

# Transposable Element Orientation Bias in the Drosophila melanogaster Genome

Asher D. Cutter,<sup>1</sup> Jeffrey M. Good,<sup>1</sup> Christopher T. Pappas,<sup>2</sup> Matthew A. Saunders,<sup>1</sup> Dean M. Starrett,<sup>3</sup> Travis J. Wheeler<sup>3</sup>

<sup>1</sup> Department of Ecology & Evolutionary Biology, University of Arizona, Tucson, AZ 85721, USA

<sup>2</sup> Department of Molecular and Cellular Biology, University of Arizona, Tucson, AZ 85721, USA

<sup>3</sup> Department of Computer Science, University of Arizona, Tucson, AZ 85721, USA

Received: 3 August 2004 / Accepted: 26 May 2005 [Reviewing Editor: Dr. Dmitri Petrov]

Abstract. Nonrandom distributions of transposable elements can be generated by a variety of genomic features. Using the full D. melanogaster genome as a model, we characterize the orientations of different classes of transposable elements in relation to the directionality of genes. DNA-mediated transposable elements are more likely to be in the same orientation as neighboring genes when they occur in the nontranscribed regions that flank genes. However, RNAmediated transposable elements located in an intron are more often oriented in the direction opposite to that of the host gene. These orientation biases are strongest for genes with highly biased codon usage, probably reflecting the ability of such loci to respond to weak positive or negative selection. The leading hypothesis for selection against transposable elements in the coding orientation proposes that transcription termination poly(A) signal motifs within retroelements interfere with normal gene transcription. However, after accounting for differences in base composition between the strands, we find no evidence for global selection against spurious transcription termination signals in introns. We therefore conclude that premature termination of host gene transcription due to the presence of poly(A) signal motifs in retroelements might only partially explain strand-specific detrimental effects in the *D. melanogaster* genome.

Correspondence to: Asher D. Cutter, Institute of Evolutionary Biology, University of Edinburgh, King's Buildings, West Mains Road, Ashworth Labs, Edinburgh EH93JT, UK; email: asher.cutter@ed.ac.uk **Key words:** Transposable elements — *Drosophila melanogaster* — Polyadenylation — Codon usage bias

# Introduction

Across eukaryote genomes, transposable elements (TEs) demonstrate strikingly nonrandom distributions (Kidwell and Lisch 2001). Because insertions can disrupt protein function, TEs are rare in coding regions; their genomic abundances also correlate with local base composition, recombination rate, and gene density depending on the species and the element type (Duret et al. 2000; Bartolome et al. 2002; Medstrand et al. 2002; Martin et al. 2002; Rizzon et al. 2002; Kaminker et al. 2002; Wright et al. 2003). These patterns have been explained by both selective and neutral processes, including biases in the sequence specificity of integration sites, selection for or against transposition into regulatory or coding regions, and selection against TE-mediated ectopic exchange (Hill and Robertson 1966; Charlesworth and Charlesworth 1983; Langley et al. 1988; Jakubczak et al. 1991; Biemont et al. 1997; Smit 1999; Wright et al. 2003; Rizzon et al. 2003).

Recently, another pattern of TE distribution was described in two mammalian genomes: some classes of TEs are oriented predominantly in the opposite direction of nearby genes, a pattern that cannot be explained by selection against ectopic recombination or by neutral insertional biases (Smit 1999; Medstrand et al. 2002; van de Lagemaat et al. 2003; Semon and Duret 2004). Selection against potential interference with local gene regulation is thought to drive orientation bias of TEs toward the noncoding strand (Smit 1999). If the transcription termination signal of a TE impedes normal expression by prematurely terminating host gene transcription, TEs oriented on the sense strand will be more deleterious than TEs oriented in the opposite direction (Smit 1999). This hypothesis was based on the observations in mammalian genomes that (1) retrotransposons that contain termination signal sequences (LINE and LTR retroelements) in introns are biased toward the noncoding orientation, and (2) no orientation bias was observed for DNA-mediated transposable elements, which lack or have only a weak termination signal sequence (Smit 1999; Medstrand et al. 2002; van de Lagemaat et al. 2003).

Although the pattern of TE orientation bias has been documented only in mammalian genomes (Smit 1999; Medstrand et al. 2002; van de Lagemaat et al. 2003), several characteristics make D. melanogaster the choice for investigating the processes that contribute to this pattern. First, D. melanogaster enjoys a rich history of studies on transposable elements that spans more than three decades. Second, the larger effective size of *D. melanogaster* populations (Kreitman 1983; Yu et al. 2004) makes it less likely that nonrandom patterns are due to chance. Third, the availability of dense genetic maps for estimating local recombination rates, a fully sequenced genome, and correlates of gene expression (e.g., microarray fluorescence levels, codon usage bias) allow analysis of genomic features that may influence TE orientation bias. Finally, the genome of D. melanogaster contains ample abundance and diversity of TEs to allow meaningful comparison among the major classes of TEs (Kaminker et al. 2002).

By characterizing the orientations of TEs across the D. melanogaster genome, we find that retroelements within introns are biased toward the noncoding strand. In contrast, DNA-mediated transposable elements in nontranscribed flanking regions occur more frequently on the sense strand relative to neighboring genes. The orientation biases are more pronounced for genes with biased codon usage, suggesting that selection for orientation bias is more effective on TEs near genes that generally respond to weak selection. Orientation biases also are stronger on the X chromosome, although this effect may be confounded by higher levels of codon usage bias and recombination rate among X-linked loci. Additionally, we searched the genome for strand biases near genes for sequences identical to poly(A) signal motifs. Based on the distributions of TEs and poly(A) signal motifs, we conclude that selection against spurious termination of transcription is unlikely to explain in entirety the phenomenon of TE orientation bias.

# Materials and Methods

## Data Acquisition

We retrieved the full sequences and genomic locations of predicted genes in the *D. melanogaster* genome from FlyBase and Ensembl (v.3.2; March 24, 2004; http://www.flybase.org, http://www.ensembl.org). Identities and locations of TEs, as reported by Kaminker et al. (2002), were obtained from the Berkeley Drosophila Genome Project (http://www.fruitfly.org/p\_disrupt/TE.html). Recombination rate estimates for genes (*R*) by Hey and Kliman (2002) were cross-referenced with other data based on gene names. Gene expression data from the microarray study of Stolc et al. (2004) were obtained from the White laboratory web page (http://genome.med.yale.edu/Lifecycle/).

### Feature Statistics

For each TE within 500 bp of a gene, we noted its strand orientation relative to that of its corresponding gene. TE positions were designated as either in introns, in 5' or 3' untranslated regions (UTRs), or in nontranscribed flanking locations up to 500 bp away. TE relationships excluded RNA genes and genes that are nested within the introns of other genes; we also removed from the analysis the few TEs that occur in exons. The presence of genes in the same intron of a given gene-TE pair was allowed, provided that all internal genes were oriented in the same direction as each other; exclusion of such cases did not qualitatively alter the results. For TEs occurring between two genes, we used only the closest gene within 500 bp. We considered three major classes of TE: LTR (long terminal repeat) retroelements, non-LTR LINE (long interspersed nuclear element)-like retroelements, and TIR (terminal inverted repeat) DNA-mediated transposable elements (Kidwell and Lisch 2001; Kaminker et al. 2002). A list of TEs and associated genes used in this study is available in Supplementary Online Table 1.

The frequency of optimal codons  $(F_{op})$  of a gene measures nonrandom codon usage from the relative incidence of codons throughout the genome (Ikemura 1985). Codon usage bias  $(F_{op})$ and G+C content at third-position silent sites for all genes were computed with CodonW (J. Peden; http://www.molbiol.ox.ac.uk/ cu/codonw.html). We used the values of recombination rate (R)estimated by Hey and Kliman (2002). These values were computed by assigning a gene of interest the value of the slope of genetic and physical positions for the nearest 8 genetically mapped loci (493 total mapped loci). Although a number of approaches to estimating recombination rate have been proposed, we chose to use R because it was the preferred metric in the analyses of Hey and Kliman (2002).  $F_{op}$  and recombination rate (R) were binned into high, middle, and low classes based on 33% and 67% percentiles with cutoffs at  $F_{\rm op}=0.45$  and 0.53 and R=0.55 and 2.80 cM/Mb (Ikemura 1985; Hey and Kliman 2002). We also calculated an index of gene expression level for 13,197 genes from microarray data by summing the fluorescence intensities among the six developmental time points sampled in the study by Stolc et al. (2004), followed by log<sub>10</sub> transformation or by 33% and 67% percentile binning. G-tests with William's correction were used to test for equal numbers of coding and noncoding orientation TEs (Sokal and Rohlf 1995). Multivariate analyses were performed using logistic multiple regression models in JMP v.5.0, in which genome features other than orientation and chromosome identity were treated as continuous variables.

To determine the potential role of selection against spurious termination signal sequences on the coding strand of transcribed regions, we compared the frequency of sequence matches to transcription termination signal motifs within or near each gene in the genome with the number on the noncoding strand. Observed numbers on each strand were calculated by scanning for the canonical polyadenylation (transcription termination) signal sequence AATAAA in exonic, intronic, UTR, and 500-bp geneflanking sequences (Zhao et al. 1999). Based on the strand-specific base frequencies  $(f_A, f_T)$  in each genic region, we computed the expected frequency of these motifs (p) as  $f_{\rm A}^5 \cdot f_{\rm T}$ , with incidence  $n \cdot p$ and variance  $n \cdot p \cdot (1 - p)$ , where *n* is the total number of bases in the region. The variance for the ratio of sense-to-antisense motif frequencies,  $var[n_s / n_a]$ , was estimated neglecting the covariance term as:  $\operatorname{var}[n_{\mathrm{s}}] \cdot n_{\mathrm{a}}^{-2} + n_{\mathrm{s}}^{2} \cdot \operatorname{var}[n_{\mathrm{a}}] \cdot n_{\mathrm{a}}^{-4}$  (Lynch and Walsh 1997). Because *n* is very large and the expected number of motifs in a region was very close to its expected variance, we make the Poisson assumption for motifs actually observed that the variance equals the observed number of motifs. We inferred significant differences between the observed and the expected sense:antisense motif ratio from nonoverlapping standard deviations.

## Results

#### Orientation Bias of TEs

We examined orientation biases for three classes of TEs (LTR and LINE-like retroelements and DNAmediated transposable elements) in association with three portions of genes (introns, UTRs, and nontranscribed sequence flanking genes). In the absence of a reasonable mechanism for a local insertional preference by TEs for a particular strand, we tested for deviations from an equal representation of TEs in the sense and antisense orientation with respect to nearby genes. Overall, retroelements (LTR + LINE) occurred less frequently on the coding strand than expected (sense:antisense, 186:228; G = 4.26, P = 0.039). When we partitioned TEs by location relative to genes, LTRs were less common on the coding strand in introns, whereas TIRs were in excess on the coding strand in flanking regions (Fig. 1). LINEs exhibited no significant patterns when divided among genic regions. Also, no significant strand bias was observed for TEs in UTRs, although sample size limits our power to detect an effect (n = 28). Summed across genic regions, LTR orientations followed the trend of bias against the coding direction of neighboring genes observed for all retroelements and for LTRs in introns (105:134, G = 3.52, P = 0.061).

Disruptive effects of TEs on gene expression should be most severe in genes expressed at high levels. Therefore, we analyzed TE orientations in relation to expression level and degree of codon bias  $(F_{op})$ , which correlates with expression level (Duret and Mouchiroud 1999). The overall orientation of LTR and LINE retroelements was directed toward the noncoding strand for genes exhibiting strong

**Fig. 1.** TE orientation bias in introns and nontranscribed flanking regions. Significantly fewer LTRs occur in the coding orientation within introns (sense:antisense, 72:102; G = 5.18, P = 0.023) and more TIRs occur on the coding strand in flanking regions (28:14; G = 4.70, P = 0.030). Numbers above bars are sample sizes. \*P < 0.05.

codon bias (LTRs, 21:57, G = 17.2, P < 0.0001; LINEs, 21:34, G = 3.07, P = 0.08), whereas the orientations of TEs were not significantly biased for genes displaying low  $F_{op}$  values (Fig. 2). The magnitude of bias was particularly strong among intronic LTRs associated with genes with high  $F_{op}$  (16:46, G = 15.0, P = 0.0001; Fig. 2). TIRs in nontranscribed flanking regions also showed stronger orientation bias when near genes with high  $F_{op}$ , but toward the sense strand (13:4, G = 4.87, P = 0.027; Fig. 2). However, we found no significant individual association of our index of gene expression with orientation bias for any TE class (all P's > 0.6).

Differences in the sign and magnitude of selection among linked loci are expected to reduce the efficacy of selection on any given locus (i.e., Hill-Robertson interference [Hill and Robertson 1966; Charlesworth et al. 1993]). Consequently, a greater response to selection should occur in regions of higher recombination, especially for targets of weak selection (as expected for the orientation of TEs). To test for an effect of recombination rate on the degree of TE orientation bias, we divided LTRs into bins according to the level of recombination (R) based on the estimates of Hey and Kliman (2002). We found no significant differences in orientation bias among genes with low to high recombination rates ( $P \ge$ 0.075), although LTRs and TIRs showed a trend in the direction of stronger orientation bias in high recombination regions (LTRs: sense:antisense low  $R \ 17:18 = 49\%$ , medium  $R \ 25:39 = 39\%$ , high R





**Fig. 2.** Orientation bias among genes with respect to codon usage bias ( $F_{op}$ ). Significantly stronger orientation bias is found for genes with strong codon usage bias ( $F_{op} > 0.53$ ; LTRs in introns, sense:antisense 16:46, G = 15.0, P = 0.0001; TIRs in nontranscribed flanking regions 13:4, G = 4.87, P = 0.027; LINEs in all regions 21:34, G = 3.07, P = 0.08). Numbers above bars are sample sizes. \*P < 0.05; \*\*P < 0.001.

24:38 = 39%; TIRs: low R 31:26 = 54%, medium R 16:12 = 57%, high R 17:11 = 61%).

The strength of the orientation bias in introns was stronger for all classes of TEs in genes on the sex chromosome than on the autosomes (Table 1). Each class of TE paralleled this X-to-autosome difference in introns, but the differences were individually nonsignificant: LTRs, 32% X:44% autosome coding orientation  $(n_{\rm X} = 37, n_{\rm autosome} = 137);$ LINEs, 31%:49% ( $n_{\rm X} = 13$ ,  $n_{\rm autosome} = 109$ ); and TIRs, 40%:46% ( $n_{\rm X} = 20, n_{\rm autosome} = 108$ ). All seven cases of TIRs in flanking regions on the X occurred in the same orientation as their neighboring genes (sign test, P = 0.016) versus the 60% coding direction bias seen on autosomes (X > autosome, G = 6.36, one-tailed P = 0.006). Complicating the X-to-autosome analysis, however, is the observation that average levels of codon bias (Wilcoxon  $\chi^2 = 15.3$ , P < 0.0001) and recombination rate (Wilcoxon  $\chi^2 = 53.7$ , P < 0.0001) also are higher on the X chromosome (N. Singh and D. Petrov, pers. commun.).

Because previous studies have demonstrated significant covariation among several of these variables (e.g., Duret and Mouchiroud 1999; Hey and Kliman 2002), we performed logistic multiple regression analyses to infer their partial effects. We constructed stepwise-reduced models from several factors and their first-order interactions for LTRs in introns and for TIRs in flanking regions: X-autosome identity, log-transformed values of  $F_{op}$ , R, expression index, and third position silent G+C content (GC3s). For LTRs in introns  $(n = 115, \chi^2 = 19.3, df = 9,$ P = 0.023), three factors contributed significantly and independently to orientation bias:  $F_{\rm op} (\chi^2 = 6.3,$ P = 0.012), GC3s ( $\chi^2 = 5.0$ , P = 0.026), and the chromosome identity × expression index interaction  $(\chi^2 = 4.6, P = 0.032)$ . The recombination rate  $\times$ GC3s interaction was marginally significant in orientation bias  $(\chi^2 = 3.7,$ explaining LTR P = 0.054), whereas the other five factors included in the reduced model showed no significant independent effects (all P's > 0.1). Fewer variables showed significant independent effects in the model constructed for TIRs in nontranscribed flanking regions (n = 32,  $\chi^2 = 22.2$ , df = 8, P = 0.005). Only recombination rate ( $\chi^2 = 5.3$ , P = 0.022) and, marginally, the GC3s × expression index interaction ( $\chi^2 = 3.7$ , P = 0.054) contributed independently to TIR orientation bias in flanking regions.

# Orientation Bias of Transcription Termination Signal Sequence Motifs

One hypothesis for the existence of orientation bias in retrotransposons predicts that selection acts against the polyadenylation signal sequences of retrotransposons because they might induce termination of host gene transcription (Smit 1999). To explore the extent to which selection might act specifically on such sequence motifs, we characterized the strand bias of the canonical transcription termination signal sequence AATAAA throughout the genome (Zhao et al. 1999). We identified a total of 197,630 matches to this poly(A) signal motif on either strand in the vicinity of genes across the D. melanogaster genome. Overall, the raw counts of these motifs showed unequal abundances on the two strands in introns and untranslated regions (Fig. 3a). However, these different regions also exhibit strand-specific base composition biases (Fig. 3b), which generate unequal expected frequencies of the AATAAA motif on each strand. After accounting for base composition differences between strands, only exons and UTRs showed evidence for departure from the expected distribution of poly(A) motifs (Fig. 3c). Qualitatively identical patterns were also found with (1) a [C/G/T]AATAAA[C/G/T] motif, (2) a GT-rich "DSE" downstream element located 10-60 bp from the motif, or (3) the less common poly(A) signal sequence motif ATTAAA (not shown). We found no significant association of the fraction of spurious termination motifs on the coding strand with either codon bias or recombination rate.

Table 1. TE orientation bias on autosomes and the sex chromosome

Chromosome	Number of intronic TEs			
	Sense oriented	Antisense oriented	Total	to gene neighbor <sup>a</sup>
autosomes	163	191	354	0.460
X	24	46	70	0.343

<sup>a</sup>X < autosomes (G = 3.34, 1-tailed P = 0.034)

# Discussion

# TE Orientation Bias in Drosophila

Transposable elements that replicate via an RNA intermediate (LTRs and LINEs) are biased against orienting in the same direction as nearby genes, principally among elements located in introns (Fig. 1). This suggests that selection against one or more features specific to retroelements leads to their disproportionate elimination from the coding strand of associated genes. While ectopic recombination is likely an important force shaping the distribution of TEs in the genome (Langley et al. 1988; Carr et al. 2002; Petrov et al. 2003), this mechanism cannot plausibly generate the orientation bias of TEs with respect to nearby genes because it operates irrespective of TE orientation. Likewise, it is possible that TEs might insert preferentially in the antisense orientation relative to nearby native genes, but it is not obvious how such a scenario could operate. Smit (1999) accounted for retroelement orientation bias in the human genome by concluding that poly(A) signal sequences of LTRs and LINEs might prematurely terminate transcription of the host gene. However, other studies of mammalian genomes have since identified the same pattern of orientation bias among Alu SINEs, which lack a strong poly(A) signal, and in untranscribed sequences more than 30 kb away from genes (Medstrand et al. 2002). These observations combined with our finding of no strand bias for poly(A) signal motifs in introns demonstrates that TE orientation bias is unlikely to be explained fully by selection against spurious termination of transcription from poly(A) signals. In addition, both mammalian and fly genomes contain some classes of TEs with significant orientation biases in the opposite direction predicted by this interference explanation (van de Lagemaat et al. 2003).

Populations with larger effective sizes should respond more efficiently to selection, especially when fitness effects are weak (Li 1997). Therefore, we expect orientation bias to be stronger in flies than in mammals if the pattern of selection is similar between them, given the larger effective sizes of *D. melanogaster* populations. However, the magnitude of TE orientation bias does not differ dramatically among these species (Smit 1999; Medstrand et al. 2002; van de Lagemaat et al. 2003). This might be due to the higher density of TEs in mammalian genomes in combination with more and longer introns in mammalian genes (Waterston et al. 2002), such that multiple sense-oriented TEs occur within a single gene more often in mammals. Hence, selection against the joint negative effects of multiple sense-oriented TEs in mammalian genes may result in their more efficient removal overall, especially if the combined fitness effects are greater than additive. Mammals and flies also differ in that the pattern of orientation bias among LINEs is weaker in D. melanogaster. This difference for LINEs might reflect a higher rate of large deletions and weaker purifying selection, as observed for Helena LINE elements in Drosophila (Petrov et al. 1996; Petrov and Hartl 1998).

# Gene Expression

The fitness consequences of disrupted gene expression should be most severe for highly transcribed genes. For example, increased usage of preferred codons among highly expressed genes reflects, in part, the ability of these genes to respond to weak selection for expression efficiency (Akashi 1994). We demonstrate that LTR and TIR orientation bias is stronger among genes with more strongly biased codon usage (Fig. 2), and interactions with an index of gene expression also correlate significantly in multivariate analyses. The disproportionate antisense orientation of retroelements in the introns of genes with strong codon usage bias suggests that sense-oriented LTRs have more severe deleterious fitness consequences in highly expressed genes. A gene's codon bias reflects its ability to respond to weak selection, due to the combined influence of expression level and recombination rate; this effect likely underlies the stronger association that TE orientation bias exhibits with codon bias than with expression level alone. The possibility that greater error is associated with estimating expression level than codon usage bias may also contribute to a stronger signal of codon bias.

In contrast to the deficit of retroelements in the sense orientation, we observe an excess of DNA-mediated



Fig. 3. A Observed and expected counts of the poly(A) signal sequence motif AATAAA among gene regions for the sense and antisense strands. B Asymmetric base composition between the sense and the antisense strands for sequence near genes. C Observed and expected strand bias of the AATAAA motif among gene regions. Exonic and UTR regions show significant differences between observed and expected bias. Bars indicate  $\pm 1$  SD.

transposable elements (TIRs) in the sense orientation among genes with strongly biased codon usage (Fig. 2). This suggests either that antisense-oriented TIRs are more strongly deleterious or that such DNAmediated transposable elements might alter gene expression in a nondeleterious manner that is favored by selection (Girard and Freeling 1999; Kidwell and Lisch 2001), implicating sense-oriented TIRs as candidates for surveys of positive selection (Supplementary Online Table 2). The mammalian TEs (LTRs rather than TIRs) that show an orientation bias toward the coding strand in transcript termini (van de Lagemaat et al. 2003) might also provide candidates for positive selection. Furthermore, multiple studies have implicated positive selection on specific TE insertions near genes in Drosophila (Pardue and DeBaryshe 1999; Bartolome et al. 2002; Daborn et al. 2002; Lerman et al. 2003; Schlenke and Begun 2004; Franchini et al. 2004)—although new TE insertions principally have deleterious effects (Montgomery et al. 1987). In at least one case, the inverted repeat of a TIR was implicated as the selective target (Bartolome et al. 2002). In general, the insertion preference of P elements for the 5' ends of genes suggests that such DNA-mediated transposable elements might be predisposed to altering gene regulation (Spradling et al. 1995). For example, additional regulatory elements associated with TIRs might enhance transcription of nearby genes.

Why is the pattern of TIR orientation bias the opposite of that observed for LTRs, given that new insertions by both classes of TE are generally deleterious? One hypothesis is that the TIR elements used in our analysis are less active and more ancient in origin than the LTR elements. The lower abundance of full-length TIR copies in the D. melanogaster genome and the greater degeneration of TIR sequences have lead to the conclusion that extant TIRs are primarily more ancient in origin than LTRs (Kaminker et al. 2002; Lerat et al. 2003), although this issue may be complicated by interactions between TE age and selective constraints or by defective TEs utilizing trans sources of transposase. If current TIRs are in fact more ancient, then the population of extant TIR elements should have conferred nearly neutral or beneficial effects that have managed to persist, because those with strongly deleterious effects will have been eliminated by selection (Li 1997).

# Recombination Rate

The efficiency of selection is reduced in regions of low recombination because of interference among linked sites that have different selection coefficients (Hill and Robertson 1966). Loci subject to weak selection appear to be particularly vulnerable. For example, interference of linkage appears to play an important role in the evolution of codon bias in *D. melanogaster* (Hey and Kliman 2002; but also see Marais et al. 2001, 2003; Kliman and Hey 2003). However, our univariate analyses did not identify a significant effect of recombination rate on TE ori-

entation bias, despite trends for stronger bias among genes associated with LTRs and TIRs in high recombination regions. Multivariate analyses of orientation bias, on the other hand, did distinguish a significant independent effect of recombination rate interaction terms involving or recombination rate. We note that a role for recombination rate as a mediator of selection efficiency might also be implicated by our finding that codon bias is a stronger correlate of TE orientation bias than is our index of expression level. In general, codon bias may reflect the ability of a locus to respond to weak selection by tracking the joint effects of recombination and expression level (Hey and Kliman 2002).

Models to explain the abundance and distribution of TEs across the genome include the "gene disruption model," in which TEs can potentially alter the coding capability or regulation of genes, and the "ectopic recombination model," whereby deleterious chromosome rearrangements are mediated by high TE copy numbers (Montgomery et al. 1987; Charlesworth and Langley 1989; McDonald et al. 1997; Petrov et al. 2003). Only the "gene disruption model" provides a reasonable explanation for the origin of bias against the sense orientation of TEs with respect to nearby genes. Consequently, the observation that TE orientation bias against the sense orientation is stronger among LTRs than LINEs suggests that disruption of gene function might be a more important force shaping the distributions of LTRs than of LINEs in the D. melanogaster genome.

# Strand Bias of Poly(A) Signal Sequence Motifs

A proposed mechanism driving orientation bias in LTR and LINE retroelements is the presence of poly(A) transcription termination signals which are absent or weak in DNA-mediated transposable elements (Smit 1999). The TE poly(A) signal sequence could disrupt host gene expression by prematurely terminating transcription due to spurious recognition of a TE poly(A) signal in a transcribed region. For this scenario to provide a general explanation, poly(A) signal motifs should be subject to negative selection within transcribed regions regardless of any association with TEs. However, we found that canonical poly(A) signal motifs in gene regions were not underrepresented on the coding strand of introns after accounting for strand-specific base composition differences. This observation indicates either that this mechanism of selection is extremely weak or that only more extensive transcription termination signal complexes than the simple AA-TAAA motif are subject to such purifying selection (Zhao et al. 1999). However, we found qualitatively

similar patterns of strand bias when we included a GT-rich downstream element (DSE) within 60 bp of the AATAAA motif.

With the exception of the sense strand in exons, all genic regions contain a greater abundance of poly(A) motifs than expected based on individual base frequencies (Fig. 3A). This observation may reflect the importance of di- and trinucleotide frequencies in DNA sequence patterns (Burge et al. 1992) and the resemblance of the poly(A) sequence motif to a microsatellite (Dieringer and Schlotterer 2003). However, our analysis comparing the relative abundance of the motifs between strands factors out the overall excess of the poly(A) sequence motif in the genome by focusing on strand bias rather than the absolute motif counts.

The strand bias for poly(A) signal motifs is greater than expected in exons and UTRs, after accounting for base composition differences between the strands. The relative rarity of the AATAAA motif on the coding strand in exons probably reflects selection against premature protein truncation via the stop codon TAA. The strand bias in UTRs is consistent with selection against precocious transcription termination signals or other spurious AT-rich regulatory motifs, although the excess in the absolute number of motifs in UTRs makes a conclusive explanation elusive.

Because it is unclear whether selection on poly(A) signal motifs is likely to account fully for TE orientation bias, we propose an additional possible mechanism. Host gene transcripts that include TEs with matched orientations could increase the production of TE-encoded proteins and lead to deleterious TE activity. This could be caused, for example, by shared regulatory motifs or accidental translation of TE genes from the host gene transcript (Errede et al. 1987; Girard and Freeling 1999). Additionally, if TEs are more active when oriented in the same direction as nearby genes, they will be targeted disproportionately by cellular mechanisms of TE silencing (Jiang 2002), potentially leading to disruption of host gene transcription. Consequently, this would result in stronger negative selection on senseoriented TEs.

In conclusion, the nonrandom distribution of TEs throughout eukaryote genomes is indicative of selection shaping their location. We have characterized orientation bias among TEs that occur in or near transcribed regions throughout the *D. melanogaster* genome. These data provide support for a role of selection in shaping differential TE patterns on the sense and antisense strands near genes in *D. melanogaster*. Further investigations on the potential for advantageous effects of TE insertion remain to be explored to more fully understand the dynamics between TEs and their host genome.

# References

- Akashi H (1994) Synonymous codon usage in *Drosophila mela-nogaster*: natural selection and translational accuracy. Genetics 136:927–935
- Bartolome C, Maside X, Charlesworth B (2002) On the abundance and distribution of transposable elements in the genome of *Drosophila melanogaster*. Mol Biol Evol 19:926–937
- Biemont C, Tsitrone A, Vieira C, Hoogland C (1997) Transposable element distribution in *Drosophila*. Genetics 147:1997–1999
- Burge C, Campbell AM, Karlin S (1992) Over-representation and under-representation of short oligonucleotides in DNAsequences. Proc Natl Acad Sci USA 89:1358–1362
- Carr M, Soloway JR, Robinson TE, Brookfield JFY (2002) Mechanisms regulating the copy numbers of six LTR retrotransposons in the genome of *Drosophila melanogaster*. Chromosoma 110:511–518
- Charlesworth B, Charlesworth D (1983) The population-dynamics of transposable elements. Genet Res 42:1–27
- Charlesworth B, Langley CH (1989) The population-genetics of Drosophila transposable elements. Annu Rev Genet 23:251–287
- Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. Genetics 134:1289–1303
- Daborn PJ, Yen JL, Bogwitz MR, et al. (2002) A single P450 allele associated with insecticide resistance in *Drosophila*. Science 297:2253–2256
- Dieringer D, Schlotterer C (2003) Two distinct modes of microsatellite mutation processes: Evidence from the complete genomic sequences of nine species. Genome Res 13:2242–2251
- Duret L, Marais G, Biemont C (2000) Transposons but not retrotransposons are located preferentially in regions of high recombination rate in *Caenorhabditis elegans*. Genetics 156:1661–1669
- Duret L, Mouchiroud D (1999) Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, Arabidopsis. Proc Natl Acad Sci USA 96:4482–4487
- Errede B, Company M, Hutchison CA (1987) Tyl sequence with enhancer and mating-type-dependent regulatory activities. Mol Cell Biol 7:258–265
- Franchini LF, Ganko EW, McDonald JF (2004) Retrotransposongene associations are widespread among *D. melanogaster* populations. Mol Biol Evol 21:1323–1331
- Girard L, Freeling M (1999) Regulatory changes as a consequence of transposon insertion. Dev Genet 25:291–296
- Hey J, Kliman RM (2002) Interactions between natural selection, recombination and gene density in the genes of *Drosophila*. Genetics 160:595–608
- Hill WG, Robertson A (1966) Effect of linkage on limits to artificial selection. Genet Res 8:269–294
- Ikemura T (1985) Codon usage and transfer-RNA content in unicellular and multicellular organisms. Mol Biol Evol 2:13–34
- Jakubczak JL, Burke WD, Eickbush TH (1991) Retrotransposable elements R1 and R2 interrupt the ribosomal-RNA genes of most insects. Proc Natl Acad Sci USA 88:3295–3299
- Jiang YW (2002) Transcriptional cosuppression of yeast Ty1 retrotransposons. Genes Dev 16:467–478

- Kaminker J, Bergman C, Kronmiller B, et al. (2002) The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. Genome Biol 3:research0084.1–20
- Kidwell MG, Lisch DR (2001) Perspective: Transposable elements, parasitic DNA, and genome evolution. Evolution 55:1–24
- Kliman RM, Hey J (2003) Hill–Robertson interference in *Drosophila melanogaster*: reply to Marais, Mouchiroud and Duret. Genet Res 81:89–90
- Kreitman M (1983) Nucleotide polymorphism at the alcoholdehydrogenase locus of *Drosophila melanogaster*. Nature 304:412–417
- Langley CH, Montgomery E, Hudson R, Kaplan N, Charlesworth B (1988) On the role of unequal exchange in the containment of transposable element copy number. Genet Res 52:223–235
- Lerat E, Rizzon C, Biemont C (2003) Sequence divergence within transposable element families in the *Drosophila melanogaster* genome. Genome Res 13:1889–1896
- Lerman DN, Michalak P, Helin AB, Bettencourt BR, Feder ME (2003) Modification of heat-shock gene expression in *Drosophila melanogaster* populations via transposable elements. Mol Biol Evol 20:135–144
- Li W-H (1997) Molecular evolution. Sinauer Associates, Sunderland, MA
- Lynch M, Walsh JB (1997) Genetics and analysis of quantitative traits. Sinauer Associates, Sunderland, MA
- Marais G, Mouchiroud D, Duret L (2001) Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. Proc Natl Acad Sci USA 98:5688–5692
- Marais G, Mouchiroud D, Duret L (2003) Neutral effect of recombination on base composition in *Drosophila*. Genet Res 81:79–87
- Martin E, Laloux H, Couette G, Alvarez T, Bessou C, Hauser O, Sookhareea S, Labouesse M, Segalat L (2002) Identification of 1088 new transposon insertions of *Caenorhabditis elegans*: A pilot study toward large-scale screens. Genetics 162:521–524
- McDonald JF, Matyunina LV, Wilson S, Jordan IK, Bowen NJ, Miller WJ (1997) LTR retrotransposons and the evolution of eukaryotic enhancers. Genetica 100:3–13
- Medstrand P, van de Lagemaat LN, Mager DL (2002) Retroelement distributions in the human genome: Variations associated with age and proximity to genes. Genome Res 12:1483–1495
- Montgomery E, Charlesworth B, Langley CH (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
- Pardue ML, DeBaryshe PG (1999) Drosophila telomeres: two transposable elements with important roles in chromosomes. Genetica 107:189–196
- Petrov DA, Lozovskaya ER, Hartl DL (1996) High intrinsic: Rate of DNA loss in *Drosophila*. Nature 384:346–349
- Petrov DA, Hartl DL (1998) High rate of DNA loss in the Drosophila melanogaster and Drosophila virilis species groups. Mol Biol Evol 15:293–302
- Petrov DA, Aminetzach YT, Davis JC, Bensasson D, Hirsh AE (2003) Size matters: Non-LTR retrotransposable elements and ectopic recombination in *Drosophila*. Mol Biol Evol 20:880–892
- Rizzon C, Marais G, Gouy M, Biemont C (2002) Recombination rate and the distribution of transposable elements in the *Dro-sophila melanogaster* genome. Genome Res 12:400–407
- Rizzon C, Martin E, Marais G, Duret L, Segalat L, Biemont C (2003) 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*. Genetics 165:1127–1135
- Schlenke TA, Begun DJ (2004) Strong selective sweep associated with a transposon insertion in *Drosophila simulans*. Proc Natl Acad Sci USA 101:1626–1631
- Semon M, Duret L (2004) Evidence that functional transcription units cover at least half of the human genome. Trends Genet 20:229–232

- Smit AFA (1999) Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr Opin Genet Dev 9:657–663
- Sokal RR, Rohlf FJ (1995) Biometry. W.H. Freeman and Company, New York
- Spradling AC, Stern DM, Kiss I, Roote J, Laverty T, Rubin GM (1995) Gene disruptions using *P* transposable elements: An integral component of the *Drosophila* genome project. Proc Natl Acad Sci USA 92:10824–10830
- Stolc V, Gauhar Z, Mason C, et al. (2004) A gene expression map for the euchromatic genome of *Drosophila melanogaster*. Science 306:655–660
- van de Lagemaat LN, Landry JR, Mager DL, Medstrand P (2003) Transposable elements in mammals promote regulatory varia-

tion and diversification of genes with specialized functions. Trends Genet 19:530-536

- Waterston RH, Lindblad-Toh K, Birney E, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562
- Wright SI, Agrawal N, Bureau TE (2003) Effects of recombination rate and gene density on transposable element distributions in *Arabidopsis thaliana*. Genome Res 13:1897–1903
- Yu N, Jensen–Seaman MI, Chemnick L, Ryder O, Li W–H (2004) Nucleotide diversity in gorillas. Genetics 166:1375–1383
- Zhao J, Hyman L, Moore C (1999) Formation of mRNA 3' ends in eukaryotes: Mechanism, regulation, and interrelationships with other steps in mRNA synthesis. Microbiol Mol Biol Rev 63:405–445