whole Alu family and showed that the sequence divergence estimated from the data is incompatible with the theoretical expectation of the transposon model. Ohta (1986) developed a transient transposon model and showed that her model can explain the sequence divergence among Alu. On the other hand, Kaplan and Hudson (1989) developed an equilibrium master copy model in which the number of copies is kept constant by the balance between duplication and deletion and showed that this model also is compatible with the Alu divergence data. However, recent advances in the studies of Alu and other SINE families showed that SINE families take subfamily structures resulting from successive waves of expansions (Britten 1988; Jurka and Smith 1988; Quentin 1988; Kido et al. 1994). This leads us to consider transient models in which expansion periods are specified. Moreover, the amount of sequence data of SINE elements increased, and it allows us to examine quantities beyond simple sequence divergence and contrast these with their expected values under different model parameters.

In this paper, I investigate the sequence evolution of SINE elements assuming the transient master copy model. Only one master copy locus is assumed. For simplicity, consequences of one episode of expansion that lasts for some period is investigated. Expected values and variances of nucleotide divergence, consensus frequency and other quantities of interests are computed. Using these theoretical results, the Sb subfamily of human Alu sequences are analyzed by first estimating the expansion period and then examining the adequacy of the model.
MODEL.

Assumptions and parameters: We assume the Wright-Fisher model. The population has a constant size $N$ through time and we measure time in unit of $2N$ generations. Complete linkage within an element and free recombination between elements are also assumed. In our model, all genes of the master copy locus become capable of retroposition at the beginning of the expansion period. We assume that the expansion period persisted for $t_0$ units with a constant rate of duplication and $t_1$ units time elapsed since the cessation of the expansion. The model is the same as that of Tachida and Iizuka (1993) in which the polymorphism with regard to the presence or absence of an element was investigated except that $t_0$ there was $t_0 + t_1$ in the present paper. Because the number of elements in a SINE family is usually very large, we assume that infinitely many pseudogene elements were created during the expansion period. All retroposed elements are assumed to be inserted into different locations (loci) of the genome and their insertions are selectively neutral. The mutations occur at a rate of $u$ per site per $2N$ generations and they are assumed to be neutral (Kimura 1968).

Moments of gene frequencies: First we investigate the homozygosity per site, $h_0$, defined as the probability of two elements having the same nucleotide at a position. Suppose that we sampled $n$ elements each of which has $m$ nucleotide sites. Then, $h_0$ is estimated by

$$h_0 = \frac{1}{m(n-1)} \sum_{i} \sum_{j} X_{ij},$$

where $X_{ij}$ takes one if the $i$th sites of $i$th- and $j$th-sampled elements have the same nucleotide and zero otherwise. If we assume the Jukes-Cantor mutation model (Jukes and Cantor 1968) in which mutation to any other nucleotides occur at the same rate of $u/4$ per $2N$ generations, then $h_0$ is expressed by the probability that the sites of the two elements are identical by descent (IBD) as

$$h_0 = \theta + \frac{1}{4}(1 - \theta) = \frac{1}{4}(1 + 3\theta).$$

(1)

Now we compute $\theta$. Because the number of elements is assumed to be infinite, the two random elements taken from the present population are from different loci. Also the probability of sampling an element from the master copy locus is low and negligible. In the computation, first note that $\theta$ is expressed using the coalescence time $T$ of the two genes as

$$\theta = E[e^{-2uT}],$$

where $E$ denotes the expectation. The coalescence time $T$ is divided into three independent parts as

$$T = T_1 + T_2 + T_3,$$

where $T_1$ is the time between the present and the end of the expansion ($T_1 = t_0$), $T_2$ is the time between the end of the expansion and the older transposition event of the two elements from the master copy, and $T_3$ is the coalescence time of two random genes at the master copy locus (see Figure 1). Since the three $T$'s are all independent,

$$\theta = E[e^{-2uT_1}]E[e^{-2uT_2}]E[e^{-2uT_3}].$$

(4)

The first term is $e^{-2u}$. The third term is $1/(1 + 2u)$ (Kimura and Crow 1964). To compute the second term, we need the density of $T_2$. Since the rate of duplication is constant through the expansion period, the time between the end of the expansion and the transposition event is distributed uniformly in $(0, t_0)$.

Thus,

$$\text{Prob}[T_2 < t] = (t/t_0)^2 (0 \leq t \leq t_0)$$

and this leads to

$$E[e^{-2uT_2}] = \frac{2}{(2ut_0)^2} \left[ 1 - (1 + 2ut_0)e^{-2ut_0} \right].$$

(5)

Putting all these terms into Equation 4, we obtain

$$\theta = \frac{2}{(2ut_0)^2} \left[ 1 - (1 + 2ut_0)e^{-2ut_0} \right] e^{-2uT_3},$$

(6)

For small $u$, $t_0u$ and $t_1u$, $\theta$ is approximately expressed as

$$\theta \approx 1 - \frac{1}{3}(ut_0) - 2(ut_1) - 2u.$$

(7)

Thus, $\theta$ decays with a rate of $4/3$ and 2 per unit increase of $ut_0$ and $ut_1$, respectively.

Higher order IBD probabilities can be obtained, in principle, in a similar manner. For example, the probability $\gamma$ that three random genes are identical by descent is expressed as

$$\gamma = \frac{[(8 + 6t_0u)e^{-3t_0u} - 9e^{-2t_0u} + 1]e^{-t_1u}}{3(t_0u)^3(1 + 2u)(1 + u)}.$$

(8)

Using (8) for $\gamma$ and (6) for $\theta$, the probability, $m_3$,
that three elements have the same nucleotide at a position is
\[ m_3 = \frac{1}{16} (6\gamma + 9\theta + 1). \]  
(9)
The probability \( m_3 \) is the sum of the third moments of the nucleotide frequencies at a site. For small \( u, t u \) and \( t_s \), \( \gamma \) is approximately expressed as
\[ \gamma \approx 1 - 2(ut_s) - 3(u_t) - 3u. \]  
(10)
Note that the approximate expressions (7) and (10) for \( \theta \) and \( \gamma \) have the same form \( 1 - aX \), where \( X = (2\theta + 3\theta + 3)u \), \( a = \frac{2}{3} \) for \( \theta \) and \( a = 1 \) for \( \gamma \). Therefore, the relationship between \( \theta \) and \( \gamma \) is the same as that in the one locus random mating model, that is,
\[ 1 - \theta = \frac{2}{3}(1 - \gamma). \]  
(11)
Also, since \( t_s \) and \( t_s \) enter in the expressions for \( \theta \) and \( \gamma \) through the parameter \( X \), it is not possible to estimate them separately from two statistics related to the two IBD measures, such as homozygosity and the third moments of nucleotide frequencies when \( u, t u \) and \( t_s u \) are small compared to one. Computations of IBD measures for more than three genes become very complex. Thus, we will compute quantities involving fourth moments by simulations described later.

**Consensus frequency:** Another quantity of interest is the consensus nucleotide frequency that is defined as the frequency of the most frequent nucleotide at a site. Although the exact treatment of this quantity is difficult, we can calculate the average consensus frequency approximately when \( \epsilon = (t_s + t_s)u \) is much smaller than one. In this approximation, the population polymorphism is ignored and the master copy is assumed to have a single lineage (the monomorphic approximation). We also assume the Jukes-Cantor mutation scheme for simplicity. Since \( \epsilon \ll 1 \), we can assume that at most one mutation occurs on each lineage. Under the Jukes-Cantor mutation model, the probability of no mutation in a lineage with length \( t \) is approximately \( 1 - 3ut/4 \) from this assumption. Let \( X_{\text{cons}} \) be the consensus frequency at the site. The consensus frequency depends on the time, \( t \), when a nucleotide change occurs on the master copy lineage. Here, the time is measured from the start of the expansion and master copy lineages are lineages of genes at the master copy locus. The density of \( t \) is \( 3u/4 \) for \( 0 \leq t \leq t_s \) and the probability of no change on the master copy lineage is approximately \( 1 - 3tu_4/4 \). We need to consider three cases, \( 0 \leq t \leq t_s/2, t_s/2 \leq t \leq t_s \) and \( t = \infty \). In the last case, the consensus nucleotide is the original one and the conditional expected consensus frequency is
\[
\frac{1}{t_s} \int_0^{t_s} \left[ 1 - \frac{3u}{4} (t_s + t_s - s) \right] ds = 1 - \frac{3}{4} \left( \frac{t_s}{2} + t_s \right) u.
\]  
Thus, the contribution to the expected consensus frequency from this event is
\[
\left( 1 - \frac{3}{4} \left( \frac{t_s}{2} + t_s \right) u \right) \left[ 1 - \frac{3}{4} \left( \frac{t_s}{2} + t_s \right) u \right]
\]  
(12)
where \( O(x) \) represents a term of the same order as that of \( x \). For \( 0 \leq t \leq t_s/2 \), the consensus nucleotide is the changed nucleotide and the conditional expected consensus frequency is \( (ts - t) / (ts + O(\epsilon)) \). Therefore, the contribution from this event is
\[
\int_0^{t_s/2} \left[ 1 - \frac{t}{t_s} + O(\epsilon) \right] \frac{3}{4} u dt = \frac{9tu}{32} + O(\epsilon^2).
\]  
(13)
Incidentally, this is the only case where the consensus nucleotide does not coincide with the original nucleotide (incongruence) and its probability is
\[
\int_0^{t_s/2} \frac{3}{4} u dt = \frac{3tu}{8}.
\]  
(14)
For \( t_s/2 \leq t \leq t_s \), the consensus nucleotide is the original nucleotide and the conditional expectation is \( t_s / (ts + O(\epsilon)) \). Thus, the contribution is
\[
\int_{t_s/2}^{t_s} \left[ \frac{t}{t_s} + O(\epsilon) \right] \frac{3}{4} u dt = \frac{9tu}{32} + O(\epsilon^2).
\]  
(15)
Combining (12), (13), and (15), we obtain
\[
E[X_{\text{cons}}] = 1 - \frac{3}{4}tu_s - \frac{3}{4}tu_s + O(\epsilon).
\]  
(16)
In analyses of SINE sequences, the divergence, \( D_r \), from the consensus and the divergence, \( D_s \), between two sequences are often computed. The expected value of \( D \) is one minus the average consensus frequency. Thus, if \( tu_s \) and \( tu_s \) are small compared to one, it is approximately expressed as
\[
E[D_r] \approx \frac{3}{4}tu_s + \frac{3}{4}tu_s.
\]  
(17)
The expected value of \( D_s \) is the heterozygosity and from (1) and (7) and if we ignore the population polymorphism, it is approximately expressed as
\[
E[D_s] \approx \frac{3}{4}tu_s + tu_s.
\]  
(18)
From these two equations, the relationship between the two divergences is
\[
E[D_s] \approx \frac{1}{2} E[D_r] + \frac{1}{16} tu_s.
\]  
(19)
Britten (1988), analyzing subfamilies of the primate Alu sequences, showed that \( D_r \) is a bit larger than twice the value of \( D_s \), and this suggests that the expansion period \( t_s \) is not zero.

**Simulation**
As mentioned above, it is difficult to compute higher order moments and the exact average consensus fre-
frequency analytically. In addition, quantities of another interest related to shared differences whose definition is explained later are determined by the whole sample of elements and it is impractical to compute such quantities analytically. Therefore, I conducted simulations.

**Method:** The method of simulation is a modification of that described in HUDSON (1990). Consider a situation where we sample \( n \) elements each with \( m \) nucleotide sites from a population. First, a genealogical tree is generated for the \( n \) elements. In the construction of the genealogical tree, the time is measured in unit of \( 2N \) generations and backward from the present (time 0). A genealogical tree consists of \( 2n - 1 \) nodes. The first \( n \) nodes correspond to \( n \) sampled elements and the latter \( n - 1 \) nodes correspond to respective coalescence events, i.e., ancestors of the sampled elements. Each node records the time of coalescence, the two descendant nodes that coalesced to yield the node and the ancestral node that is the next coalescence event for the current node. In the period \((0, t_\text{c})\), no coalescences of elements occur. Respective elements start to participate in the coalescence process in the master copy locus from \( t_\text{c} + T \) according to the algorithm described in HUDSON (1990) where \( T \) is an uniform random variable in \([0, t_\text{c}]\). Thus, after \( t_\text{c} \), a coalescence occurs at a rate of \( \eta_\text{c}(n_\text{e} - 1)/2 \), where \( n_\text{e} \) is the number of the ancestral elements of the sample that were at the master copy locus at the time. In the end, all elements coalesce to a common ancestor (designated as node A) and the genealogy is complete. Next, mutations are assigned to branches of the genealogical tree with mutation rate \( u \). The nucleotides of the sampled elements at a site are determined according to the JUKES-CANTOR mutation model starting from node A and going down the tree to respective sampled elements. This is repeated \( m \) times with the same tree to determine the nucleotide sequence of the sampled \( n \) elements. From the generated sequences, quantities of interests are computed. The whole process is repeated 1000 times to compute the mean and variance of the quantities. Mutation rates used are \( u = 0.002667 \) and \( u = 0.000667 \) that, under the JUKES-CANTOR model, give nucleotide diversities of 0.004 and 0.001, respectively, typical values found in primates (TAKAHATA 1993).

**Variance of homozygosity:** The first quantity examined is the variance of estimated homozygosity among sites. Since mutations that occur in the ancestral elements at the master copy locus result in changes of nucleotide in present multiple elements, such mutations affect nucleotide frequencies more than those that occurred after the retroposition. Thus, the variance is expected to increase as \( t_\text{c} \) increases because the number of such mutations increases. This is indeed the case as shown in Figure 2 where \( t_\text{c} \) is changed keeping \( t_\text{c} + t_r = 30 \). By changing \( t_\text{c} \) from zero to 30, the variance of homozygosity is almost tripled. In the figure, the standard deviations of the estimated variances of homozygosity when the sample size is \( n = 100 \) and the number of sites is \( m = 200 \) are also indicated by error bars. The standard deviations are generally large and this limits the utility of the variance of homozygosity as a statistics to estimate \( t_\text{c} \).

**Consensus frequency:** Next we examined the average consensus frequency and compared it with the approximate result (16) obtained above. I also computed the congruence probability that the consensus nucleotide frequency, \( X_{\text{cons}} \), is the frequency, \( X_0 \), of the original nucleotide of the common ancestral element (\( \text{Prob}(X_{\text{cons}} = X_0) \)) and the approximate value 1 - (14). Table 1 shows the result when \( t_\text{c} \) is changed, keeping \( t_\text{c} + t_r = 30 \). As \( t_\text{c} \) increases, the average consensus frequency decreases and the congruence probability decreases. As shown in the table, values computed from (16) and (14) are fairly close to those obtained by simulations. Slight overestimations of the approximate expressions are probably due to the population polymorphism that will increase variation at each site. In addition to these quantities, the variance of the consensus frequency among sites is computed. The variance is computed for each sample set (\( n \) samples each with \( m \) sites) and its mean and standard deviation are shown. As the variance of homozygosity, the variance of consensus frequency increases as \( t_\text{c} \) increases. Again its standard deviation is very large even when \( m = 200 \).

**Shared difference:** In the analysis of SINE elements, shared differences (SDs) are frequently used to identify sites that changed in the lineage of master copies. SDs are defined as nonconsensus nucleotides that have high frequencies (more precise definition will be given later). Consider a pseudogene element sampled. Mutations that occurred after retroposition of the element can never be inherited by any other elements. On the other hand, mutations that occurred before retroposition can be inherited to other elements because they
TABLE 1

Expected consensus frequencies and probabilities of congruence

<table>
<thead>
<tr>
<th>$t_b$</th>
<th>$t_c$</th>
<th>E[$X_{cons}$] Simulated</th>
<th>Approximate</th>
<th>Var[$X_{cons}$] $^*$</th>
<th>Prob[$X_{cons} = X_c$] Simulated</th>
<th>Approximate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>30</td>
<td>0.939 ± 0.004</td>
<td>0.942</td>
<td>0.00115 ± 0.00074</td>
<td>0.997 ± 0.004</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>24</td>
<td>0.943 ± 0.003</td>
<td>0.945</td>
<td>0.00155 ± 0.00087</td>
<td>0.992 ± 0.007</td>
<td>0.994</td>
</tr>
<tr>
<td>12</td>
<td>18</td>
<td>0.946 ± 0.003</td>
<td>0.948</td>
<td>0.00225 ± 0.00110</td>
<td>0.987 ± 0.009</td>
<td>0.988</td>
</tr>
<tr>
<td>18</td>
<td>12</td>
<td>0.949 ± 0.004</td>
<td>0.951</td>
<td>0.00307 ± 0.00130</td>
<td>0.981 ± 0.010</td>
<td>0.982</td>
</tr>
<tr>
<td>24</td>
<td>6</td>
<td>0.952 ± 0.005</td>
<td>0.954</td>
<td>0.00368 ± 0.00148</td>
<td>0.974 ± 0.012</td>
<td>0.976</td>
</tr>
<tr>
<td>30</td>
<td>0</td>
<td>0.955 ± 0.005</td>
<td>0.957</td>
<td>0.00464 ± 0.00164</td>
<td>0.970 ± 0.013</td>
<td>0.971</td>
</tr>
</tbody>
</table>

Other parameters are $n = 100$ (sample size), $m = 200$ (number of sites) and $u = 0.002667$ (mutation rate). One standard deviation is also shown for quantities obtained by simulation.

$^*$ The variance of estimated $X_{cons}$ among sites is shown.

occurred on the master copy lineages. Therefore, if $(t_b + t_c)u$ is small so that the expected nonconsensus frequency is not high (say, less than several percents), some of master copy mutations result in exceptionally higher frequencies of nonconsensus nucleotides. This is observed in our simulation. Figure 3 shows the frequency spectrum of the nonconsensus nucleotide for $t_bu = 0.011$, $t_cu = 0.045$. If all mutations are independent, the distribution is expected to be approximately Poisson with the same mean and it is also shown for comparison. Apparently, some master copy mutations contribute to higher nonconsensus frequencies. Here, I find SD by examining whether the nonconsensus frequency at a site is significantly different from that averaged over all other sites. More concretely, for each site, I use the $\chi^2$ test to test whether the ratio of the nonconsensus frequency to the consensus frequency at the site is significantly different from that averaged over all other sites. If the test gives significance, the site is regarded as a SD-carrying site, and the second frequent nucleotide at the site is considered as a SD. Since there are $m$ sites, the critical value of $\chi^2$ should be modified according to the Bonferroni procedure to avoid falsely finding SDs. By this procedure, if there is no mutation on the master copy lineage, we will find a SD site with the specified significance level. Note that this procedure is conservative in finding SDs because the Bonferroni procedure gives the critical value for the largest nonconsensus frequency among the whole sites. The number of SDs is expected to increase if the expansion period, $t_b$, increases because then the number of mutations on the master copy lineage increases.

The average numbers of SDs are computed by simulations and they are shown as functions of $u_b$ for $u = 0.00267$ and $u = 0.00067$ in Figure 4 when $(t_b + t_c)u = 0.08$ and the number of sites is 200. They increase almost linearly as functions of $u_b$. If we use the monomorphic approximation, the expected number of mutations in the lineage of the master copy is $m \times X_{cons}$, so the linearly increases. The average number of SDs would differ from this expected value for two reasons. First, not all master copy mutations are detected because of statistical noise from mutations after retroposition. Second, the sum of the master copy lineages that is esti-

![Figure 3](image-url)  
**FIGURE 3.**—Frequency spectrum of nonconsensus frequency. Observed values are from simulations with $t_b = 4$, $t_c = 16.9$, $u = 0.002667$, $n = 100$. In total 200,000 sites are examined. A Poisson distribution with the same mean, 4.10, as that of the observed distribution is also shown for comparison.

![Figure 4](image-url)  
**FIGURE 4.**—The average number of SDs. Two mutation rates are used. $t_b$ and $t_c$ are changed keeping $(t_b + t_c)u = 0.08$. The number of sites is $m = 200$. 
Correlated change: In the master copy model, if there is no population polymorphism (or if $2N$ is one) as considered in the monomorphic approximation, there is only a single lineage of the master copy. Thus, elements that have a mutation that occurred later in the master copy lineage are expected to have all the mutations that occurred before that mutation in the lineage. In the sampled sequences, this is observed as correlated changes of SD-carrying sites. Correlated changes are often used to determine subfamilies of a SINE family. In such cases, the SD-carrying sites are called diagnostic sites (BRITTEN 1988). To quantify the successive occurrences of SDs, I computed a new quantity, $r$. Consider a pair of SD-carrying sites, $A$ and $B$. In the calculation of $r$, it is important to discriminate SDs and shared changes (SCs). SCs are defined as the variation that occurred later in the master copy lineages. They are not necessarily the second frequent nucleotides at SD-carrying sites because if mutations occur early in the master copy lineages, SC is the most frequent nucleotide at the site. Note that SCs are not recognizable by just looking at the frequencies of nucleotides at the site in a subfamily. However, we may be able to determine them by inferring the nucleotide of the common ancestor of the subfamily, for example, from the sequences of older subfamilies. Let $X(A)$, $X(B)$, and $X(AB)$ be the numbers of elements that have SC at the site $A$, site $B$, and at both $A$ and $B$ sites, respectively. Then, $r_{AB}$ is defined as

$$r_{AB} = \frac{X(AB)}{\min[X(A), X(B)]} .$$

The average, $r$, of $r_{AB}$ over all pairs of SD-carrying sites is calculated to evaluate the successiveness of SD mutations on the master copy lineages. If there is no polymorphism in the master copy locus and mutations after retroposition is ignored, $r$ is expected to become one. However, in real situations, neither condition holds true. If there are multiple lineage of the master copy, mutations do not necessarily occur successively in one lineage. If mutations occur after retroposition, SC in some elements may be changed. Both contribute to reduce $r$ and thus $E[r]$ is always $<1$.

In the simulation, the number of SD-carrying sites are not always $>1$, especially for small $ut_0$. In these cases, $r$ was not computed and the averages were calculated excluding these cases. Table 2 shows the expected values and standard deviations of $r$ under two different values of $u$ when $t_0$ and $t_1$ are changed keeping $(t_0 + t_1)u = 0.08$. $E[r]$ rapidly approaches one as $ut_0$ is increased from zero. This is especially evident in the case with small $u$. Thus, the reduction of $r$ is largely caused by the initial polymorphism of the master copy locus, and most SC mutations that occurred in the master copy lineages after the start of retroposition are considered to be successive. Standard deviation of $r$ is rather large when $E[r]$ is small but is small for $E[r] \geq 0.9$. Thus, if we obtain $r \geq 0.9$, this would be a good indicator showing large $ut_0$.

<table>
<thead>
<tr>
<th>$ut_0$</th>
<th>$ut_1$</th>
<th>$u = 0.002667$</th>
<th>$u = 0.000667$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.000</td>
<td>0.080</td>
<td>0.390 ± 0.304</td>
<td>0.349 ± 0.343</td>
</tr>
<tr>
<td>0.016</td>
<td>0.064</td>
<td>0.612 ± 0.294</td>
<td>0.838 ± 0.205</td>
</tr>
<tr>
<td>0.032</td>
<td>0.048</td>
<td>0.776 ± 0.216</td>
<td>0.904 ± 0.115</td>
</tr>
<tr>
<td>0.048</td>
<td>0.032</td>
<td>0.859 ± 0.148</td>
<td>0.926 ± 0.073</td>
</tr>
<tr>
<td>0.064</td>
<td>0.016</td>
<td>0.901 ± 0.108</td>
<td>0.940 ± 0.062</td>
</tr>
<tr>
<td>0.080</td>
<td>0.000</td>
<td>0.926 ± 0.081</td>
<td>0.952 ± 0.052</td>
</tr>
</tbody>
</table>

Other parameters are $n = 100$ (sample size) and $m = 200$ (number of sites). One standard deviation is also shown.

ANALYSIS OF ALU SB SUBFAMILY

Based on the theoretical result of the previous sections, I analyzed the Sb subfamily of Alu sequences (JURKA and MILOSAVLJEVIC 1991). The reasons for choosing this subfamily is that its size is not small and that it is fairly young and thus noise due to mutation is considered to be small (JURKA and MILOSAVLJEVIC 1991; BRITTEN 1994b). Sequence data are taken from the human Alu database compiled by JURKA and maintained in the National Center for Biotechnology Information. From this database, sequences with the subfamily name Sb and whose lengths are $\geq 250$ were extracted and aligned without including identical sequences. A complete Alu element has 281 bases. A very young subfamily Sb1 was not included. For the multiple alignment, the malign program developed by HEIN in the package ODEN (INA 1994) was used with manual adjustments. From this alignment, consensus nucleotides were determined at respective sites. In the consensus sequence, 25 CpG dinucleotides are identified. Since these sites are known to evolve rapidly after retroposition (LABUDA and STRIKER 1989; BATZER et al. 1990), they are not included in the following analysis. Total number of nucleotide differences of each element from the consensus sequence is counted and its distribution is examined (data not shown). Since there are two separate peaks, one around eight and the other around 17 suggesting inclusion of another subfamily, I removed

\[e^{t_0} - e^{-t_0} = 0.064\text{ (comparing } e^{t_0} - e^{-t_0}\text{ with } 0.064)\]
those elements that are separated by >14 differences from the consensus. Thus, I analyzed 153 elements of the Sb subfamily at 231 sites. Only nucleotide changes were considered and changes due to indels were excluded from the analysis. In the following, first, I estimate \( t_\mu \) and \( t_\sigma \) using statistics related to moments of frequencies. Then, from this information, I examine the validity of the master copy model for this data set utilizing the number of SD and the measure of correlated change, \( r \).

In the analysis of data, sites whose consensus nucleotide are G or C (GC sites) and sites whose consensus are A or T (AT sites) should be discriminated since mutation rates at the two types of sites in noncoding regions of primates are known to differ (Bailey et al. 1992; Blake et al. 1992; Britten 1994a). Table 3 shows the estimates of the average heterozygosity \( \langle h \rangle \), variance of heterozygosity among sites \( \text{Var}(\langle h \rangle) \), average consensus frequency \( \langle X_{cm} \rangle \) and variance of consensus frequency \( \text{Var}(\langle X_{cm} \rangle) \) separately for the GC and AT sites. Because GC sites have more mutations and thus contain more information, I use the data for GC sites \((m = 128)\) in the following analysis. There are four statistics available and the number of parameters to be estimated is three \((\langle h \rangle, t_\mu, \text{and} u)\). Therefore, we can, in principle, estimate them by methods of moments or the maximum likelihood method. Here, I do not take this approach because some of the statistics are correlated and there is a better estimate of \( u \) from other regions of genomes. Li and Sadler (1991) compared alleles of 49 loci in human and estimated the nucleotide diversity, \( \pi \), in noncoding regions of the nuclear genome to be 0.001. Expected value of \( \pi \) is \( 3u/2 \) in our notation. Takahata (1993) estimated the ratio of \( \pi \) in chimpanzee to that in human to be about four. I regard \( \pi \) of chimpanzee to be typical in primate evolution and use \( u = 4 \times 0.001 = 0.002667 \). Next, I obtained the relationship between \( t_\mu \) and \( t_\sigma \) from the average homozygosity since an analytical expression of the average homozygosity is available. From (1) and (6), \( t_\mu \) is expressed in terms of \( t_\mu \) and \( u \) as

\[
 t_\mu = - \frac{1}{2u} \ln \left[ \frac{4E[\langle h \rangle] - 1}{6[1 - (1 + 2ut_\mu)e^{-2ut_\mu}]^2} \right]. 
\]

I calculated \( t_\mu \) from this formula changing \( t_\mu \) and used them to compute means of various statistics and their standard deviations by simulation. The standard deviations are those when the number of sites is 128 and the sample size is 153. The result is shown in Table 4. Because \( t_\mu \) is adjusted so that \( E[\langle h \rangle] \) equals the observed value, other three statistics, \( \text{Var}(\langle h \rangle) \), \( \langle X_{cm} \rangle \) and \( \text{Var}(\langle X_{cm} \rangle) \) can be used for the estimation of \( t_\mu \) and \( t_\sigma \).

Although the expected values of all these three statistics are monotone functions of \( t_\mu \), \( \text{Var}(\langle h \rangle) \) seems to be best for estimating \( t_\mu \) because the ratio of its one standard deviation to the difference between the two extreme values, 0.0019 and 0.0052, when \( ts \) are 0 and 20, respectively, is the smallest. Since the observed value of \( \text{Var}(\langle h \rangle) \) is 0.00231, \( t_\mu \) and \( t_\sigma \) are estimated to be 4 and 16.9, respectively. The observed values of \( \langle X_{cm} \rangle \) and \( \text{Var}(\langle X_{cm} \rangle) \) are also close to the expected values with these \( t_\mu \) and \( t_\sigma \). However, note that since the standard deviations are all very large reliability of the estimate is not very high.

Next, I examined the validity of the master copy model. The SD-carrying sites were determined by the procedure described in the previous section. The 5% critical value of the \( \chi^2 \) is 12.53 when the number of sites is 128. Six SD-carrying sites were found in the GC sites. From the consensus sequence of the older families (Jurka and Milosavljevic 1991), the original nucleotide at these sites are inferred and SCs were determined. All six SD-carrying sites have C or G as original nucleotides and the measure of correlated change, \( r \), in these GC sites was estimated to be 0.30. The theoretical expectation of the number of SD and \( r \) are also shown in Table 4. With \( t_\mu = 4 \) and \( t_\sigma = 16.9 \), the number of SD-carrying sites and \( r \) are 2.18 and 0.49, respectively. Although the observed \( r \) was within one standard deviation from the theoretical expectation, the observed number of SD-carrying sites was much larger than the expected value. Indeed the observed number of cases where the number of SD sites is larger than 5 was 33 out of 1000 replications in the simulation \((P = 0.0033)\) when \( t_\mu = 4 \), \( t_\sigma = 16.9 \) (data not shown). This suggests that the present single master copy model may not be appropriate for the Sb subfamily, although there are uncertainties, for example, in the estimation of \( t_\mu \) and \( t_\sigma \), and the basis of the suggestion might not be very strong.

**DISCUSSION**

**Assumptions of the model:** In our model, mutations are assumed to be all selectively neutral (Kimura 1983). This assumption seems to hold after retroposition of elements since elements are considered to become pseudogenes without function after retroposition (but see Britten 1994a for other possibilities). However, there may be some constraints when elements were at the master copy locus. If the number of such sites are small compared to the total number of sites, this does not affect the consequence much. But if the number of such sites is not negligible, quantities examined in this paper will be changed. Especially,
The retroposition rate per genome is assumed to be constant through the expansion period. There is no biological justification of this assumption except that if the rate is constant per master copy, the rate per genome is constant because the number of master copy locus is one throughout time. If changes of the rate are small, the behavior may be approximated by the constant rate model. However, if the rate drastically changes, somewhat different behavior may be observed. For example, consider an extreme case where a proportion \( p \) of the elements retroposed at time \( t_1 \) instantaneously and the rest (a proportion \( 1 - p \)) retroposed at time \( t_2 \). In this case, most of the master copy mutations occur between \( t_1 \) and \( t_2 \) and thus there would be a peak at \( \min(p, 1 - p) \) in the spectrum of the nonconsensus frequencies. This spectrum is quite different from that of the constant rate model where a flat pattern is observed (see Figure 3). Such a drastic change of the retroposition rate may provide an explanation for the discrepancy between the Alu Sb subfamily data and the expectations from the single master copy model as discussed later.

The population size is assumed to be constant through time in our analysis. Human populations are suggested to have experienced expansions, extinctions and recolonization (Takahata 1994), and such fluctuation of the population size is expected in the primate evolution. However, in our model, only the population size during the expansion period \( t_e \) influences the results. The effect of the population size is most evident on the number of SD-carrying sites when \( t_e \) is small (see Figure 4). So unless the size changes occur in a short period, the result will not change much. Even if changes occur rapidly, arguments relying on the notion of effective size may be applied but further studies are necessary to ascertain this conjecture.

Finally, we assumed that all genes at the master copy locus become capable of retroposition at the beginning of the expansion period. This assumption was made partly because of simplicity. Such a model is applicable,
Master Copy Model

for example, if some external agent makes the master copy locus capable of retroposition. However, the expansion may start because a mutant capable of retroposition by itself appears at the master copy locus. A mutant may be neutral or selectively advantageous. If the mutant is neutral, it takes about 4N generations for the mutant gene to be fixed and retroposition occurs in this period with the rate proportional to the contemporary gene frequency. Since the average sojourn time in one frequency class of a mutant gene to be fixed is about 4Ngenerations for the mutant is neutral, it takes about 4N generations for the consensus at the CONSBI positions. Thus, nucleotide differences from the consensus sequence, and his estimates for the Sb subfamily are estimated to be 26 and 21 million years, respectively. BRITTEN (1994b) obtained estimates of the expansion periods of the Alu Sb subfamily and obtained tM = 4 and tL = 16.9, assuming u = 0.002667 per 2N generations. Recall that tM and tL are measured in units of 2N generations. To estimate the expansion period in terms of years, we need to know 2Ng where g is the generation time. Note that \( \frac{3}{4} \times u = 2Ng \cdot \frac{1}{2} \), where \( \frac{1}{2} \) is the mutation rate to the other nucleotides per year (\( \frac{3}{4} \) is the correction factor for the Jukes-Cantor mutation scheme). Thus, if we assume that mutation rate is 0.16% per million years (BAILEY et al. 1991), 2Ng is estimated to be 1.25 million years. So the beginning and end of the expansion of the Sb subfamily are estimated to be 26 and 21 million years ago, respectively. BRITTEN (1994b) obtained estimates of the expansion periods for Alu subfamilies fitting a linear combination of two Poisson distributions and to the distribution of the number of differences from the consensus sequence, and his estimates for the Sb subfamily are 11 and 24 million years. Considering large statistical errors expected in the two estimation methods, the difference between his and the present estimates is not surprising. In addition, there are two other reasons that may have caused differences of estimates due to differences of the models adopted. First, BRITTEN (1994b) used sites that are considered to be conserved in the master copy lineage throughout the Alu evolution (the CONSBI positions). Thus, nucleotide differences from the consensus at the CONSBI positions are assumed to have occurred after retropositions. In our model, all differences are assumed to be neutral and free to change driven by mutation. In fact, some of the SD-carrying sites (sites 1, 144, 217) detected in the present analysis are CONSBI sites, suggesting that mutations occurred at these sites before retroposition. Second, BRITTEN (1994b) fitted a linear combination of two Poisson distributions and thus the two estimates correspond to the times of two instantaneous expansion events each with a duration zero. In our model, continuous expansion with a duration of tM is assumed. It is difficult to discriminate these two modes of expansion. This is illustrated in Figure 5 where the distribution (circle) expected from a continuous expansion \( (tM = 10, tL = 10) \) is fitted by a combination (straight line) of two Poisson distributions with means 11.1 and 16.7, respectively. For comparison, a Poisson distribution with the same mean (15) as the other two distributions is shown. Evidently, the Poisson distribution is different from the two other distributions, but the distribution of continuous expansion and that of the linear combination are similar. To discriminate the two modes of expansion, measures other than the difference from the consensus might be necessary to be examined.

Validity of the model and alternatives: If the expansion period is short compared to the time since the expansion started \( (tL > tM) \) as estimated here, the number of SD-carrying sites becomes small under the present master copy model (see Table 4). However, the number of SD-carrying sites in the GC sites of the Sb subfamily was more than twice that expected from the theoretical model. Since we estimated tM from the variance estimator of homozygosity, this suggests that the number of SD-carrying sites is larger than that expected from the homozygosity variance. Such situations occur when there are more sites whose nonconsensus fre-

**FIGURE 5.**—The distribution of distance between SINE element and the master copy at the CONSBI sites. ○ distribution under the continuous expansion model \( (tM = 10, tL = 10) \). The straight curve is obtained by fitting a linear combination of two Poisson distributions to that of the continuous expansion. The dotted curve is the Poisson distribution with the mean 15.
quences are higher than the critical values of the \( \chi^2 \) test to find SD but lower than those expected from the single master copy model. In the single master copy model, the nonconsensus frequency spectrum is rather flat (Figure 3). Some alternative models may produce larger numbers of SD-carrying sites with the same amount of the variance of homozygosity. For example, if balancing selection keeps several alleles at the master copy locus for a long time, separate lineages are created and the nonconsensus frequencies at SD-carrying sites will be decreased. Similarly if there are multiple master copy loci, the probability of the nonconsensus frequency being in the middle range will decrease. The existence of multiple master copies is suggested in the youngest SI1 subfamily (MATERA et al. 1990; LEFFLAG et al. 1992; HAMMER 1994). Also drastic changes of the retroposition rate result in a nonflat spectrum of the nonconsensus frequency as discussed earlier. Although these models seem to explain the phenomenon qualitatively, we need to investigate these and other models in a similar manner as has been done in the present paper to know whether the data fit to the values expected from respective models or not and to estimate relevant parameters.

I thank B. WALSH and two anonymous reviewers for valuable comments. This research was partially supported by a grant-in-aid from the Ministry of Education, Science and Culture of Japan.

LITERATURE CITED


BRITTEN, R. J., 1994b Evidence that most human Alu sequences were inserted in a process that ceased about 30 million years ago. Proc. Natl. Acad. Sci. USA 91: 6148–6150.


BROOKFIELD, J. F. Y., 1993 The generation of sequence similarity in SINEs and LINEs. Trends Genet. 9: 38.


INA, Y., 1994 ODEN: a program package for molecular evolutionary analysis and database search of DNA and amino acid sequences. CABIOS 10: 11–12.


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


Communicating editor: N. TAKAHATA.