J Mol Evol (2000) 50:123–130
DOI: 10.1007/s002399910014

JOURNAL OF MOLECULAR EVOLUTION

Disparate Evolution of Paralogous Introns in the *Xdh* Gene of *Drosophila*

Francisco Rodríguez-Trelles, Rosa Tarrío, Francisco J. Ayala

Department of Ecology and Evolutionary Biology, 321 Steinhaus Hall, University of California, Irvine, CA 92697-2525, USA

Received: 26 July 1999 / Accepted: 21 August 1999

Abstract. *Drosophila* nuclear introns are commonly assumed to change according to a single rate of substitution, yet little is known about the evolution of these non-coding sequences. The hypothesis of a uniform substitution rate for introns seems to be at odds with recent findings that the nucleotide composition of introns varies at a scale unknown before, and that their base content variation is correlated with that of the adjacent exons. However, no direct attempt at comparing substitution rates in introns seems to have been addressed so far. We have studied the rate of nucleotide substitution over a region of the *Xdh* gene containing two adjacent short, constitutively spliced introns, in several species of *Drosophila* and related genera. The two introns differ significantly in base composition and substitution rate, with one intron evolving at least twice as fast as the other. In addition, the substitution pattern of the introns is positively associated with that of the surrounding coding regions, evidencing that the molecular evolution of these introns is impacted by the region in which they are embedded. The observed differences cannot be attributed to selection acting differently at the level of the secondary structure of the pre-mRNA. Rather, they are better accounted for by locally heterogeneous patterns of mutation.

Key words: *Drosophila — Xanthine dehydrogenase —* Intron–exon evolution — Within-gene nonuniform mutation pattern — Substitution rate heterogeneity

Correspondence to: Francisco Rodríguez-Trelles c/o Francisco J. Ayala; *e-mail:* ibge2@blues.uab.es

Introduction

Drosophila nuclear introns are commonly assumed to evolve at one universal rate, but the rate of evolution in introns has not been examined carefully (Li 1997, p. 182). The hypothesis of a single rate of substitution for introns can be traced to early work on the base composition of the genome of *Drosophila melanogaster*. It was found that (i) the average composition of introns is similar to that of the genome as a whole (Shields et al. 1988); (ii) the intron base composition is random within and over lineages when one excludes the short sequences required for splicing and processing, enhancers, and other functional motifs (Moriyama and Yartl 1993); and (iii) there is no correlation between the base composition of introns and that of exons for a given gene (Shields et al. 1988; Moriyama and Hartl 1993; Carulli et al. 1993). A correlation between the base composition of introns and their adjacent exons has, however, been detected in more recent studies (Kliman and Hey 1994), which have also uncovered compositional heterogeneity among introns of the same and different genes (Kliman and Eyre-Walker 1998). The relationship between base composition and substitution rate is, nonetheless, complex, and it is generally not feasible to infer variation in substitution rates from information on base composition (Kliman and Hey 1994).

The hypothesis of a universal substitution rate for introns conflicts with evidence that natural selection can constrain intron evolution in a number of ways, including, for example, intron length (Guo et al. 1993) or the overall stability of intron secondary structure (Stephan and Kirby 1993; Kirby et al. 1995; Leicht et al. 1995). This hypothesis is also at odds with the discovery that the local chromatin environment can impact the patterns of

mutation along the genome [e.g., transcribed vs nontranscribed regions, linker vs nucleosomal core, etc. (Boulikas 1992; Holmquist and Filipsky 1994)]. Regional variation in the mutation pattern has been invoked to account in part for the compositional heterogeneity of the *Drosophila* genome (Kliman and Eyre-Walker 1998).

Here we present evidence that proximal introns within the same gene can evolve at substantially different substitution rates. Our experimental system consists of two small (≈55-bp), constitutively spliced introns that are located near each other (about 600 bp apart) within the *Xdh* gene. One intron (intron 2) is pervasive in the *Drosophila* genus and other dipterans (and has a homologous position as an intron found in humans and other diverse organisms). The other intron (intron B) has recently been acquired by duplication of intron 2 and has a phylogenetic distribution confined to the lineages of *D. willistoni* and *D. saltans* (Tarrío et al. 1998). Hence, introns 2 and B are replicated sequences that have evolved within two local molecular environments of the *Xdh* gene, for at least some 30 million years [the estimated time for the divergence of the *saltans* and *willistoni* lineages (Powell and DeSalle 1995)]. Our observations are best accounted for by invoking mutation patterns that are locally heterogeneous.

Materials and Methods

The introns and exons considered in this study belong to the *Xdh* gene and are investigated in 22 species of Dipterans (shown in Figs. 1A and B), including 19 species of *Drosophila*, plus *Scaptodrosophila lebanonensis*, *Chymomyza amoena*, and *Ceratitis capitata* (GenBank accessions AF058984, AF093217, AF093218). *Drosophila* species include representatives of the *Sophophora* subgenus belonging to the *melanogaster* group [*D. melanogaster* (Keith et al. 1987); Y00307], *obscura* group [*D. pseudoobscura* (Riley 1989); M33977], *willistoni* group (*D. equinoxialis, D. paulistorum, D. willistoni, D. tropicalis, D. insularis, D. sucinea, D. capricorni*, and *D. nebulosa;* GenBank accessions AF093207, AF093208, AF093206, AF093209, AF093210, AF093211, AF093212, AF093213), and *saltans* group (*D. saltans, D. prosaltans, D. neocordata, D. emarginata, D. sturtevanti*, and *D. subsaltans;* GenBank accessions AF058978, AF058979, AF058982, AF058981, AF058983, AF058980) and of the *Drosophila* subgenus (*D. virilis, D. pictiventris,* and *Zaprionus tuberculatus;* GenBank accessions AF093215, AF093214, AF093216).

The *Xdh* region we have sequenced consists of 2085 coding bp plus an intron [positions 940 to 3306 in *D. melanogaster* (Keith et al. 1987)] and comprises about half of exon II (1113 bp), intron 2 (ranging from 55 to 72 bp, with the exception of *D. melanogaster*, 281 bp), and most of exon 3 (972 bp). We have presented evidence that intron B (52 to 66 bp long, located 606 bp upstream intron 2) has been acquired by duplication of intron 2 within exon II. Its phylogenetic distribution is restricted to the species of the *saltans* and *willistoni* groups (Tarrío et al. 1998). Two additional recently inserted introns (A and C), also presumably originated by duplication, are not considered in this study since they are found in very few species and thus cannot lead to reliable inferences (see Tarrío et al. 1998). Details on the amplification and sequencing primers and strategy are given by Tarrío et al. (1998).

Substitution rates in introns and exons are estimated by maximum-likelihood methods under the model of Hasegawa et al. (1985; HKY85) [as implemented in the program PAML, vs 1.3 (Yang 1997)]. Tests for uniformity of substitution rates among regions are conducted using the maximum-likelihood ratio statistic, which, under the null model, is asymptotically distributed as a χ^2 with 1 degree of freedom. This approach requires prior choice of the regions to be compared (Gaut and Weir 1994), which can be biologically ambiguous when there is no clear reason to choose regions (Hartmann and Golding 1998). This is not a problem with introns 2 and B, for they can clearly be ascribed into two different categories of rates. In the case of the coding regions, the potential drawback of prior choice is circumvented by first using the test of Hartmann and Golding (1998). This method uses a random permutation test for comparing maximum-likelihood scores and allows the detection of any rate heterogeneity present in the sequences by means of a sliding window (Hartmann and Golding 1998).

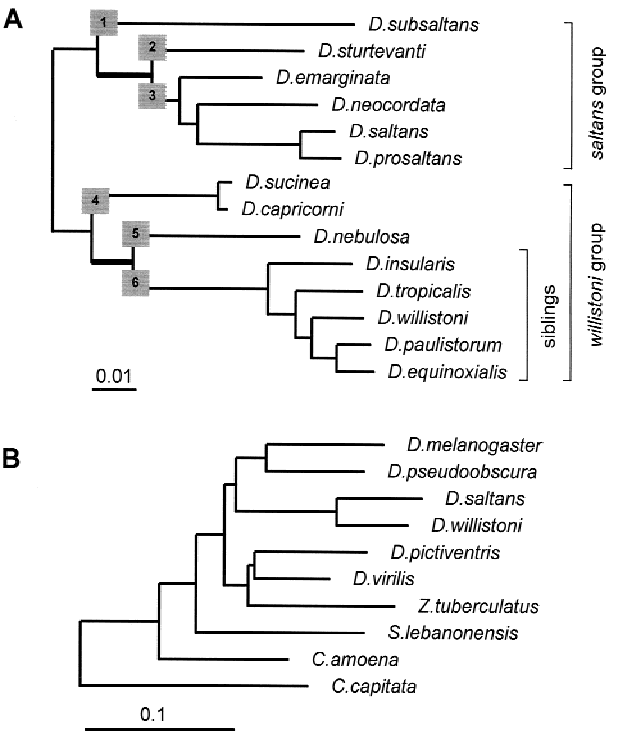

Fig. 1. The tree topologies used in this study **(A)** for 14 species of the *saltans* and *willistoni* groups and **(B)** for the remaining species plus *D. saltans* and *D. willistoni.* Branch lengths were obtained from the *Xdh* nucleotide sequences with the reversible model [Yang 1994, 1997 (program PAML 1.3)], allowing different nucleotide frequencies and transition/transversion rate ratios and assuming gamma-distributed rates among sites (eight rate categories) and different fixed rates at codon positions.

The species investigated in this study are part of a hierarchically structured phylogeny, thereby treating them as statistically independent observations may lead to overestimation of the significance level in hypothesis testing (Felsenstein 1985). The variable degree of relatedness among species seemingly has little importance for sequence similarity in the case of introns: except for a few comparisons involving the most closely related taxa (e.g., the *willistoni* siblings; see Fig. 1A). Phylogenetic differences among species are, anyhow, apparent in the coding regions. In order to alleviate phylogenetic inertia, tests for regional differences in base composition are conducted by means of a two-way analysis of variance (ANOVA) including the main effects of region (2 and B) and species group (*saltans* and *willistoni*), with the GC content as the dependent variable. By factoring out the species group, we control for the fraction of the total variation ascribable to phylogenetic differences between species groups. In addition, we use, within species groups, supraspecific taxonomic categories as follows (see Fig. 1A): within the *saltans* group we consider the GC content of *D. subsaltans* and *D. sturtevanti* and the GC content across *D. emarginata, D.*

Table 1. GC content (%) of introns 2 and B and their surrounding coding regions (25 nt each side) in 14 species of *Drosophila*[a]

	Intron		Coding region 2				Coding region B			
Species	2	B	1st	2nd	3rd	GC_4[b]	1st	2nd	3rd	GC_4[c]
willistoni group										
D. willistoni	30.2	21.7	46.0	44.0	46.0	37.6	76.0	32.0	48.0	36.9
D. tropicalis	35.5	30.8	42.0	44.0	50.0	56.3	74.0	32.0	54.0	52.6
D. paulistorum	30.5	23.4	44.0	44.0	42.0	37.6	74.0	32.0	46.0	42.2
D. equinoxialis	30.6	29.1	44.0	42.0	42.0	37.6	72.0	34.0	38.0	42.2
D. insularis	31.8	27.6	42.0	44.0	36.0	31.3	74.0	32.0	42.0	36.9
D. sucinea	35.4	26.8	46.0	44.0	52.0	50.0	76.0	34.0	52.0	52.6
D. capricorni	35.4	28.6	46.0	44.0	50.0	50.0	76.0	32.0	52.0	52.6
D. nebulosa	25.5	24.6	48.0	42.0	50.0	37.5	78.0	30.0	50.0	36.8
Saltans group										
D. saltans	27.8	12.2	44.0	44.0	46.0	31.3	78.0	30.0	34.0	21.0
D. prosaltans	27.7	12.2	44.0	44.0	42.0	31.3	76.0	32.0	38.0	21.0
D. emarginata	27.3	12.3	42.0	44.0	54.0	50.1	78.0	32.0	36.0	42.1
D. neocordata	28.1	9.5	42.0	44.0	50.0	31.3	76.0	30.0	34.0	31.6
D. sturtevanti	32.8	14.8	46.0	44.0	50.0	37.6	72.0	32.0	38.0	36.9
D. subsaltans	29.0	14.8	42.0	44.0	50.0	31.3	70.0	36.0	40.0	36.9

[a] 1st, 2nd, and 3rd refer to nucleotide positions in codons.
[b] GC content estimates based on 16 fourfold positions.
[c] GC content estimates based on 19 fourfold positions.

neocordata, *D. saltans*, and *D. prosaltans* (branches 1, 2, and 3 in Fig. 1A); within the *willistoni* group we consider the average GC content of *D. sucinea* and *D. capricorni*, the GC content of *D. nebulosa*, and the average GC content across the five *willistoni* siblings (branches 4, 5, and 6 in Fig. 1A, respectively). In this way, the original 14-species data set is reduced to six taxa; the three taxonomic units defined for each species group can be assumed to have evolved independently (i.e., according to a star-like phylogeny), except for the two thicker branches shown in Fig. 1A. These branches are comparatively short, thereby they are expected to have little influence on the tests results. ANOVAs were conducted with the ANOVA module of the STATISTICA 5.0 (1996) package. Frequency data were angularly transformed before the analyses.

Association between the GC content of introns and their surrounding coding regions is investigated by means of Felsenstein's (1985) pairwise independent contrast test. Contrast tests take into account correlations among sequences induced by phylogenetic differences across species. Contrast tests are performed with the Contrast program in the computer package Phylip 2.5 (Felsenstein 1993). Because they are highly conserved across species, the first six and the last two nucleotides of the intron sequences are not included in analyses concerning base composition.

For the likelihood-ratio tests and Felsenstein's (1985) contrast tests, we use two trees shown in Fig. 1: one for the analyses including the species of the *saltans* and *willistoni* groups (Fig. 1A) and another for the analyses including the remaining species plus *D. saltans* and *D. willistoni* (Fig. 1B). The phylogenetic hypotheses are supported by data of several sorts [reviewed by Powell (1997); see Rodríguez-Trelles et al. (1999*a*) for the species of the *saltans* group]. Branch lengths and rate parameter estimates are obtained using the maximum-likelihood method on the *Xdh* sequences, under the general reversible model with different substitution rate and nucleotide frequency parameters for first, second, and third codon positions and discrete gamma-distributed (with eight categories of rates) substitution rates across sites.

Predicted folding free energies are calculated with the RNAstructure program [version 2.51 (Mathews and Burkard 1997, and references therein)]. Folding stability of pre-mRNA structures is examined by comparison of observed free energies with those obtained from 100 random permutation of the same sequences. Specific searchers for putative stems in introns were conducted using the BESTSTEM program (Parsch et al. 1998). Each intron sequence is converted to a reverse complement and slided along the unconverted sequence to find the strongest RNA pairing stem. Each pairing score (see Parsch et al. 1998) is then compared with the distribution of pairing scores found between the unconverted sequence and 500 random permutations of its corresponding converted sequence.

Results

Introns 2 and B. Table 1 shows the GC content (in percent) of introns 2 and B in the 14 species of the *saltans* (6 species) and *willistoni* (8 species) groups. As inferred from two-way ANOVA, intron 2 has a significantly higher average GC content (29.4%) than intron B (19.3) [$F(1,8) = 34.20$, $p < 0.001$], and the *willistoni* group has a higher GC content in both introns than the *saltans* group [$F(1,8) = 20.17$, $p \approx 0.002$]. The interaction term, intron × species group, is significant [$F(1,8) = 16.11$, $p \approx 0.004$], reflecting that the extent of the difference in average GC content between species groups is larger in intron B (26.2 vs 10.0 for the *willistoni* and *saltans* species, respectively) than in intron 2 (30.9 vs 27.5). Variances are homogeneous (Levene's test, $p \approx 0.46$), meaning that the degree of the GC content variation is about the same across species groups and introns.

Larger compositional differences between species groups in intron B than in intron 2 suggest that intron B has changed faster than intron 2 since the time of the duplication. In order to test this, we performed a log-likelihood ratio test to the null model that the two introns evolve at equal rates (see Gaut and Weir 1994; Yang 1994). For this test a minimum of three sequences is required (Yang 1994). We conducted the test on the five sibling species of the *willistoni* subgroup (see Fig. 1A), because this is the only data subset including more than

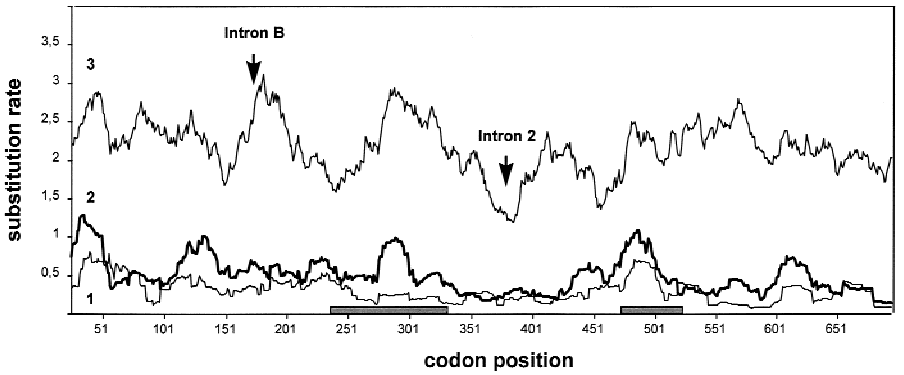

Fig. 2. Estimated substitution rates for sites in the first, second, and third codon positions along 695 codons of the *Xdh* gene. Estimates are obtained by maximum likelihood assuming the topology shown in Fig. 1B under the general reversible model [Yang 1994, 1997 (program PAML 1.3)], assuming gamma-distributed rates among sites (eight categories of rates) and allowing different nucleotide frequencies and transition biases for codon positions. *Arrows* indicate the locations of introns 2 and B. *Gray bars* at the *bottom* indicate the known position of binding domains for the *Xdh* enzyme.

two sequences for which it is possible to obtain a reliable alignment for each intron; the HKY85 (Hasegawa et al. 1985) model and the topology shown in Fig. 1A are assumed. According to the test, intron B is evolving almost twice as fast as intron 2 (1.79 ± 0.52 vs 1, respectively; $-2\log\Lambda = 4.04$, $p < 0.05$). Aside from the *willistoni* siblings, it is possible to obtain a reliable alignment of the intron sequences only for the closely related *D. saltans* and *D. prosaltans* (see Fig. 1A). A similar pattern emerges: these two species have diverged about two times faster in intron B than in intron 2 (nucleotide distances based on the HKY85 correction are 0.1563 vs 0.0492, for introns B and 2, respectively).

Coding Regions Surrounding Introns 2 and B. Table 1 displays the percentage GC content of the first, second, and third codon positions and fourfold degenerate sites of the 50 codons (25 on each side) region surrounding intron 2 (denoted "coding region 2"; CR2), and intron B ("coding region B"; CRB). Similarly as with the introns, average GC content differences between species groups are larger in CRB than in CR2; differences are detected by ANOVA only for third codon positions [interaction CR × species group, $F(1,8) = 13.64$, $p \approx 0.006$], which can be accounted by stronger functional constraints operating on first and second codon positions, and in the case of the fourfold degenerate class, because of the fewer number of sites included in the analysis (16 and 19 for CR2 and CRB, respectively). Also, as for the introns, CR2 has significantly greater average GC content than CRB in third codon positions [$F(1,8) = 9.20$, $p \approx 0.016$].

Figure 2 shows the substitution rates along the *Xdh* region for first, second, and third codon positions. The general reversible model (Tavaré 1986) with discrete gamma rates for sites is used, rates are obtained using the phylogeny shown in Fig. 1B, and increased depth of the phylogeny is expected to improve the power of the analysis. Vertical arrows indicate the location of introns B and 2. Clearly, the coding region around the point where intron B has been inserted occupies a peak of substitution rate, whereas intron 2 is placed in a local valley, where substitution rates are slow.

In order to test the statistical significance of this result, we first conducted a "blind" search for rate heterogeneity with the test of Hartmann and Golding (1998) separately for first, second, and third codon positions; in this test, the substitution rate for regions within a sliding window is compared against the average substitution rate for the whole sequence. According to this test, the region comprising 12 codons upstream and downstream of intron B (i.e., from codon 157 to codon 181) is evolving the fastest ($p > 0.10$), and the region stretching 15 codons upstream and 17 codons downstream of intron 2 is evolving the slowest ($p < 0.06$), in third codon positions. Also, CR2 is identified as the second slowest evolving (codons 317 to 422; $p < 0.09$) in first positions and CRB as the fastest evolving (16 to 240; $p < 0.07$) in second positions.

The test of Hartmann and Golding (1998) is particularly suited when an a priori hypothesis for the regional rate variation of the sequences is not available; otherwise, the likelihood-ratio test is more powerful, especially for the less variable first and second codon positions (Gaut and Weir 1994). Likelihood-ratio tests that CR2 and CRB evolve at the same rate as the average sequence (i.e. the substitution rate for all the *Xdh* coding sequence except CR2 or CRB) are conducted separately for each codon position; we use regions previously identified by the test of Hartmann and Golding (1998) in third codon positions, and the HKY85 (Hasegawa et al. 1985) model is assumed. In all three positions, CR2 evolves significantly slower than the average sequence ($-2\log\Lambda = 6.62$ and $p \sim 0.01$, $-2\log\Lambda = 5.42$ and $p < 0.05$, and $-2\log\Lambda = 25.82$ and $p \ll 0.001$ for the first, second, and third codon positions, respectively), and CRB evolves faster, but in this case the differences are significant only for third codon positions ($-2\log\Lambda = 17.76$, $p \ll 0.001$). Maximum likelihood-estimated averages of the number of substitutions per site in CRB relative to CR2 (CRB:CR2), shown graphically in Fig. 3, are in the proportions 2.22:1, 1.32:0.39, and 22.36:6.39, for

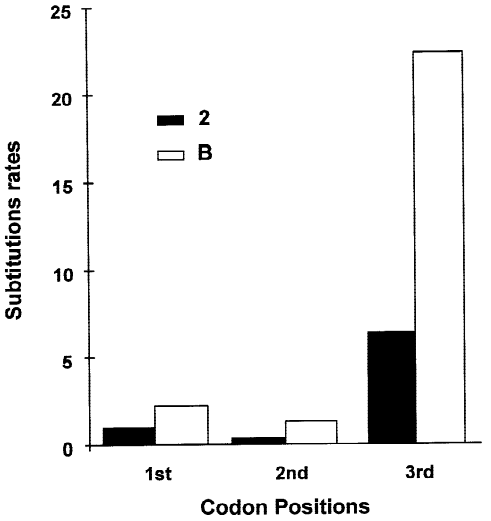

Fig. 3. Estimated substitution rates for the first, second, and third codon positions in the coding regions surrounding introns 2 (*open bars*) and B (*filled bars*). Taxa are the same as in Fig. 2. Estimates are for the regions identified with the test of Hartmann and Golding (1998) in third codon positions: 12 codons upstream and downstream of intron B and 15 codons upstream and 17 downstream of intron 2. Rates are obtained by maximum likelihood under the Hasegawa et al. (1985) model [program PAML 1.3 (Yang 1997)] assuming six categories of rates (for the first, second, and third codon position of each codon).

the first, second, and third codon positions, respectively (estimates obtained simultaneously under the HKY85 model with six rate parameters). These results remain qualitatively similar removing *D. saltans* and *D. willistoni,* which means that the coding region surrounding intron B was already highly variable before the insertion of this intron.

Correlated Patterns in Introns and Exons. Despite the different rates of evolution and base composition, the GC content variation of intron 2 is positively correlated with the GC content variation of intron B [Felsenstein's (1985) contrast test, $r_C = 0.63$, $p = 0.02$; Spearman's nonparametric correlation, $r_S = 0.76$, $p \approx 0.001$].

In spite of the relatively short time elapsed (~30 My) since the origin of the *saltans* and *willistoni* lineages, the GC content in fourfold degenerate sites of the CR2 and CRB is correlated with the GC content of their corresponding introns ($r_C = 0.62$, $p = 0.02$, and $r_C = 0.61$, $p = 0.03$, for the comparison CR2–intron 2 and CRB–intron B, respectively; Spearman's correlations, $r_S = 0.53$, $p \approx 0.05$, and $r_S = 0.74$, $p \approx 0.002$, respectively). Similar analyses comparing the first, second, and third codon positions of CR2 and CRB with their respective introns are nonsignificant (except for the Spearman's coefficient for the comparison between intron B and third codon positions of CRB: $r_S = 0.77$, $p \approx 0.001$). Presumably, this lack of correlation is the case because of the lower level of variation of codon positions. Indeed, a strong correlation between the GC content of intron 2 and CR2 emerges when computing the correlation over a wider phylogenetic spectrum (i.e., considering the 10-species data set in Fig. 1B, $r_C = 0.77$ and $p \approx 0.02$, $r_C = 0.71$ and $p \approx 0.03$, $r_C = 0.76$ and $p \approx 0.02$, and $r_C = 0.68$ and $p \approx 0.04$ for the first, second, and third codon positions and fourfold degenerate sites; $r_S = 0.62$ and $p \approx 0.05$, $r_S = 0.62$ and $p \approx 0.05$, $r_S = 0.85$ and $p \approx 0.001$, and $r_S = 0.66$ and $p \approx 0.03$, respectively).

Discussion

We have shown that two introns can display disparate substitution patterns even when they are located rather proximally to each other within the same gene. In addition, the evolutionary dynamics of introns 2 and B exhibits a striking parallelism with the surrounding coding regions. Heterogeneous base composition among introns within genes, correlated with the composition of adjacent exons, has been detected in the genome of *D. melanogaster* (Kliman and Hey 1994; Kliman and Eyre-Walker 1998), although in these previous studies it was not determined whether the substitution rates vary between introns.

Intron sequences are most frequently assumed to evolve in neutral or nearly neutral fashion (Moriyama and Hartl 1993). Still, there are ways that intron sequences may be functionally constrained. For example, the overall size of introns might be of functional relevance (Leicht et al. 1993). In fact, introns 2 and B are both short in the 14 representatives of the *saltans* and *willistoni* lineages, with sizes fluctuating within a narrow range (52 to 66 bp). It is, however, unlikely that size constraints on neighboring introns that are about the same length can lead to such disparate rates of substitution; in addition, the size of intron 2, whose sequence evolves the slowest, changes greatly outside *saltans* and *willistoni* [up to 527 bp in *D. subobscura* (Comeron and Aguadé 1995)].

Another way introns might be constrained is owing to the presence of regulatory motifs (Leicht et al. 1995). Conservation of regulatory signals is especially evident in genes that are alternatively spliced in different tissues (Leicht et al. 1993, 1995; Clark et al. 1996). This might be the case for intron 2, for its sequence is fairly conserved. Yet, apart from the 5′ and the 3′ splice sites, it is not possible to line up intron 2 sequences outside the closely related *willistoni* siblings; also, intron 2 has been lost in the moth *Calliphora viccina* (Houde et al. 1989), suggesting that it is not essential for the appropriate functioning of the gene. In any case, there is not any evidence of either alternative splicing or regulatory features that would account for the different evolutionary rates of introns 2 and B.

Introns might also be constrained in order to maintain the RNA secondary structure (Stephan and Kirby 1993; Kirby et al. 1995). There is growing evidence that sec-

Table 2. Folding free energy analysis for introns 2 and B

Species	$LG_O - LG_E$[a] Intron 2	Intron B
willistoni group		
D. paulistorum	+1.1	+3.3*
D. equinoxialis	+0.4	+1.4
D. willistoni	+1.9	+1.0
D. tropicalis	−0.8	+1.8
D. insularis	+2.0	+3.1*
D. sucinea	+0.9	+0.5
D. capricorni	−1.1	+1.9
D. nebulosa	+0.2	+2.4
saltans group		
D. saltans	+4.0*	+2.6
D. prosaltans	+2.7	−2.0
D. neocordata	+1.3	0.0
D. emarginata	+2.8	−2.3
D. sturtevanti	+4.6*	+0.1
D. subsaltans	−0.5	+1.5

[a] The difference between the observed free folding energy value of each intron sequence (LG_O) and the corresponding expected value (LG_E) averaged over 100 random permutations of the actual intron sequence.

* Statistically significant at the 0.05 level.

ondary structure of the pre-mRNA plays a role in the selection of splice sites (e.g., D'Orval et al. 1991; Goguel and Rosbash 1993). For short introns like the ones in the present study, RNA folding can likely give two types of structural configurations: (i) hairpins, created by the pairing of complementary sequences (stems); and (ii) open, less tightly folded structures. Hairpins relevant functionally can be maintained by epistatic selection for the retention of the form of the stems of these structures (Stephan and Kirby 1993). Thus, mutations occurring in a pairing region of a stem are individually deleterious if they destabilize the structure. But fitness can be restored when a compensatory mutation occurs that reestablishes the pairing potential (Stephan and Kirby 1993; Kirby et al. 1995). Accordingly, regions within stems are expected to show a reduced rate of evolution relative to unpaired regions (Stephan 1996). This model has been proposed to explain the retention of a stem structure in the introns of the *Adh* gene of *Drosophila* (Stephan and Kirby 1993). Similarly, the presence of a conserved hairpin in intron 2 might account for the reduced substitution rate of this intron.

In order to examine this possibility we have determined the predicted folding free energy of each intron in the 14 representatives of the *saltans* and *willistoni* groups, using the RNAstructure program (see Materials and Methods) (Mathews and Burkard 1997). Differences between observed free energies and the expected average for a random sequence are shown in Table 2. Overall, introns 2 and B exhibit a tendency to avoid stable stem structures (11 species of 14 show a higher free energy than expected in intron 2, and 12 of 14 in intron B); however, there are not obvious differences in the way each intron departs from the random expectation. This observation is not affected by GC content differences among the introns, since the randomization procedure does not alter the base composition of the RNA sequences.

Because a hairpin is a local phenomenon, it can go undetected by the folding analysis. Therefore, we have made a more specific search for putative stems using the BESTSTEM program (Parsch et al. 1998) (see Materials and Methods). We have conducted searches for introns 2 and B of the five representatives of the *willistoni* species subgroup, the same for which different substitution rates between introns were inferred. In accordance with the excessively unpaired structure of the introns, evidence for putative pairing stems cannot be detected. This conclusion is likely to be correct given that introns are very short and that the splice sites, including the 5′, the 3′, and the branchpoint, show a tendency to elude the pairing regions of hairpins, something probably related to the accession by the spliceosome (Solnik 1985; Eperson et al. 1988); both circumstances greatly circumscribe the positions available for a stem, increasing the likelihood of its detection if any were present.

The excess of free energy coupled to the virtual absence of stems suggests that natural selection has maintained introns 2 and B in a fairly open structure. The fact that both introns are quite different, also varying widely across species, indicates that this constraint is not reflected on a site-by-site basis but, rather, is a global property of whole sequences. A similar structural behavior of the RNA has been reported for the introns of the sarcomeric myosin alkali light gene of *Drosophila* (Leicht et al. 1993, 1995; Clark et al. 1996). Current models of DNA evolution are unable to account for this class of constraint. However, as long as introns 2 and B behave similarly in this respect, it is unlikely that natural selection acting differently on the pre-mRNA secondary structure can account for the different rates of evolution of the introns, nor for their different average GC contents. The latter could be explained if we assume that, instead of having originated by duplication of intron 2, intron B was originally richer in AT and has gained an extra AT rich string in the common ancestor of the *saltans* lineage (see Table 1). If this were true, however, one would not expect to find a similar compositional pattern in the surrounding coding region. Ultimately, this scenario would leave unexplained why intron B changes faster, at least within the *willistoni* lineage.

Natural selection does not appear to be responsible for the differences in substitution rates observed between the coding regions surrounding the introns. In *Xdh* it is known that the number of amino acid replacement substitutions shows a relative reduction in most regions where the binding domains of the enzyme have been located (Comeron and Aguade 1996); both coding re-

gions, CR2 and CRB, nonetheless fall outside any putative functional domain (Fig. 2). Also, the possibility that CR2 and CRB are subject to different constraints at the protein level can not explain the differences in third codon positions. These could be due to changes in codon usage related to translational accuracy and/or efficiency. However, first, codon bias is weak in *Xdh*, so that codon usage is not expected to vary much along the sequence (Riley 1989), and second, synonymous and nonsynonymous substitution rates are not correlated with codon usage in this gene (Comeron and Aguade 1996). Alternatively, differences in third codon positions could be mediated by differences in the secondary structure of the mRNA. But it is not obvious how these differences could lead to a pattern of covariation with introns. Equally unlikely seems the hypothesis that natural selection affects the fraction of neutral mutations in first, second, and third codon positions proportionally in CR2 and CRB (Fig. 3).

The inference that the coding region in the vicinity of intron B was already changing rapidly before the insertion of the intron strongly suggests that the increased substitution rate of intron B responds to causes other than the peculiarities of the intron sequence itself. Most likely these causes are related to systematic differences in the pattern of point mutation along the *Xdh* gene. This hypothesis can account for the low substitution rate in region 2; the relative increase in region B in the first, second, and third codon positions and the intron; and, also, the correlated patterns of base composition between introns and exons in both regions. As expected for relatively unconstrained sites, this correlation is more apparent for fourfold degenerate sites of CR2 and CRB and emerges for the first, second, and third codon positions after increasing the depth of the phylogenetic analysis. This hypothesis might also explain previously reported heterogeneous synonymous substitution rates along the *Xdh* region (Comeron and Aguade 1996). The heterogeneity was attributed to different selective constrains on the secondary structure of the mRNA, but the authors did not take into account intron data (see Comeron and Aguade 1996).

Evidence of heterogeneity in the mutation pattern over the short segment involved in our study (~600 bp) was previously unknown in *Drosophila*. It cannot be attributed to differences in the GC content of the regions. It is well known that the rate of mutation from G/C to A/T is greater than that from A/T to G/C (Li et al. 1984), so more mutations are expected in GC-rich sequences than in GC-poor sequences. Higher substitution rates should therefore be observed in region CR2, which, to the contrary, changes more slowly than CRB. Other factors must, therefore, be important. There is a growing body of evidence indicating that the local environment of the chromatin can critically influence the rates of DNA damage and its repair, thus leading to variable patterns of mutation along the genome (Boulikas 1992; Holmquist and Filipski 1994). For example, DNA in the nucleosome core seems to be more protected from damaging agents than in the linker or totally naked DNA (Boulikas 1992); also, there is experimental evidence that the nuclear DNA of eukaryotes is nonuniformly protected by proteins maintaining chromosome structure (Kladde and Simpson 1994) and that the DNA sequence position in the nucleus can affect the efficiency of reparation (Holmquist and Filipski 1994). It might, therefore, be reasonable to speculate that CRB had a less protected, more exposed configuration than CR2 already long before intron B was acquired. This very circumstance might indeed have increased the chance for the intron to be inserted. Once this event took place, the greater propensity for damage of the region impacted the evolution of the intron B sequence, leading to the pattern currently observed. Conversely, CR2 might be embedded into a more protected region, which also might explain why this intron has persisted within this *Xdh* region from a remote ancestor into mammals and flies.

It might be noted that, for the arguments above to be valid, it is not necessary to assume that intron B originated by duplication of intron 2. Yet the scenario depicted by this study makes more plausible the duplication hypothesis. For example, it provides an answer to the question of why the sequence of intron B drifted toward a higher AT content after it originated by duplication from intron 2. Interestingly, we have recently shown that the GC mutation pressure has experienced a shift toward AT in the common ancestor of the *saltans* and *willistoni* lineages, presumably intensified in *saltans* (Rodríguez-Trelles et al. 1999*b*,*c*). A trivial answer to the question above may, thus, be that introns 2 and B are equally constrained, but intron B has accumulated more AT because it mutates faster.

Two general questions arise. First, since they were first discovered, introns have been placed into the black box of unconstrained, fast-changing sequences used as reference when investigating selection in other regions of the genome. If, however, weak selection generally holds for introns, it may be profoundly misleading to view intron sequences as fast evolving. In the case at hand, a rather different picture of the evolution of the coding region around intron B would emerge, depending on whether one used intron 2 or intron B for the comparison. Second, whereas our results appear better to fit a model of locally heterogeneous mutation patterns, this by no means implies that selection is unimportant. In fact, a general correlation between the location of genes and that of blocks with different mutation patterns remains a possibility to explore in the future.

Acknowledgments. We are grateful to B.S. Gaut, R.R. Hudson, C.A. Machado, M. Santos, and A. Tatarenkov for suggestions and critical discussion. F.R.-T. has received support from Ministerio de Educacion y Cultura (Spain) (Contrato de Reincorporación) and Grant PB96-1136 to A. Fontdevila. This work was supported by National Institutes of Health Grant GM42397 to F.J.A.

References

Boulikas T (1992) Evolutionary consequences of nonrandom damage and repair of chromatin domains. J Mol Evol 35:156–180

Carulli JP, Krane DE, Hartl DL, Ochman H (1993) Compositional heterogeneity and patterns of molecular evolution in the *Drosophila* genome. Genetics 134:837–845

Casane D, Boissinot S, Chang BH-J, Shimmin LC, Li W-H (1997) Mutation pattern variation among regions of the primate genome. J Mol Evol 45:216–226

Clark AG, Leicht BG, Muse SV (1996) Length variation and secondary structure of introns in the mlc1 gene in six species of *Drosophila.* Mol Biol Evol 13:471–482

Comeron JM, Aguadé M (1996) Synonymous substitutions in the *Xdh* gene of *Drosophilia:* Heterogeneous distribution along the coding region. Genetics 144:1053–1062

D'Orval BC, Carafa YD, Sirand-Pugnet P, Gallego M, Brody E, Marie J (1991) RNA secondary structure repression of a muscle-specific exon in Hela cell nuclear extracts. Science 252:1823–1828

Eperon LP, Graham IR, Griffiths AD, Eperon IC (1988) Effects of RNA secondary structure on alternative splicing of pre-mRNA: Is folding limited to a region behind the transcribing RNA polymerase? Cell 65:393–401

Eyre-Walker A (1994) DNA mismatch repair and synonymous codon evolution in mammals. Mol Biol Evol 11:88–98

Felsenstein J (1985) Phylogenies and the comparative method. Am Nat 125:1–15

Felsenstein J (1993) PHYLIP, Phylogeny inference package, v. 3.5c. University of Washington, Seattle

Filipski J (1987) Correlation between molecular clock ticking, codon usage, fidelity of DNA repair, chromosome banding and chromatin compactness in germline cells. FEBS Lett 217:184–186

Filipski J (1988) Why the rate of silent substitutions is variable within the a vertebrate's genome. J Theor Biol 134:159–164

Gaut BS, Weir BS (1994) Detecting substitution-rate heterogeneity among regions of a nucleotide sequence. Mol Biol Evol 11:620–629

Goguel V, Rosbash M (1993) Splice site choice and splicing efficiency are positively influenced by pre-mRNA intramolecular base pairing in yeast. Cell 72:893–901

Guo M, Lo PCH, Mount SM (1993) Species-specific signals for the splicing of a short intron in vitro. Mol Cell Biol 13:1104–1118

Hartmann M, Golding GB (1998) Searching for substitution rate heterogeneity. Mol Phylogenet Evol 9:64–71.

Hasegawa M, Kishino H, Yano T (1985) Dating the human-ape splitting by a molecular clock of mitochondria DNA. J Mol Evol 22: 160–174

Holmquist GP (1994) Chromatin self-organization by mutation bias. J Mol Evol 39:436–438

Holmquist GP, Filipski J (1994) Organization of mutations along the genome: A prime determinant of genome evolution. Trends Ecol Evolut 9:65–69

Houde M, Tiveron M, Bregegere F (1989) Divergence of the nucleotide sequences encoding xanthine dehydrogenase in *Calliphora vicina* and *Drosophila melanogaster*. Gene 85:391–402

Keith TP, Riley MA, Kreitman M, Lewontin RC, Curtis D, Chambers G (1987) Sequences of the structural gene for xanthine dehydrogenase (rosy locus) in Drosophila melanogaster. Genetics 116:67–73

Kirby DA, Muse SV, Stephan W (1995) Maintenance of pre-mRNA secondary structure by epistatic selection. Proc Natl Acad Sci USA 92:9047–9051

Kladde MP, Simpson RT (1994) Positioned nucleosomes inhibit *dam* methylation in vivo. Proc Natl Acad Sci USA 91:1361–1365

Kliman RM, Eyre-Walker A (1998) Patterns of base composition within the genes of *Drosophila melanogaster*. J Mol Evol 46:534–541

Kliman RM, Hey J (1994) The effects of mutation and natural selection on codon bias in the genes of *Drosophila.* Genetics 137:1049–1056

Leicht BG, Lyckegaard EMS, Benedict CM, Clark AG (1993) Conservation of alternative splicing and genomic organization of the Myosin Alkali Light-Chain *Mlc1* gene among *Drosophila* species. Mol Biol Evol 10:769

Leicht BG, Muse SV, Hanczye M, Clark AG (1995) Constrains on intron evolution in the gene encoding the myosin alkali light chain in *Drosophila.* Genetics 139:299–308

Li W-H (1997) Molecular evolution. Sinauer Associates, Sunderland, MA

Li W-H, Wu C-I, Luo C-C (1984) Nonrandomness of point mutations reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J Mol Evol 21:58–71

Mathews DH, Burkard ME (1997) RNA structure *vs.* 2.51. Distributed by the authors at http://www.rna.chem.rochester.edu/RNAstructure.html

Moriyama EN, Hartl DL (1993) Codon usage bias and base composition of nuclear genes in *Drosophila.* Genetics 134:847–858

Parsch J, Stephan W, Tanda S (1998) Long-range base pairing in *Drosophila* and human mRNA sequences. Mol Biol Evol 15:820–826

Powell JR, DeSalle R (1995) *Drosophila* molecular phylogenies and their uses. Evol Biol 28:87–138

Riley M (1989) Nucleotide sequence of the *Xdh* region in *Drosophila pseudoobscura* and an analysis of the evolution of synonymous codons. Mol Biol Evol 6:33–52

Rodríguez-Trelles F, Tarrío R, Ayala FJ (1999a) Molecular evolution and phylogeny of the Drosophila saltans species group inferred from the Xdh gene. Mol Phylogenet Evol (in press)

Rodríguez-Trelles F, Tarrío R, Ayala FJ (1999b) Switch in codon bias and increased rates of amino acid substitution in the *Drosophila saltans* species group. Genetics 153:339–350

Rodríguez-Trelles F, Tarrío R, Ayala FJ (1999c) Fluctuating mutation bias and the evolution of base composition in *Drosophila.* J Mol Evol (in press)

Shields DC, Sharp PM, Higgins DG, Wright F (1988) "Silent" sites in *Drosophila* genes are not neutral: Evidence of selection among synonymous codons. Mol Biol Evol 5:704–716

Stephan W (1996) The rate of compensatory evolution. Genetics 144: 419–426

Stephan W, Kirby DA (1993) RNA folding in *Drosophila* shows a distance effect for compensatory fitness interactions. Genetics 135: 97–103

Solnick D (1985) Alternative splicing caused by RNA secondary structure. Cell 43:667–676

StatSoft Inc. (1996) STATISTICA for Windows. Tulsa, OK

Tarrío R, Rodríguez-Trelles F, Ayala FJ (1998) New *Drosophila* introns originate by duplication. Proc Natl Acad Sci USA 95:1658–1662

Tavaré S (1986) Some probabilistic and statistical problems on the analysis of DNA sequences. In: Lectures in mathematics in the life sciences, Vol 17, pp 57–86

Yang Z (1994) Estimating the pattern of nucleotide substitution. J Mol Evol 39:105–111

Yang Z (1997) PAML: A program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–556