

# Different Regulatory Mechanisms Underlie Similar Transposable Element Profiles in Pufferfish and Fruitflies

Daniel E. Neafsey, Justin P. Blumenstiel, and Daniel L. Hartl

Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts

Comparative analysis of recently sequenced eukaryotic genomes has uncovered extensive variation in transposable element (TE) abundance, diversity, and distribution. The TE profile in the sequenced pufferfish genomes is more similar to that of *Drosophila melanogaster* than to human or mouse, in that pufferfish TEs exhibit low overall abundance, high family diversity, and localization in the heterochromatin. It has been suggested that selection against the deleterious effects of ectopic recombination between TEs has structured the TE profile in *Drosophila* and pufferfish but not in humans. We test this hypothesis by measuring the sample frequency of 48 euchromatic TE insertions in the genome of the green spotted pufferfish (*Tetraodon nigroviridis*). We estimate the strength of selection acting on recent insertions by analyzing the site frequency spectrum using a maximum-likelihood approach. We show that in contrast to *Drosophila*, euchromatic TE insertions in *Tetraodon* are selectively neutral and that the low copy number and compartmentalized distribution of TEs in the *Tetraodon* genome must be caused by regulation by means other than purifying selection acting on recent insertions. Inference of regulatory processes governing TE profiles should take into account factors such as effective population size, incidence of inbreeding/outcrossing, and other species-specific traits.

## Introduction

Transposable elements (TEs) are virtually ubiquitous tenants of animal and plant genomes. Differences in the abundance, distribution, and diversity of TEs in recently sequenced eukaryotic genomes suggests that TEs have evolved a variety of “leases” with their diverse hosts. For example, retrotransposable elements, which propagate by inserting reversed-transcribed DNA copies of their RNA transcripts into the genome, exhibit dramatically different profiles in the human and *Drosophila melanogaster* genomes (Eickbush and Furano 2002). More than a million retrotransposon insertions constitute more than 30% of the human euchromatin (International Human Genome Sequencing Consortium 2001). Most of these insertions are ancient and fixed in the population and represent only three families of long terminal repeat (LTR) elements and three families of non-LTR elements. In contrast, there are approximately 1,400 retrotransposon insertions in the *Drosophila* genome, making up less than 5% of the euchromatic DNA (Bartolome, Maside, and Charlesworth 2002). This small number of insertions belies the great diversity of *Drosophila* retrotransposons. There are approximately 40 families of non-LTR elements and 20 families of LTR elements in *Drosophila*, with many families demonstrating evidence of recent activity in the form of dimorphic insertions (Nuzhdin 1999).

A large body of empirical and theoretical work exists on TE regulation in *Drosophila* (for review see Nuzhdin [1999]). Classical models of TE regulation posit selective control of TE copy number caused by the disruptive effects of insertions on local genes (Finnegan 1992), costs to the host resulting from TE transcription (Nuzhdin, Pasyukova, and Mackay 1996) and/or translation (Brookfield and Badge 1997), or an increased likelihood of ectopic recombination between nonhomologous insertions (Goldberg et al. 1983;

Langley et al. 1988; Montgomery et al. 1991), which can lead to chromosomal disruptions. All of these mechanisms likely contribute to control of TE copy number in *Drosophila*, but the importance of ectopic recombination was recently underscored by evidence of greater purifying selection opposing the fixation of long TE insertions relative to short ones, as well as greater selection against insertions belonging to families with high copy number in the genome (Petrov et al. 2003). Both of these characteristics would increase the likelihood of an insertion participating in an ectopic recombination event (Montgomery, Charlesworth, and Langley 1987; Dray and Gloor 1997).

The recently sequenced *Takifugu rubripes* (fugu) and *Tetraodon nigroviridis* pufferfish genomes exhibit TE profiles more similar to that of *Drosophila* than that of human in many respects. These similarities have naturally invited suggestions that ectopic recombination plays a greater role in structuring the TE profile in the pufferfish genomes than in the human genome (Volf et al. 2003; Furano, Duvernell, and Boissinot 2004). TE profiles in the sequenced pufferfish genomes fulfill two important predictions of a system regulated by selection against ectopic recombination. First, they exhibit high TE family diversity and low copy number within families, which limits the opportunity for any individual element to recombine with a nonhomologous insertion from the same family. Six families of non-LTR retrotransposons have been described in the fugu and *Tetraodon* genomes, as well as 15 families of LTR elements. Copy numbers of elements belonging to these families are at least an order of magnitude smaller than copy numbers of human TE families and range from eight to 1,670 per haploid genome (Volf et al. 2003). Second, as in *Drosophila* (Bartolome, Maside, and Charlesworth 2002), the TE distribution in the pufferfish genome is localized to areas that presumably undergo low rates of recombination, such as centromeric heterochromatin (Dasilva et al. 2002). A low rate of recombination reduces the likelihood that a local insertion will participate in an ectopic recombination event, presumably relaxing selection against TE insertions and permitting fixation in the population.

Key words: Transposable elements, population genetics, *Tetraodon*, ectopic recombination, *Drosophila*.

E-mail: neafsey@oeb.harvard.edu.

*Mol. Biol. Evol.* 21(12):2310–2318. 2004

doi:10.1093/molbev/msh243

Advance Access publication September 1, 2004

We tested whether an ectopic recombination model of TE regulation is sufficient to explain the TE distribution and diversity in *Tetraodon* by measuring the frequency of 48 recent euchromatic TE insertions in a sample of 24 noninbred fish. If deleterious ectopic recombination or some other insult to host fitness causes selection against TE insertions in regions of the pufferfish genome that undergo a normal rate of recombination, this selection should be evident in an insertion site frequency spectrum that is skewed towards low frequency variants. Alternatively, if the distribution and diversity of TEs in the pufferfish genome is determined by regulatory factors preceding insertion, then individual euchromatic TE insertions should be segregating and fixing according to neutral expectations. In contrast to *Drosophila*, we found evidence of abundant fixed euchromatic TE insertions in *Tetraodon*. Analysis of 36 of the most recent TE insertions failed to reject a neutral diffusion model of fixation. We suggest that selection opposing the fixation of insertions to prevent ectopic recombination does not structure the pufferfish TE profile and that host regulation of transposition rates may be a more important factor. Thus, the similar TE profiles in *Drosophila* and pufferfish may be generated by dissimilar regulatory mechanisms.

## Methods

### Samples

A sample of 24 *Tetraodon nigroviridis* pufferfish specimens from Thailand was acquired through a pet-trade importer. The exact collection locality is unknown, as is the collection locality of the fish used for genome sequencing (C. Fischer, personal communication). To attempt to establish whether the fish in our sample were collected from the same population as the fish used for genome sequencing, we PCR-amplified a 412-bp fragment of the highly variable mitochondrial D-loop region using primers DL-F (TCCTGGCATTGGTTTCCTAC) and DL-R2 (GGGGGTTTGCAGGATAATAAG). All DNA extractions for PCR template were performed on small (1 to 3 mm<sup>2</sup>) pieces of caudal fin tissue using NucleoSpin columns (BD Biotek), according to the manufacturer's protocol. PCR was carried out in 20 µl volumes containing 1 µl of template DNA, 0.1 µl taq polymerase (Roche), 2 µl 10× PCR buffer (Roche), and final concentrations of 50 µM for each dNTP and 0.5 µM for each primer. Amplifications were performed as follows: initial denaturation at 94°C for 90 s, followed by 35 cycles of denaturation at 94°C for 30 s, annealing at 54°C for 30 s, and extension at 72°C for 1 min. Amplification products were analyzed on 1.5 % agarose gels. Successful amplifications were sequenced in both directions using PCR primers with BigDye3 (ABI) on an Applied Biosystems 3100 sequencer.

Sequences were aligned by hand in BioEdit (Hall 1999) to the corresponding D-loop fragment in the genome assembly. We estimated the population genetics parameter  $\theta = 4N\mu$  using the  $\pi$  and  $\theta_w$  estimators, where  $N$  is the effective population size and  $\mu$  is the neutral mutation rate.  $\pi$  is calculated as the average number of pairwise differences between two random sequences (Tajima 1989), and  $\theta_w$  is derived from the number of segregating sites

(Watterson 1975). Population genetic analyses were conducted using DnaSP version 3.50 (Rozas and Rozas 1999). A haplotype network was constructed by hand to analyze divergence among haplotypes in our sample relative to the haplotype in the genome assembly.

### Selection of TE Insertions for Screening

We downloaded publicly available consensus/archetype sequences for four described families of *Tetraodon* retrotransposable elements. Families with a range in copy number size were chosen from the LTR and non-LTR groups. Copy number data, taken from Volff et al. (2003), were derived by Blast analyses (Altschul et al. 1990) against the publicly available whole-genome shotgun (WGS) sequence database. *Babar/Rex1* (accession number AJ312227; 3,528 bp) is a non-LTR element with the largest copy number (1,670) of any single retrotransposon family in the *Tetraodon* genome. *Rex3* (accession number AJ312226; 1,352 bp) is a non-LTR element with 690 copies in the genome. *Zebulon* (accession number AJ496734; 2,339 bp [Bouneau et al. 2003]) is a non-LTR element with 290 copies. Finally, *TnDIRS1* (accession number AF442732; 6,183 bp [Goodwin and Poulter 2001]) is an LTR element with an estimated 55 copies in the sequenced *Tetraodon* genome.

These representative TE sequences were used to Blast against the *Tetraodon nigroviridis* genome assembly version 6.0. All hits less than 2% divergent from the query sequence were retained, as well as a small selection of more divergent elements from each family. *Tetraodon* contigs containing hits of interest were used to Blast against the fugu genome assembly as a conservative test of their euchromatic identity. Insertions residing on *Tetraodon* contigs that unambiguously matched fugu scaffolds were retained for PCR screening. Nonmatching *Tetraodon* contigs were discarded as possibly heterochromatic.

Primers to detect the presence or absence of individual TE insertions were designed using Primer3 software (Supplementary Material online; Rozen and Skaletzky [2000]). When possible, primers were designed to anneal to genomic sequence flanking both sides of an insertion to provide codominant indication of allelic state. Presence or absence of an insertion was indicated in this case by long or short PCR products, respectively. The *Tetraodon* genome assembly is highly fragmented, however, and many contigs terminate with TE insertions because of the shotgun sequencing protocol employed. To screen these terminal TE insertions, one primer was designed to anneal to the TE sequence, and a complementary primer was designed to anneal to proximal genomic sequence. With these “overlapping” primers, fish that were heterozygous or homozygous for an insertion yielded identical PCR products, and fish that were homozygous for absence of an insertion failed to produce a PCR product. When a TE insertion screened with overlapping primers tested negatively in all samples, PCR screening was repeated with a redesigned second pair of overlapping primers to reduce the chance of mistaking PCR failure for an absence of the insertion from the sample. All PCR reactions were carried out with an annealing temperature of 58°C according to the protocol described above.

## Data Analysis

The sample frequency of individual TE insertions was directly determined from codominant screens that used flanking primer pairs. The sample frequency of insertions screened with dominant overlapping primer pairs was inferred using maximum likelihood. Under Hardy-Weinberg equilibrium, we can assume that  $x \approx 2pq + q^2$ , where  $x$  is the proportion of a large sample testing positive for a dominant marker,  $p$  is the population frequency of the marker, and  $q = 1 - p$ . Substituting the observed frequency of an insertion for  $x$  and solving for  $p$  yields the most likely sample frequency of the insertion. The 95% confidence intervals of the frequency estimates were determined by setting the definite integral of the density function for population frequency equal to 0.025 and 0.975, and then solving for the correct integration boundary. The density function  $F[p]$  for population frequency ( $p$ ) given the number of dominant markers ( $i$ ) observed in a sample of 24 individuals is as follows:

$$F[p] = \frac{\binom{24}{i} (2p - p^2)^i (1 - p)^{(24-i)}}{\int_0^1 \left( \binom{24}{i} (2p - p^2)^i (1 - p)^{(24-i)} \right) dp}$$

In addition to computing the sample frequency for each insertion, we calculated the percent divergence of each insertion from the consensus/archetype sequence. We ignored gaps and only counted nucleotide substitutions, as these occur at a faster rate and provide a more reliable indication of the age of an insertion. The youngest insertions are expected to exhibit the least divergence. We considered TE insertions less than 2% divergent to be “recent” and useful for estimation of selection intensity.

Our rationale for selection of this subset derives from a coalescent-based estimation of the probability that a mutation that occurred at time  $t$  in the past and has been identified in a single genome sample is represented in  $i$  copies in a sample of size  $n$ . This probability is a function of (1) the probability distribution of the number of ancestors  $j$  at time  $t$  of a sample of size  $n$  and (2) the probability that the single lineage that has the mutation is represented  $i$  times in the sample given  $j$  ancestors. For example, the probability that the mutant is represented in all samples is 1 if the number of ancestors  $j$  at time  $t$  is 1. If  $j \geq 2$  at time  $t$ , the probability that the mutant is represented in all samples (fixed) is 0. The probability of  $j$  ancestors to a sample of size  $n$  being present at time  $t$  is given by

$$P(j | t, n) = \sum_{k=j}^n \rho_k(t) \frac{(2k-1)(-1)^{k-j} j_{(k-1)} n_{[k]}}{j!(k-j)!n_{(k)}},$$

$$2 \leq j \leq n$$

$$P(j | t, n) = 1 - \sum_{k=2}^n \rho_k(t) \frac{(2k-1)(-1)^k n_{[k]}}{n_{(k)}}, \quad j = 1 \quad (1)$$

Where  $\rho_k(t) = e^{-k(k-1)t/2}$ ,  $a_{(j)} = a(a+1) \cdots (a+j-1)$ ,  $a_{[j]} = a(a-1) \cdots (a-j+1)$  and  $t$  is in units of 2N generations (Tavaré 1984).

Given  $j$  ancestors, the probability that a single lineage is represented by  $i$  copies in a sample of size  $n$  is given by

$$P(i | j, n) = \frac{(j-1)(n-i-1)!(n-j)!}{(n-1)!(n-j-i+1)!} \quad (2)$$

(Feller 1957; Felsenstein 1992). As noted by Sherry et al. (1997), the probability distribution of  $i$  is uniform when  $j=2$  for  $i=1$  to  $n-1$  and is equal to 1 for  $i=1$  when  $j=n$ .

The probability that a mutation that occurred at time  $t$  is represented in  $i$  out of  $n$  samples is simply the probability of  $i$  given  $j$  ancestors, summed over all possible  $j$ . An assumption of this model is that there is no back mutation at this time scale. Using equations (1) and (2):

$$P(i | t, n) = \sum_{j=1}^n P(j | t, n) P(i | j, n) \quad (3)$$

From this function, 90% of TEs that inserted 6.6N generations ago and are present in the genome sequence will be fixed in a sample size of 48, and 95% of TEs that inserted 8.1N generations ago will be fixed. *Tetraodon* is estimated to have an effective population size of approximately  $10^5$  from the incidence of polymorphisms observed in the genome sequence (H. Roest Crolius, personal communication). If we assume the average vertebrate neutral mutation rate of  $2.3 \times 10^{-8}$  /bp/generation (Drake et al. 1998), then a TE sequence will be expected to have diverged by approximately 1.5% after 6.6N generations and 1.9% after 8.1N generations. Consideration of only those elements less than 2% divergent for selection analysis should, therefore, capture virtually all insertions still in a dynamic, dimorphic state without biasing the sample with an excessive amount of older, fixed insertions. A more stringent analysis of only those elements less than 1.5% divergent should still capture most dimorphic insertions and include even fewer fixed insertions.

We analyzed these two subsets of recent insertion sites for the impact of natural selection using a sojourn time density function based on a diffusion approximation (Ewens 1979; Nagylaki 1974) (see Petrov et al. [2003] for a detailed description of this type of analysis). Briefly, this method computes the likelihood of observing a set of site frequencies given a selection coefficient ( $s$ ), effective population size ( $N$ ), and a dominance coefficient ( $h$ ). The maximum-likelihood estimate of  $s$  is determined by maximizing the probability of the data given fixed parameters for  $N$  and  $h$ . This method assumes an infinite number of possible insertion sites in the genome, a large number of elements in the population, and a state of transposition-selection equilibrium, and it accounts for ascertainment bias inherent in screening a population sample for insertions observed in a sequenced genome. We conducted our analyses assuming semi-dominance ( $h = 0.5$ ) and complete dominance ( $h = 1$ ).

## Results

### D-Loop Polymorphism Analysis

A haplotype network of mitochondrial D-loop sequences (GenBank accession numbers AY599628 to

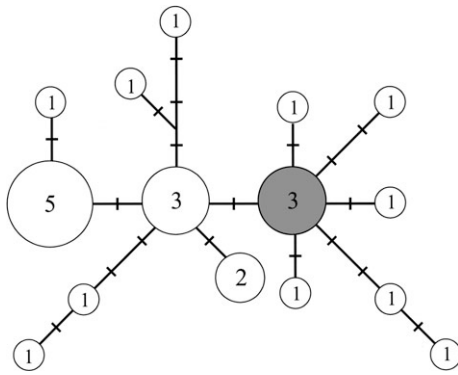


FIG. 1.—Haplotype network constructed from 430 bp of mitochondrial D-loop sequence from 24 *T. nigroviridis* specimens. Numbers inside circles indicate haplotype count. Tick marks on connecting bars indicate single-nucleotide substitutions. Shaded haplotype is identical to the haplotype of the specimen used for genome sequencing.

AY599651) from our sample is illustrated in figure 1. Tick marks on connecting bars represent one nucleotide substitution. The network shows no evidence of deep vicariance, with three clusters of the most common haplotypes separated by only a single substitution. This indicates that the samples were likely collected from one population, rather than two or more vicariant populations. The greatest span between any two haplotypes is seven substitutions. The D-loop haplotype of the fish used for genome sequencing is identical to a haplotype present in three copies in our sample, indicating it is not inappropriate for us to consider our sample and the sequenced fish as members of the same population. We observed 16 segregating sites in our sample. The population genetics test Tajima's  $D$  statistic, which detects departure from equilibrium conditions using differences in the  $\pi$  and  $\theta_w$  estimators of  $\theta$  (Tajima 1989), was not significant ( $D = -1.41$ ;  $P > 0.10$ ). This further establishes the suitability of our sample for measuring the intensity of selection on TE insertions.

#### Sample Frequency of TE Insertions

Table 1 lists the sample frequencies of 48 TE insertions screened with primers in a dominant or codominant fashion. We observed 11 TE insertions segregating at intermediate frequencies. All TE insertions exhibiting greater than 1.5% sequence divergence from the consensus/archetype sequence were fixed in the sample. This is in sharp contrast to *Drosophila*, where fixed TEs in the euchromatin are uncommon (Nuzhdin 1999), except in regions of low recombination, such as the fourth chromosome (Bartolome and Maside 2004). Of the set of 36 "recent" TEs exhibiting less than 2% divergence, 13 were fixed and 12 were undetected in our sample. For these recent insertions, we observed no significant correlation between insertion age and frequency (figure 2; Spearman's  $r_s = -0.199$ ; 1-tailed  $P > 0.20$ ). There was no significant difference in site frequency spectra between the four TE families (Kruskal-Wallis rank test,  $P < 0.812$ ).

Five of the TE insertions determined to be dimorphic were detected with codominant flanking primers. We tested the proportion of samples heterozygous and

homozygous for these insertions for conformity to Hardy-Weinberg expectations. All fit Hardy-Weinberg expectations except for the *TnDIRSI-3* insertion ( $\chi^2$  test,  $P < 0.01$ ). In our sample, 19 fish were heterozygous for the *TnDIRSI-3* insertion, three fish were homozygous for the insertion, and two fish were homozygous for absence of the insertion. The significant excess of fish heterozygous for the *TnDIRSI-3* insertion raises the possibility of balancing selection at this locus caused by overdominance or some other selective mechanism. The *TnDIRSI-3* insertion resides in an intron of the *T-complex protein 1* gene, which encodes the CCT- $\alpha$  subunit of the group II chaperonins. Interestingly, this gene is extremely polymorphic in zebrafish, with two classes of haplotypes estimated to have diverged as long as 3.5 MYA (Takami et al. 2000). We performed analyses of selection intensity on data sets including and excluding the *TnDIRSI-3* insertion to prevent bias caused by a possibly unique selection mechanism operating at this locus.

The large proportion of recent TE insertions that are fixed in our sample (13 out of 36) could be explained as a result of one of more recent population bottlenecks, rather than genetic drift or selection under equilibrium conditions. The high level of polymorphism observed at the mitochondrial D-loop locus could be masking a recent bottleneck, because genetic diversity at this locus is expected to rebound more quickly from such an event than a nuclear locus caused by a higher mutation rate. We believe that population bottlenecks are unlikely to explain the fixation of recent TE insertions for two reasons. First, all 12 of the TEs we screened that were more than 2% divergent from the consensus were fixed in our sample, indicating that the fixation of TEs in the euchromatin of *Tetraodon* is likely a continuous process rather than a purely recent phenomenon. Second, serial bottlenecks are unlikely to explain the historic fixation of TEs in this genome because the genome-wide effective population size of *Tetraodon* is estimated to be approximately  $10^5$  (H. Roest Crollius, personal communication), which is quite large by vertebrate standards (Lynch and Conery 2003). Such a large effective population size would be unlikely to be maintained in a vertebrate population subject to frequent expansion and contraction.

#### Estimation of Selection Intensity

We performed our analyses using 36 "recent" insertions with less than 2% divergence from the consensus/archetype sequence as well as a smaller subset of 29 insertions exhibiting less than 1.5% divergence. Figure 3 illustrates the likelihood plot of each data set given  $N_s$  according to the sojourn time-density function, with  $h = 0.5$  and *TnDIRSI-3* included. Setting  $h$  to 1 for a completely dominant selective model or excluding *TnDIRSI-3* did not change our conclusions. For both data sets,  $|N_s|$  at the likelihood peak is less than 1 (2%:  $N_s = -0.152$ ; 1.5%:  $N_s = 0.148$ ), indicating that the fate of TE insertions is determined primarily by drift and not natural selection. The maximum-likelihood estimate of  $N_s$  does not differ significantly from 0 for either data set (likelihood ratio test; 2%,  $P < 0.823$ ; 1.5%,  $P < 0.884$ ).

**Table 1**  
**Transposable Element Insertions Screened in a *T. nigroviridis* Population Sample**

Name	Contig	Size (bp)	%Divergence Relative to Consensus	Sample Frequency	Expected Population Frequency	(95% CI) <sup>a</sup>
<i>Babar/Rex1-1</i>	FS_CONTIG_23043_1	590	2.71	24/24	100	(72.7–100)
<i>Babar/Rex1-2</i>	FS_CONTIG_3573_1	578	1.21	24/24	100	(72.7–100)
<i>Babar/Rex1-3</i>	FS_CONTIG_20417_1	628	3.82	24/24	100	(72.7–100)
<i>Babar/Rex1-4</i>	FS_CONTIG_395_4	569	1.93	0/48	0	
<i>Babar/Rex1-5</i>	FS_CONTIG_15141_1	698	5.30	24/24	100	(72.7–100)
<i>Babar/Rex1-6</i>	FS_CONTIG_12_7	528	3.03	24/24	100	(72.7–100)
<i>Babar/Rex1-7</i>	FS_CONTIG_5716_1	470	1.70	0/24	0	
<i>Babar/Rex1-8</i>	FS_CONTIG_4238_2	315	0.95	0/48	0	
<i>Babar/Rex1-9</i>	FS_CONTIG_5528_2	281	0.71	0/48	0	
<i>Babar/Rex1-10</i>	FS_CONTIG_23363_1	268	0.37	17/48	35.4	
<i>Babar/Rex1-11</i>	FS_CONTIG_660_1	261	0.77	43/48	89.6	
<i>Babar/Rex1-12</i>	FS_CONTIG_12406_1	441	1.81	24/24	100	(72.7–100)
<i>Babar/Rex1-13</i>	FS_CONTIG_47266_1	423	0.95	24/24	100	(72.7–100)
<i>Babar/Rex1-14</i>	FS_CONTIG_4089_2	362	1.38	24/24	100	(72.7–100)
<i>Babar/Rex1-15</i>	FS_CONTIG_10607_1	303	1.32	0/48	0	
<i>Babar/Rex1-16</i>	FS_CONTIG_4286_1	289	1.38	24/48	50	
<i>Babar/Rex1-17</i>	FS_CONTIG_4179_1	290	1.72	0/48	0	
<i>Babar/Rex1-18</i>	FS_CONTIG_8440_1	429	0.70	24/24	100	(72.7–100)
<i>Babar/Rex1-19</i>	FS_CONTIG_49209_1	357	0.84	22/24	71.2	(51.3–86.7)
<i>Babar/Rex1-20</i>	FS_CONTIG_12152_1	38	1.18	24/24	100	(72.7–100)
<i>Rex3-4</i>	FS_CONTIG_10926_2	587	3.24	24/2	100	(72.7–100)
<i>Rex3-5</i>	WI_TRUEWICONTIG_22737_1	351	0.57	24/24	100	(72.7–100)
<i>Rex3-6</i>	FS_CONTIG_10007_1	336	0.60	4/24	8.3	(3.6–20.5)
<i>Rex3-7</i>	FS_CONTIG_5852_3	343	1.46	0/48	0	
<i>Rex3-8</i>	WI_CONTIG_291297_2	237	1.69	48/48	100	
<i>Rex3-9</i>	FS_CONTIG_29055_1	379	0.53	22/24	71.1	(51.3–86.7)
<i>Rex3-10</i>	WI_TRUEWICONTIG_129661_1	341	0.59	24/24	100	(72.7–100)
<i>Rex3-11</i>	FS_CONTIG_29612_1	430	1.86	24/24	100	(72.7–100)
<i>Zebulon-1</i>	FS_CONTIG_2430_4	513	0.97	0/48	0	
<i>Zebulon-2</i>	FS_CONTIG_18443_1	622	4.50	24/24	100	(72.7–100)
<i>Zebulon-3</i>	FS_CONTIG_2173_1	538	2.97	24/24	100	(72.7–100)
<i>Zebulon-4</i>	FS_CONTIG_36030_1	474	0.84	0/48	0	
<i>Zebulon-5</i>	FS_CONTIG_168_3	720	7.64	24/24	100	(72.7–100)
<i>Zebulon-6</i>	FS_CONTIG_1270_1	428	0.23	24/24	100	(72.7–100)
<i>Zebulon-7</i>	WI_TRUEWICONTIG_56422_1	431	0.93	24/24	100	(72.7–100)
<i>Zebulon-8</i>	FS_CONTIG_50368_1	495	3.43	24/24	100	(72.7–100)
<i>Zebulon-9</i>	FS_CONTIG_26949_2	401	1.00	1/24	2	(0.50–11.0)
<i>Zebulon-10</i>	FS_CONTIG_12780_1	324	0.00	2/24	4.2	(1.30–12.9)
<i>Zebulon-11</i>	FS_CONTIG_8768_2	299	0.67	20/24	50	(33.9–66.8)
<i>TnDIRS1-1</i>	FS_CONTIG_9470_2	219	0.91	48/48	100	
<i>TnDIRS1-2</i>	FS_CONTIG_1424_2	1105	12.7	24/24	100	(72.7–100)
<i>TnDIRS1-3</i>	FS_CONTIG_352_5	90	1.11	25/48	52.1	
<i>TnDIRS1-4</i>	FS_CONTIG_632_1	89	7.87	48/48	100	
<i>TnDIRS1-5</i>	FS_CONTIG_4985_1	123	12.2	48/48	100	

<sup>a</sup> Confidence interval (CI) not estimated for codominant screens.

## Discussion

Similarities between the TE profiles of *Drosophila melanogaster* and *Tetraodon* do not extend to the site frequency spectrum of recent insertions. Relative to *Drosophila*, TE insertions that have fixed or are segregating at intermediate to high frequencies are common in the *Tetraodon* euchromatin. This finding is incompatible with a model of TE copy number regulation in *Tetraodon* dependent on removal of individual insertions by natural selection caused by deleterious effects of ectopic recombination. Further, this result implies that the concentration of TEs in the heterochromatic regions of the *Tetraodon* genome (Dasilva et al. 2002) may be a result of insertion site preferences or recombinational excision rather than selective pruning of insertions in the euchromatin.

It should not be surprising that similarities between the TE profiles of organisms with very different biology

may arise by different means. TEs in the *Arabidopsis* genome also exhibit notable similarities to the case in *Drosophila*: high family diversity, low copy number within families, and low overall abundance (Surzycki and Belknap 1999; Le et al. 2000). The *Arabidopsis* TE profile differs from that of *Drosophila*, however, in that TE density is not inversely correlated with recombination rate, despite an accumulation of TEs near centromeres. Rather, TE density in *Arabidopsis* negatively correlates with gene density, suggesting that disruption of local genes rather than ectopic recombination fuels natural selection against TEs in this organism (Wright, Agrawal, and Bureau 2003). The high homozygosity of self-pollinating plants such as *Arabidopsis* may obviate a selective response against TEs driven by ectopic recombination, as there is evidence that heterozygosity may be an important factor promoting ectopic exchange (Charlesworth and Charlesworth 1995; Morgan 2001; Wright et al. 2001).

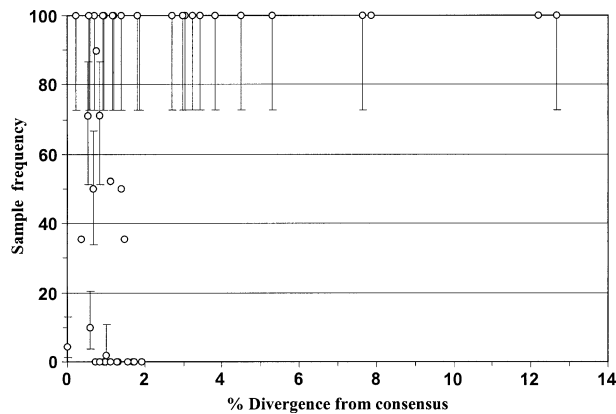


FIG. 2.—Relationship between element age, as measured by percent divergence from consensus sequence, versus sample frequency for 48 TE insertions. Vertical error bars surround points screened with dominant primer pairs and indicate the 95% confidence interval of the expected sample frequency.

Although TEs are much less prevalent in the *Tetraodon* genome than other vertebrate genomes, the neutral site frequency spectrum of recent *Tetraodon* TE insertions appears to be a common vertebrate feature. TEs have been observed segregating at intermediate to high frequencies in the genomes of pupfish (genus *Cyprinodon* [Duvernell and Turner 1999]), salmonids (Takasaki et al. 1997), and humans (Batzer and Deininger 2002). One feature most vertebrates share in common regardless of their TE load is a smaller effective population size ( $10^4$  to  $10^5$  [Lynch and Conery 2003]). The purifying selection against TE insertions that has been repeatedly observed in *Drosophila* (Montgomery and Langley 1983; Charlesworth and Lapid 1989; Petrov et al. 2003) may be contingent upon the comparatively large effective population size ( $10^5$  to  $10^6$ ) of this insect (Nuzhdin 1999).

Several reviews have recently suggested that there may be a reduced susceptibility to ectopic recombination during meiosis in humans relative to *Drosophila* (Eickbush and Furano 2002; Furano, Duvernell, and Boissinot 2004) or pufferfish (Volf et al. 2003) because of the abundance of TEs in human euchromatin. Comparative experimental data to test this hypothesis unfortunately remain lacking because of the difficulty of directly measuring the incidence of rare ectopic crossovers. Most studies of ectopic recombination rely on measurements of mitotic gene conversion in cell culture (e.g., Richard et al. 1994), characterization of breakpoints at deletion hotspots (McNaughton et al. 1998), or other undemonstrated correlates of meiotic ectopic crossing over. Although it is true that the tens of thousands of *L1* and *Alu* sequences in the human genome may offer many more opportunities for ectopic recombination than do the comparatively few TEs in *Drosophila* or pufferfish, prospects for unequal exchanges between TEs are likely to be much smaller than might be implied by the sheer number of insertions. High sequence homology and sequence length have been demonstrated to be important requirements for ectopic recombination in mammals (Metzenberg et al. 1991; Cooper, Schimenti, and Schimenti 1998), decreasing the likelihood that ancient TE insertions that have been subject to extensive mutations would

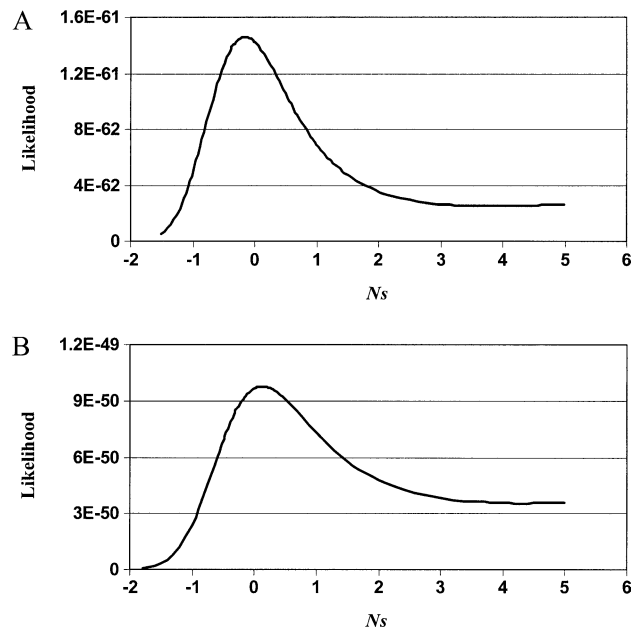


FIG. 3.—Likelihood plot of  $Ns$  for subsets of TE insertions less than 2% (A) and less than 1.5% (B) divergent from the consensus/archetype sequences. For both data subsets the likelihood of  $Ns = 0$  does not differ significantly from the local peak likelihood (likelihood ratio test; 2%,  $P < 0.823$ ; 1.5%,  $P < 0.884$ ), indicating that the TE insertions are not subject to natural selection.

participate in ectopic recombination events. Although the human genome still certainly harbors a greater number of TEs than *Drosophila* with sufficient homology to participate in an ectopic recombination event, the likelihood of a euchromatic human TE insertion undergoing intralocus recombination (ectopic or not) is 30% less than a *Drosophila* insertion of equal length, because of differences in the ratio between the physical and genetic maps of these species (*Drosophila* = 603 kb/cM [von Wettstein 1984]; human = 903 kb/cM [Morton 1991]). This would further diminish the need to infer a reduced susceptibility to ectopic recombination in humans, assuming a positive correlation between the rate of normal recombination and the risk of ectopic recombination.

Medical evidence indicates that the lack of a detectable selective response against individual TE insertions in vertebrate genomes is not caused by an absence of ectopic recombination. Deininger and Batzer (1999) have documented 33 cases of germline genetic disease and 16 cases of somatic cancer caused by ectopic recombination between short ( $\sim 300$  bp) *Alu* elements in humans. Ectopic homologous recombination between *L1* elements has also been implicated in several cases of human disease (Burwinkel and Kilimann 1998; Segal et al. 1999).

If most TE insertions are free to potentially fix by genetic drift in vertebrate genomes despite weakly deleterious fitness effects, the dynamic equilibrium of TE copy number must be maintained by regulation at a higher level (preinsertion). Self-regulation of hybrid-dysgenic DNA transposons has been well characterized in several cases (Bucheton et al. 1976; Pelisson and Bregliano 1987; Hartl, Lohe, and Lozovskaya 1997), but self-regulation of transposition rates is not expected to occur in retroelement

systems where the deleterious consequences of transposition cannot be linked to the mother copy (Charlesworth and Langley 1986).

Several mechanisms of host regulation of transposition have been suggested. Cosuppression is proposed to induce transcriptional and posttranscriptional regulation of TEs through small interfering RNAs (siRNA) and has been observed in *S. cerevisiae* (Jiang 2002; Garfinkel et al. 2003), *Drosophila* (Jensen, Gassama, and Heidmann 1999), and *C. elegans* (Sijen and Plasterk 2003). Methylation has been suggested to suppress TE activity (Bestor 2003), but the details of this mechanism remain unclear. Demethylation has been correlated with increased expression of retrotransposons in mouse (Walsh, Chaillet, and Bestor 1998) and increased transposition in interspecific wallaby hybrids (O'Neill, O'Neill, and Graves 1998). Relative to autonomous DNA transposons, retrotransposons are thought to be particularly susceptible to host regulation of transposition, given their reliance on host machinery for replication. In support of this, Boissinot and Furano (2001) have documented adaptive evolution in mammalian L1 retroelements at the coiled-coil domain, which mediates protein-protein interactions.

It might be suggested that ectopic recombination provides the selective force moderating transposition rates in the pufferfish genomes. For an organism with a small effective population size, regulation of TEs caused by any of the deleterious consequences of their activity might be more likely to take place preinsertion rather than postinsertion. This is because the deleterious consequences of a multitude of TE insertions can be summed to generate a fitness decrement large enough to incur a selective regulatory response in a small population. Ectopic recombination may be mediating selection at this level in *Tetraodon*, but its significance is likely diluted or superseded by selection against other deleterious consequences of TEs. Evidence of selection opposing the fixation of full-length L1 retroelements in the human genome does exist in the form of discrepancies in TE abundance between autosomes and sex chromosomes, but this selection can be attributed to the deleterious consequences of retrotransposition in the germline (Boissinot, Entezam, and Furano 2001). For truncated human L1 insertions less than 500 bp in length, similar to those profiled in this experiment, there are no data to suggest fixation does not come about according to neutral expectations. It may be impossible to dissect the relative importance of the various deleterious consequences of TEs to their regulation in organisms with small effective population sizes or even to distinguish selective modulation of TE load from neutral expansions or contractions in TE copy numbers.

Although much has been learned from studying the regulation of TEs in *Drosophila melanogaster*, the analytical techniques devised in this service may ultimately be more portable to other model organisms than any specific conclusions about regulatory processes. Indeed, the mode of TE regulation employed in *D. melanogaster* may not even be shared by all members of its genus. *Drosophila algonquin* has a similarly small genome as *D. melanogaster*, but its TE profile is primarily composed of only four dominant families with copy numbers higher than are

generally observed in *D. melanogaster* (Hey 1989). Further, most insertions were found to be segregating at middle or high frequencies in a population sample, rather than as rare variants. None of these features suggest that ectopic recombination influences the TE profile in *D. algonquin*. A better understanding of the range of deleterious effects of TEs, when integrated with information about life history and effective population size, should permit elucidation of the forces structuring the TE profile in the increasing number of sequenced eukaryotic genomes.

## Acknowledgments

We thank David Haig, John Wakeley, and two anonymous reviewers for helpful comments on the manuscript. We thank Aaron Hirsh for assistance with the analysis of selection coefficients. This project was funded by a student dissertation grant to D.N. from the Department of Organismic and Evolutionary Biology at Harvard University and a National Science Foundation Graduate Fellowship.

## Literature Cited

- Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215**:403–410.
- Bartolome, C., and X. Maside. 2004. The lack of recombination drives the fixation of transposable elements on the fourth chromosome of *Drosophila melanogaster*. *Genet. Res. Camb.* **83**:91–100.
- Bartolome, C., X. Maside, and B. Charlesworth. 2002. On the abundance and distribution of transposable elements in the genome of *Drosophila melanogaster*. *Mol. Biol. Evol.* **19**:926–937.
- Batzer, M. A., and P. L. Deininger. 2002. *Alu* repeats and human genomic diversity. *Nat. Rev. Genet.* **3**:370–379.
- Bestor, T. H. 2003. Cytosine methylation mediates sexual conflict. *Trends Genet.* **19**:185–190.
- Boissinot, S., A. Entezam, and A. V. Furano. 2001. Selection against deleterious LINE-1-containing loci in the human lineage. *Mol. Biol. Evol.* **18**:926–935.
- Boissinot, S., and A. V. Furano. 2001. Adaptive evolution in LINE-1 retrotransposons. *Mol. Biol. Evol.* **18**:2186–2194.
- Bouneau, L., C. Fischer, C. Ozouf-Costaz, A. Froschauer, O. Jaillon, J. P. Coutanceau, C. Korting, J. Weissenbach, A. Bernot, and J. N. Volff. 2003. An active non-LTR retrotransposon with tandem structure in the compact genome of the pufferfish *Tetraodon nigroviridis*. *Genome Res.* **13**:1686–1695.
- Brookfield, J. F., and R. M. Badge. 1997. Population genetics models of transposable elements. *Genetica* **100**:281–294.
- Bucheton, A., J. M. Lavigne, G. Picard, and P. L'Heritier. 1976. Non-mendelian female sterility in *Drosophila melanogaster*: quantitative variations in the efficiency of inducer and reactive strains. *Heredity* **36**:305–314.
- Burwinkel, B., and M. W. Kilimann. 1998. Unequal homologous recombination between LINE-1 elements as a mutational mechanism in human genetic disease. *J. Mol. Biol.* **277**: 513–517.
- Cooper, D. M., K. J. Schimenti, and J. C. Schimenti. 1998. Factors affecting ectopic gene conversion in mice. *Mamm. Genome* **9**:355–360.

- Charlesworth, D., and B. Charlesworth. 1995. Transposable elements in inbreeding and outbreeding populations. *Genetics* **140**:415–417.
- Charlesworth, B., and C. H. Langley. 1986. The evolution of self-regulated transposition of transposable elements. *Genetics* **112**:359–383.
- Charlesworth, B., and A. Lapid. 1989. A study of ten families of transposable elements on X chromosomes from a population of *Drosophila melanogaster*. *Genet. Res.* **54**:113–125.
- Dasilva, C., H. Hadji, C. Ozouf-Costaz, S. Nicaud, O. Jaillon, J. Weissenbach, and H. Roest Crolius. 2002. Remarkable compartmentalization of transposable elements and pseudogenes in the heterochromatin of the *Tetraodon nigroviridis* genome. *Proc. Natl. Acad. Sci. USA* **99**:13636–13641.
- Deininger, P. L., and M. A. Batzer. 1999. *Alu* repeats and human disease. *Mol. Genet. Metab.* **67**:183–193.
- Drake, J. W., B. Charlesworth, D. Charlesworth, and J. F. Crow. 1998. Rates of spontaneous mutation. *Genetics* **148**:1667–1686.
- Dray, T., and G. B. Gloor. 1997. Homology requirements for targeting heterologous sequences during *P*-induced gap repair in *Drosophila melanogaster*. *Genetics* **147**:689–699.
- Duvernell, D. D., and B. J. Turner. 1999. Variation and divergence of Death Valley pupfish populations at retrotransposon-defined loci. *Mol. Biol. Evol.* **16**:363–371.
- Eickbush, T. H., and A. V. Furano. 2002. Fruit flies and humans respond differently to retrotransposons. *Curr. Opin. Genet. Dev.* **12**:669–674.
- Ewens, W. H. 1979. *Mathematical population genetics*. Springer-Verlag, New York.
- Feller, W. 1957. *An introduction to probability theory and its applications*. 2nd edition. John Wiley and Sons, New York.
- Felsenstein, J. 1992. Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. *Genet. Res.* **59**:139–147.
- Finnegan, D. J. 1992. Transposable elements. Pp. 1096–1107 in D. L. Lindsley and G. Zimm, eds. *The genome of Drosophila melanogaster*. Academic Press, New York.
- Furano, A. V., D. D. Duvernell, and S. Boissinot. 2004. L1 (LINE-1) retrotransposon diversity differs dramatically between mammals and fish. *Trends Genet.* **20**:9–14.
- Garfinkel, D. J., K. Nyswaner, J. Wang, and J. Y. Cho. 2003. Post-transcriptional cosuppression of Ty1 retrotransposition. *Genetics* **165**:83–99.
- Goldberg, M. L., J.-Y. Sheen, W. J. Gehrubg, and M. M. Green. 1983. Unequal crossing-over associated with asymmetrical synapses between nomadic elements in the *Drosophila melanogaster* genome. *Proc. Natl. Acad. Sci. USA* **80**:5017–5021.
- Goodwin, T. J., and R. T. Poulter. 2001. The *DIRS1* group of retrotransposons. *Mol. Biol. Evol.* **18**:2067–2082.
- Hall, T. A. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* **41**:95–98.
- Hartl, D. L., A. R. Lohe, and E. R. Lozovskaya. 1997. Regulation of the transposable element mariner. *Genetica* **100**:177–184.
- Hey, J. 1989. The transposable portion of the genome of *Drosophila algonquin* is very different from that in *D. melanogaster*. *Mol. Biol. Evol.* **6**:66–79.
- International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**:860–921.
- Jensen, S., M. P. Gassama, and T. Heidmann. 1999. Co-suppression of *I* transposon activity in *Drosophila* by I-containing sense and antisense transgenes. *Genetics* **153**:1767–1774.
- Jiang, Y. W. 2002. Transcriptional cosuppression of yeast Ty1 retrotransposons. *Genes Dev.* **16**:467–478.
- 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, Q. H., S. Wright, Z. Yu, and T. Bureau. 2000. Transposon diversity in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA* **97**:7376–7381.
- Lynch, M., and J. S. Conery. 2003. The origins of genome complexity. *Science* **302**:1401–1404.
- McNaughton, J. C., D. J. Cockburn, G. Hughes, W. A. Jones, N. G. Laing, P. N. Ray, P. A. Stockwell, and G. B. Petersen. 1998. Is gene deletion in eukaryotes sequence-dependent? A study of nine deletion junctions and nineteen other deletion breakpoints in intron 7 of the human dystrophin gene. *Gene* **222**:41–51.
- Metzenberg, A. B., G. Wurzer, T. H. Huisman, and O. Smithies. 1991. Homology requirements for unequal crossing over in humans. *Genetics* **128**:143–161.
- Montgomery, E., and C. H. Langley. 1983. Transposable elements in Mendelian populations. II. Distribution of three copia-like elements in a natural population of *Drosophila melanogaster*. *Genetics* **104**:473–483.
- Montgomery, E., B. Charlesworth, and C. H. Langley. 1987. A test for the role of natural selection in the stabilization of transposable element copy number in a population of *Drosophila melanogaster*. *Genet. Res.* **49**:31–41.
- Montgomery, E. A., S. M. Huang, C. H. Langley, and B. H. Judd. 1991. Chromosome rearrangement by ectopic recombination in *Drosophila melanogaster*: genome structure and evolution. *Genetics* **129**:1085–1098.
- Morgan, M. T. 2001. Transposable element number in mixed mating populations. *Genet. Res.* **77**:261–275.
- Morton, N. E. 1991. Parameters of the human genome. *Proc. Natl. Acad. Sci. USA* **88**:7474–7476.
- Nagylaki, T. 1974. The moments of stochastic integrals and the distribution of sojourn times. *Proc. Natl. Acad. Sci. USA* **71**:746–749.
- Nuzhdin, S. V. 1999. Sure facts, speculations, and open questions about the evolution of transposable element copy number. *Genetica* **107**:129–137.
- Nuzhdin, S. V., E. G. Pasyukova, and T. F. C. Mackay. 1996. Positive association between *copia* transposition rate and copy number in *Drosophila melanogaster*. *Proc. R. Soc. Lond. B Biol. Sci.* **263**:823–831.
- O'Neill, R. J., M. J. O'Neill, and J. A. Graves. 1998. Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. *Nature* **393**:68–72.
- Pelisson, A., and J. C. Bregliano. 1987. Evidence for rapid limitation of the *I* element copy number in a genome submitted to several generations of *I-R* hybrid dysgenesis in *Drosophila melanogaster*. *Mol. Gen. Genet.* **207**:306–313.
- 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.
- Richard, M., A. Belmaaza, N. Gusew, J. C. Wallenburg, and P. Chartrand. 1994. Integration of a vector containing a repetitive LINE-1 element in the human genome. *Mol. Cell. Biol.* **14**:6689–6695.
- Rozas, J., and R. Rozas. 1999. DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* **15**:174–175.



- Rozen, S., and H. Skaletsky. 2000. Primer3 on the WWW for general users and for biologist programmers. *Meth. Mol. Biol.* **132**:365–386.
- Segal, Y., B. Peissel, A. Renieri, M. de Marchi, A. Ballabio, Y. Pei, and J. Zhou. 1999. LINE-1 elements at the sites of molecular rearrangements in Alport syndrome-diffuse leiomyomatosis. *Am. J. Hum. Genet.* **64**:62–69.
- Sherry, T. S., H. C. Harpending, M.A. Batzer, and M. Stoneking. 1997. *Alu* evolution in human populations: using the coalescent to estimate effective population size. *Genetics* **147**:1977–1982.
- Sijen, T., and R. H. Plasterk. 2003. Transposon silencing in the *Caenorhabditis elegans* germline by natural RNAi. *Nature* **426**:310–314.
- Surzycki, S. A., and W. R. Belknap. 1999. Characterization of repetitive DNA elements in Arabidopsis. *J. Mol. Evol.* **48**:684–691.
- Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**:585–595.
- Takami, K., F. Figueroa, W. E. Mayer, and J. Klein. 2000. Ancient allelism at the cytosolic chaperonin-alpha-encoding gene of the zebrafish. *Genetics* **154**:311–322.
- Takasaki, N., T. Y. Amaki, M. Hamada, L. Park, and N. Okada. 1997. The salmon *SmaI* family of short interspersed repetitive elements (SINEs): interspecific and intraspecific variation of the insertion of SINEs in the genomes of chum and pink salmon. *Genetics* **146**:369–380.
- Tavaré, S. 1984. Line-of-descent and genealogical processes, and their applications in population genetics models. *Genetics* **123**:585–595.
- Volff, J. N., L. Bouneau, C. Ozouf-Costaz, and C. Fischer. 2003. Diversity of retrotransposable elements in compact pufferfish genomes. *Trends Genet.* **19**:674–678.
- von Wettstein, D. 1984. The synaptonemal complex and genetic segregation. *Symp. Soc. Exp. Biol.* **38**:195–231.
- Walsh, C. P., J. R. Chaillet, and T. H. Bestor. 1998. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nat. Genet.* **20**:116–117.
- Watterson, G. A. 1975. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**:256–276.
- Wright, S. I., N. Agrawal, and T. E. Bureau. 2003. Effects of recombination rate and gene density on transposable element distributions in Arabidopsis thaliana. *Genome Res.* **13**:1897–1890.
- Wright, S. I., Q. H. Le, D. J. Schoen, and T. E. Bureau. 2001. Population dynamics of an Ac-like transposable element in self- and cross-pollinating Arabidopsis. *Genetics* **158**:1279–1288.

Pierre Cappy, Associate Editor

Accepted August 23, 2004