Patterns of Selection Against Transposons Inferred From the Distribution of Tc1, Tc3 and Tc5 Insertions in the *mut-7* Line of the Nematode *Caenorhabditis elegans*

Carène Rizzon,* Edwige Martin,† Gabriel Marais,* Laurent Duret,* Laurent Ségalat† and Christian Biémont*,[1]

**Biométrie, Biologie Evolutive, UMR 5558, Université Lyon 1, 69622 Villeurbanne Cedex, France and* †*Centre de Génétique Moléculaire et Cellulaire, UMR 5534, Université Lyon 1, 69622 Villeurbanne Cedex, France*

Manuscript received March 31, 2003
Accepted for publication July 25, 2003

ABSTRACT

To identify the factors (selective or mutational) that affect the distribution of transposable elements (TEs) within a genome, it is necessary to compare the pattern of newly arising element insertions to the pattern of element insertions that have been fixed in a population. To do this, we analyzed the distribution of recent mutant insertions of the Tc1, Tc3, and Tc5 elements in a *mut-7* background of the nematode *Caenorhabditis elegans* and compared it to the distribution of element insertions (presumably fixed) within the sequenced genome. Tc1 elements preferentially insert in regions with high recombination rates, whereas Tc3 and Tc5 do not. Although Tc1 and Tc3 both insert in TA dinucleotides, there is no clear relationship between the frequency of insertions and the TA dinucleotide density. There is a strong selection against TE insertions within coding regions: the probability that a TE will be fixed is at least 31 times lower in coding regions than in noncoding regions. Contrary to the prediction of theoretical models, we found that the selective pressure against TE insertions does not increase with the recombination rate. These findings indicate that the distribution of these three transposon families in the genome of *C. elegans* is determined essentially by just two factors: the pattern of insertions, which is a characteristic of each family, and the selection against insertions within coding regions.

TRANSPOSABLE elements (TEs) are not uniformly distributed along chromosomes, but tend to accumulate more frequently in some genomic regions than in others. This nonrandom distribution can be explained by mutational factors (the rate of TE insertions and the rate of TE loss—by either deletion or accumulation of point mutations) or by selective pressures acting on these TEs. The relative contributions of these evolutionary forces have not yet been elucidated, and various, nonmutually exclusive hypotheses have been proposed. Natural selection plays an important part in determining the distribution of TEs within genomes but how this selection process works is still a matter of debate (Biémont 1992; Biémont *et al.* 1997; Charlesworth *et al.* 1997; Nuzhdin 1999; Blumenstiel *et al.* 2002; Carr *et al.* 2002). It is proposed that selection may act either directly on TE insertions that are deleterious for the genome (*e.g.*, within genes or regulatory elements—the "gene disruption model") or indirectly by eliminating chromosomal rearrangements due to ectopic recombination between TE copies (the "ectopic recombination model"). Both hypotheses predict a negative relationship between the recombination rate along the chromosomes and the insertion copy frequency. TEs can therefore be expected to accumulate in regions with low recombination rates, where selection against deleterious effects is less efficient (the Hill-Robertson effect; Hill and Robertson 1966) and ectopic recombination is less frequent (Charlesworth *et al.* 1997). (Note, however, that a full theoretical analysis would be necessary to quantify the expected impact of the Hill-Robertson effects.) In *Drosophila melanogaster*, the data are generally consistent with the predictions of these selective models: TEs clearly accumulate in regions with very low recombination rates, such as the chromocenter and pericentromeric regions (Charlesworth *et al.* 1992a,b; Bartolome *et al.* 2002), but there is no clear negative relationship between recombination rates along the chromosomes and TE frequency in data from natural populations (Hoogland and Biémont 1996; Biémont *et al.* 1997). In the sequenced Drosophila genome, however, the overall frequency of transposons (DNA-based TEs) in euchromatic regions is weakly negatively correlated with recombination rate, but the density of retrotransposons (LTR and non-LTR RNA-based elements) is not (Rizzon *et al.* 2002). The absence of any relationship between the recombination rate and the LTR and non-LTR retrotransposon insertions and the fact that the density in TEs is negatively correlated to recombination rate for some TE families, but not for others, indicate

[1]*Corresponding author:* Biométrie, Biologie Evolutive, UMR CNRS 5558, Université Lyon 1, 69622 Villeurbanne Cedex, France. E-mail: biemont@biomserv.univ-lyon1.fr

that recombination is not the main factor shaping TE distribution in the genome of Drosophila.

The distribution of TEs according to recombination rate observed in Drosophila is not universal. Indeed, in the genome sequence of the nematode *Caenorhabditis elegans*, most DNA-transposable element families (the transposons) tend to accumulate preferentially in regions of high recombination rate (Duret *et al.* 2000). Since this pattern conflicts with the prediction of the selective models, it has been suggested that the distribution of the transposons in the nematode genome reflects their preferential insertion in regions with high recombination rates, possibly because they use the recombination machinery for their own transposition (Duret *et al.* 2000). To determine more directly which factors (selective or mutational) affect the distribution of TEs within a given genome, it is therefore necessary to compare the pattern of newly arising TE insertions to the pattern of older element insertions that have already been fixed in the genome. For this purpose, we compared the distribution of recent insertions of the Tc1, Tc3, and Tc5 elements in a *mut-7* background of the nematode *C. elegans* and the distribution of copies (presumably fixed) of these three elements within the complete sequenced genome (strain N2). Differences in the insertion patterns and selective pressure on TEs were analyzed in coding and noncoding regions on the basis of the rate of recombination and the gene density.

MATERIALS AND METHODS

Sequence data: Full-length sequences of the six *C. elegans* chromosomes along with gene annotations were retrieved from the WormBase release WS62 (2002). Data available totaled 100.25 Mb, corresponding to 99.9% of the whole genome sequence (release WS62, 2002, WormBase, http://www.wormbase.org/).

Detection and localization of Tc1, Tc3, and Tc5 insertions in the sequenced genome: Tc1, Tc3, and Tc5 sequences in the sequenced genome were retrieved using the RepeatMasker program (A. F. A. Smit and P. Green, unpublished data; RepeatMasker is available at http://repeatmasker.genome.washington.edu/cgi-bin/RM2_req.pl) and a database of reference sequences for each of the known transposon families [accession numbers in GenBank: Tc1, K01135 from position 46 to 1655; Tc5, Z35400; Repbase Update database (Jurka 2000) for the following families: IR-1, IR-2, IR-3, IR-4, IR-5, Tc2, Tc4, Tc6, Tc7]. The reference sequence available for Tc3 (M77697 from position 14870 to 15907 in GenBank) did not correspond to a complete copy. To define complete copies, regions of ~5 kb overlapping the RepeatMasker matches retrieved from the genome with this reference sequence were aligned with CLUSTALW. The inverted terminal repeat regions of the copies were defined using BLAST to determine the ends of the complete copies. A total of 19 complete copies were thus retrieved (for example, from 12509999 to 12512340 in chromosome IV and from 17098031 to 17100372 in chromosome V) as in the previous study of Tu and Shao (2002). A 2341-bp consensus sequence of these 19 copies was used as a reference to search for Tc3 copies in the genome. All the segments matching a given TE sequence were manually checked (1) to determine whether neighboring matches corresponded to different fragments of a single TE or to independent insertion events and (2) to discriminate copies of Tc1 and Tc5 families from the copies of the associated nonautonomous families (Oosumi *et al.* 1996). To determine whether these elements were inserted in coding or noncoding regions, their location was compared to the annotations of protein-coding regions (CDS) in WormBase. However, because TE coding regions are sometimes annotated as protein genes, we discriminated host genes from TE coding regions. CDSs, described in the annotations, were thus compared using BLAST to the nematode TE reference sequences of the Repbase Update database (Jurka 2000) and to Tc3, which is not referenced in the Repbase Update database. Only the CDSs that did not match any known TE over their entire length were classified as host genes.

Detection and localization of Tc1, Tc3, and Tc5 mutant insertions: A total of 1080 independent mutant Tc1, Tc3, and Tc5 insertions were obtained by propagating independent lines carrying the *mut-7* mutation of *C. elegans* (Martin *et al.* 2002) and analyzed by the transposon display technique (Wicks *et al.* 2000; Martin *et al.* 2002). The flanking sequences of each insertion were sequenced. The insertions were thus precisely localized on the *C. elegans* chromosomes using the sequenced genome (release wS56, 2001, WormBase: http://www.wormbase.org/) and we checked whether they were inserted in a host gene coding region by using WormBase annotations as described above. Only 1049 insertions that were accurately annotated in the genome were used in this study.

Estimation of the recombination rate: The rate of recombination along the chromosomes was determined using a procedure similar to that described by Kliman and Hey (1993). We used the 780 markers of the wS56 release. The recombination rate was estimated for each chromosome arm by taking the derivative of the best-fitting polynomial function of the genetic distance *vs.* the nucleotide coordinate in the genomic sequence. These polynomial curves fitted the data set well for all chromosome arms ($r^2 > 0.99$). For each 100-kb genomic fragment, the recombination rate was estimated from the value of the derivative of the polynomial curve at the middle position of the fragment (Duret *et al.* 2000). The rate of recombination varies more than sixfold within the genome of *C. elegans* from a mean value of 0.7 cM/Mb in the class with the lowest recombination rate (the central region of the autosomes) to a mean value of 4.7 cM/Mb in the class with the highest recombination rate (the two arms flanking the central region).

The chromosome sequences of the genome were split into 100-kb fragments and were analyzed with respect to the amount and distribution of Tc1, Tc3, and Tc5 copies. The amounts of host coding regions and TA dinucleotides were calculated for each genomic fragment. Each fragment was attributed to one of the four recombination rate classes, each defined as containing 25% of the total number of genomic fragments, as follows: <1.1 cM/Mb (very low), <2.5 cM/Mb (low), <4.2 cM/Mb (moderate), and >4.2 cM/Mb (high).

Statistical tests: The distribution of the transposon copies in the four classes of recombination rate was compared to the expected numbers by χ^2 tests, assuming that the distribution of the total number of copies paralleled the total amount of DNA in each class. The same method was used with the seven classes of TA dinucleotide amount, each defined as corresponding to at least 4 kb of the genome as follows: $[4.5 \times 10^3$ TA/kb; 5×10^3 TA/kb[, $[5 \times 10^3$ TA/kb; 5.5×10^3 TA/kb[, $[5.5 \times 10^3$ TA/kb; 6×10^3 TA/kb[, . . . , $[7.5 \times 10^3$ TA/kb; 8×10^3 TA/kb[.

The genome was split into 100-kb fragments within which the coding amount was calculated. The values of the coding region percentages of the genome fragments were divided into four classes, each corresponding to 25% of the total

number of genome fragments, as follows: [0%; 18.72%[(very low), [18.72%; 24.40%[(low), [24.40%; 30.71%[(moderate), [30.71%; 52.27%[(high). The distribution of the intergenic transposon copies in these four classes was compared to the expected numbers by χ^2 tests, assuming that the total number of copies was distributed according to the total amount of noncoding DNA in each class.

RESULTS

To determine the pattern of newly arising TE insertions in the genome of *C. elegans*, we analyzed the distribution of the 1049 independent mutations, among which 597 corresponded to Tc1 insertions, 246 corresponded to Tc3, and 206 corresponded to Tc5. These insertions were recovered in a *mut-7* strain of *C. elegans*, which is characterized by a high rate of germ-line transposition, in the course of a project to create a large collection of mutants for functional genomic experiments in the nematode (Martin *et al.* 2002). These TEs have been localized within the genome by sequencing their flanking regions. These mutants are generally heterozygous and have undergone at most 10 generations (Martin *et al.* 2002), which left very little time for selection to operate except in the case of strongly deleterious dominant mutations. The distribution of these elements can therefore be expected to reflect the pattern of insertion mutations before selection has had any effect. These 1049 insertions will hereafter be referred to as "recent insertions."

In the sequenced genome, we identified 171 TE sequences for the three families: 33 for Tc1, 24 for Tc3, and 114 for Tc5. These TEs are presumed to be fixed or at least to have been subject to selection for a very long period of time (see the discussion), and hence, for the sake of simplicity, they will hereafter be called "fixed insertions." The ratio of fixed-to-recent insertions should therefore reflect the intensity of the selection acting on these TEs.

Frequency of recent TE insertions according to recombination rate: We first analyzed the distribution of recent TE insertions according to recombination rate in the whole genome and in coding or noncoding regions taken separately. We considered four classes of recombination rate (very low, low, moderate, high), each one covering 25% of the whole genome. The relationship between the frequency of recent insertions and recombination rate was assessed in each compartment (whole genome, noncoding region, coding region) by a χ^2 test, taking as the null hypothesis that insertions are distributed in the four classes of recombination according to the amount of DNA in each class.

For the Tc1 family, we found in the whole genome that insertions were significantly more frequent in the highest class of recombination rate ($P < 0.0001$; Table 1). Because the recombination rate in *C. elegans* is negatively correlated with gene density (Barnes *et al.* 1995; see also Table 1), we analyzed the Tc1 insertions within noncoding regions. We observed the same pattern ($P < 0.001$) as in the whole genome, which indicates that the relationship between Tc1 insertions and recombination cannot simply be attributed to a possible selection against dominant effects of highly deleterious insertions in coding regions. When only coding regions were considered, the Tc1 insertions showed a nonsignificant ($P = 0.07$) tendency to accumulate in classes with high recombination rates. In Duret *et al.* (2000), no clear relationship was found between recombination rate and the density of Tc1 insertions of the genome. The difference between the two studies likely results from the improvement of the genetic map [225 genetic markers in the map used in Duret *et al.* (2000), 780 in the present study] and from the distinction of Tc1 copies from copies of associated nonautonomous families (see materials and methods). The Tc3 and Tc5 families showed a different pattern: χ^2 tests, when possible, were not significant, suggesting a random distribution unrelated to the recombination rate. Only the Tc3 insertions in coding regions showed a tendency to accumulate in the class with the highest recombination rate ($\chi^2 = 8.86$; $P = 0.03$). However, given the multiplicity of tests that were performed, this tendency cannot be considered to be significant (Bonferroni correction: $\alpha/3 = 0.017$).

TE insertions in TA-rich regions: Tc1 and Tc3 are known to insert in TA dinucleotides (Rosenzweig *et al.* 1983; Mori *et al.* 1988; van Luenen and Plasterk 1994; Ketting *et al.* 1997) and Tc5 in TNA trinucleotides (Collins and Anderson 1994). We therefore tested whether the frequency of TE insertions depended on the amount of TA dinucleotides. There was a significant deficit in the number of Tc1 in the classes with the lowest TA amount ($\chi^2 = 24.3$; $P < 0.0001$), whereas Tc3 ($\chi^2 = 0.7$; $P = 0.94$) and Tc5 ($\chi^2 = 6.7$; $P = 0.15$) showed no significant difference from a random distribution. Moreover, we found that the TA dinucleotide amount was negatively correlated with the recombination rate (Figure 1), which was in agreement with the findings of Barnes *et al.* (1995). The accumulation of the Tc1 inserts in the class with the highest recombination rate cannot therefore be explained simply in terms of a relationship with the TA dinucleotide amount. All these results thus globally suggest that the Tc1, Tc3, and Tc5 distributions according to recombination rate were not biased by the TA dinucleotide amount.

Strong selection against TE insertions within coding regions: The most striking difference between recent and fixed insertions is the pattern of their distribution within coding regions. There were 1049 recent TE insertions, 18.3% of which were located within the coding region of a host gene, compared to only 0.6% (1/171) of the fixed insertions (Table 2). The probability that newly arising insertions will be fixed is therefore at least ~31 times lower for those located in a coding region than for those located in a noncoding region. This ten-

TABLE 1

Fixed and recent TE insertions in coding and noncoding regions and in the whole genome of *C. elegans*, according to recombination rate

Genomic compartment		Class of recombination rate						
		Total	Very low	Low	Moderate	High	χ^2 value[a]	*P* value
		DNA amount (Mb)						
Whole genome		100.2	25.0	25.1	25.2	25.9	—	—
Noncoding regions		75.3	17.8	18.3	19.4	19.8	—	—
Coding regions		24.9	7.2	6.8	5.8	5.2	—	—
		No. of insertions						
Tc1								
Whole genome	Fixed	33	5	4	11	13	7.07	*
	Recent	597	140	135	124	198	22.75	***
Noncoding regions	Fixed	33	5	4	11	13	5.84	NS
	Recent	486	117	106	98	165	18.14	***
Coding regions	Fixed	0	0	0	0	0	—	—
	Recent	111	23	29	26	33	7.00	NS
Tc3								
Whole genome	Fixed	24	2	5	9	8	4.23	NS
	Recent	246	56	72	55	63	3.02	NS
Noncoding regions	Fixed	23	2	5	9	7	3.21	NS
	Recent	207	50	61	48	48	3.55	NS
Coding regions	Fixed	1	0	0	0	1	—	—
	Recent	39	6	11	7	15	8.86	*
Tc5								
Whole genome	Fixed	114	23	25	27	39	5.53	NS
	Recent	206	43	49	51	63	4.17	NS
Non-coding regions	Fixed	114	23	25	27	39	3.75	NS
	Recent	164	35	40	38	51	2.23	NS
Coding regions	Fixed	0	0	0	0	0	—	—
	Recent	42	8	9	13	12	4.31	NS

NS, not significant. *$P < 0.05$; ***$P < 0.001$.

[a] Comparison of the number of insertions observed in each class of recombination rate to the number expected according to the hypothesis that insertions are distributed in the four classes of recombination according to the amount of DNA in each class.

dency was confirmed for every chromosome and for the entire genome.

Recombination rate and selection against TE insertions: As mentioned in the Introduction, selective models (ectopic recombination or direct effect of TEs) predict a positive correlation between the rate of recombination and the strength of selection against TE insertions. According to these models, the ratio of fixed-to-recent insertions should therefore decrease with recombination rate. As shown in Table 1, the distribution of fixed TE insertions within the genome according to recombination rate appears to be similar to that of recent insertions (except for fixed insertions in coding regions, for which there were not enough data to perform any test): fixed Tc1 insertions showed a slight tendency to accumulate in regions of high recombination rate, whereas the Tc3 and Tc5 insertions showed no statistically significant difference from the random distribution.

To determine directly whether the intensity of selection against TE insertions varied with the recombination rate, we compared the distribution of fixed and recent insertions in the noncoding region. For this, we performed a χ^2 test of the distribution of fixed insertions in the four recombination classes under the null hypothesis that this distribution matches that of recent insertions. The numbers of fixed insertions of the three families in the different recombination classes showed no significant departure from the expected values of recent insertions (Table 3). There was thus no evidence that the selective pressure against TE insertions was stronger in regions with a high recombination rate.

Gene density and selection against TE insertions in noncoding regions: TE insertions in noncoding DNA in the vicinity of genes are expected to be counterselected because they can affect the proper expression of genes. To analyze the relationship between gene density and the intensity of selection against the insertions of the three transposons in noncoding DNA, we split the genome into four classes of protein-coding DNA (CDS) density (very low, low, moderate, and high), correspond-

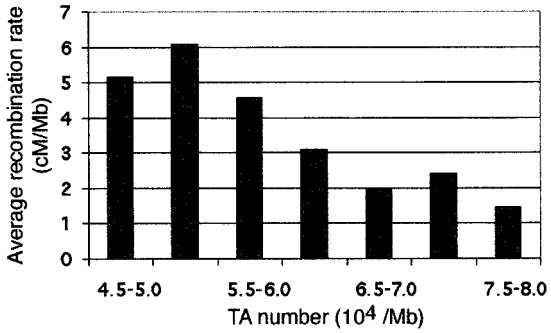

FIGURE 1.—Average recombination rate according to the density of TA dinucleotides.

ing to 25% increments of the percentage of the whole genome. The average gene density varied from 14.8% coding in the lowest class to 36.0% coding in the highest class. We first checked whether gene density affected the distribution of recent insertions within noncoding regions. The relationship between the frequency of recent insertions and CDS density was assessed by a χ^2 test, with the null hypothesis that insertions were distributed in the four classes of CDS density according to the amount of noncoding DNA in each class. Table 4 shows that the distributions of recent insertions of Tc1 and Tc5 did not significantly depart from the null hypothesis of random insertions in noncoding DNA, whereas recent Tc3 insertions showed a significant excess in the higher classes of CDS density ($\chi^2 = 10.33$; $P = 0.02$). Hence, to determine directly whether the intensity of selection against transposon insertions in noncoding regions varied with gene density, we compared the distribution of fixed and recent insertions. To do this, we performed a χ^2 test of the distribution of fixed insertions in the four classes of CDS density, with the null hypothesis that this distribution matched that of recent insertions. Although we observed a weak tendency in regions of high gene density for the number of transposons fixed in noncoding regions to be generally lower than expected in the light of the distribution of recent insertions (Table 4), the χ^2 test was not significant for any of the three transposon families. To circumvent a possible problem of low data number, we pooled the observed and expected values for the three transposon families. Although the χ^2 value was significant ($\chi^2 = 7.93$; $P = 0.047$; Table 4), its P value was not strong, suggesting only a slight tendency for selection against transposon insertions in the vicinity of genes.

DISCUSSION

Natural selection acting against Tc1, Tc3, and Tc5 insertions: To analyze the factors that affect the strength of the selection acting on TEs, we compared the distribution of 1049 recent insertions recovered in the *mut-7* line of *C. elegans* to the distribution of all insertions

TABLE 2

Proportion of fixed and recent Tc1, Tc3, and Tc5 insertions in CDSs

	Fixed insertions			Recent insertions		
	Total no.	No. in CDS	% in CDS	Total no.	No. in CDS	% in CDS
Chromosomes						
I	32	1	3.12	154	23	14.94
II	42	0	0	212	39	18.40
III	14	0	0	104	18	17.31
IV	38	0	0	139	29	20.86
V	26	0	0	244	60	24.59
X	19	0	0	196	23	11.73
Transposon family						
Tc1	24	0	0	597	111	18.59
Tc3	33	1	3.03	246	39	15.85
Tc5	114	0	0	206	42	20.39
Total	171	1	0.58	1049	192	18.30

from the same TE families found in the coding and noncoding compartments of the sequenced genome. Our study depends on two essential assumptions: first, we assume that the pattern of new insertions observed in the *mut-7* line of *C. elegans* reflects that of the wild type. The *mut-7 (pk204)* mutation has the property of promoting mobilization of various transposons (DNA transposable elements) in the germ line (KETTING *et al.* 1999). This mutation, which affects an RNAseD homolog, acts by derepressing transposition by a mechanism that remains to be elucidated. In nonmutant strains, however, transposition depends on transposase activity and is largely independent of host-specific factors, which, for example, renders the Tc1-*mariner* superfamily ubiquitous in eukaryotic genomes (ROBERTSON and LAMPE 1995; REZSOHAZY *et al.* 1997). Moreover, it should be noted that in noncoding regions, the distributions of recent and fixed insertions are very similar (Tables 3 and 4). If the pattern of insertion were different in *mut-7* mutants compared to that in the wild type, there would have been no obvious reason to expect this similarity between the two distributions.

The second assumption is that most of the TEs that are found in the genome sequence have been fixed in the population or have at least been subject to selection for a long period of time. Whereas germ-line transposition is active in some natural isolates of *C. elegans*, only somatic (nonheritable) transposition has been described in the laboratory strain N2, the genome of which has been sequenced (PLASTERK 1993; KETTING *et al.* 1999). This means that germ-line transposition is very rare and, therefore, that most of these 171 insertions are ancient. TE insertions may be deleterious, neutral, or advantageous. Advantageous insertions are probably rare, and most deleterious insertions are generally elimi-

TABLE 3

Fixed insertions within noncoding regions according to recombination rate and comparison to the neutral expectation

	Class of recombination rate						
TE data	Very low	Low	Moderate	High	Total	χ^2 value	*P* value
Tc1							
Observed	5	4	11	13	33		
Expected[a]	7.95	7.20	6.65	11.20	33	5.64	NS
Tc3							
Observed	2	5	9	7	23		
Expected[a]	5.56	6.78	5.33	5.33	23	5.78	*
Tc5							
Observed	23	25	27	39	114		
Expected[a]	24.33	27.80	26.42	35.45	114	0.72	NS
Pooled							
Observed	30	34	47	59	170		
Expected[a]	37.83	41.78	38.40	51.99	170	5.94	NS

NS, not significant. *$P < 0.05$.

[a] Number of fixed insertions expected according to the hypothesis that their distribution is identical to that of recent insertions (*i.e.*, assuming that all recent insertions have the same probability of fixation).

nated by negative selection. Hence, most of the TE copies that are present in the line used to sequence the genome of *C. elegans* ("fixed" TEs) probably correspond to selectively neutral or slightly deleterious insertions fixed by genetic drift. The ratio of the number of TEs observed in the genome sequence to the number of recent insertions in the *mut-7* strain should thus reflect, although underestimate, the intensity of the selection acting on these TEs.

As expected, we observed a strong selection against TE insertions within coding regions. For two reasons, the strength of selection against such insertions is certainly underestimated: first, it is possible that the number of host genes containing a fixed insertion is overestimated, because some of these genes could correspond to misannotated genes or pseudogenes. Second, the number of newly arising insertions within coding regions could be underestimated. Indeed, coding regions that account for 25% of the whole genome contain only 18.3% of the 1049 recent insertions. This suggests that although the mutations that we analyzed were very recent and generally heterozygous, their distribution in coding regions did not fully reflect the pattern of newly arising insertions, probably because of selection against dominant, strongly deleterious mutations.

TE insertions in the vicinity of genes are also expected to be counterselected because they can hinder the proper expression of genes either by disrupting a gene regulatory element (promoter, enhancer, etc.) or by the fact that their own regulatory elements may interfere with those of the flanking genes (Tomilin 1999; Borie *et al.* 2000). Such an effect could account for the higher TE density observed in regions of low gene density in various species. However, the real impact of this selective pressure on the genomic distribution of TEs has not yet been quantified, and it has been shown recently that many TEs may have provided their human host with novel regulatory sequences (Jordan *et al.* 2003). For the three transposon families analyzed here, selection against insertions in the vicinity of genes appears to be weak and is not an important determinant of the distribution of fixed insertions in noncoding DNA.

We investigated the relationship between recombination rate and the selective pressure against TE insertions. For the three Tc1, Tc3, and Tc5 families, we observed a weak tendency: the number of fixed insertions in regions of moderate or high recombination rate was higher than expected according to the neutral model. We can thus clearly conclude that, contrary to the prediction of the selective models, the selective pressure against transposon insertions did not increase with recombination rate in the genome of the nematode. Therefore, whereas TE distribution in the Drosophila genome is globally consistent with these selective models for transposons (see the Introduction), this does not appear to be the case for the nematode *C. elegans*. This difference between the two species could be due to their different modes of reproduction. *C. elegans* is a hermaphrodite species and is likely to be highly self-fertilizing and homozygous in nature. Its effective recombination rate is therefore predicted to be lower than that in outbreeding species like Drosophila (Nordborg

TABLE 4

Distribution of fixed and recent insertions within noncoding regions according to CDS density

	Class of CDS density					χ^2 value		*P* value	
	Total (75.0)	Very low (21.2)	Low (19.7)	Moderate (18.1)	High (16.0)	Under $H0_1$[a]	Under $H0_2$[b]	Under $H0_1$	Under $H0_2$
Tc1 inserts									
Fixed	33	9	9	6	9	1.06	1.07	NS	NS
Expected fixed[b]	33	7.92	9.16	8.34	7.58				
Recent	483	116	134	122	111	4.41	—	NS	—
Tc3 inserts									
Fixed	24	5	6	9	4	2.49	1.16	NS	NS
Expected fixed[b]	24	4.87	5.68	7.30	6.15				
Recent	207	42	47	65	53	10.33	—	*	—
Tc5 inserts									
Fixed	113	27	38	29	19	4.33	5.70	NS	NS
Expected fixed[b]	113	22.46	30.88	31.58	28.08				
Recent	161	32	44	45	40	6.00		NS	
Pooled									
Fixed	170	41	53	44	32	3.35	7.93	NS	*
Expected fixed[b]	170	35.26	45.72	47.22	41.80				
Recent	851	90	227	230	204				

The noncoding DNA amounts (in megabases) are in parentheses. NS, not significant. $*P < 0.05$.

[a] χ^2 test under the hypothesis $H0_1$: comparison of the distribution of insertions in each class of gene density to the distribution expected according to the amount of noncoding DNA in each class.

[b] χ^2 test under the hypothesis $H0_2$: comparison of the distribution of fixed insertions in each class of recombination rate to the distribution expected according to the observed pattern of recent insertion (*i.e.*, assuming that all recent insertions have the same probability of fixation).

2000). This means that Hill-Robertson effects and the link between natural selection and the meiotic recombination rate are reduced (Charlesworth and Wright 2001; Morgan 2001; Wright *et al.* 2001; Bartolome *et al.* 2002). Hence, because weakly deleterious mutations caused by TE insertions are expected to be mostly recessive, selection should be stronger in this highly homozygous species than in Drosophila and poorly or not affected by the local recombination rate (Bartolome *et al.* 2002). Moreover, because homozygosity reduces opportunity for ectopic exchange, selection against ectopic recombinations between TE insertions in *C. elegans* should be weaker than that in the outbreeding *D. melanogaster* species (Bartolome *et al.* 2002). These points could explain why we did not detect any significant relationship between recombination rate and the intensity of selection on transposon insertions in the *C. elegans* genome.

Relationship between the Tc1, Tc3, and Tc5 insertion patterns and the recombination rate: In a previous analysis of the complete genome sequence, it was shown that for 9 of 12 transposon families, insertions tended to accumulate preferentially in regions with a high recombination rate (Duret *et al.* 2000). Because this distribution conflicts with the predictions of the selective models, it was proposed that this distribution reflects the preferential insertion of these transposons in regions of high recombination rate rather than selective effects. The analysis of the recent insertions in the *mut-7* strain confirmed this interpretation for the Tc1 family, which clearly inserts more frequently in regions of high recombination rate. The pattern, however, was not observed for all transposon families in the present study. Indeed, as in the previous analysis (Duret *et al.* 2000), we did not find any clear relationship between recombination rate and the density of recent and fixed insertions for Tc3 and Tc5 (Table 1). The hypothesis that transposons use the recombination machinery for their own transposition into the nematode (Duret *et al.* 2000) cannot therefore be extrapolated to all transposon families.

How can we explain the difference between the insertion pattern of Tc1 and those of Tc3 and Tc5? Tc1 and Tc3 belong to the Tc1-*mariner* family of transposable elements (Collins *et al.* 1989; Rezsohazy *et al.* 1997), which transpose by a cut-and-paste mechanism and are related to the *IS 630* family of bacterial transposons (Capy *et al.* 1997). They both have target-site specificity in the TA sequence (Mori *et al.* 1988; Ketting *et al.* 1997). We would thus expect the insertion patterns of these two transposons to be similar. However, their patterns of insertions along the chromosomes are in fact completely different (van Luenen and Plasterk 1994),

which could be explained by differing recognitions of the TA flanking sequences (Ketting *et al.* 1997). Unlike Tc1 and Tc3, Tc5 was not found to contain a DDE motif on its putative transposase, suggesting a distant relationship to the Tc1-*mariner* family (Collins and Anderson 1994). Moreover, Tc5 recognizes the target site CTNAG (Collins and Anderson 1994; Smit and Riggs 1996). Hence, specific target sites may be one of the key factors determining transposon insertion patterns in the nematode. Although Tc1 may really take advantage of recombination for its own transposition and insertion, we cannot rule out the possibility that the correlation between its insertions and recombination is indirect. We can, however, reject the hypothesis that Tc1 insertions parallel the recombination rate because Tc1 inserts in TA sequences. Indeed, the recombination rate was positively correlated to the GC amount (Barnes *et al.* 1995; Marais *et al.* 2001) and thus negatively correlated with the TA content (Figure 1). However, Tc1 accumulated in regions with high recombination rates rather than in those with low recombination rates, as would be expected if the density in TA dinucleotide were the key factor determining the distribution of insertions.

In conclusion, the results presented here indicate that the distribution of the Tc1, Tc3, and Tc5 families in the genome of *C. elegans* is determined essentially by just two factors: the initial pattern of primary insertions and the selection against the insertions within coding regions.

We thank Manolo Gouy for comments and Monika Ghosh for reviewing the English text. This work was supported by the Centre National de la Recherche Scientifique (UMR 5558, UMR 5534, GDR 2157 on transposable elements).

LITERATURE CITED

Barnes, T. M., Y. Kohara, A. Coulson and S. Hekimi, 1995 Meiotic recombination, noncoding DNA and genomic organization in *Caenorhabditis elegans*. Genetics **141:** 159–179.

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.

Biémont, C., 1992 Population genetics of transposable elements. A Drosophila point of view. Genetica **86:** 67–84.

Biémont, C., A. Tsitrone, C. Vieira and C. Hoogland, 1997 Transposable element distribution in Drosophila. Genetics **147:** 1997–1999.

Blumenstiel, J. P., D. L. Hartl and E. L. Lozovsky, 2002 Patterns of insertion and deletion in contrasting chromatin domains. Mol. Biol. Evol. **19:** 2211–2225.

Borie, N., C. Lœvenbruck and C. Biémont, 2000 Developmental expression of *412* retrotransposon in natural populations of *D. melanogaster* and *D. simulans*. Genet. Res. **76:** 217–226.

Capy, P., C. Bazin, D. Higuet and T. Langin, 1997 *Dynamic and Evolution of Transposable Elements*. R. G. Landes, Austin, TX.

Carr, M., J. R. Soloway, T. E. Robinson and J. F. Brookfield, 2002 Mechanisms regulating the copy numbers of six LTR retrotransposons in the genome of *Drosophila melanogaster*. Chromosoma **110:** 511–518.

Charlesworth, B., and S. I. Wright, 2001 Breeding systems and genome evolution. Curr. Opin. Genet. Dev. **11:** 685–690.

Charlesworth, B., A. Lapid and D. Canada, 1992a 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, B., A. Lapid and D. Canada, 1992b The distribution of transposable elements within and between chromosomes in a population of *Drosophila melanogaster*. II. Inferences on the nature of selection against elements. Genet. Res. **60:** 115–130.

Charlesworth, B., C. H. Langley and P. Sniegowski, 1997 Transposable element distributions in Drosophila. Genetics **147:** 1993–1995.

Collins, J. J., and P. Anderson, 1994 The Tc5 family of transposable elements in *Caenorhabditis elegans*. Genetics **137:** 771–781.

Collins, J., E. Forbes and P. Anderson, 1989 The Tc3 family of transposable genetic elements in *Caenorhabditis elegans*. Genetics **121:** 47–55.

Duret, L., G. Marais and C. Biémont, 2000 Transposons but not retrotransposons are found preferentially in regions of high recombination rate in *Caenorhabditis elegans*. Genetics **156:** 1661–1669.

Hill, W. G., and A. Robertson, 1966 The effect of linkage on the limit to artificial selection. Genet. Res. **8:** 269–294.

Hoogland, C., and C. Biémont, 1996 Chromosomal distribution of transposable elements in *Drosophila melanogaster*: test of the ectopic recombination model for maintenance of insertion site number. Genetics **144:** 197–204.

Jordan, I. K., I. B. Rogozin, G. V. Glazko and E. V. Koonin, 2003 Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet. **19:** 68–72.

Jurka, J., 2000 Repbase update: a database and an electronic journal of repetitive elements. Trends Genet. **16:** 418–420.

Ketting, R. F., S. E. Fischer and R. H. Plasterk, 1997 Target choice determinants of the Tc1 transposon of *Caenorhabditis elegans*. Nucleic Acids Res. **25:** 4041–4047.

Ketting, R., T. Haverkamp, H. G. van Luenen and R. H. Plasterk, 1999 *Mut-7* of *C. elegans*, required for transposon silencing and RNA interference, is a homolog of Werner syndrome helicase and RNaseD. Cell **99:** 133–141.

Kliman, R. M., and J. Hey, 1993 Reduced natural selection associated with low recombination in *Drosophila melanogaster*. Mol. Biol. Evol. **10:** 1239–1258.

Marais, G., D. Mouchiroud and L. Duret, 2001 Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. Proc. Natl. Acad. Sci. USA **98:** 5688–5692.

Martin, E., H. Laloux, G. Couette, T. Alvarez, C. Bessou *et al.*, 2002 Identification of 1088 new transposon insertions of *Caenorhabditis elegans*: a pilot study toward large-scale screens. Genetics **162:** 521–524.

Morgan, M. T., 2001 Transposable element number in mixed mating populations. Genet. Res. **77:** 261–275.

Mori, I., G. M. Benian, D. G. Moerman and R. H. Waterston, 1988 Transposable element Tc1 of *Caenorhabditis elegans* recognizes specific target sequences for integration. Proc. Natl. Acad. Sci. USA **85:** 861–864.

Nordborg, M., 2000 Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. Genetics **154:** 923–929.

Nuzhdin, S. V., 1999 Sure facts, speculations, and open questions about the evolution of transposable element copy number. Genetica **107:** 129–137.

Oosumi, T., B. Garlick and W. R. Belknap, 1996 Identification of putative nonautonomous transposable elements associated with several transposon families in *Caenorhabditis elegans*. J. Mol. Evol. **43:** 11–18.

Plasterk, R. H., 1993 Molecular mechanisms of transposition and its control. Cell **74:** 781–786.

Rezsohazy, R., H. G. van Luenen, R. M. Durbin and R. H. Plasterk, 1997 Tc7, a Tc1-hitch hiking transposon in *Caenorhabditis elegans*. Nucleic Acids Res. **25:** 4048–4054.

Rizzon, C., G. Marais, M. Gouy and C. Biémont, 2002 Transposable elements distribution in relation to recombination rate in the *Drosophila melanogaster* genome. Genome Res. **12:** 400–407.

Robertson, H. M., and D. J. Lampe, 1995 Recent horizontal transfer of a mariner transposable element among and between Diptera and Neuroptera. Mol. Biol. Evol. **12:** 850–862.

Rosenzweig, B., L. W. Liao and D. Hirsh, 1983 Sequence of the

C. elegans transposable element Tc1. Nucleic Acids Res. **11:** 4201–4209.

Smit, A. F. A., and A. D. Riggs, 1996 *Tiggers* and other transposon fossils in the human genome. Proc. Natl. Acad. Sci. USA **93:** 1443–1448.

Tomilin, N. V., 1999 Control of genes by mammalian retroposons. Int. Rev. Cytol. **186:** 1–48.

Tu, Z., and H. Shao, 2002 Intra- and inter-specific diversity of Tc3-like transposons in nematodes and insects and implications for their evolution and transposition. Gene **282:** 133–142.

van Luenen, H. G., and R. H. Plasterk, 1994 Target site choice of the related transposable elements Tc1 and Tc3 of *Caenorhabditis elegans*. Nucleic Acids Res. **22:** 262–269.

Wicks, S., C. de Vries, H. van Luenen and R. H. Plasterk, 2000 CHE-3, a cytosolic dynein heavy chain, is required for sensory cilia structure and function in *Caenorhabditis elegans*. Dev. Biol. **221:** 295–307.

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.

Communicating editor: T. H. Eickbush