Population Genetics and Molecular Evolution of DNA Sequences in Transposable Elements.

I. A Simulation Framework

T. E. Kijima and Hideki Innan¹

Graduate University for Advanced Studies, Hayama, Kanagawa 240-0193, Japan

ABSTRACT A population genetic simulation framework is developed to understand the behavior and molecular evolution of DNA sequences of transposable elements. Our model incorporates random transposition and excision of transposable element (TE) copies, two modes of selection against TEs, and degeneration of transpositional activity by point mutations. We first investigated the relationships between the behavior of the copy number of TEs and these parameters. Our results show that when selection is weak, the genome can maintain a relatively large number of TEs, but most of them are less active. In contrast, with strong selection, the genome can maintain only a limited number of TEs but the proportion of active copies is large. In such a case, there could be substantial fluctuations of the copy number over generations. We also explored how DNA sequences of TEs evolve through the simulations. In general, active copies form clusters around the original sequence, while less active copies have long branches specific to themselves, exhibiting a star-shaped phylogeny. It is demonstrated that the phylogeny of TE sequences could be informative to understand the dynamics of TE evolution.

TRANSPOSABLE elements (TEs) are ubiquitous and one of the major components in the genomes of many organisms, although the proportion of TEs in a genome varies across species [*Homo sapiens*, 40% (Lander *et al.* 2001); *Drosophila melanogaster*, 15% (Adams *et al.* 2000); *Zea mays*, 70–85% (Craig *et al.* 2002)]. TEs are classified into two major categories based on the transposition machinery: DNA-mediated transposons (DNA transposons) and RNA-mediated transposons (RNA transposons or retrotransposons). Retrotransposons are further classified into three superfamilies: long-terminal repeat (LTR) retrotransposons, long interspersed nuclear elements (LINEs), and short interspersed nuclear elements (SINEs). Our major interest is in RNA transposons that can increase the copy number through replicative transposition.

Integration of new TEs in a genome has a direct effect on the fitness of the host genome. Insertion of a TE in the coding region of a functional gene will cause a loss of the function. If a TE is inserted near a gene, it potentially changes the

expression pattern. Insertions of TEs in junk regions might not be harmful for the host genome, but they can have indirectly deleterious effects because deletion, duplication, inversion, and translocation likely occur through ectopic recombination between TEs. Thus, to understand the mechanisms of genome evolution, it is crucial to know their potential roles, and genomic sequence data have tremendous amounts of information on how TEs have behaved in the past.

Population genetics provide a theoretical framework to understand the evolutionary behavior of any genetic variation, including TEs. In the 1980s, many population genetic models of transposons were developed (Charlesworth and Charlesworth 1983; Langley *et al.* 1983; Charlesworth and Langley 1989). The authors focused on the evolutionary changes of the copy number of TEs in natural populations and found that the transposition–selection balance is the major factor to determine the equilibrium of the TE copy number in a genome. As mentioned above, because TEs usually have harmful effects on the host genome to some extent, almost all of these models consider selection that operates to decrease the copy number of TEs. Models were further extended to explore the potential roles of other factors, such as breeding system (Hickey 1982; Charlesworth and Charlesworth 1995; Wright and Schoen 1999; Morgan 2001), migration between

Copyright © 2013 by the Genetics Society of America
doi: 10.1534/genetics.113.150292

Manuscript received February 11, 2013; accepted for publication August 9, 2013
Supporting information is available online at <http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.150292/-DC1>.

¹Corresponding author: Graduate University for Advanced Studies, Hayama, Kanagawa 240-0193, Japan. E-mail: innan_hideki@soken.ac.jp

subpopulations (Deceliere *et al.* 2005), competition between subfamilies (Le Rouzic and Capi 2006), and activity decay by point mutations (Le Rouzic and Capi 2006; Le Rouzic *et al.* 2007). However, these models focus only on the copy number dynamics and did not consider the sequences of TEs. To the best of our knowledge, the most notable exception is the retrotransposon model of Navarro-Quezada and Schoen (2002) (but see Ohta 1985, Slatkin 1985, Brookfield 1986, and Hudson and Kaplan 1986 for fundamental numerical approaches and Quesneville and Anxolabéhère 2001 and Cerveau *et al.* 2011 for simple models of DNA transposons). Navarro-Quezada and Schoen (2002) modified the simulation model of Charlesworth and Charlesworth (1983) such that each TE has a DNA sequence. Their model assumed that the TE sequence has nonsynonymous and synonymous sites and that point mutations at nonsynonymous sites decrease the transposition rate. With this model, they investigated the pattern of the accumulation of nonsynonymous and synonymous mutations. Le Rouzic *et al.* (2007) also incorporated the decay of activity into the model of Charlesworth and Charlesworth (1983), but this model did not consider DNA sequences of TEs.

We here extend the model of Navarro-Quezada and Schoen (2002) with a few modifications. First, the genome is assigned to functionally important regions and junk regions. It is assumed that insertion of TEs in the former regions is deleterious, but not that in the latter. Second, we use a more realistic function of the relationship between the transposition rate and point mutations according to recent data of *Alu* elements (Bennett *et al.* 2008). Then, we simultaneously investigate the evolutionary behavior of the copy number of TEs and their sequence evolution.

Model

Our model basically follows that of Charlesworth and Charlesworth (1983), except that each TE has a DNA sequence, which is assumed to determine the transpositional activity of the TE. It is also assumed that the genome has functionally important regions and junk regions. The evolutionary dynamics of TEs are simulated in a random-mating population with N diploids. A diploid has a pair of haploid genomes, and each haploid consists of two chromosomes with equal lengths. A single haploid genome has in total T sites, which can be occupied by TEs. These T sites are located on the chromosomes with equal length of intervals. It is assumed that $p \times T$ sites are in functionally important regions so that TE insertion is deleterious at these sites. At the remaining $(1 - p) \times T$ sites, there would be no direct selective force against TE insertion. The allocation of the T sites to the neutral and selected classes is random. The selection parameters for $p \times T$ sites are determined before a simulation run, and they are assumed to be constant over generations.

Each generation of the evolutionary simulation consists of (i) transposition and excision of TEs, (ii) recombination (chromosomal crossing over), (iii) point mutations and indels in the DNA sequences of TEs, and (iv) construction of a new

population for the next generation through random mating and selection.

Detailed descriptions for each step are as follows.

(i) Transposition and excision of TEs

Transposition of an active TE creates a completely identical copy, which will randomly occupy one of the empty sites. The transposition rates of TEs depend on their DNA sequences. The transposition rate for the i th TE is given by $u_i = u_0 \times a_i$ per generation, where u_0 is the maximum transposition rate and a_i is the relative activity of the TE that is determined by its DNA sequence (see below for details). TEs are randomly deleted from the genome at rate v per generation, which is referred to as the excision rate.

(ii) Recombination

TEs are located on recombining chromosomes, and the rate of recombination between adjacent sites is assumed to be c per generation. That is, the number of chiasmata per meiosis is a random variable from a Poisson distribution with mean $c \times T$, and the chromosomal locations are also randomly determined. We define the population recombination parameter as $\rho = 4Nc$.

(iii) Point mutations and indels in the DNA sequences of TEs

We assume that each TE has a DNA sequence that consists of L nucleotides. This sequence is supposed to encode an essential *cis*-acting factor for transposition machinery, and the sequence determines the transpositional activity. We conventionally assume that the sequence consists of $L/3$ codons and that the first and second positions of the codon are nonsynonymous and the third position is synonymous. We use a simple two-allele model, in which only two nucleotides, 0 and 1, are allowed and recurrent mutation occurs between 0 and 1. Nucleotide changes at nonsynonymous positions can potentially affect the activity. The sequence is also subject to indels. We assume that any indel causes a frameshift, immediately resulting in a complete “deactivation” of the transpositional activity [*i.e.*, “pseudogene effect” (McAllister and Werren 1997)]. Thus, once a TE experiences an indel, it will have no chance to create a new copy, while it still occupies the site and accumulates point mutations. The rates of these two types of mutations (point mutations and indels) are denoted by μ and $\phi \times \mu$ per site per generation, respectively, where ϕ represents the relative rate of indels to point mutations. The population mutation parameter is defined as $\theta = 4N\mu$.

In contrast to indels, the effect of nonsynonymous mutations is rather mild. It is assumed that the original sequence at the initiation of a simulation run has the full activity; that is, $a = 1$. Then, as nonsynonymous mutations accumulate, the activity reduces by following a certain function. Let x_i be the proportion of nucleotide differences at nonsynonymous sites between the i th TE and the original sequence. Then, the activity of the i th TE is given by

Figure 1 Relationship between the activity and DNA sequence of human *Alu* elements (data from Bennett *et al.* 2008). The horizontal axis represents the nucleotide identity between *Alu* elements and their consensus sequence. In this work, four curves are used to approximate the relationship. See text for details.

$$\begin{cases} a_i = \left(\frac{x_i - \alpha}{1 - \alpha}\right)^\beta & \text{for } x_i > \alpha \\ a_i = 0 & \text{for } x_i \leq \alpha. \end{cases} \quad (1)$$

This equation produces a monotonically decreasing function from $(a_i, x_i) = (1, 1)$ to $(0, \alpha)$ (see Figure 1). The transposition rate is used in step i.

Thus, the model assumes that the activity is 100% determined by a *cis*-acting factor. One might imagine a sequence like the coding region of reverse transcriptase as a *cis*-acting factor. For such a region, the nonsynonymous and synonymous sites in our model can be considered as functionally important sites as a *cis*-acting factor and as other less functional sites, respectively. Alternatively, for SINEs such as *Alu*, the entire sequence may be important as a *cis*-acting factor because the RNA secondary structure should determine the ability of transposition. Thus, we assume that the activity of a TE is determined by its sequence. The environments provided by *trans*-acting factors are assumed to be uniform for all TEs.

(iv) Construction of a new population for the next generation through random mating and selection

Selection operates against TEs at the diploid level, which is incorporated in this step to create N diploid individuals in the next generation. In this process, a pair of haploid genomes is randomly chosen from the current generation, which becomes a candidate of the diploid offspring that constitute the next generation. Let w be the fitness of this candidate diploid. w is determined by two modes of selection: genomic position-dependent (GP) selection and copy number-dependent (CN) selection.

Then, a random value between 0 and 1 is generated, and if it is smaller than w , this individual is accepted as a member of the next generation. The process is repeated until we have N diploids accepted.

GP selection works, depending on what part of the genome the TE is inserted in. As described above, a haploid genome has in total $p \times T$ sites in a functionally important region, where TEs are deleterious. The selection coefficient against TE insertion at the j th site ($j = \{1, 2, 3, \dots, T\}$) is denoted by $s_{GP(j)}$. If it is in a functionally important region, we assume that $s_{GP(j)}$ follows an exponential distribution with mean λ , while $s_{GP(j)} = 0$ if it is in a junk region. $s_{GP(j)}$'s for all T sites are determined before a simulation run and they are assumed to be constant over generations. The total selection coefficient against TE insertions for the k th individual is given by assuming the selection effect is additive,

$$s_{GP,k} = 1 - \prod_{j=1}^T (1 - z_{(j,k)} s_{GP(j)}), \quad (2)$$

where

$$z(j, k) = \begin{cases} 0 & \text{if TE is absent} \\ 1 & \text{if TE is heterozygous} \\ 2 & \text{if TE is homozygous} \end{cases}$$

at the j th site. This model is similar to the deleterious insertion model (Charlesworth 1991; Tsitrone *et al.* 1999).

CN selection works on the number of TEs (whether active or not) in the diploid individual, which is denoted by n . The coefficient of CN selection is denoted by s_{CN} , which we assume follows the function proposed by Charlesworth and Charlesworth (1983),

$$s_{CN,k} = \xi n_k^\tau, \quad (3)$$

where n_k is the number of TEs in the k th diploid individual, and ξ and τ (>1) are coefficients that specify the function. This function is considered to represent the situation where the deleterious effect of TEs is produced by ectopic recombination (e.g., Langley *et al.* 1988; Capy *et al.* 1996). As the number of TEs increases, the chance of ectopic recombination dramatically increases (Petrov *et al.* 2003), and this effect is characterized by ξ and τ . Then, adding the effects of the two modes of selection, the fitness of the k th individual is given by

$$w_k = (1 - s_{GP,k})(1 - s_{CN,k}). \quad (4)$$

If $w_k < 0$, we assume that the fitness is zero according to Le Rouzic *et al.* (2007).

Simulations

The parameter values used in the simulations are summarized in Table 1. Two values of the population sizes ($N = 500$ and 1000) were considered. It was assumed that a diploid genome has $2T = 4000$ sites, which TEs can occupy. These sites were distributed on the genome with equal lengths of intervals, and the recombination rate (c) between adjacent sites was set such that the population recombination parameter $\rho = 4Nc = 200$. This recombination rate is quite high so that TEs can behave almost independently. It was assumed that $p = 75\%$ of the

Table 1 Summary of the parameters used in this study

Symbol	Description	Parameter values
N	Diploid population size	500, 1000
ρ	Population recombination parameter ($4Nc$)	200
T	No. sites in a haploid genome	2000
p	Proportion of functionally important sites	0.75
λ	Mean selection coefficient against TEs at functionally important sites	1.0×10^{-4} , 1.0×10^{-3}
(τ, ξ)	Parameters to specify the intensity of CN selection	$(1.5, 1.0 \times 10^{-5})$, $(1.5, 1.0 \times 10^{-4})$, $(1.5, 1.0 \times 10^{-3})$
L	Sequence length of TE (bp)	300
θ	Population mutation parameter in TEs' DNA sequences ($4N\mu$)	0.01
ϕ	Relative rate of indels to point mutations	0.2
u_0	Maximum transposition rate	1.0×10^{-2}
ν	Excision rate	0, 1.0×10^{-6} , 1.0×10^{-4}
(α, β)	Parameters to characterize the activity function (Equation 1)	$(0.70, 6)$, $(0.75, 12)$, $(0.80, 24)$, $(0.85, 48)$

2000 sites were in functionally important regions, and we considered two values of the deleterious selection parameter, $\lambda = \{1.0 \times 10^{-4}, 1.0 \times 10^{-3}\}$, representing weak and strong GP selection cases, respectively. For CN selection, we fixed $\tau = 1.5$, which was the value used in many previous simulation works following the original one by Charlesworth and Charlesworth (1983). We used three values of $\xi = \{1.0 \times 10^{-5}, 1.0 \times 10^{-4}, 1.0 \times 10^{-3}\}$, representing weak, moderate, and strong CN selection cases.

The maximum rate of transposition was fixed to be $u_0 = 1.0 \times 10^{-2}$, which is roughly in agreement with both empirical estimations (Charlesworth *et al.* 1992; Nuzhdin and Mackay 1995; Suh *et al.* 1995; Maside *et al.* 2000) and those used in previous simulation studies (Morgan 2001; Navarro-Quezada and Schoen 2002; Le Rouzic and Capy 2006; Le Rouzic *et al.* 2007). We used three different rates of excision, $\nu = \{0, 1.0 \times 10^{-6}, 1.0 \times 10^{-4}\}$. We did not consider cases with $\nu > u_0$ because the population cannot maintain TEs with such an unrealistically high excision rate.

For point mutations that occur in TE sequences, we set the population mutation parameter $\theta = 4N\mu = 0.01$, which is within a typical range of nucleotide diversities in eukaryotes (Hartl and Clark 2006). The relative rate of indels to point mutations was assumed to be $\phi = 0.2$, which is according to the observation in polymorphism data in humans (Chen *et al.* 2009).

The two parameters (α and β) involved in Equation 1 were determined based on the human *Alu* data (Bennett *et al.* 2008). The data are shown in Figure 1; *Alu* elements of high identity to the consensus sequence have generally high activities, and the activity decreases with decreasing identity. We found that the solid line with $(\alpha, \beta) = (0.80, 24)$ in Figure 1 approximately agrees with the data, which was used in the simulation. In addition, we considered three pairs $(\alpha, \beta) = (0.70, 6)$, $(0.75, 12)$, and $(0.85, 48)$. The first two pairs represent very small and small degeneration effects of point mutations (*i.e.*, the activity is less sensitive to mutations than the *Alu*-fitted case represented by the solid line in Figure 1), while the last one $(0.85, 48)$ was for a large degeneration effect (*i.e.*, the activity dramatically drops with a small number of mutations).

Note that we here used the *Alu* data because they should somehow reflect the relationship between the TE's sequence divergence and the transpositional activity. Although an *Alu* element does not encode a protein, it forms a secondary structure at the RNA level, which should be related to the activity level. Therefore, as a proxy, we used the observed relationship between the *Alu* activity and sequence divergence in the entire *Alu* region to represent the selective constraint on nonsynonymous (or functionally important) sites in our simulation.

For each of all possible combinations of these parameters (listed in Table 1), we performed simulations for 50×10^3 generations. We assumed the initial condition for each replication such that a single copy of a TE appears in a single haploid individual in the population. With this setting, TEs completely disappear in the population in a very short time in a number of replications. Excluding such replications, we obtained data to 50×10^3 generations for five replications in all parameter sets. In each replication, we scored the following statistics every $N/5$ generations: $\bar{n}_{\text{total}}$, the total number of TEs per diploid individual; and $\bar{n}_{\text{act}}$, the total number of active copies ($a > 0$) per diploid individual.

These statistics were computed for all diploid individuals in the population and the averages were obtained. In addition, using 1% of representative individuals in the population, we obtained more detailed information every $N/5$ generations. For each TE, we scored genomic position and transpositional activity and nucleotide divergence from the original copy at synonymous and nonsynonymous sites, denoted by d_N and d_S , respectively.

Results

Average copy number

Throughout this work, the transposition rate was fixed, and we explored the effects of all other parameters. We show the results for $N = 500$ in the main text, and the results for $N = 1000$ are in Supporting Information, Figure S1, because the general patterns for $N = 1000$ are very similar to those for $N = 500$.

Figure 2 summarizes the effects of the parameters on the total copy number per individual ($\bar{n}_{\text{total}}$). The copy number was calculated every $N/5$ generations from the five replications of simulation data and the average is shown (the initial 10×10^3 generations were excluded because $\bar{n}_{\text{total}}$ is unstable and depends on the initial state). We found that the total copy number is strongly affected by the intensity of CN selection. $\bar{n}_{\text{total}}$ is very small when selection is strong ($\bar{n}_{\text{total}} < 30$), while for the cases of moderate and weak CN selection, $\bar{n}_{\text{total}}$ is relatively large, $\bar{n}_{\text{total}} \sim 400$ for moderate CN selection and $\bar{n}_{\text{total}} \sim 2000$ for weak CN selection. GP selection also decreases $\bar{n}_{\text{total}}$ but the effect is not visible in the comparison between the results for $\lambda = 10^{-4}$ and 10^{-3} (note that if λ was increased to 10^{-2} , in almost all simulation runs, TEs are virtually eliminated from the population soon after the initiations of the simulations). Note that the total number of sites per diploid genome is 4000, so that approximately half of them are occupied in the simulation runs with $\bar{n}_{\text{total}} \sim 2000$.

It seems that $\bar{n}_{\text{total}}$ is less affected by the decay rate of transpositional activity by point mutations [characterized by (α, β)]. In Figure 2, A–D are placed such that the degeneration effect increases from top to bottom [i.e., A (0.70, 6) $\rightarrow$ B (0.75, 12) $\rightarrow$ C (0.80, 24) $\rightarrow$ D (0.85, 48)]. The effect of the excision rate on the total number ($\bar{n}_{\text{total}}$) is also small.

Proportion of active copies

The effect of the parameters on q_{act} is also seen in Figure 2, in which the open boxes represent the number of active copies. q_{act} is also affected by the two modes of selection; as the selective pressure increases, q_{act} generally increases. This would be because highly active copies have more chances to survive especially under strong selective pressure.

There is quite a positive correlation between the excision rate and q_{act} , which can be understood as follows. The potential copy number of TEs is mainly determined by the strength of the two modes of selection, as demonstrated above. In the meantime, a number of turnovers of TEs occur. Excisions are the major event to cause turnovers. Each event makes room, in which one of the TEs in the genome can generate a daughter copy. It is obvious that a more active copy has more chance to contribute. Therefore, when turnovers occur frequently (i.e., the excision rate is high), highly active copies increase, resulting in a large q_{act} . This applies particularly when $\bar{n}_{\text{total}}$ is small or competition for empty sites among TEs is stiff due to strong selective pressure. Indeed, q_{act} is close to 1.0 in all cases with $\bar{n}_{\text{total}} < 400$ in Figure 2.

There is little effect of (α, β) on q_{act} . This is simply because we define active copies as those with nonzero activity ($a > 0$). α and β determine only how much point mutations decrease the activity (a), and, in any case, a very large number of mutations are required to make a completely nonactive copy. It is suggested that the contribution of point mutations to make a completely nonactive copy is much smaller than that of indels.

Figure 2 Total copy number of TEs ($\bar{n}_{\text{total}}$) when $(\alpha, \beta) = (0.7, 6)$, $(0.75, 12)$, $(0.8, 24)$, and $(0.85, 48)$. Active and nonactive copies are shown by open and shaded bars, respectively. The averages over five replications are shown.

Volatility of the copy number of TEs and their activities through generations

The changes of the copy number of TEs ($\bar{n}_{\text{total}}$) and their activities were investigated through the simulation runs. Some of the results for $N = 500$ are shown in Figure 3 and

Figure 3 Behavior of the copy number of TEs along simulations in the weak GP selection case. (A) $(\alpha, \beta) = (0.7, 6)$; (B) $(\alpha, \beta) = (0.80, 24)$. The levels of activity are shown in color from red (highly active) to purple (deactivated). It should be noted that when $(\alpha, \beta) = (0.80, 24)$, the effect of a single mutation on the activity level is larger so that the activity level decreases with a relatively large “gap”. As a consequence, we see a limited number of colors to represent the activity level. For example, one mutation can decrease the activity level from 1 to 0.54, which changes the color from red to green (skipping yellow).

Figure 4; see Figure S2 and Figure S3 for full results with $N = 500$ and 1000, respectively. Figure 3 shows the results of eight parameter sets in the weak GP selection case with relative large copy numbers. Let us first focus on Figure 3, A and B, left, that is, the moderate CN selection cases. The copy number seems to be stable at $\bar{n}_{\text{total}} \sim 400$ and there are a substantial number of active TEs. α and β are the major factors to affect the distribution of the activity level because they determine how quickly a TE’s activity is degenerated.

In the weak CN selection cases (Figure 3, A and B, right), the total copy number is very large (~ 2000). When the excision rate is zero, a large proportion of TEs are less active (shown in purple) because many TEs can survive for a long time under less pressure of direct excision. A low excision rate causes slow turnovers of TEs, resulting in a longer time to reach an equilibrium, as seen in Figure 3B, iv. Note that

the cases of no excision are unique in that, at each site, once a TE is fixed, it persists forever. That is, the ultimate state of each site is the fixation of TEs that will be eventually deactivated by point mutations; therefore, there will be no true equilibrium. This should be the reason why the proportion of deactivated copies gradually increases in the zero excision cases (especially Figure 3, A, iv and Figure B, iv).

Figure 4 shows the results for the strong CN selection cases. The total number of TEs is very small, and most of them are highly active as mentioned earlier. The major difference from the patterns in Figure 3 is that the total copy number is not stable and fluctuates greatly over generations.

Evolution of TE sequences

Our simulations allow exploring the relationship between the behavior of copy number and that of sequence evolution.

Figure 4 Behavior of the copy number of TEs along simulations in the strong CN selection cases. (A) $(\alpha, \beta) = (0.7, 6)$; (B) $(\alpha, \beta) = (0.80, 24)$. The levels of activity are shown in color from red (highly active) to purple (deactivated).

Through a single run of simulation, every $N/5$ generations, we constructed a neighbor-joining (NJ) tree by using all TEs in a 1% sample of the population. We also computed the d_N/d_S ratios and nucleotide distances from the original sequence. The results for the weak GP selection case are summarized in Figure 5. Figure 5, A and B, respectively shows the results for $(\alpha, \beta) = (0.70, 6)$ and $(0.80, 24)$ with zero excision rate. The results for two cases (moderate and weak CN selection cases) are shown, which correspond to Figure 3, A, iii and Figure B, iv, respectively. In each case, the NJ tree, the density distribution of d_N/d_S , the plot of d_N/d_S against the distance from the original sequence, and the frequency spectrum of TE insertions are presented. They are based on the data at $t = 50 \times 10^3$ generations because the patterns are quite stable over generations for these cases (but see below).

The patterns for $(\alpha, \beta) = (0.70, 6)$ and $(0.80, 24)$ are quite similar. In Figure 5, in the moderate CN selection case with $\bar{n}_{\text{total}} \sim 400$, most TEs are highly active (from orange to

red) and they form clusters in the NJ tree, which distribute around the original sequence (presented by an asterisk). There are some nonactive copies, and they locate quite far from the original sequence. This pattern is also seen in the density distribution of d_N/d_S and the plot of d_N/d_S against the distance from the original sequence. The distances for most TEs are <0.10 , while there are several TEs with distances >0.10 . The former are highly active copies as is obvious from the low d_N/d_S ratio (generally <0.2), while the latter are inactive copies with d_N/d_S ratios close to unity. The frequency spectra show that the frequencies of TEs at almost all sites are $<50\%$, indicating fixations of TEs are very rare and they are polymorphic.

In Figure 5, in the weak CN selection case with $\bar{n}_{\text{total}} \sim 2000$, most of the TEs are almost nonactive (from blue to purple), and they exhibit a star-shaped phylogeny, indicating these nonactive copies have independently accumulated point mutations. Active copies form clusters around the original

A $\alpha = 0.70, \beta = 6$

Moderate CN Selection

Weak CN Selection

B $\alpha = 0.80, \beta = 24$

Moderate CN Selection

Weak CN Selection

Figure 5 NJ tree, density distribution of d_N/d_S , plot of d_N/d_S as a function of the distance from the original sequence, and the frequency spectrum of TE insertions. (A) $(\alpha, \beta) = (0.70, 6)$; (B) $(\alpha, \beta) = (0.80, 24)$. Results for two parameter sets are shown to represent the strong and weak CN selection cases. These cases correspond to Figure 3, A and B, iii and iv. All results are based on the data at $t = 50 \times 10^3$ generations. In the NJ trees, the size and color of each circle represent the copy number and activity level, respectively. The original sequence is presented by an asterisk.

sequence. The density distribution of d_N/d_S is bimodal with two peaks; one is a sharp peak that consists of highly active copies with very small d_N/d_S . In contrast, the other peak consists of less active copies that broadly distribute around $d_N/d_S = 1$. As is shown in the plot of d_N/d_S against the distance, active copies make a cluster along the horizontal axis at $d_N/d_S = 0$, while nonactive copies distribute over a wide range. The frequency spectra show that the frequencies of TEs distribute in a wide range from very low (unique) to 1 (fixed). Fixed copies likely have very low activity (or are

completely deactivated), consistent with the observation of Alu (Batzer and Deininger 2002).

As shown in Figure 4, there are substantial fluctuations of the copy number when CN selection is strong. Such cases provide opportunities to explore how TE sequences evolve through fluctuations. In Figure 6, we show how NJ trees change along fluctuations. As an example, we use the data between $t = 10 \times 10^3$ and 23.5×10^3 generations in Figure 4A, ii, which are redrawn in Figure 6A. NJ trees are shown at five time points with stars in Figure 6A (a, b, c, d and e,

corresponding to $t = \{11.5, 14.5, 17, 19.5, 22\} \times 10^3$ generations, respectively).

From time a to c, the copy number of highly active copies increases dramatically, and Figure 6B shows that the most active copies presented by the red circle with an arrowhead in the tree at time a greatly contribute to this increase. The copy number increases to ~ 35 from time a to c (the corresponding copies are shown by a red circle with an arrowhead). At times d and e, highly active copies decrease, and then, with accumulation of point mutations, they are degenerated, resulting in several clusters presented by circles from orange to green. Thus, analyses of TE sequences can enhance our understanding of the evolutionary mechanism of TEs.

Discussion

By taking advantage of recent drastic improvement in computational power, we performed a large-scale forward simulation of TEs. The model basically follows the classic model by Charlesworth and Charlesworth (1983). In this work, we added a number of factors to their model, and the most novel part is that each TE has DNA sequence information that defines its transpositional activity. Almost all previous simulation studies have been performed with no DNA sequence information, except for very few works including that of Navarro-Quezada and Schoen (2002). They performed a simulation of TE sequences based on the model of Charlesworth and Charlesworth (1983) to explore the behavior of nucleotide changes at synonymous and nonsynonymous sites. Unfortunately, this sequence model has not been fully explored for a decade.

In addition to incorporating TE sequences, our modifications include that two modes of selection (*i.e.*, CN and GP selection) are incorporated and that the transpositional activity can be degenerated by mutations. We used a more realistic function to determine the transpositional activity, according to the human *Alu* data.

Furthermore, our simulation is different from that of Navarro-Quezada and Schoen (2002) in that we performed simulations on greater ranges for various parameters, including the genome size, the population size, and the number of simulated generations. Our extensive simulations allowed us to investigate the evolutionary dynamics of TE sequences not only at a specific time point but also through generations.

We first explored the effect of selection on the total copy numbers ($\bar{n}_{\text{total}}$) and the proportion of active copies (q_{act}). Our results show that the combination of the two modes of selection determines the number of TEs that can be maintained in the genome. When selection is weak, the genome can maintain a large number of TEs but the proportion of active copies should be small. In contrast, when selection is moderate, the genome can maintain a relatively small copy number of TEs, most of which are likely highly active. When the total copy number is restricted to be very low by strong selective pressure, the copy number is unstable and fluctuates over generations.

Figure 6 (A and B) Change of NJ trees in a cycle of fluctuation of the TE copy number. An example from Figure 4A, i is presented. NJ trees at (a) $t = 11.5 \times 10^3$, (b) 14.5×10^3 , (c) 17×10^3 , (d) 19.5×10^3 , and (e) 22×10^3 generation are shown. The size and color of each circle represent the copy number and activity level, respectively. The original sequence is presented by an asterisk.

We observed a clear trend that when the total copy number is smaller due to strong selection, more copies are active (Figure 2). This trade-off should be directly related to the competition among copies. When selection is strong, the

capacity to maintain TEs in the genome is limited. In such a situation, if some occupiable sites become available (mainly by excision), extensive competition among TEs is expected, and more active copies have more chance to occupy.

We also explored how DNA sequences of TEs evolve in the simulations. As we have shown in Figure 5 and Figure 6, the phylogeny of TE sequences could be informative to understand the dynamics of TE evolution. A general trend revealed by our simulations is that active copies form clusters located close to the original sequence. In contrast, non- and less active copies usually have long branches specific to each, exhibiting a star-shaped phylogeny. This is because they accumulate independent point mutations without increasing copy number. Such an accumulation causes a growth of the branch until the copy is removed from the genome. As a consequence, the average length of external branches to non- and less active copies reflects how often turnovers of TEs occur. We also showed dynamic change of phylogenetic trees of TEs when the copy number changes substantially (Figure 6).

The pattern of TE sequence evolution is similar to that of protein-coding genes under purifying selection (McAllister and Werren 1997; Navarro-Quezada and Schoen 2002). That is, the d_N/d_S ratio is low for active copies, and once the activity is reduced, it increases and exhibits saturation around 1.0. It is indicated that strong competitions among TEs result in selective pressure to maintain the function of transposition.

Thus, our simulation framework provides a useful tool to understand population genetics and molecular evolution of TEs, and further simulations will reveal more detailed behavior of TEs in genome evolution. It could also provide deep insights into the evolutionary dynamics of TEs in a genome, such as the formation of subfamilies and their competitions.

Acknowledgments

We thank two anonymous reviewers for their helpful comments. This work is in part supported by grants from the Japan Society for the Promotion of Science (JSPS) (to H.I.). T.E.K. is a JSPS predoctoral fellow.

Literature Cited

- Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne *et al.*, 2000 The genome sequence of *Drosophila melanogaster*. *Science* 287: 2185–2195.
- Batzler, M. A., and P. L. Deininger, 2002 Alu repeats and human genomic diversity. *Nat. Rev. Genet.* 3: 370–379.
- Bennett, E. A., H. Keller, R. E. Mills, S. Schmidt, J. V. Moran *et al.*, 2008 Active Alu retrotransposons in the human genomes. *Genome Res.* 18: 1875–1883.
- Brookfield, J. F. Y., 1986 A model for DNA sequence evolution within transposable element families. *Genetics* 112: 393–407.
- Capy, P., D. Anxolabéhère, and T. Langin, 1996 *Dynamics and Evolution of Transposable Elements (Molecular Biology Intelligence Unit)*, Ed. 1. Springer-Verlag, New York.
- Cerveau, N., S. Leclercq, E. Leroy, D. Bouchon, and R. Cordaux, 2011 Short- and long-term evolutionary dynamics of bacterial insertion sequences: insights from *Wolbachia* endosymbionts. *Genome Biol. Evol.* 3: 1175–1186.
- Charlesworth, B., 1991 Transposable elements in natural populations with a mixture of selected and neutral insertion sites. *Genet. Res.* 57: 127–134.
- Charlesworth, B., and D. Charlesworth, 1983 The population dynamics of transposable elements. *Genet. Res.* 42: 1–27.
- Charlesworth, B., and C. H. Langley, 1989 The population genetics of *Drosophila* transposable elements. *Annu. Rev. Genet.* 23: 251–287.
- Charlesworth, B., A. Lapid, and D. Canada, 1992 The distribution of transposable elements within and between chromosomes in a population of *Drosophila melanogaster*. I. Element frequencies and distribution. *Genet. Res.* 60: 103–114.
- Charlesworth, D., and B. Charlesworth, 1995 Transposable elements in inbreeding and outbreeding populations. *Genetics* 140: 415–417.
- Chen, J.-Q., Y. Wu, H. Yang, J. Bergelson, M. Kreitman *et al.*, 2009 Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria. *Mol. Biol. Evol.* 26: 1523–1531.
- Craig, N. L., R. Craigie, M. Gellert, and A. Lambowitz (Editors), 2002 *Mobile DNA II*, Ed. 2. American Society for Microbiology, Washington, DC.
- Deceliere, G., S. Charles, and C. Biémont, 2005 The dynamics of transposable elements in structured populations. *Genetics* 169: 467–474.
- Hartl, D. L., and A. G. Clark, 2006 *Principles of Population Genetics*, Ed. 4. Sinauer Associates, Sunderland, MA.
- Hickey, D. A., 1982 Selfish DNA: a sexually-transmitted nuclear parasites. *Genetics* 101: 519–531.
- Hudson, R., and N. Kaplan, 1986 On the divergence of members of a transposable element family. *J. Math. Biol.* 24: 207–215.
- Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody *et al.*, 2001 Initial sequencing and analysis of the human genome. *Nature* 409: 860–921.
- Langley, C. H., J. F. Y. Brookfield, and N. Kaplan, 1983 Transposable elements in Mendelian populations. I. A theory. *Genetics* 104: 457–471.
- Langley, C. H., E. Montgomery, R. Hudson, N. Kaplan, and B. Charlesworth, 1988 On the role of unequal exchange in the containment of transposable element copy number. *Genet. Res.* 52: 223–235.
- Le Rouzic, A., and P. Capy, 2006 Population genetics models of competition between transposable element subfamilies. *Genetics* 174: 785–793.
- Le Rouzic, A., T. S. Boutin, and P. Capy, 2007 Long-term evolution of transposable elements. *Proc. Natl. Acad. Sci. USA* 104: 19375–19380.
- Maside, X., S. Assimacopoulos, and B. Charlesworth, 2000 Rates of movement of transposable elements on the second chromosome of *Drosophila melanogaster*. *Genet. Res.* 75: 275–284.
- McAllister, B. F., and J. H. Werren, 1997 Phylogenetic analysis of a retrotransposon with implications for strong evolutionary constraints on reverse transcriptase. *Mol. Biol. Evol.* 14: 69–80.
- Morgan, M. T., 2001 Transposable element number in mixed mating populations. *Genet. Res.* 77: 261–275.
- Navarro-Quezada, A., and D. J. Schoen, 2002 Sequence evolution and copy number of Ty1-copia retrotransposons in diverse plant genomes. *Proc. Natl. Acad. Sci. USA* 99: 268–273.
- Nuzhdin, S. V., and T. F. Mackay, 1995 The genomic rate of transposable element movement in *Drosophila melanogaster*. *Mol. Biol. Evol.* 12: 180–181.
- Ohta, T., 1985 A model of duplicative transposition and gene conversion for repetitive DNA families. *Genetics* 110: 513–524.

- Petrov, D. A., Y. T. Aminetzach, J. C. Davis, D. Bensasson, and A. E. Hirsh, 2003 Size matters: non-LTR retrotransposable elements and ectopic recombination in *Drosophila*. *Mol. Biol. Evol.* 20: 880–892.
- Quesneville, H., and D. Anxolabéhère, 2001 Genetic algorithm-based model of evolutionary dynamics of class II transposable elements. *J. Theor. Biol.* 213: 21–30.
- Slatkin, M., 1985 Genetic differentiation of transposable elements under mutation and unbiased gene conversion. *Genetics* 110: 145–158.
- Suh, D. S., E. H. Choi, T. Yamazaki, and K. Harada, 1995 Studies on the transposition rates of mobile genetic elements in a natural population of *Drosophila melanogaster*. *Mol. Biol. Evol.* 12: 748–758.
- Tsitrone, A., S. Charles, and C. Biémont, 1999 Dynamics of transposable elements under the selection model. *Genet. Res.* 74: 159–164.
- Wright, S. I., and D. J. Schoen, 1999 Transposon dynamics and the breeding system. *Genetica* 107: 139–148.

Communicating editor: L. M. Wahl

GENETICS

Supporting Information

<http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.150292/-/DC1>

Population Genetics and Molecular Evolution of DNA Sequences in Transposable Elements. I. A Simulation Framework

T. E. Kijima and Hideki Innan

Figure S1 Total copy number of TEs ($\bar{n}_{total}$) for $N = 1,000$. See the legend of Figure 2 for details

Figure S2 Full simulation results on the behavior of the copy number of TEs along simulations for $N = 500$.

Figure S3 Full simulation results on the behavior of the copy number of TEs along simulations for $N = 1,000$.