How Intron Splicing Affects the Deletion and Insertion Profile in *Drosophila melanogaster*

Susan E. Ptak[1] and Dmitri A. Petrov

Department of Biological Sciences, Stanford University, Stanford, California 94305

Manuscript received April 5, 2002
Accepted for publication August 5, 2002

ABSTRACT

Studies of "dead-on-arrival" transposable elements in *Drosophila melanogaster* found that deletions outnumber insertions ~8:1 with a median size for deletions of ~10 bp. These results are consistent with the deletion and insertion profiles found in most other Drosophila pseudogenes. In contrast, a recent study of *D. melanogaster* introns found a deletion/insertion ratio of 1.35:1, with 84% of deletions being shorter than 10 bp. This discrepancy could be explained if deletions, especially long deletions, are more frequently strongly deleterious than insertions and are eliminated disproportionately from intron sequences. To test this possibility, we use analysis and simulations to examine how deletions and insertions of different lengths affect different components of splicing and determine the distribution of deletions and insertions that preserve the original exons. We find that, consistent with our predictions, longer deletions affect splicing at a much higher rate compared to insertions and short deletions. We also explore other potential constraints in introns and show that most of these also disproportionately affect large deletions. Altogether we demonstrate that constraints in introns may explain much of the difference in the pattern of deletions and insertions observed in Drosophila introns and pseudogenes.

MUTATIONAL biases are an important factor shaping genetic evolution. Relative frequencies of transitions/transversions, inversions and duplications, GC- and AT-enriching mutations, deletions, and insertions of different sizes may all affect the evolution of genes and genomes in profound ways. In this article we focus in particular on the distribution of insertions and deletions (or indels) in Drosophila. Several recent studies reported that Drosophila has a strong deletion bias. Studies of 5′-truncated, dead-on-arrival non-LTR elements (Petrov *et al.* 1996; Petrov and Hartl 1998), several *bona fide* pseudogenes (Pritchard and Schaeffer 1997; Ramos-Onsins and Aguadé 1998; Petrov and Hartl 2000; Robin *et al.* 2000), and a nuclear insertion of mitochondrial DNA (Petrov 2002a) all suggest that among small (1–400 bp) indels, deletions are substantially more frequent and on average much longer.

As is generally the case, the inference of the mutational spectra from the pattern of substitutions is difficult. Most sequences are subject to selective constraints and the pattern of substitutions observed in such sequences reflects both mutational and selective biases. Because pseudogenes are nonfunctional, it is tempting to assume that the pattern of indels observed in pseudogenes is unbiased by selection. This inference is less straightforward than may appear, because natural selection may act in a variety of ways beyond selection for gene expression (Charlesworth 1996; Robin *et al.* 2000). However, in the case of Drosophila, several lines of evidence suggest that the indel spectrum in pseudogenes is indeed a fair approximation of the mutational indel spectrum (Petrov and Hartl 2000; Petrov 2002a).

A high rate of deletions in Drosophila may be important in explaining the small size of the Drosophila genome and its compactness at all levels of organization, such as the small size of introns, paucity of pseudogenes, and low density of transposable elements in the euchromatin compared to many other organisms. The fact that the negative correlation between the strength of the deletion bias at small scale and the genome size extends to several insects and mammals (Graur *et al.* 1989; Petrov *et al.* 1996, 2000; Robertson and Martos 1997; Robertson 2000; Bensasson *et al.* 2001; Petrov 2002b) further underscores the importance of this parameter.

In contrast to the results obtained from Drosophila pseudogenes of different kinds, Comeron and Kreitman (2000) demonstrated a virtual parity of deletions and insertions (1.35 deletions per insertion) segregating in the *D. melanogaster* introns. In addition, the vast majority (84%) of deletions is <10 bp. One possible explanation for the discrepancy between intron and pseudogene results is that introns are not truly neutral, and thus the distribution of indels in introns is the result of both mutation and selection. Comeron and Kreitman (2000) consider and reject several versions of this scenario. First they examine the possibility that indels in their sample are subject to strong selection. In this case new indels should either quickly sweep to fixation or disappear

[1]*Corresponding author:* Max Planck Institute for Evolutionary Anthropology, Inselstr. 22, Leipzig 04103, Germany.
E-mail: ptak@eva.mpg.de

from the population. Thus there should be little or no length polymorphism present. Because introns display a substantial amount of length polymorphism, this hypothesis can be rejected. Comeron and Kreitman also examine the possibility that the sampled indels are subject to weak selection. In such a case there should be an observable difference in the distribution of indels in areas of high and low recombination since recombination increases the efficacy of selection. Since Comeron and Kreitman observe no such difference, they reject the hypothesis that indels are subject to weak selection.

Such an analysis does indeed demonstrate that indels segregating within introns are almost neutral. However, this fact alone does not ensure that the indels segregating in introns have not been affected by selection. It is possible that a subset of indels is subject to strong purifying selection and the remaining indels are (nearly) selectively neutral. Only the neutral or nearly neutral indels will persist long as indel polymorphisms, whereas the strongly deleterious or advantageous indels will be quickly eliminated and not observed. If some indels, for example long deletions, are more likely to be subject to strong purifying selection, they will be underrepresented in the intron sample of indel polymorphisms.

In this article we attempt to test and quantify this possibility. To not alter the protein structure, introns must be completely excised from the pre-mRNA. Thus, any mutation that severely alters the intron's ability to be spliced will be subject to strong purifying selection in most cases. We ask first which indels are subject to strong selection due to splicing constraints and what bias this introduces into the distribution of indels observed as polymorphisms. We find that splicing constraints do significantly affect the deletion spectrum, eliminating deletions, especially longer ones, to a much greater degree than they do insertions. This fact is clearly a part of the needed explanation for the difference between pseudogene and intron estimates of indel biases. Second, since it has been hypothesized that some introns contain functional regulatory sequences or play other functional roles in the cell, we also ask how these nonsplicing constraints affect the distribution of indels observed as polymorphisms. We demonstrate that in general such nonsplicing constraints also disproportionately affect long deletions. We argue that putting all of the intron constraints together may explain most of the difference in the indel spectra between pseudogenes and introns.

METHODS

Splicing constraints: There are several components needed to excise spliceosomal-dependent introns from pre-mRNA. The basic components and pathway are as follows. After transcription, the resulting RNA is modified, including removal of introns by the spliceosome, a protein-RNA complex. The spliceosome recognizes the exon-intron boundaries by conserved sequences at the 3′ and 5′ ends of introns, as well as by an internally conserved sequence, the branchpoint. Furthermore, there appear to be length constraints on the intron, presumably due to spatial requirements of the spliceosome (Upholt and Sandell 1986; Hwang and Cohen 1997). For this article, we focus on the following splicing constraints, where the values for these constraints come from Drosophila (for summary see Green 1986; Padgett *et al.* 1986; Mount *et al.* 1992).

1. Exon spillage: Insertions and deletions do not spill into the exon.
2. Essential sites on edge: There are 2 bases absolutely required at each end, with 9 bases present in 70% of all introns. Some introns also have a sequence rich in pyrimidines that adds up to 10 additional bases at the 3′ end.
3. Internal branchpoint: The internal branchpoint occurs some 15–50 bases upstream of the 3′ end. One base is absolutely required, with 6 bases present in 70% of all individuals.
4. Length constraints: Most introns are >45 bases, where the subsequence from the 5′ end to the branchpoint is >38 bases and the subsequence from the branchpoint to the 3′ end is >15 bases and <50 bases. For some introns, there is evidence of a maximum length constraint of ~350 bases.

Non-splicing constraints: Some introns have other functional constraints besides splicing. These include introns that are alternatively spliced or introns that contain functional elements (see, *e.g.*, Bergman and Kreitman 2001; Mattick 2001). These phenomena place constraints on the intron sequence. There are also various hypotheses about other roles that introns might play in the cell. First, introns might regulate the amount of recombination between the flanking exons (Comeron and Kreitman 2000) or serve as locations for nonhomologous recombination that would allow for exon shuffling (de Souza *et al.* 1996; Patthy 1996). This would place constraints on intron length and possibly also on sequences of local homology. Length of introns also affects the rate and cost of transcription. Finally, intron length also appears to be correlated with genome size (Moriyama *et al.* 1998; Vinogradov 1999), and genome size is correlated with a wide range of ecological and cellular processes (Gregory and Hebert 1999; Knight and Ackerly 2002). The mechanism behind these correlations is not known, but it is possible that some of these processes place constraints on intron length.

Analytical approach: For each of the various constraints, we calculate the percentage of insertions and the percentage of deletions that maintain the original exons (% ok) as a function of indel size and intron length. We assume that each base within the intron is equally likely to be the location of the mutation. For deletions, this base is, with equal probability, either the beginning or the end of the deletion. For insertions, the insertion occurs following this base. We use this value of % ok to calculate the following five statistics, where $p(S)$

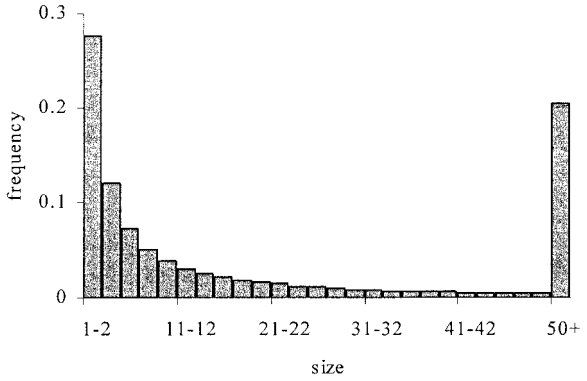

FIGURE 1.—The mutational distribution of indel sizes. The probability that an indel is 10 bases or fewer is 0.561. This distribution comes from fitting a lognormal to the distribution of deletion sizes in PETROV and HARTL (1998).

is the probability of an indel of size S and the distribution of indel sizes is as shown in Figure 1. Thus we assume the same size distribution for deletions and for insertions.

1. What fraction of deletions does not affect splicing?

$$\%\ \text{deletion} = \text{no. of deletions in mutated sequences with preserved exon/total no. of deletions}$$

$$\%\ \text{deletion} = \sum_S p(S) \times \%\ \text{ok} \quad (1)$$

2. What fraction of insertions does not affect splicing?

$$\%\ \text{insertion} = \text{no. insertions in mutated sequences with preserved exon/total no. of insertions}$$

$$\%\ \text{insertion} = \sum_S p(S) \times \%\ \text{ok} \quad (2)$$

3. What fraction of deletions contains 10 or fewer bases?

$$\%\ \leq 10 \text{ for deletions} = \text{no. deletions} \leq 10 \text{ bases/total no. of deletions}$$

$$\%\ \leq 10 \text{ for deletions} = \sum_{S=1}^{S=10} p(S) \times \%\ \text{ok}/\%\ \text{deletion} \quad (3)$$

4. What fraction of insertions contains 10 or fewer bases?

$$\%\ \leq 10 \text{ for insertions} = \text{no. insertions} \leq 10 \text{ bases/total no. of insertions}$$

$$\%\ \leq 10 \text{ for insertions} = \sum_{S=1}^{S=10} p(S) \times \%\ \text{ok}/\%\ \text{insertion} \quad (4)$$

5. What is the ratio of deletions to insertions?

$$\text{del/ins} = \text{no. deletions/no. insertions}$$

$$\text{del/ins} = (\text{del/ins ratio among all mutations}) \times (\%\ \text{deletion}/\%\ \text{insertion}). \quad (5)$$

GENSCAN: To confirm our analytical results we utilize GENSCAN as a splicing proxy since it relaxes many of the assumptions we need to make in the analytical approach. GENSCAN is a computer algorithm developed by BURGE and KARLIN (1997) to predict, among other things, intron-exon boundaries. We first obtained *Drosophila melanogaster* sequences by downloading multi_exon_GB.dat.gz from http://www.fruitfly.org/sequence/Drosophila-datasets.html, which contains *D. melanogaster* multiexonic gene sequences. This database of sequences was initially compiled by D. Kulp and M. G. Reese to train GENIE (KULP *et al.* 1996; REESE *et al.* 1997), another gene predictor algorithm. From this database of sequences we randomly selected sequences on the basis of three criteria: The sequences are scattered about the database, GENSCAN accurately predicts the correct exon-intron boundaries with probability >0.50 for each boundary, and there are equal numbers of long and short introns (where short introns have <80 bases). From GenBank, we also downloaded the sequences of three transposable elements (all *Copia*). GENSCAN correctly found no exon-intron boundaries in these three sequences. The accession numbers and intron lengths of these sequences are listed in Table 1.

TABLE 1

Information about sequences used with GENSCAN

Sequence	Accession no.	Length of introns	Source
Seq1	AF022540	67, 61	*D. melanogaster*
Seq2	X55887	303, 59, 121, 61	*D. melanogaster*
Seq3	X16715	121	*D. melanogaster*
Seq4	X71866	136, 52, 63	*D. melanogaster*
Seq5	X04695	483	*D. melanogaster*
Seq6	X72921	256	*D. melanogaster*
Seq7	X70838	71	*D. melanogaster*
Seq8	Y10276	76, 152, 118	*D. melanogaster*
Seq9	D37788	63	*D. melanogaster*
Seq10	L41867	72, 62, 364	*D. melanogaster*
Seq11	U00145	3520, 97	*D. melanogaster*
SeqA	X04456		*Copia* transposable element
SeqB	M11240		*Copia* transposable element
SeqC	X02599		*Copia* transposable element

We then subjected these sequences to 10,000 indels within each intron. Again, we assume that each base within the intron is equally likely to be the location of the mutation. For deletions, this base is, with equal probability, either the beginning or the end of the deletion. For insertions, the insertion occurs following this base. The distribution of indel sizes is shown in Figure 1. The sequence for the insertion comes from three possible sources. For one entire set of runs, the insertion is a random sequence with all 4 bases equally likely. For another set of runs, the insertion is a duplication of the downstream sequence for insertions smaller than a certain cutoff size and is one of the three transposable elements if the insertion is larger than the cutoff size. We examine

TABLE 2

Variables used in the analysis

Variable	Description
S	Size of indel
L	Total intron length
L_L	No. of bases between the 5′ edge of the intron and the branchpoint sequence
L_R	No. of bases between the 3′ edge of the intron and the branchpoint sequence
E	Total no. of edge essential sites (no. of conserved bases at 5′ and 3′ ends)
E_L	No. of edge essential sites at 5′ end (no. of conserved bases at 5′ end)
E_R	No. of edge essential sites at 3′ end (no. of conserved bases at 3′ end)
E_I	Total no. of internal essential sites (no. of conserved bases in branchpoint)
N	Counter for the number of subsequences
MinSize	Minimum size tolerated for the entire intron
MinSize_L	Minimum size tolerated for the subsequence from the 5′ end to the branchpoint
MinSize_R	Minimum size tolerated for the subsequence from the 3′ end to the branchpoint
MaxSize	Maximum size tolerated for the entire intron
MaxSize_L	Maximum size tolerated for the subsequence from the 5′ end to the branchpoint
MaxSize_R	Maximum size tolerated for the subsequence from the 3′ end to the branchpoint

a cutoff size of 100 bases and 1000 bases. We then run these mutated sequences through GENSCAN. The executable of GENSCAN can be downloaded from http://genes.mit.edu/GENSCAN.html.

We compare the location of exons in the original sequence to the location of exons in the mutated sequence and classify the mutated sequences into three groups: those in which the original exons are completely preserved and unaltered; those in which the locations of the original exons are shifted by 3, 6, or 9 bases (thus, the protein has an insertion or deletion of one to three amino acids); and those in which the original exons are not maintained. We then calculate the same statistics described above. To compare these statistics between all mutations and mutations that create sequences that preserve the original exon, we use the Wilcoxon two-sample test. Initial results suggested an effect due to the size of the intron, so we also compare these statistics for small introns (<80 bases) and large introns (>80 bases), using the Wilcoxon two-sample test.

RESULTS

Splicing constraints—analytical approach: See Table 2 for a list of all variables used in the analysis that follows.

Exon spillage: The most basic requirement is that indels do not directly affect the exon. Insertions that occur within the intron are contained within the intron. Thus the percentage of insertions that maintain the original exons, % ok, is

$$\text{Insertions: \% ok} = 1.0. \quad (6)$$

However, deletions, since they have two breakpoints, can spill out of the intron and into the exon. The percentage of deletions that are completely contained within the intron depends upon the length of the intron, L, and the size of the deletion, S. There are L possible deletions, since there are L possible sites where a deletion can begin or end within the intron. These sites are chosen with equal frequencies. Deletions that range from sites $\{1, 1 + (S - 1)\}$ to $\{L - (S - 1), L\}$ are completely contained within the intron; therefore there are $(L - (1 + (S - 1)) + 1)$ or $(L - S + 1)$ deletions that do not spill into the exons. Thus the percentage of deletions that maintain the original exons, % ok, is

$$\text{Deletions: \% ok} = \text{if } (L - S \geq 0) \text{ then } \frac{L - S + 1}{L} \text{ else } 0. \quad (7)$$

As seen in Figure 2, as intron length increases, % deletion increases while % ≤10 for deletions decreases. A higher proportion of deletions is contained within an intron the larger the intron is, as well as a higher proportion of larger deletions.

Minimum length: There is considerable evidence for a minimum size constraint on introns (Hawkins 1988; Mount *et al.* 1992; Carvalho and Clark 1999; Deutsch and Long 1999). This minimum size is probably not absolute, where by absolute we mean that all introns smaller than this size are never spliced correctly. However, as a first approximation let us treat it as such. Again, this constraint will not affect insertions. Thus,

$$\text{Insertions: \% ok} = 1.0. \quad (8)$$

All sequences <MinSize are not properly spliced, and therefore all deletions $>L -$ MinSize are not tolerated. Thus the percentage of deletions that maintain the original exons, % ok, is

$$\text{Deletions: \% ok} = \text{if } (L - S \geq \text{MinSize}) \text{ then } \frac{L - S + 1}{L} \text{ else } 0. \quad (9)$$

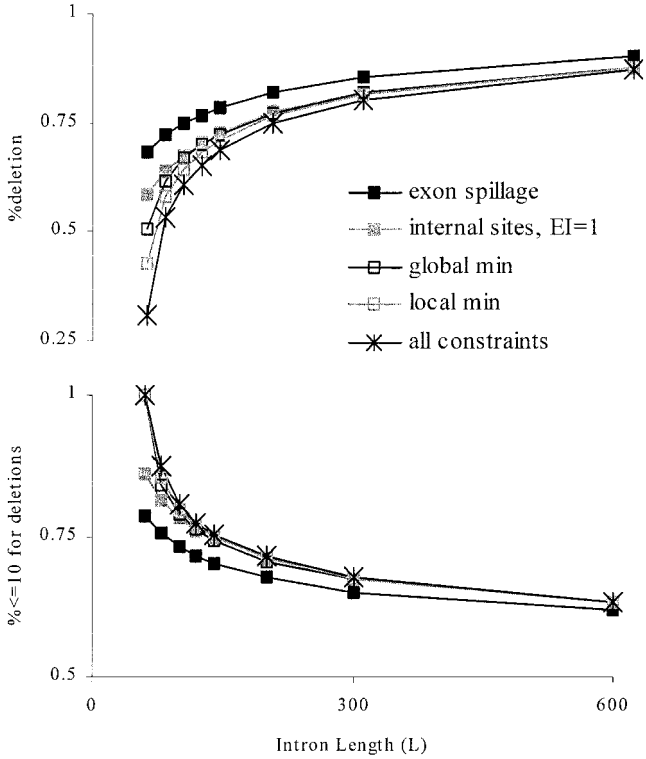

FIGURE 2.—Analytical calculation of % deletion and % ≤10 for deletions (Equations 1 and 3) for constraints of exon spillage (Equation 7), internal essential sites (Equation 15), local (Equation 17) and global (Equation 19) minimum length constraints, and combination of all splicing constraints (Equation 21). $E = 4$, $E_I = 1$, MinSize = 50, $\text{MinSize}_L = 35$, and $\text{MinSize}_R = 14$. $L_L = L - L_R - E_I$ and $L_R = 14 + L/20$.

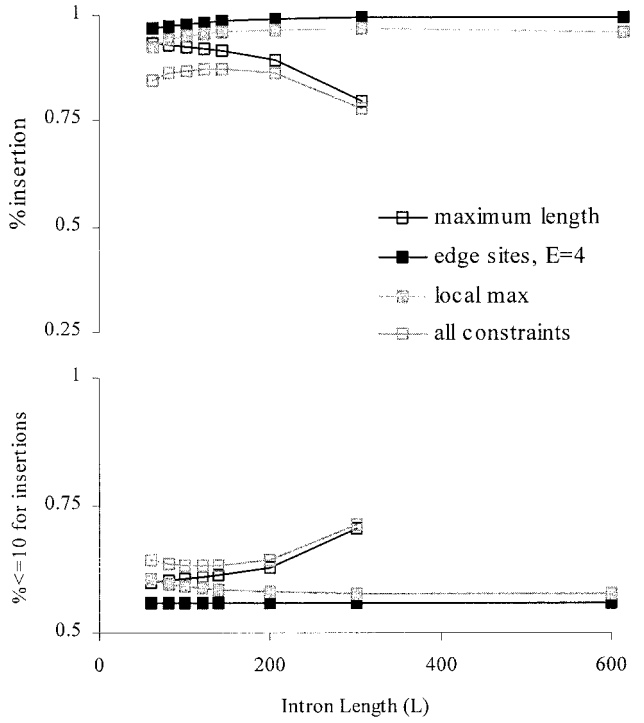

FIGURE 3.—Analytical calculation of % insertion and % ≤10 for insertions (Equations 2 and 4) for constraints of maximum length (Equation 10), essential sites on edges (Equation 12), local maximum length constraints (Equation 16), and combination of all constraints (Equation 20). $E = 4$, $E_I = 1$, MaxSize = 350, $\text{MaxSize}_R = 50$. $L_L = L - L_R - E - E_I$ and $L_R = 14 + L/20$.

Thus, this constraint is identical to that of exon spillage except that the length minus the deletion size needs to be greater than some positive value rather than greater than zero. Thus, the constraint of minimum size is more restrictive than exon spillage but similar in flavor.

Maximum length: There is some evidence that there is also a maximum size constraint at least for some introns (TALERICO and BERGET 1994; BERGET 1995; ROMFO *et al.* 2000). As a first approximation, let us also treat this as an absolute criterion and assume it is present for all introns. Therefore all sequences >MaxSize are not properly spliced and all insertions >MaxSize − L are not tolerated. Thus the percentage of insertions that maintain the original exons, % ok, is

$$\text{Insertions: \% ok} = \text{if } (L + S \leq \text{MaxSize}) \text{ then 1 else 0.} \quad (10)$$

However, this constraint does not affect deletions. Thus,

$$\text{Deletions: \% ok} = 1.0. \quad (11)$$

As seen in Figure 3, as intron length increases, % insertion decreases and % ≤10 for insertions increases. Larger introns are closer to the MaxSize and thus less tolerable of large insertions. The effect is slight if MaxSize is much greater than the intron length.

5′ and 3′ sequence: The signal for splicing includes a sequence at both the 5′ and 3′ ends of the intron and in some introns a polypyrimidine sequence (GREEN 1986; PADGETT *et al.* 1986; MOUNT *et al.* 1992). Let us assume there are a total of E essential sites on the edge of the intron that must be present for the intron to be properly spliced. Insertions can occur between any 2 bases except those between the essential sites. There are $L - 1$ possible insertion sites within the intron, $E - 2$ of which are between the essential sites. So there are $(L - 1) - (E - 2)$ insertions that preserve splicing. Thus the percentage of insertions that maintain the original exons, % ok, is

$$\text{Insertions: \% ok} = \frac{(L - 1) - (E - 2)}{(L - 1)}. \quad (12)$$

Deletions must not spill into essential sites on the edge. Since the essential sites are on the edges, the criterion of having the deletions not spill into the essential sites encompasses the criterion of having the deletions not spill into the exon. In effect, essential sites on the edges shorten the intron by E sites in terms of the number of possible sites for deletions that maintain the original exons and otherwise this criterion is identical to that of exon spillage. Furthermore, this criterion is restrictive only if the number of essential sites is large;

if E/L is small, then the effect on deletions is small beyond that encompassed by exon spillage.

$$\text{Deletions: \% ok} = \text{if } (L - E - S \geq 0) \text{ then } \frac{L - E - S + 1}{L} \text{ else } 0. \quad (13)$$

As seen in Figure 3, as length increases, % insertion increases slightly since the probability that an insertion falls between two essential sites decreases as the ratio of essential sites to nonessential sites decreases. However, the presence of essential sites has no effect on % ≤10 for insertions since insertions are either deleterious or not regardless of the size of the insertion.

Branchpoint sequence: The signal for splicing also includes an internal sequence that serves as the branchpoint for the lariat structure before the intron is excised (Green 1986; Mount *et al.* 1992). As a first approximation let us assume there are a total of E_I internal essential sites that are adjacent. These internal essential sites divide the intron into two regions: Let L_L be the length of the intron upstream from the internal essential sites and L_R be the length downstream, where $L_L + L_R + E_I = L$. As with essential sites on the edge, insertions cannot fall between two essential sites. Thus, the equation for internal essential sites is identical to the equation for edge essential sites, except that the essential sites form one block rather than two blocks, and thus there is one more possible insertion site for edge essential sites. Thus,

$$\text{Insertions: \% ok} = \frac{(L - 1) - (E_I - 1)}{(L - 1)}. \quad (14)$$

Deletions cannot cross these internal essential sites and are thus contained to either the sequence on the left or the sequence on the right. Let us continue to include the restriction of exon spillage (for the sake of comparison and since it is the most basic restriction). Thus, there are $(L_L - S + 1)$ possible sites upstream from the branchpoint and $(L_R - S + 1)$ downstream, where the probability of a deletion occurring upstream is (L_L/L) and downstream is (L_R/L). Thus, the percentage of deletions that maintain the original exons, % ok, is

$$\text{Deletions: \% ok} = \text{if } (L_L - S \geq 0) \text{ then } \left[\frac{L_L}{L}\left(\frac{L_L - S + 1}{L_L}\right)\right] \text{ else } 0 + \text{if } (L_R - S \geq 0) \text{ then } \left[\frac{L_R}{L}\left(\frac{L_R - S + 1)}{L_R}\right)\right] \text{ else } 0 + \left[\frac{E_I}{L}\left(\frac{0}{E_I}\right)\right]. \quad (15)$$

As seen in Figure 2, as length increases, % deletion increases, while % ≤10 for deletion decreases. This pattern is partially due to a decrease in the ratio of essential sites to nonessential sites as length increases and due to the constraints of exon spillage as seen above. However, the effect of internal essential sites on deletions beyond that encompassed by exon spillage is significant, especially for small introns. More deletions (especially long deletions) are deleterious with internal essential sites than with essential sites on the edge, since internal essential sites divide the possible sites for deletions into two subsequences that are shorter than the entire sequence.

Branchpoint with constraints on minimum and maximum length: There is evidence of length restrictions on the two subsequences surrounding the branchpoint (Green 1986; Padgett *et al.* 1986; Mount *et al.* 1992). Let MinSize_L be the minimum length of the upstream subsequence and MinSize_R the minimum length of the downstream subsequence. Likewise, MaxSize_L is the maximum length of the upstream subsequence and MaxSize_R is the maximum length of the downstream subsequence. Adding these new constraints now divides the possible insertion sites also into two regions, with probability $L_L/(L - 1)$ the insertion will occur upstream of the branchpoint and with probability $L_R/(L - 1)$ the insertion will occur downstream of the branchpoint. Thus, the percentage of insertions that maintain the original exons, % ok, is

$$\text{Insertions: \% ok} = \text{if } (L_L + S \leq \text{MaxSize}_L) \text{ then } \frac{L_L}{L - 1} \text{ else } 0 + \text{if } (L_R + S \leq \text{MaxSize}_R) \text{ then } \frac{L_R}{L - 1} \text{ else } 0. \quad (16)$$

Likewise, the percentage of deletions that maintain the original exons, % ok, is

$$\text{Deletions: \% ok} = \text{if } (L_L - S \geq \text{MinSize}_L) \text{ then } \left(\frac{L_L - S + 1}{L}\right) \text{ else } 0 + \text{if } (L_R - S \geq \text{MinSize}_R) \text{ then } \left(\frac{L_R - S + 1}{L}\right) \text{ else } 0. \quad (17)$$

Let us contrast the results from including a branchpoint with local length constraints to the results with global length constraints. Note that $(L - 1) - (E_I - 1) = L - E_I = L_L + L_R$. Thus the equation for % ok for insertions presented here agrees with the equation for % ok presented in the previous section except for the addition of maximum length restriction. Thus,

$$\text{Insertions: \% ok} = \text{if } (L + S \leq \text{MaxSize}) \text{ then } \frac{L_L}{L - 1} + \frac{L_R}{L - 1} \text{ else } 0. \quad (18)$$

$$\text{Deletions: \% ok} = \text{if } (L - S \geq \text{MinSize})$$
$$\text{then } \{\text{if } (L_L - S \geq 0) \text{ then } \left(\frac{L_L - S + 1}{L}\right) \text{ else } 0$$
$$+ \text{ if } (L_R - S \geq 0) \text{ then } \left(\frac{L_R - S + 1}{L}\right) \text{ else } 0\}$$
$$\text{else } 0. \quad (19)$$

As seen in Figure 2, for both local and global constraints on length, as length increases, % deletion increases while % ≤10 for deletions decreases. For deletions, minimum constraints on the two subsequences are more restrictive than global minimum constraints (albeit only slightly for % ≤10), presumably since the former constraint breaks the sequence into two smaller regions, each subject to minimum size constraints.

When $E_I = 1$ an internal branchpoint places no constraints on insertions, and thus the constraints for global maximum with and without a branchpoint are identical. In contrast, as seen in Figure 3 for local constraints on length, as length increases, % insertion increases (until length exceeds 300) while % ≤10 for insertions decreases slightly. This occurs since as L increases, L_L increases disproportionately compared to L_R such that L_L/L_R increases. The two sides do not increase proportionally, since there is evidence that the sequence downstream of the branchpoint has a limit of 20–50 bases. That L_L/L_R increases as L increases is not a factor for global constraints, since insertions are tolerated or not regardless of on which side of the branchpoint they fall. For most intron lengths, global constraints are more restrictive on insertions than local constraints, since the local constraints used here include only a constraint on L_R and not on L_L.

Combination of constraints: With the combination of all factors, $L = L_L + L_R + E + E_I$. With essential sites on the edges, there is one more possible insertion site in each subregion as compared to the equations for just internal essential sites. Thus

$$\text{Insertions: \% ok} = \text{if } (L + S \leq \text{MaxSize})$$
$$\text{then } \{\text{if } (L_L + S \leq \text{MaxSize}_L) \text{ then } \frac{L_L + 1}{L - 1} \text{ else } 0$$
$$+ \text{ if } (L_R + S \leq \text{MaxSize}_R) \text{ then } \frac{L_R + 1}{L - 1} \text{ else } 0\}$$
$$\text{else } 0. \quad (20)$$

$$\text{Deletions: \% ok} = \text{if } (L - S \geq \text{MinSize})$$
$$\text{then } \{\text{if } (L_L - S \geq \text{MinSize}_L) \text{ then } \left(\frac{L_L - S + 1}{L}\right) \text{ else } 0$$
$$+ \text{ if } (L_R - S \geq \text{MinSize}_R) \text{ then } \left(\frac{L_R - S + 1}{L}\right) \text{ else } 0\}$$
$$\text{else } 0. \quad (21)$$

As seen in Figure 2, as length increases % deletion increases while % ≤10 for deletions decreases. As seen in Figure 3, as length increases % insertion initially increases until length approaches the maximum constraints and then % insertion decreases, while % ≤10 for insertions follows the opposite trend. Unlike for deletions where each constraint leads to the same pattern as a function of length, for insertions the constraints on global maximum length oppose the other constraints, as seen in Figure 3.

Relative strengths of the various splicing criteria: To summarize the various constraints and compare the relative strength of each on the various statistics, in Table 3 we list the values for the various statistics for each constraint and for two intron sizes (60 and 300) corresponding to small and large introns.

For deletions, as seen in the statistics % deletion and % ≤10 for deletions, for small introns the constraints with the greatest effect are exon spillage, various length constraints, and internal essential sites. These constraints either divide the intron into regions within which the deletion must be contained or further limit the size of the deletion through length constraints. In large introns, where these regions are much larger, only exon spillage has a large effect. Large introns are far from their length constraints, unlike small introns. For all criteria, however, deletions are more likely to disrupt splicing in small introns than in large introns.

For insertions, as seen in the statistics % insertion and % ≤10 for insertions, for small introns no one criterion has a large effect. In large introns, however, the global length constraint does have a large effect. Large introns are much closer to the maximum length constraint than are small introns. However, local maximum constraints do not have a significant effect, since the constraint is only on the right subsequence, and most of the increase in intron length takes place in the left subsequence. For some criteria, insertions have a greater effect in small introns, whereas for other criteria insertions have a greater effect in large introns.

As seen in the del/ins ratio, in small introns for all criteria, deletions are more likely to alter splicing than insertions. The criteria with the greatest effect are, again, exon spillage, length constraints, and internal essential sites. However, for large introns, insertions are more likely than deletions to alter splicing in the presence of global length constraints. Again, large introns are close to their maximum length constraints and thus sensitive to insertions, whereas small introns are close to their minimum length constraints and thus sensitive to deletions. Thus, for large introns, the constraints with a large effect are exon spillage (toward deletions) and global length constraints (toward insertions).

In conclusion, length constraints have a large effect if the intron length is close to those constraints but not if the intron length is far from the constraints. Exon spillage places a strong constraint on deletions (but none on insertions), which diminishes only somewhat as the length of the intron increases. Finally, essential

TABLE 3

Comparison of splicing constraints for small and large introns

	% deletion		% <10 for deletions		Del/ins	
	Small	Large	Small	Large	Small	Large
Exon spillage	0.68	0.86	0.79	0.65	5.95	7.44
Min and max length	0.54	0.85	1.00	0.65	5.02	9.33
Edge sites, $E = 4$	0.63	0.84	0.79	0.65	5.66	7.37
Internal sites, $E_I = 1$	0.59	0.82	0.86	0.67	5.11	7.12
Local min and max	0.43	0.82	1.00	0.68	4.00	7.32
Global min and max	0.51	0.82	1.00	0.67	4.71	8.94
All constraints	0.31	0.80	1.00	0.68	3.17	8.96
	% insertion		% <10 for insertions			
	Small	Large	Small	Large		
Exon spillage	1.00	1.00	0.56	0.56		
Min and max length	0.93	0.80	0.60	0.71		
Edge sites, $E = 4$	0.97	0.99	0.56	0.56		
Internal sites, $E_I = 1$	1.00	1.00	0.56	0.56		
Local min and max	0.93	0.97	0.61	0.58		
Global min and max	0.93	0.80	0.60	0.71		
All constraints	0.84	0.78	0.64	0.72		

Small introns are size 60 and large introns are size 300. The values of the various parameters are the same as used in graphs 3–9. The mutational del/ins ratio is 8.7:1 and the mutational % ≤10 for deletions and insertions is 0.56.

sites do not place a large constraint on indels in introns, presumably since essential sites reduce the number of potential sites for the location of the indel by a small proportion only. Furthermore, the indel will disrupt splicing irrespective of its size; so essential sites do not alter the mutational size distribution of indels. The one exception is for deletions. If the essential site occurs in the middle of the intron rather than at the edge, then the essential site has the added effect of dividing the intron into two smaller areas that must completely contain the deletion.

GENSCAN: To lend support to these analytical results, we use GENSCAN as an alternative proxy for splicing, since it relaxes many of the stringent criteria we needed to use above. The caveat to the GENSCAN results, which we discuss below, is that GENSCAN uses not only the known splicing requirements but also statistical properties of introns and exons that presumably the spliceosome does not use.

As mentioned in METHODS, for GENSCAN we explore different insertion schemes, none of which qualitatively alter the results except where noted. We show here the results for when insertions consist of duplicating the downstream sequence with a cutoff of 100 bases and otherwise the insertion is one of three transposable elements. Likewise, there is very little difference between the two sets of statistics comparing all mutations to either (1) those mutations that result in sequences that completely maintain the original exons and (2) those mutations that result in sequences that alter the original exons by at most a change of 3, 6, or 9 bases (the values differ by 3% or less, except for one difference at 6%, and the significance of the statistical tests is unaltered). Consequently only the statistics for the former are included.

Since % deletion is lower for small introns than large introns ($W = 121$, $p < 0.002$, $n = 11, 11$), small introns are more sensitive to deletions than long introns. Furthermore, since % ≤10 for deletions is higher for small introns than large introns ($W = 120$, $p < 0.002$, $n = 11, 11$), small introns are more sensitive to long deletions than are long introns. Also, small deletions are more likely to maintain the exons than large deletions for all introns, since % ≤10 for deletions is higher for mutations that preserve exons compared to all mutations ($W = 483$, $t_W = 5.66$, $p < 0.002$, $n = 22, 22$). These results are identical to the patterns observed using analysis.

Since % insertion is lower for small introns than large introns ($W = 110$, $p < 0.002$, $n = 11, 11$), small introns are more sensitive to insertions than large introns. However, % ≤10 for insertions is roughly the same regardless of the size of the intron ($W = 62$, $p > 0.2$, $n = 11, 11$). (Note that this is significant at $p = 0.02$ using a cutoff of 1000, presumably since the splice site is more likely to be duplicated with large insertions and small introns.) For all introns, small insertions are slightly more likely

to preserve exons than large insertions, since % ≤10 for insertions is higher for mutations that preserve the exons than for all mutations ($W = 353$, $t_W = 2.605$, $p < 0.02$, $n = 22, 22$). These results are identical to the patterns observed using analysis with no maximum size restriction, since GENSCAN recognizes large introns (Burge and Karlin 1997).

The del/ins ratio is slightly elevated in large introns as compared to small introns ($W = 95$, $p < 0.05$, $n = 11, 11$). The effect is slight since both % deletion and % insertion increase as the size of the intron increases and thus mostly cancel each other. However, the del/ins ratio is significantly lower among mutated sequences that maintain the exon than among all mutated sequences ($W = 462$, $t_W = 5.16$, $p < 0.002$), since fewer deletions are tolerated than insertions. These too are the same patterns as found for the analytical results.

To identify intron-exon boundaries, GENSCAN uses not only the splicing signal that cells presumably use but also statistical patterns about introns and exons, including the difference in GC content. Thus mutations that alter these statistical patterns may not interfere with splicing but may have interfered with GENSCAN's ability to identify the exon-intron boundary. There is evidence that this does not dramatically alter the results. The difference among random insertion, duplication, and addition of transposable elements is small. Also the insertion of random sequence and the insertion of a transposable element, which most alter the GC composition of the intron, are intermediate in effect as compared to duplication of sequence. So the use of GENSCAN is a crude test but still useful since it relaxes many of the assumptions we needed to make above. The requirements for splicing are not as rigid as we have used in the analytical approach. GENSCAN tolerates small alterations, including using nearby cryptic splice sites, more flexible size constraints, similar but not exact matches to the splicing sequence, and small insertions and deletions to the exon. Using GENSCAN allows us to examine the effect of including this more realistic flexibility. Since the results from the two methods largely agree with one another, it seems that the more simplistic approach is a reasonable approximation. Thus for the remaining results we utilize only the analytical approach.

Nonsplicing constraints: The various nonsplicing constraints mentioned in methods take a similar form to the constraints mentioned above: length constraints or additional internally conserved blocks. Length constraints can be due to constraints on the time or cost of transcription and effects on recombination. Conserved blocks are most likely due to the presence of regulatory sequences such as enhancers within introns. These can easily be incorporated by altering the values used for MinSize and MaxSize or adding additional subequations for the additional subsequences.

$$\text{Insertions: \% ok} = \text{if } (L - S \le \text{MaxSize})$$
$$\text{then } \{\textstyle\sum_{N=1}^{N=\text{Number of subsequences}} \text{if } (L_N - S \le \text{MaxSize}_N)$$
$$\text{then } \frac{L_N + 1}{L - 1} \text{ else } 0\} \text{ else } 0. \quad (22)$$

$$\text{Deletions: \% ok} = \text{if } (L - S \ge \text{MinSize})$$
$$\text{then } \{\textstyle\sum_{N=1}^{N=\text{Number of subsequences}} \text{if } (L_N - S \le \text{MinSize}_N)$$
$$\text{then } \left(\frac{L_N - S + 1}{L_N}\right) \text{ else } 0\} \text{ else } 0. \quad (23)$$

When $N = 2$, these equations reduce to the equations for all constraints combined. As the number of blocks increases or as the number of bases within each block increases, fewer indels are tolerated. As above, the effect is much smaller for insertions than for deletions.

Analysis of the empirical data in Drosophila introns: To approximate the extent of the selective constraint that splicing places on population summaries, we collapse the analytical equations developed above into a single population number by incorporating the length distribution of introns. We then evaluate this number for Drosophila, using the length distribution of introns found in Figure 1 in Comeron and Kreitman (2000), the size distribution of indels in Figure 1, and the splicing constraints in Drosophila as detailed in methods. To calculate a population figure for % deletion, % insertion, % ≤10 for deletions, and % ≤10 for insertions we substitute (21) into (1) and (3) and substitute (20) into (2) and (4). Based on this,

$$\text{Population value of statistic} = \sum_{L=1}^{\infty} p(L) f_{\#}(L), \quad (24)$$

where $p(L)$ is probability of intron of length L and $f_{\#}(L)$ is the appropriate function from Equations 1–4. Since Figure 1 in Comeron and Kreitman (2000) bins the intron length, we use the midpoint of each bin as an approximation. We use the parameters defined in Figures 2 and 3 except $E = 9$ and $E_I = 6$ (for justification of these parameters see methods). As is typical, there are a few rare introns below the generally accepted minimum size. We assume these introns cannot tolerate further deletions, but can tolerate insertions. Likewise, many of the larger introns are greater than the maximum size used here. We assume these introns are exon defined and thus not subject to a global maximum size constraint. Using these parameters, we find an overall del/ins ratio of 4.28 where 83.7% of deletions are <10 bases and 73.2% of insertions are <10 bases. Although there is evidence that at least some introns have maximum length constraints, not all do. If we completely remove the maximum length constraints then the del/ins ratio decreases from 4.28 to 4.25.

If the data from pseudogenes are representative of the mutational distribution of indels and the data from Comeron and Kreitman (2000) are representative of

the distribution in introns, then the known splicing constraints bring us from a deletion to insertion ratio of 8:1 to 4:1, still far from the 1.35:1 ratio observed in introns. For the percentage of deletions 10 bases or fewer, the constraints bring us from 50% to >80%, near the 77% observed in introns. These results suggest that the known splicing constraints, as modeled here, help explain the discrepancy between the intron data and the pseudogene data, but they are not sufficient. This suggests that there are additional constraints on introns.

One possibility is additional constraints on intron length. To further lower the del/ins ratio, these additional constraints should be primarily on the minimum length. It seems unlikely that the minimum length restriction could be much higher since there are many introns at the current limit. Another possibility is that different-sized introns have different minimum lengths. However, to bring the del/ins ratio down to the level observed in introns (using the lower confidence interval of Petrov and Hartl's data and the upper confidence interval of Comeron and Kreitman's data) requires that the minimum size for each intron be 99.8% of the intron length. Furthermore, altering the minimum length alters both the del/ins ratio and the percentage of deletions that are 10 bases or fewer.

Increasing the number of essential sites on the edge will alter the del/ins ratio but not the percentage of deletions that are 10 bases or fewer. However, this alters both % insertion and % deletion. It seems that the change to % insertion as the number of essential sites increases is greater than the change to % deletion, so that the del/ins ratio actually increases as the number of essential sites increases. Increasing the number of internal essential sites leads to a similar pattern. Finally, there is the possibility of additional internal blocks of conserved bases due to regulatory or other functional elements. Bergman and Kreitman (2001) investigate conserved regions in long introns and introns that are known transcription enhancers and find that on average there are 10.7 blocks per 1000 bases, where the median block length is 19. They find that the blocks are approximately regularly spaced (C. M. Bergman, personal communication). Using these average results with Equations 22–24 lowers the del/ins ratio to 2.95 where % ≤10 for deletions rises to 90.3%. Note that in this calculation introns <94 bases do not have any additional blocks beyond the splicing signals.

DISCUSSION

Discrepancy between mutational spectra as found in pseudogenes and introns: The pattern of deletions and insertions in the mutational spectrum may be an important parameter in genome evolution. The distribution of insertions and deletions in *D. melanogaster* has now been investigated in a variety of pseudogene-like sequences (see the Introduction). There is an overall agreement among these studies that deletions are significantly more frequent and longer than insertions, producing strong mutational pressure toward DNA loss. Because of the variety of the studied pseudogenes—euchromatic and heterochromatic, transposable and nontransposable, repetitive and unique—it is tempting to suggest that a high rate of DNA loss affects most sequences in the Drosophila genome. However, these results contrast with the pattern of polymorphic indels segregating in *D. melanogaster* introns (Comeron and Kreitman 2000). There is almost parity in the numbers of deletions and insertions (1.35:1), with deletions being significantly shorter than those found in the pseudogene studies.

Possibility of differential mutational processes: It is possible that mutational processes operating in introns are different from those operating in pseudogenes. Introns are parts of genes and may have a different chromatin organization and may also undergo transcription in the germline. Transcription, in particular, appears to be a good candidate, because it is often associated with DNA repair (Hanawalt 2001). However, in Drosophila, unlike most other organisms, transcription-coupled DNA repair has not been detected (de Cock *et al.* 1992; van der Helm *et al.* 1997; Sekelsky *et al.* 2000) and the importance of transcription in determining mutational biases in Drosophila remains unknown. Nevertheless, this scenario in which genes, especially those active in the germline, have a different mutational pattern remains a live possibility.

Possibility of selection acting on pseudogenes: It is also possible that natural selection affects indels differently in introns and pseudogenes. Pseudogenes, especially upon creation, may be expressed at the level of RNA and/or protein. Such expression may often be harmful and any mutation that eliminates it would be beneficial. Pseudogenes may also affect expression of neighboring genes via enhancers or other regulatory sequences carried by pseudogenes or via changes in the distances among promoters and enhancers. On average, pseudogene effects on the expression of neighboring genes are likely to be detrimental, and mutations that are more likely to alleviate such effects are likely to be beneficial. Because deletions, especially large ones, are more likely to disrupt any functional sequence, they may be advantageous on average.

There are several reasons to doubt that selection against detrimental effects of pseudogenes has notably affected pseudogene estimates of deletion/insertion biases. First of all, pseudogenes with significantly deleterious effects are unlikely to fix in natural populations. The majority of fixed pseudogenes are probably neutral or nearly neutral at the inception. In addition, selective effects are likely to be pseudogene specific and, moreover, likely to operate only at some points in the evolution of particular pseudogenes. This is inconsistent with the similarity of the indel spectra in a variety of pseu-

dogenes, including pseudogenes with and without original function and pseudogenes of different ages.

Possibility of selection acting on introns: We now consider the possibility of selection biasing the pattern of indels in introns. Comeron and Kreitman consider and reject the possibility of either strong or weak selection acting on the polymorphic indels in their sample. Their arguments show that the polymorphic indels in Drosophila introns are indeed (almost) neutral. However, there still remains a possibility that some indels are strongly deleterious and never reach sufficiently high frequencies to be included in the polymorphism sample. For instance, those indels that severely affect splicing will presumably be subject to strong purifying selection. If indels of different types vary in their propensity to disrupt splicing, the pattern of polymorphic indels would not faithfully reflect the mutational deletion/insertion spectrum.

In this article we explore the effect that known splicing constraints have on deletions and insertions. We found that indels that preserve the original exons tend to be shorter (especially for deletions) and include a lower ratio of deletions to insertions. This finding is consistent with the overall difference between the pseudogene and intron indel data. We argue on the basis of these results that the length polymorphisms observed in population surveys of introns are primarily neutral but involve only a subset of all possible indels within introns. Furthermore, based on the results presented here, this observable subset of indels is biased and not representative of the mutational spectrum of indels. Although the strength of the bias depends upon the size distribution of indels and the relative importance of the various splicing criteria, we feel that the general direction of the bias is fairly robust. Since large deletions are more likely to interfere with splicing than small deletions under almost all of the splicing criteria, as long as the size distribution of deletions includes a range of sizes, there will be a shift toward smaller deletions. For insertions, the effect is not as strong, since for many splicing constraints large and small insertions are equally likely to disrupt splicing. Also, most of the various splicing criteria place more restrictions on deletions than on insertions, mostly by dividing the intron into smaller sections where an increasing percentage of deletions spill into essential sites (including the exon). Furthermore, in *D. melanogaster* (and many other organisms) introns are much closer to minimum size constraints than to maximum size constraints. For these reasons, it seems likely that whatever the exact nature of splicing is, deletions will more likely be deleterious than insertions.

Can we explain all of the discrepancy? Our analysis demonstrates that splicing constraints by themselves can explain some of the discrepancy between the indel distributions in pseudogenes and introns. However, we also show that the known splicing constraints, as modeled in this article, do not seem to account for this discrepancy fully. On the basis of the pseudogene data and the known splicing constraints, we predict the del/ins ratio of 4.28, compared with 1.35 seen in introns. There are several, nonmutually exclusive, possible explanations for this difference.

The first possibility is that if we consider all possible sources of error in the estimates of the indel distributions, the discrepancy would disappear. For example, if we recalculate the population statistics using the lower bound on both the observed del/ins ratio and $\% \leq 10$ in pseudogenes and assume that all insertions are 10 bases or fewer (as indicated by the admittedly limited available data) rather than using the same distribution as deletions, we predict a ratio of 3.10 for del/ins ratio and 81.6% for $\% \leq 10$. The upper boundary on the intron data is 1.53 for del/ins ratio and 79.7% for $\% \leq 10$. Note also that this takes into consideration only sampling error. Other errors, such as in classifying indels as either insertions or deletions in introns, would tend to widen the confidence intervals and would tend to move the del/ins ratio closer to 1:1.

Another possibility is that there are additional constraints on introns that further bias the indel spectrum observed in introns. In particular there is good evidence that introns frequently contain blocks of conserved sites probably corresponding to regulatory sequences (Bergman and Kreitman 2000). If we also take these internal sites into account, our prediction of del/ins ratio in introns drops to 2.04 (the lower boundary), which is very close indeed to the observed ratio of 1.53 (the upper boundary).

Conclusions: The known splicing and nonsplicing constraints in Drosophila introns bias the insertion/deletion spectrum in favor of insertions and against longer deletions. Although it is clear that these constraints explain a good portion of the discrepancy between insertion/deletion spectra observed in pseudogenes and introns, it is as yet unclear whether they are sufficient to explain all of the difference. Additional studies on the insertion/deletion mutational spectra and on the constraints operating in introns should help resolve this issue.

We thank M. Lachmann for helpful comments especially in developing the analytical approach, as well as the Feldman and Petrov labs. We thank M. W. Feldman for critical comments on the manuscript. S.E.P. was supported by a Howard Hughes Medical Institute predoctoral fellowship and by National Institutes of Health Grant GM28016 to M. W. Feldman.

LITERATURE CITED

Bensasson, D., D. A. Petrov, D. X. Zhang, D. L. Hartl and G. M. Hewitt, 2001 Genomic gigantism: DNA loss is slow in mountain grasshoppers. Mol. Biol. Evol. **18:** 246–253.

Berget, S. M., 1995 Exon recognition in vertebrate splicing. J. Biol. Chem. **270:** 2411–2414.

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.

Burge, C., and S. Karlin, 1997 Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. **268:** 78–94.

Carvalho, A. B., and A. G. Clark, 1999 Intron size and natural selection. Nature **401:** 344.

Charlesworth, B., 1996 The changing sizes of genes. Nature **384:** 315–316.

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., E. C. Klink, W. Ferro, P. H. Lohman and J. C. Eeken, 1992 Neither enhanced removal of cyclobutane pyrimidine dimers nor strand-specific repair is found after transcription induction of the beta 3-tubulin gene in a Drosophila embryonic cell line Kc. Mutat. Res. **293:** 11–20.

de Souza, S. J., M. Long and W. Gilbert, 1996 Introns and gene evolution. Genes Cells **1:** 493.

Deutsch, M., and M. Long, 1999 Intron-exon structure of eukaryotic model organisms. Nucleic Acids Res. **27:** 3219–3228.

Graur, D., Y. Shuali and W. H. Li, 1989 Deletions in processed pseudogenes accumulate faster in rodents than in humans. J. Mol. Evol. **28:** 279–285.

Green, M. R., 1986 Pre-mRNA splicing. Annu. Rev. Genet. **20:** 671–708.

Gregory, T. R., and P. D. Hebert, 1999 The modulation of DNA content: proximate causes and ultimate consequences. Genome Res. **9:** 317–324.

Hanawalt, P. C., 2001 Controlling the efficiency of excision repair. Mutat. Res. **485:** 3–13.

Hawkins, J. D., 1988 A survey on intron and exon lengths. Nucleic Acids Res. **16:** 9893–9908.

Hwang, D. Y., and J. B. Cohen, 1997 U1 small nuclear RNA-promoted exon selection requires a minimal distance between the position of U1 binding and the 3′ splice site across the exon. Mol. Cell. Biol. **17:** 7099–7107.

Knight, C. A., and D. D. Ackerly, 2002 Variation in nuclear DNA content across environmental gradients: a quantile regression analysis. Ecol. Lett. **5:** 66–76.

Kulp, D., D. Haussler, M. G. Reese and F. H. Eeckman, 1996 *A Generalized Hidden Markov Model for the Recognition of Human Genes in DNA in ISMB-96*. AAAI/MIT Press, St. Louis.

Mattick, J. S., 2001 Non-coding RNAs: the architects of eukaryotic complexity. EMBO Rep. **21:** 986–991.

Moriyuma, E. N., D. A. Petrov and D. L. Hartl, 1998 Genome size and intron size in *Drosophila*. Mol. Biol. Evol. **15:** 770–773.

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.

Padgett, R. A., P. J. Grabowski, M. M. Konarska, S. Seiler and P. A. Sharp, 1986 Splicing of messenger RNA precursors. Annu. Rev. Biochem. **55:** 1119–1150.

Patthy, L., 1996 Exon shuffling and other ways of module exchange. Matrix Biol. **15:** 301–310.

Petrov, D. A., 2002a DNA loss and evolution of genome size in Drosophila. Genetica **115:** 81–91.

Petrov, D. A., 2002b Mutational equilibrium model of genome size evolution. Theor. Popul. Biol. **61:** 531–544.

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

Petrov, D. A., and D. L. Hartl, 2000 Pseudogene evolution and natural selection for a compact genome. J. Hered. **91:** 221–227.

Petrov, D. A., E. R. Lozovskaya and D. L. Hartl, 1996 High intrinsic rate of DNA loss in Drosophila. Nature **384:** 346–349.

Petrov, D. A., T. A. Sangster, J. S. Johnston, D. L. Hartl and K. L. Shaw, 2000 Evidence for DNA loss as a determinant of genome size. Science **287:** 1060–1062.

Pritchard, J. K., and S. W. Schaeffer, 1997 Polymorphism and divergence at a Drosophila pseudogene locus. Genetics **147:** 199–208.

Ramos-Onsins, S., and M. Aguadé, 1998 Molecular evolution of the Cecropin multigene family in Drosophila: functional genes *vs.* pseudogenes. Genetics **150:** 157–171.

Reese, M. G., F. H. Eeckman, D. Kulp and D. Haussler, 1997 Improved splice site detection in Genie. J. Comput. Biol. **4:** 311–323.

Robertson, H. M., 2000 The large *srh* family of chemoreceptor gene in *Caenorhabditis* nematodes reveals processes of genome evolution involving large duplications and deletions and intron gains and losses. Genome Res. **10:** 192–203.

Robertson, H. M., and R. Martos, 1997 Molecular evolution of an ancient mariner transposon Hsmar1 in human genome. Gene **205:** 219–228.

Robin, G. C., R. J. Russell, D. J. Cutter and J. G. Oakeshott, 2000 The evolution of an alpha-esterase pseudogene inactivated in the Drosophila melanogaster lineage. Mol. Biol. Evol. **17:** 563–575.

Romfo, C. M., C. J. Alvarez, W. J. van Heeckeren, C. J. Webb and J. A. Wise, 2000 Evidence for splice site pairing via intron definition in *Schizosaccharomyces pombe*. Mol. Cell. Biol. **20:** 7955–7970.

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.

Talerico, M., and S. M. Berget, 1994 Intron definition in splicing of small *Drosophila* introns. Mol. Cell. Biol. **14:** 3434–3445.

Upholt, W. B., and L. J. Sandell, 1986 Exon/intron organization of the chicken type II procollagen gene: intron size distribution suggests a minimal intron size. Proc. Natl. Acad. Sci. USA **83:** 2325–2329.

van der Helm, P. J., E. C. Klink, P. H. Lohman and J. C. Eeken, 1997 The repair of UV-induced cyclobutane pyrimidine dimers in the individual genes Gart, Notch and white from isolated brain tissue of *Drosophila melanogaster*. Mutat. Res. **383:** 113–124.

Vinogradov, A. E., 1999 Intron-genome size relationship on a large evolutionary scale. J. Mol. Evol. **49:** 376–384.

Communicating editor: M. A. F. Noor