DOI: 10.1534/genetics.104.037689

Insertion/Deletion and Nucleotide Polymorphism Data Reveal Constraints in *Drosophila melanogaster* Introns and Intergenic Regions

Lino Ometto, Wolfgang Stephan and David De Lorenzo[1]

Section of Evolutionary Biology, Biocenter, University of Munich, D-82152 Planegg-Martinsried, Germany

Manuscript received October 18, 2004
Accepted for publication December 5, 2004

ABSTRACT

Our study of nucleotide sequence and insertion/deletion polymorphism in *Drosophila melanogaster* noncoding DNA provides evidence for selective pressures in both intergenic regions and introns (of the large size class). Intronic and intergenic sequences show a similar polymorphic deletion bias. Insertions have smaller sizes and higher frequencies than deletions, supporting the hypothesis that insertions are selected to compensate for the loss of DNA caused by deletion bias. Analysis of a simple model of selective constraints suggests that the blocks of functional elements located in intergenic sequences are on average larger than those in introns, while the length distribution of relatively unconstrained sequences interspaced between these blocks is similar in intronic and intergenic regions.

NONCODING DNA constitutes a considerable fraction of the genome of eukaryotes. Despite being often referred to as "junk DNA," there is mounting evidence for its potential functions. Introns can play a role in alternative splicing and exon shuffling (Sharp 1994; Hanke *et al.* 1999) and—in some cases—their pre-mRNA secondary structure can affect gene expression (Chen and Stephan 2003; Hefferon *et al.* 2004). Regulatory elements are present in the immediate 5′ neighborhood of genes (*i.e.*, TATA and CG boxes), but they can also modulate gene expression from a greater distance to the target gene (*i.e.*, enhancers and transcription-factor binding sites). Regulatory elements can also reside in introns (*e.g.*, Bergman and Kreitman 2001). Indeed, evidence for selective constraints in noncoding DNA has been found in whole-genome comparisons in Caenorhabditis (*e.g.*, Shabalina and Kondrashov 1999), mammals (*e.g.*, Dermitzakis *et al.* 2002), and Drosophila (Bergman and Kreitman 2001). Matrix attachment regions and *cis*-regulatory elements have also been recognized as targets of purifying selection (Ludwig and Kreitman 1995; Glazko *et al.* 2003).

A recent analysis of polymorphic insertions and deletions in *Drosophila melanogaster* noncoding DNA revealed an overall ratio of deletion-to-insertion events of 1.35 (referred to as polymorphic deletion bias or PDB; Comeron and Kreitman 2000). The authors hypothesized that this deletion bias must be compensated by selection to maintain minimum intron length and generally favor longer introns to enhance recombination. The polymorphism data they used to substantiate their claim were from 31 genomic regions (with very different recombination rates), from multiple sources (generated in various labs by restriction mapping, SSCP, and DNA sequencing) and multiple sampling locations (with very different sample sizes).

A broad range of PDB estimates is found in the literature. In a survey of sequence length diversity in the *Adh* region of *D. pseudoobscura*, Schaeffer (2002) observed a PDB of 0.83 for all insertion/deletion (indel) types (including repetitive ones such as microsatellites), and of 1.89 for nonrepetitive indels (calculated from his Table 1). Similarly, Parsch (2003) reported a ratio of fixed deletions to insertions of 1.66 in a comparison of orthologous introns among species of the *D. melanogaster* subgroup. On the other hand, studies of "dead-on-arrival" non-LTR retrotransposons in Drosophila (Petrov and Hartl 1998; Blumenstiel *et al.* 2002) found deletion-to-insertion ratios ranging from ~4 to 8. The differences among the polymorphic deletion bias estimates are most likely due to different samples, sequences, and methods used in these studies. However, disagreements may also derive from the way repetitive indels are treated. Only Schaeffer (2002) distinguished between repetitive and nonrepetitive indels.

In this study, we used nucleotide sequence data from a single population of *D. melanogaster* from Africa to revisit the various hypotheses concerning deletion bias and its consequences. Our data consist of short fragments (introns and intergenic sequences) from regions of normal recombination on the X chromosome. These fragments are of similar length (~500 bp); *i.e.*, the introns belong to the large size class (>90 bp; see Mount *et al.* 1992; Stephan *et al.* 1994). They were previously analyzed for patterns of nucleotide diversity (generally using a sample of 12 chromosomes) and divergence (to

[1]*Corresponding author:* Section of Evolutionary Biology, Biocenter, University of Munich, Grosshaderner Strasse 2, D-82152 Planegg-Martinsried, Germany. E-mail: delorenzo@lmu.de

a single *D. simulans* line; Glinka *et al.* 2003). This analysis suggested that the African population is close to equilibrium between mutational forces and genetic drift. For these reasons, this sample is particularly suitable for analyzing the selective constraints in introns and intergenic regions (which are expected to fall into the realm of weak selection).

MATERIALS AND METHODS

Drosophila data set: To reduce the possible constraints due to the presence of complex transcription-factor binding sites, we use here only the intergenic regions from the original data set that are at least 1 kb away from the 5′-UTR of an annotated gene (based on FlyBbase 3.0 release, retrieved by the Apollo tool; http://www.flybase.org). Similarly, to avoid potential problems due to the specific location of the fragments within introns (*e.g.*, presence *vs.* absence of splicing elements), we excluded partial introns. The data set meeting the above criteria consists of 22 intergenic regions and 54 introns with average lengths (standard error, SE) of 561.1 bp (61.0) and 492.1 bp (128.4), respectively (excluding deletions and insertions; sample size and fragment lengths are available in the online supplementary Table 3 at http://www.genetics.org/supplemental/).

Analysis of insertion and deletion variation: Insertions and deletions segregating in *D. melanogaster* were polarized according to the state observed in *D. simulans*. Only indels for whom the reconstruction of the ancestral state was unambiguous (*i.e.*, those in which one of the two *D. melanogaster* variants was also present in *D. simulans*) were used in this study. Insertions and deletions were classified into two categories (modified from Schaeffer 2002): (i) nonrepetitive and (ii) repetitive (duplications and mononucleotide and microsatellite repeats). Indels containing repeated DNA sequences have been treated separately, as their expansion/contraction dynamics may produce homoplasy and different numbers of repeats may be added (deleted) at the same location in separate events. We follow here Schaeffer's (2002) suggestion, since the discrepancies among the PDB estimates may derive from the definition of indels. Only Schaeffer (2002) classified indels in different categories (repetitive and nonrepetitive), while Comeron and Kreitman (2000) grouped complex indels (*i.e.*, repetitive ones) and counted them as one event. Nucleotide and indel diversity π (Tajima 1983) and Tajima's *D* (Tajima 1989) statistic were estimated using the program NeutralityTest, kindly provided by H. Li (available at http://hgc.sph.uth.tmc.edu/neutrality_test). Divergence was analyzed using DnaSP 4.0 (Rozas *et al.* 2003).

Modeling of selective constraints: To understand how the distribution of selectively constrained regions in intergenic and intronic sequences can relate to the observed pattern of insertions and deletions, we analyzed simple models of sequence constraints. We assume that a sequence consists of subsequences delimited by functionally constrained blocks (*i.e.*, exons, transcription-factor binding sites, or regulatory regions). In this way, the model can apply to both introns and intergenic regions. Deletions and insertions are considered neutral if they do not alter the block structure (*i.e.*, if they do not fall into a functionally important region) and, because of their size, if they are meeting the spacing constraints between consecutive blocks (Figure 1). Otherwise, deletions and insertions are subjected to strong purifying selection and thus eliminated from the population very shortly after they appear.

We used an approach similar to that described in Ptak and Petrov (2002) to calculate the following statistics: (i) the fraction of deletions and insertions that do not interfere with the functional constraints, (ii) the fraction of these deletions and insertions ≤10 bp, and (iii) the resulting deletion-to-insertion ratio. These values were calculated as a function of the length L of a given subsequence and of its maximum (L_{max}) and minimum (L_{min}) lengths tolerated (reflecting spacing constraints). Then, the fraction of insertions of length S, $f_{ins}(S)$, that do not interfere with the constraints is

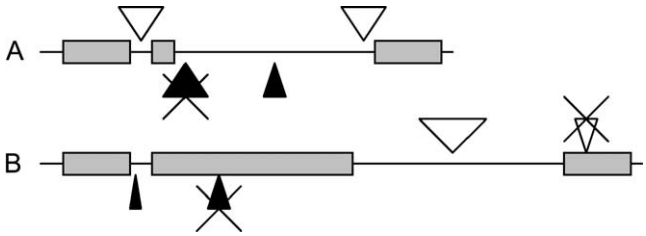

Figure 1.—Schematic of the model of selective constraints considered in the analysis. Subsequences are delimited by blocks (shaded boxes) of coding (exons) or noncoding functional DNA (*e.g.*, regulatory regions or splicing elements). Deletions (solid triangles) are deleterious when they overlap with constrained blocks (crossed-out triangles), while both insertions (open triangles) and deletions may be subjected to purifying selection if they alter spacing constraints (*i.e.*, length of subsequence).

$$f_{ins}(S) = \begin{cases} 1, & \text{if } L + S \le L_{max} \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

Similarly, for deletions we have

$$f_{del}(S) = \begin{cases} \dfrac{L - S + 1}{L}, & \text{if } L - S \ge L_{min} \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

To vary length (spacing) constraints, we define

$$L_{min} = L(1 - \gamma) \quad \text{and} \quad L_{max} = L(1 + \delta),$$

where $0 \le \gamma, \delta < 1$.

It is evident that the smaller L is, the fewer indels will be neutral; moreover, the closer L_{max} and L_{min} are to L (*i.e.*, the more that spacing constraints are present), the higher will be the fraction of small indels.

In applying this model to our data we have to take into account that our fragments may contain subsequences of different lengths, each with possibly specific spacing constraints. For simplicity, we consider only two length classes of subsequences, "short" and "long" ones, and we compute the indel statistics on the basis of the fraction of short *vs.* long subsequences (thus varying sequence composition). Let F_{short} be the proportion of short sequences in the total sequence ($0 < F_{short} < 1$) and let $f_{indel,s}(S)$ and $f_{indel,l}(S)$ be the fractions of indels of size S that do not interfere with the constraints of short and long sequences, respectively. The fraction of indels of size S that does not interfere with any sequence constraint is then given as

$$f_{indel}(S) = F_{short} f_{indel,s}(S) + (1 - F_{short}) f_{indel,l}(S),$$

where we substitute for $f_{indel,s}(S)$ and $f_{indel,l}(S)$ the right-hand sides of Equations 1 and 2 for insertions and deletions, respectively.

The statistics are then computed using Equations 1–5 of Ptak and Petrov (2002), based on the indel size distributions of Petrov and Hartl (1998). Here we rely on the assumption that the size distributions of deletions and insertions of

TABLE 1

Analysis of polymorphic insertions (ins) and deletions (del) in noncoding DNA of *D. melanogaster*

	Introns					Intergenic regions				
	n[a]	PDB[b]	Av. size (SE)[c]	Av. freq. (SE)[d]	% ≤10 bp[e]	n[a]	PDB[b]	Av. size (SE)[c]	Av. freq. (SE)[d]	% ≤10 bp[e]
Nonrepetitive:	62	2.00	8.94	0.244	0.73	26[g]	2.17	10.00	0.219	0.56
del			(1.13)	(0.033)				(1.19)	(0.044)	
DNA indels:	31[f]	(1.06–2.05)	6.32	0.354	0.81	12	(0.62–2.38)	5.33	0.421	0.83
ins			(1.54)	(0.047)				(2.09)	(0.103)	
Wilcoxon test										
Z			−2.122	0.304				−2.823	0.274	
P			0.034	0.761				0.005	0.784	
All indels										
del	108	0.92	6.06	0.268	0.83	41[g]	0.69	6.83	0.248	0.71
			(0.60)	(0.024)				(1.00)	(0.038)	
ins	118[f]	(0.62–1.91)	3.33	0.382	0.94	59	(0.52–1.72)	3.10	0.483	0.95
			(0.58)	(0.027)				(0.52)	(0.040)	
Wilcoxon test										
Z			2.988	−1.515				2.975	−2.779	
P			0.003	0.130				0.003	0.005	

[a] Number of polymorphic events.
[b] Polymorphic deletion bias, ratio between the number of observed deletions and insertions. The minimum and maximum values observed per fragment are given in parentheses. Note that these were calculated only when at least one insertion and one deletion were available.
[c] Average size in base pairs; standard error is given in parentheses.
[d] Average frequency of the indel event; standard error is given in parentheses.
[e] Fraction of indels ≤10 bp.
[f] One insertion of 132 bp was excluded.
[g] One deletion of 113 bp was excluded.

Petrov and Hartl (1998) are the result of neutral processes. Finally, it should be noted that this analysis refers to the data set as a whole rather than to a single fragment. As Table 1 indicates, the values of PDB across fragments may be rather different.

RESULTS AND DISCUSSION

Introns and intergenic regions show a similar polymorphic deletion bias: When all indels are considered, the values of PDB are <1 for both introns and intergenic regions, in agreement with Schaeffer (2002; Table 1). For the nonrepetitive indels we find PDB values of 2.00 and 2.17 for introns and intergenic regions, respectively, in line with Schaeffer (2002). The lower value (1.35) obtained by Comeron and Kreitman (2000) is most likely the result of the way repetitive indels were counted in their study.

Insertions have smaller sizes and higher frequencies than deletions: Deletions are significantly larger than insertions (Figure 2 and Table 1). If we exclude very large indels (one insertion in an intergenic fragment and one deletion in an intron, both >100 bp), nonrepetitive deletions are larger than insertions in both intergenic regions and introns (Wilcoxon test, $P = 0.005$ and $P = 0.034$, respectively; unless indicated, this test is used in all comparisons). Including these two indels, deletions are still significantly larger than insertions in intergenic regions, but not in introns (data not shown). When repetitive indels are included, the difference is even more significant ($P < 0.005$ for both comparisons).

A consequence of both the higher rate and larger size of deletions is that, in the absence of other forces, a spontaneous loss of DNA should occur. Is this loss compensated? When we average the frequency of each independent indel in the sample, we note that insertions are in higher frequency than deletions (Table 1). In intergenic regions, this difference is significant when all indels are considered ($P = 0.005$). Similarly, in introns, insertions tend to have higher average frequencies than deletions ($P = 0.162$). These results suggest that insertions in both introns and intergenic regions have a higher probability of fixation than deletions, to compensate for the deletion bias by favoring longer regions of noncoding DNA. This agrees with Parsch (2003), who proposed that large insertions are positively selected to restore the optimal intron length.

Estimates of indel and nucleotide sequence variation: We estimated the average indel diversity π and divergence per nucleotide site, considering indels as binary characters of length 1 bp (*i.e.*, presence *vs.* absence of the derived state; for polarization, see above). To estimate divergence, we used the fixed indels observed between

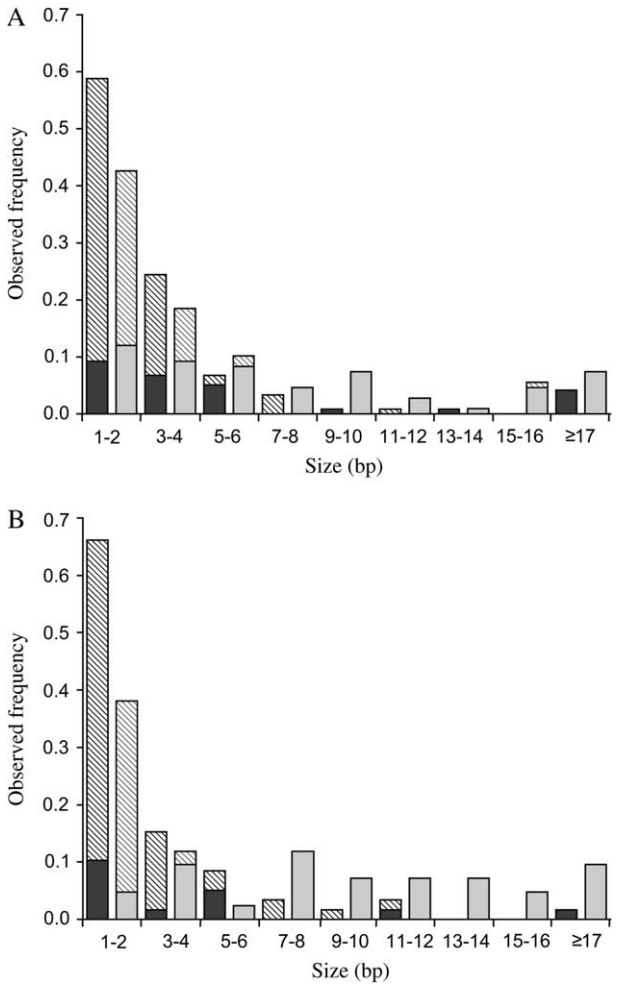

FIGURE 2.—Size distribution of insertions (solid bars) and deletions (shaded bars) in (A) introns and (B) intergenic regions. The solid portions correspond to nonrepetitive indels.

the two species. Introns and intergenic regions show similar values for both nonrepetitive and all indels, except that divergence is higher in introns than in intergenic regions (Table 2). There are considerable differences in average nucleotide diversity π between introns and intergenic regions. Intergenic regions are less polymorphic and diverged than introns although these differences are not significant (Table 2). This is in line with recent observations by KERN and BEGUN (2005). Furthermore, the frequencies (SE) of derived variants at polymorphic nucleotide sites are significantly higher in introns than in intergenic regions: 0.291 (0.009) and 0.261 (0.013), respectively ($P = 0.02$).

Introns, but not intergenic sequences, are larger in *D. melanogaster* than in *D. simulans*: We observed a significant excess of introns that are longer in *D. melanogaster* than in *D. simulans* (39 *vs.* 15, $P = 0.0015$; two-tailed sign test); to be conservative, two introns with equal lengths in both species were counted as if they were smaller in *D. melanogaster*. In intergenic regions, however, no difference was found (12 *vs.* 10, $P = 0.832$). Both observations agree with COMERON and KREITMAN's (2000) analysis.

The observed differences between introns and intergenic regions may be due to either different mutational patterns or different selective pressures. Indeed, some studies provide evidence of transcription-coupled repair mechanisms and transcription-associated mutations (TAM) that could lead to specific mutational patterns in introns. This effect is well known in bacteria and yeast (AGUILERA 2002). In higher eukaryotes, it has been observed only in genes transcribed in mammalian germline cells, where a bias in base composition rather than in substitution rate is observed (GREEN *et al.* 2003; COMERON 2004). In Drosophila, no evidence has been found for transcription-coupled repair (DE COCK *et al.* 1992; SEKELSKY *et al.* 2000), although TAM has been recently proposed as a possible cause of compositional bias observed in introns (KERN and BEGUN 2005).

The following argument suggests, however, that the observed length differences of introns (but not intergenic regions) between *D. melanogaster* and *D. simulans* are probably due to selection rather than mutation. First, introns have a higher (nonrepetitive) indel divergence than intergenic regions (Table 2). This means that either more insertions have been fixed in introns of *D. melanogaster* or more deletions are in those of *D. simulans*. Second, PDB estimates for introns and intergenic regions are comparable (Table 1). Therefore, something other than mutation must have caused the observed difference in fixed indel divergence between intronic and intergenic sequences.

Analysis of selective constraints: The presence of functional elements and/or specific spacing constraints can severely affect polymorphism and divergence patterns. For example, enhancers contain several transcription-factor binding sites separated by spacers with strong length constraints (*e.g.*, LUDWIG *et al.* 1998). Furthermore, PTAK and PETROV (2002) suggested that the large difference between PDB observed in introns and in dead-on-arrival non-LTR retrotransposons was due to splicing constraints in introns, causing many deletions (particularly the larger ones) to be deleterious and be removed by purifying selection. Hence, our finding that intergenic regions show a similar PDB value to that of introns indicates that our intergenic fragments may contain a considerable number of regulatory elements under selective constraints. Several putative transcription-factor binding sites were indeed identified using TRANSFAC (WINGENDER *et al.* 2000) and MatInspector (QUANDT *et al.* 1995) tools. Their density (number of hits per base pair) does not differ from that of introns (data not shown).

To characterize these constraints and relate them to the observed insertion/deletion pattern, we modeled sequences with a certain proportion of functional non-

coding DNA (*e.g.*, exons, regulatory regions; see Figure 1) and calculated the resulting equilibrium deletion and insertion profiles. We assumed that our sequences consist of subsequences delimited by functionally constrained blocks. Preliminary analyses indicated that subsequences of equal (or similar) length are not compatible with our data, independent of the amount of constraints (some examples are provided in online supplementary Figure 4 available at http://www.genetics.org/supplemental/). This suggests the presence of short and long subsequences with variable length constraints in our fragments.

To model spacing constraints, we considered two contrasting scenarios, in which the short subsequences have either strong (str) or relaxed (rel) spacing constraints, while only relaxed constraints are present in long subsequences. For the analyses presented here, we assume in the str scenario $\delta = 0.1$ and $\gamma = 0$ for the short subsequence and $\delta = \gamma = 0.3$ for the long subsequence. In the rel scenario, $\delta = \gamma = 0.2$ for both subsequences (for the definition of these parameters, see MATERIALS AND METHODS). We chose these parameters according to the results reported in supplementary Figure 4, to obtain theoretical results in close agreement with the observed indel profile. Using $\delta = \gamma \geq 0.2$ in both subsequences or $\gamma = 0$ in the short ones results in indel profiles equivalent to the rel and str scenarios, respectively.

As shown in Figure 3A, the theoretical results differ according to both sequence composition (*i.e.*, the fraction of short *vs.* long subsequences) and spacing constraints. Depending on whether the short subsequences are under relaxed or strong length constraints, we obtain remarkably contrasting patterns in PDB and the fraction of deletions ≤10 bp. When ~85% of the subsequences are short and have strong constraints, we obtain theoretical values close to those observed in both introns and intergenic regions (see Table 1). The indel profiles obtained using short sequences of length ≤50 bp and long sequences ≥100 bp are similar to those presented. This suggests that the majority of the subsequences in our fragments are indeed short and have strong length constraints.

Our theoretical results also provide evidence that the number of functional elements should not be considered as a direct measure of the amount of constraints. Rather, it is the combined effect of spacing constraints and the proportion of the functional DNA (*i.e.*, the number and spatial extension of the functional elements) that limits the number of neutral mutations (Figure 1). The presence of spacing constraints poses a limit to the number of indels (but not nucleotide substitutions) that can accumulate in the subsequence. Figure 3B gives the proportion of indels that contribute to the polymorphic indel profile, *i.e.*, the expected indel diversity. Since we observed similar indel polymorphism π in intergenic

TABLE 2

Nucleotide and indel diversity in intergenic regions and introns of *D. melanogaster*

	Nucleotide data			Indel data					
				Nonrepetitive indels				All indels	
	π (SE)	Divergence (SE)	Tajima's D (SE)	π (SE)	Divergence[a]	Tajima's D (deletions)[a]	Tajima's D (insertions)[a]	π (SE)	Divergence[a]
Intergenic regions	0.010 (0.001)	0.052 (0.005)	−0.744 (0.110)	0.0009 (0.0001)	0.0062	−0.822	−0.297	0.0026 (0.0004)	0.0129
Introns	0.012 (0.001)	0.064 (0.004)	−0.526 (0.065)	0.0011 (0.0002)	0.0082	−0.527	−0.359	0.0027 (0.0003)	0.0132

Unless indicated, the average (SE) across loci is given.
[a] Fragments were lumped before analysis.

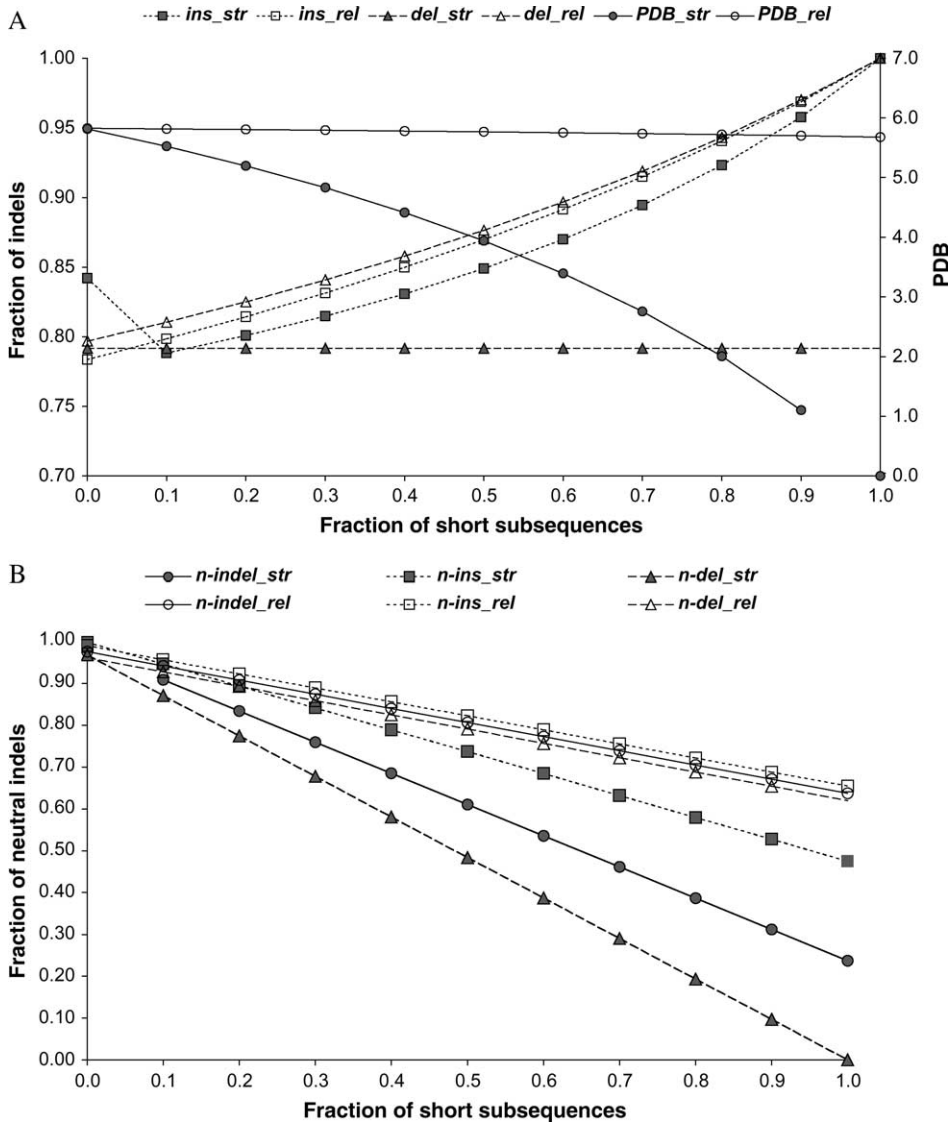

FIGURE 3.—Modeling the insertion and deletion profile in the presence of varying functional constraints. (A) Theoretical results for the fraction of insertions (ins) and deletions (del) ≤10 bp and the polymorphic deletion bias (PDB). (B) Fraction of insertion (*n*-ins), deletion (*n*-del), and deletion and total indel (*n*-indel) events that do not alter functional DNA blocks and spacing constraints. We assume that under neutrality the ratio of deletions to insertions is 6:1 and that there are equal size distributions for insertions and deletions (PETROV and HARTL 1998; BLUMENSTIEL *et al.* 2002). The short and long subsequences have lengths of 30 and 200 bp, respectively, and are subjected to relaxed (rel) or strong (str) spacing constraints (see text for details).

and intronic sequences, spacing constraints seem to be comparable in the two genomic regions.

The low nucleotide sequence diversity and divergence observed in intergenic regions can be understood by noting that the number and spatial extension of functional elements are sources of distinct constraints. In introns, the branch point (which mediates the formation of the lariat structure during splicing) is—strictly defined—only 1 nucleotide long and defines two subsequences, including a short one of 20–30 bp that is under strong spacing constraints (MOUNT *et al.* 1992; *e.g.*, Figure 1A). On the other hand, a large regulatory element can determine two equivalent subsequences, separated by a large functionally important sequence (*e.g.*, Figure 1B). While the indel profile is similar in the two situations, the different proportions of functional DNA may affect the number and pattern of nucleotide substitutions and may result in contrasting diversity values. Thus, because our intronic and intergenic regions have similar PDB values and similar fractions of small indels, they may have similar subsequence structures. In contrast, our nucleotide sequence data (Table 2) suggest that intergenic regions host a larger proportion of constrained DNA, *i.e.*, larger functional elements.

Our simple model of sequence constraints is based on the assumption that a subsequence is completely unconstrained, yet delimited by sequence blocks under very strong purifying selection. However, the following observations suggest that this model needs to be used with care. First, we found evidence that compensatory insertions are under weak positive selection to maintain the proper spacing and structure of regulatory elements, which in turn are often negatively affected by the large and numerous deletions. Second, the observed pattern of Tajima's *D* values also suggests that the sequences are under weak selection pressures. *D* is more negative for both single-nucleotide polymorphisms and deletions in intergenic regions than in introns (Table 2). While the observed excess of rare indels and nucleotide variants, leading to an overall negative Tajima's *D*, is likely

the result of population expansion (Glinka *et al.* 2003), the more negative value observed for deletions (than for nucleotide variation) may reflect the action of purifying selection. On the other hand, the less negative Tajima's *D* value for insertions is consistent with weak positive selection (discussed above). This pattern is more pronounced in intergenic regions than in introns. The introns analyzed belong to the large size class (Mount *et al.* 1992; Stephan *et al.* 1994), very different from the small and most common length class of 61 ± 10 bp (Yu *et al.* 2002). Our observations suggest that these introns evolve in a (nearly) neutral fashion.

We thank J. Parsch for valuable discussions, the reviewers for constructive comments on a previous version of this article, and the Deutsche Forschungsgemeinschaft for funding (grant STE 325/6).

LITERATURE CITED

Aguilera, A., 2002 The connection between transcription and genomic instability. EMBO J. **21:** 195–201.

Bergman, C. M., and M. Kreitman, 2001 Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences. Genome Res. **11:** 1335–1345.

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

Chen, Y., and W. Stephan, 2003 Compensatory evolution of a precursor messenger RNA secondary structure in the *Drosophila melanogaster Adh* gene. Proc. Natl. Acad. Sci. USA **100:** 11499–11504.

Comeron, J. M., 2004 Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence. Genetics **167:** 1293–1304.

Comeron, J. M., and M. Kreitman, 2000 The correlation between intron length and recombination in Drosophila: dynamic equilibrium between mutational and selective forces. Genetics **156:** 1175–1190.

de Cock, J. G., A. van Hoffen, J. Wijnands, G. Molenaar, P. H. Lohman *et al.*, 1992 Repair of UV-induced (6–4) photoproducts measured in individual genes in the *Drosophila* embryonic Kc cell line. Nucleic Acids Res. **20:** 4789–4793.

Dermitzakis, E. T., A. Reymond, R. Lyle, N. Scamuffa, C. Ucla *et al.*, 2002 Numerous potentially functional but non-genic conserved sequences on human chromosome 21. Nature **420:** 578–582.

Glazko, G. V., E. V. Koonin, I. B. Rogozin and S. A. Shabalina, 2003 A significant fraction of conserved noncoding DNA in human and mouse consists of predicted matrix attachment regions. Trends Genet. **19:** 119–124.

Glinka, S., L. Ometto, S. Mousset, W. Stephan and D. De Lorenzo, 2003 Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: a multi-locus approach. Genetics **165:** 1269–1278.

Green, P., B. Ewing, W. Miller, P. J. Thomas, NISC Comparative Sequencing Program *et al.*, 2003 Transcription-associated mutational asymmetry in mammalian evolution. Nat. Genet. **33:** 514–517.

Hanke, J., D. Brett, I. Zastrow, A. Aydin, S. Delbrük *et al.*, 1999 Alternative splicing of human genes: More the rule than the exception? Trends Genet. **15:** 389–390.

Hefferon, T. W., J. D. Groman, C. E. Yurk and G. R. Cutting, 2004 A variable dinucleotide repeat in the *CFTR* gene contributes to phenotype diversity by forming RNA secondary structures that alter splicing. Proc. Natl. Acad. Sci. USA **101:** 3504–3509.

Kern, A. D., and D. J. Begun, 2005 Patterns of polymorphism and divergence from noncoding sequences of *Drosophila melanogaster* and *D. simulans*: evidence for nonequilibrium processes. Mol. Biol. Evol. **22:** 51–62.

Ludwig, M. Z., and M. Kreitman, 1995 Evolutionary dynamics of the enhancer region of *even-skipped* in *Drosophila*. Mol. Biol. Evol. **12:** 1002–1011.

Ludwig, M. Z., N. H. Patel and M. Kreitman, 1998 Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change. Development **125:** 949–958.

Mount, S. M., C. Burks, G. Hertz, G. D. Stormo, O. White *et al.*, 1992 Splicing signals in *Drosophila*: intron size, information content, and consensus sequences. Nucleic Acids Res. **20:** 4255–4262.

Parsch, J., 2003 Selective constraints on intron evolution in Drosophila. Genetics **165:** 1843–1851.

Petrov, D. A., and D. L. Hartl, 1998 High rate of DNA loss in the *Drosophila melanogaster* and *Drosophila virilis* species groups. Mol. Biol. Evol. **15:** 293–302.

Ptak, S. E., and D. A. Petrov, 2002 How intron splicing affects the deletion and insertion profile in *Drosophila melanogaster*. Genetics **162:** 1233–1244.

Quandt, K., K. Frech, H. Karas, E. Wingender and T. Werner, 1995 MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. Nucleic Acids Res. **23:** 4878–4884.

Rozas, J., J. C. Sánchez-DelBarrio, X. Messeguer and R. Rozas, 2003 DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics **19:** 2496–2497.

Schaeffer, S. W., 2002 Molecular population genetics of sequence length diversity in the *Adh* region of *Drosophila pseudoobscura*. Genet. Res. **80:** 163–175.

Sekelsky, J. J., M. H. Brodsky and K. C. Burtis, 2000 DNA repair in *Drosophila*: insights from the *Drosophila* genome sequence. J. Cell Biol. **150:** F31–F36.

Shabalina, S. A., and A. Kondrashov, 1999 Pattern of selective constraint in *C. elegans* and *C. briggsae* genomes. Genet. Res. **74:** 23–30.

Sharp, P. A., 1994 Split genes and RNA splicing. Cell **77:** 805–815.

Stephan, W., V. S. Rodriguez, B. Zhou and J. Parsch, 1994 Molecular evolution of the metallothionein gene *Mtn* in the *melanogaster* species group: results from *Drosophila ananassae*. Genetics **138:** 135–143.

Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105:** 437–460.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.

Wingender, E., X. Chen, R. Hehl, H. Karas, I. Liebich *et al.*, 2000 TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Res. **28:** 316–319.

Yu, J., Z. Yang, M. Kibukawa, M. Paddock, D. Passey *et al.*, 2002 Minimal introns are not "junk". Genome Res. **12:** 1185–1189.

Communicating editor: P. J. Oefner