The Correlation Between Intron Length and Recombination in Drosophila: Dynamic Equilibrium Between Mutational and Selective Forces

Josep M. Comeron and Martin Kreitman

Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637

Manuscript received August 2, 1999
Accepted for publication July 14, 2000

ABSTRACT

Intron length is negatively correlated with recombination in both *Drosophila melanogaster* and humans. This correlation is not likely to be the result of mutational processes alone: evolutionary analysis of intron length polymorphism in *D. melanogaster* reveals equivalent ratios of deletion to insertion in regions of high and low recombination. The polymorphism data do reveal, however, an excess of deletions relative to insertions (*i.e.*, a deletion bias), with an overall deletion-to-insertion events ratio of 1.35. We propose two types of selection favoring longer intron lengths. First, the natural mutational bias toward deletion must be opposed by strong selection in very short introns to maintain the minimum intron length needed for the intron splicing reaction. Second, selection will favor insertions in introns that increase recombination between mutations under the influence of selection in adjacent exons. Mutations that increase recombination, even slightly, will be selectively favored because they reduce interference among selected mutations. Interference selection acting on intron length mutations must be very weak, as indicated by frequency spectrum analysis of Drosophila intron length polymorphism, making the equilibrium for intron length sensitive to changes in the recombinational environment and population size. One consequence of this sensitivity is that the advantage of longer introns is expected to decrease inversely with the rate of recombination, thus leading to a negative correlation between intron length and recombination rate. Also in accord with this model, intron length differs between closely related Drosophila species, with the longest variant present more often in *D. melanogaster* than in *D. simulans*. We suggest that the study of the proposed dynamic model, taking into account interference among selected sites, might shed light on many aspects of the comparative biology of genome sizes including the *C* value paradox.

NUCLEAR spliceosomal introns are nucleotide sequences that are transcribed but spliced out of the precursor mRNA and they generally do not encode any other polypeptide (Berget *et al.* 1977; Chow *et al.* 1977; Sambrook 1977). Beyond essential splicing signals and regulatory elements, many introns are fast evolving, indicating a general lack of function (Li and Graur 1991; Hughes and Yeager 1997). Transcription in a variety of eukaryotes, including yeast and Drosophila, proceeds at a rate of between 18 and 25 nucleotides (nt)/sec (Irvine *et al.* 1991; Izban and Luse 1992). It follows that the transcription time for intron-containing genes can increase from a few seconds to several minutes, thus suggesting an energetic cost to maintain introns. Yet, the majority of introns are ancient and many, perhaps even the majority, have persisted individually throughout eukaryotic evolution (Shah *et al.* 1983; Marchionni and Gilbert 1986; Kersanach *et al.* 1994; see de Souza *et al.* 1996 for a review).

Might there be a universal selective benefit that could explain the evolutionary persistence of introns? The recent discovery of a correlation between intron length and recombination rate in *Drosophila melanogaster* (Carvalho and Clark 1999) provided the first indication that selection might play a role in governing intron length. These authors proposed that long introns have a deleterious effect but that this selection is overwhelmed by genetic drift in regions of low recombination, thus leading to the observed correlation between intron length and recombination rate. This hypothesis is not likely to be a satisfactory explanation, however, because natural mutational tendencies also appear to be biased toward the production of shorter introns. Deletion bias (the deletion-to-insertion mutational events ratio) seems to be a general feature of the mutational process, being detected in species as divergent as mammals (Graur *et al.* 1989; Saitou and Ueda 1994; Ogata *et al.* 1996; Ophir and Graur 1997) and Drosophila (Petrov *et al.* 1996; Petrov and Hartl 1998). Selection that favors deletions (and/or opposes insertions) will reinforce rather than oppose the spontaneous mutation bias that acts in the same direction, and either process alone should be sufficient to cause the rapid collapse of intron lengths.

Assuming the mutational deletion bias to be true, the presence of long introns can only mean that selection must favor (and preserve) their lengths in some in-

Corresponding author: Josep M. Comeron, Department of Ecology and Evolution, University of Chicago, 1101 E. 57th St., Chicago, IL 60637. E-mail: jcomeron@midway.uchicago.edu

stances, in opposition to the natural mutational bias. The discovery of the negative correlation between intron length and recombination rate further indicates that the strength or efficacy of this selection must be in some way recombination rate dependent. There is no evidence for strong selective constraints in long intron sequences, a possible explanation for the persistence of their long lengths. In fact, in Drosophila, short introns (defined as <80–90 bp; Mount *et al.* 1992; Stephan *et al.* 1994) whose lengths are close to the minimum required for proper splicing (Upholt and Sandell 1986; Tsurushita and Korn 1987; Mount *et al.* 1992) tend to be conserved in size, whereas longer introns are less constrained with respect to length and are evolutionarily more variable (Stephan *et al.* 1994).

D. melanogaster's genome exhibits a wide range of recombination rates and the level of silent nucleotide polymorphism is known to vary across the genome and to be positively correlated with local recombination rates (Begun and Aquadro 1992; Aguadé and Langley 1994; Aquadro *et al.* 1994). This empirical observation is compatible with theories of selection when genetic linkage and hitchhiking are taken into account (Maynard Smith and Haigh 1974; Kaplan *et al.* 1989; Stephan *et al.* 1992; Charlesworth *et al.* 1993; Charlesworth 1994; Hudson 1994; Hudson and Kaplan 1995). These models are embedded in the more general Hill-Robertson effect (Hill and Robertson 1966; Felsenstein 1974), which describes the interaction between selection, linkage, and drift. The effect of interference is an increment of the variance in reproductive success, and hence of genetic drift, and can be viewed as being equivalent to a reduction in a species' effective population size (Hill and Robertson 1966; Felsenstein 1974).

Models of strong selection and linkage have been invoked to explain the correlation between silent polymorphism and recombination, but an effect can also be produced with very weak selection acting on many sites, both for complete linkage (Li 1987; Comeron *et al.* 1999; Tachida 2000) and for recombination rates found across the Drosophila genome (Comeron *et al.* 1999). Interference among weakly selected mutations, which are numerous and segregate at high frequency, across the entire range of recombination rates for Drosophila, raises the possibility that insertions and deletions (indels) in introns may be subject to natural selection as modifiers of recombination. In Drosophila, weakly selected mutations could include synonymous mutations (Shields *et al.* 1988; Sharp and Li 1989; Kliman and Hey 1993; Moriyama and Hartl 1993; Akashi 1995, 1996; Powell and Moriyama 1997; Zeng *et al.* 1998; Comeron *et al.* 1999) and replacement mutations (Ohta 1993; Eanes *et al.* 1996; Takano 1998; Zeng *et al.* 1998; Schmid *et al.* 1999), as well as mutations in regulatory regions (Ludwig and Kreitman 1995; Ludwig *et al.* 2000).

Theoretical studies of Barton (1995) and Otto and Barton (1997) have shown that in regions of very low recombination, neutral modifiers that increase recombination between selected loci can be selectively beneficial and will be preferentially fixed. Hey (1998) has further elaborated on this finding by showing that it pertains not only to neutral modifiers of recombination but also to both beneficial and deleterious modifiers that increase and reduce recombination, respectively. This effect, however, requires tight linkage between selected mutations or a high rate of advantageous mutation.

In this article, we investigate whether the preferred fixation of small insertions over deletions in introns as enhancers of recombination can explain the existence of longer introns in regions of low recombination in *D. melanogaster*. We also investigate whether relative mutation rates for insertions and deletions are the same in genomic regions of high and low recombination. Recombination rate heterogeneity in this species allows us to explore predictions of selection and interference influencing intron length. Comparison of *D. melanogaster* with both *D. simulans* and humans allows us to further test our predictions by taking advantage of their very different effective population sizes (N_e).

MATERIALS AND METHODS

***D. melanogaster* sequences:** We obtained all completely sequenced nuclear coding regions in *D. melanogaster* from FlyBase (1998; v.2.8.5; 2546 genes; http://flybase.bio.indiana.edu/), which do not contain genes related to transposon and repetitive elements. Only one sequence per gene was used, and for genes with multiple splicing forms only the longest one was analyzed. We later selected those sequences with DNA as source by comparison to GenBank features, with complete coding regions and uninterrupted sequence, and with accurate cytological position in FlyBase. Out of the genes meeting these criteria, coding sequence annotation from three genes (FBgn0011297, FBgn0005674, and FBgn0015269) was manually corrected based on GenBank annotations. This allowed us to obtain a set of 620 complete genes with accurate intron information; 447 (72.1%) of the genes contained one or more introns, with a total of 1345 introns. The average length of all coding regions was 1598.65 bp, and the average number of introns was 2.17 (ranging up to 25 in the Rya-r44F locus), with an overall average intron length of 298.02 bp among all genes.

To avoid the inclusion in our dataset of introns whose lengths were influenced by the insertion of transposable elements (TE), we screened introns >200 bp for the putative presence of TEs. We conducted BLAST2 (Tatusova and Madden 1999) homology searches using all FlyBase entries annotated as transposons in *D. melanogaster*. BLAST searches failed in detecting homologies between TE and intron sequences using the default parameters. We can be reasonably confident, therefore, that our intron length data have not been influenced by recent TE insertions.

Polymorphism analyses: Insertion and deletion polymorphic events were studied in 31 genomic regions of *D. melanogaster*. The regions analyzed, the sample sizes, the cytological position, and estimated recombination rates (see below) are given in Table 1. Nucleotide sequences were available for most of the analyzed regions; information from high-resolution restriction fragment length polymorphisms (4-cutter technique; Kreitman and Aguadé 1986) was used for regions *Pgd*, *su(f)*, and *y-ac-sc*. For most regions, we assigned a polymorphism as an insertion event (IE) or deletion event (DE) by comparison to the homologous sequence, or sequences, in *D. simulans*; for the rest of the indel polymorphisms, the least frequent variant was assumed to be the newly arisen (*i.e.*, derived) mutation. When possible, the homologous sequence or sequences in *D. yakuba* were also used to classify indels. Nucleotide sequences were aligned using ClustalX (Thompson *et al.* 1997) and complex indels (*i.e.*, microsatellites) were grouped and counted as one indel polymorphism with the average indel length.

Recombination rates: Following Kliman and Hey (1993), the recombination rate for a given cytological map position in *D. melanogaster* was estimated after obtaining polynomial curves as a function of the quantity of DNA in each division along each chromosome (Sorsa 1988) *vs.* the change of the cytogenetic (FlyBase) map position (see Comeron *et al.* 1999 for details). Recombination rates for humans were equivalently obtained, based on radiation hybrid (RH) mapping (GB4) and the genetic map of sequence-tagged site markers (Deloukas *et al.* 1998).

Estimation of selection coefficients on synonymous mutations based on the synonymous codon usage in *D. melanogaster*: The scaled selection coefficient ($\sigma = 4N_e s$) on synonymous mutations was estimated for the 620 genes studied in *D. melanogaster*. The analysis focuses on the two-fold degenerate amino acids because in this case a symmetry argument allowed us to assign identical selection coefficients (but of opposite sign) to mutations to preferred ($+s$) or to unpreferred ($-s$) codons. The estimation of selection coefficients for the different synonymous codons for three-, four-, and sixfold degenerate amino acids depends on the selective model associated with the different codons (*i.e.*, the different relative contribution that each codon makes to fitness), which has not been yet established. Following Li (1987), σ on synonymous mutations was estimated from the frequency (P_2) of the advantageous codon (preferred codon; Akashi 1995) in the sequence and the expected fraction of G + C nucleotides due to mutational biases among nucleotides (fGC). fGC was set to 0.35 on the basis of polymorphism data (Comeron *et al.* 1999) and nucleotide composition (Moriyama and Hartl 1993) in noncoding regions, and patterns of point mutations among *Helena* retroposon elements (Petrov and Hartl 1999).

Confidence intervals for the number of replacement substitutions per site (K_a) between *D. melanogaster* and *D. simulans*: The average K_a estimate between *D. melanogaster* and *D. simulans* for regions of low and high recombination is $K_a = 0.0164$ (five genes) and $K_a = 0.0095$ (eight genes), respectively (Takano 1998). We used *K*-estimator v5.3 program (Comeron 1999) to obtain the confidence intervals of these estimates, taking into account the number of sites under analysis, the number of nucleotide substitutions (Poisson distributed), and the variance generated by multiple hits at a site. The results indicate nonoverlapping 95% confidence intervals for K_a estimates in regions of low and high recombination [0.0122–0.0214 (3224 nonsynonymous sites) and 0.0075–0.0117 (7550 nonsynonymous sites), respectively]. As expected, the ratio K_a/K_s is significantly higher in genes located in regions of low recombination than in those genes in regions of high recombination (95% confidence intervals of [0.1506–0.3100] and [0.00753–0.1276], respectively).

Computer simulations to study intron length under a mutation-selection-drift model: We studied a model of selection in which mutations that increment the length of the transcripts are weakly deleterious with semidominant effects. The model imposes a minimum intron length ($L_{\min}$), with mutations that decrease the length below $L_{\min}$ being strongly selected against. We generated and studied the equilibrium for a population of $N = 500$ diploid individuals and $N\mu = 0.001$, where μ is the indel mutation rate per base pair. Each generation was obtained by randomly choosing N individuals of the previous generation, with a probability (Pw) proportional to their relative fitness, according to the equations $w_{i,k} = 1 - s(L_{i,k} - L_{\min})$, where s is the absolute deleterious selection coefficient per base pair, and $L_{i,k}$ is the length of the kth chromosome (k = 1, 2) of the ith individual, and

$$Pw_i = \overline{w}_i / \sum_{j=1}^{N} \overline{w}_j,$$

where $\overline{w}_i$ is the average fitness of the two chromosomes of the ith individual. The $2N$ genomes (lengths in this case) were then randomly paired to generate the new generation of N diploid individuals. The analyses of simulated data began after a minimum of $250N$ generations to ensure equilibrium. This was confirmed by visual inspection of each trajectory between $500N$ and $2500N$ generations. Deleterious selection coefficients (s) per base pair included $\sigma(4Ns)$ of 0, 0.04, 0.08, and 0.2, unless when $L_{i,k} < L_{\min}$, in which case $w_{i,k} = 0$. In all simulations, $L_{\min} = 60$ bp and $L_{i,k} \ll 1/s$. All statistical analyses were carried out using Statistica for Windows 5.1 (1997).

RESULTS

Intron length distribution in *D. melanogaster*: Mount *et al.* (1992) detected a dimorphic distribution of intron length in *D. melanogaster* on the basis of 209 complete introns with one class of introns <80–90 bp and a discontinuous distribution for longer introns (also see Hawkins 1988). The larger number of introns used in the present study allowed us to confirm the highly asymmetrical intron length distribution. But with a larger sample, it can be seen that the distribution of intron lengths is continuous, with no clear boundaries between two intron classes (Figure 1). The analysis of the 1345 introns yields an intron length average of 402.54 bp, with a mode of 58 bp and a median of 77 bp.

Longer introns in regions of low recombination in *D. melanogaster*: If intron length is a result of neutral mutational processes alone, then intron length should be independent of recombination rate. On the other hand, if indel mutations in Drosophila introns are weakly selected, either as modifiers of recombination or transcription time, then intron length and recombination rates could be correlated. Recombination rates in *D. melanogaster* correlate negatively with individual intron length (Spearman's rank correlation, $R = -0.1822$, $P < 1 \times 10^{-6}$; $n = 1345$), average intron length ($R = -0.2048$, $P = 1.3 \times 10^{-5}$; $n = 447$), and total intron length ($R = -0.2163$, $P = 4 \times 10^{-6}$, $n = 447$; see Figure 2A). A Kruskall-Wallis ANOVA test of data divided into three recombination rate groups and using the same three measures of intron length also reveals that introns in regions of low recombination are significantly longer than in regions of high recombination ($H = 40.28$, $H = 21.12$, and $H = 20.82$, respectively; $P < 1 \times 10^{-5}$ for all cases; see Figure 2B).

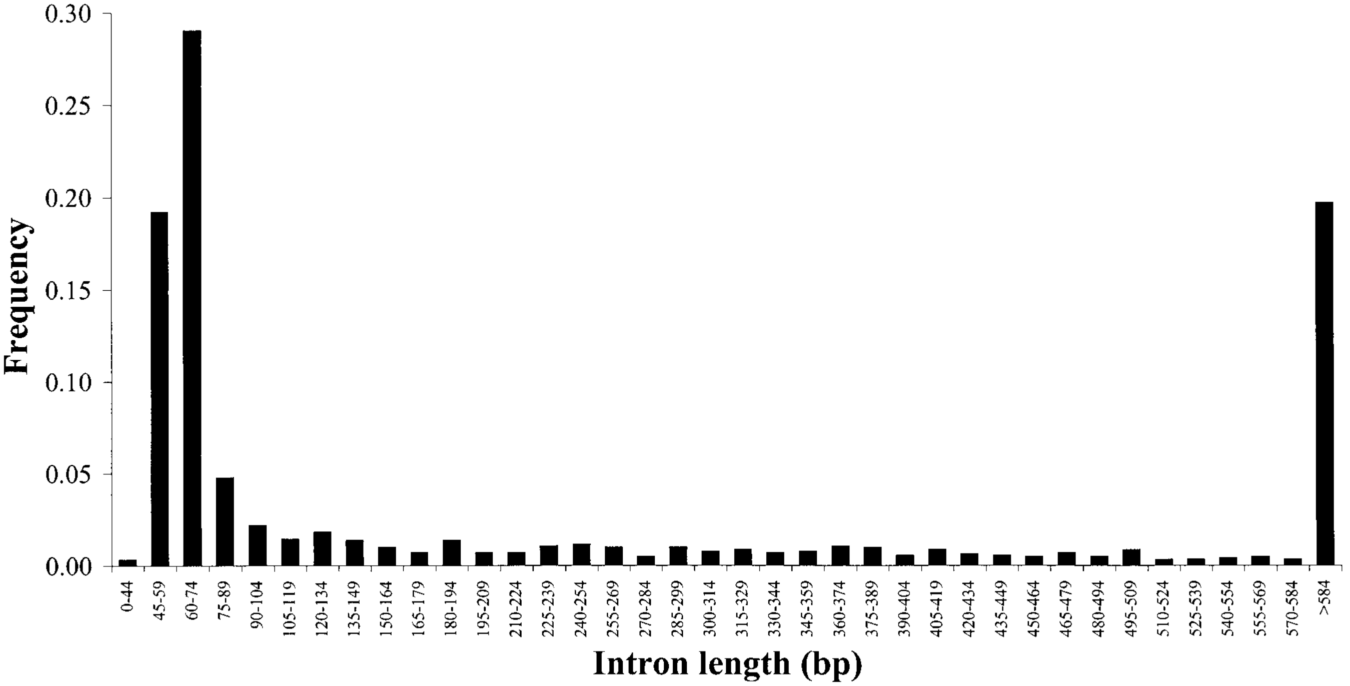

Figure 1.—Intron length distribution in *D. melanogaster*. A total of 1345 introns from 620 completely sequenced nuclear genes with accurate cytological position were studied (see materials and methods for details).

Relationship between recombination in *D. melanogaster* and both the length of the coding region and intron presence: The total lengths of coding regions and the number of exons are both weakly negatively correlated with recombination rates ($R = -0.1047$, $P = 0.0091$, $n = 620$; and $R = -0.085$, $P = 0.034$, $n = 620$, respectively). Since intron length is negatively correlated with recombination rate, we also detect a positive and significant correlation between total coding region length and total intron length ($R = -0.3833$ and $R = -0.4766$ for all genes and for genes with introns, respectively; $P < 10^{-6}$ in both cases). Similarly, average intron length correlates positively with total coding region length ($R = 0.2589$, $P < 10^{-6}$; $n = 447$) as well as with average exon length ($R = 0.1682$, $P = 0.00036$; $n = 447$). These correlations suggest at least two possibilities: (1) a recombination-sensitive mutational mechanism that affects both the introns and exons of a gene or (2) weak selection for smaller total gene length, including introns and exons, which is sensitive to the recombinational environment of a gene. Further consideration is given to these possibilities in the following section and in the discussion.

Similar mutational bias in regions of high and low recombination: One possible explanation for a negative correlation between intron length and recombination rates is a change in the deletion bias, *i.e.*, the deletion/insertion ratio, with recombination rate. To investigate this possibility we analyzed available data on polymorphic insertion and deletion events in *D. melanogaster* in introns and noncoding regions, with the aim of capturing the mutational deletion bias. Because the number of segregating sites (involving length polymorphisms) is expected to be less affected by selection than their frequency in the population (Crow and Kimura 1970), we have focused attention on the polymorphic presence of IE and DE in sampled populations. From studies of polymorphism based on either nucleotide sequencing or high resolution restriction fragment length polymorphism mapping, we identified a total of 31 genomic regions in which essentially every indel mutation in a sample could be detected (see materials and methods and Table 1).

With a total of 256 polymorphic indel events, we identified a significant excess of DEs compared to IEs (147 DEs *vs.* 109 IEs; $G = 5.66$, $P = 0.017$), with an overall polymorphic deletion bias (PDB) of 1.35. As shown in Table 1 there is no indication that the ratio of IEs to DEs differs between regions of high and low recombination ($G = 0.289$, $P = 0.591$). Neither intergenic regions nor introns show signs of a DE/IE bias that changes in relation to recombination rate (intergenic regions, $G = 0.020$, $P = 0.888$; introns only, $G = 0.006$, $P = 0.938$). The DE/IE ratios are similar for intergenic regions and introns ($G = 2.479$, $P = 0.115$) and the ratios do not change when the data are further subdivided between low and high recombination regions ($G = 0.718$, $P = 0.397$; and $G = 1.568$, $P = 0.210$, respectively). We conclude, therefore, that the negative correlation between intron length (or coding region length) and recombination is not caused by a mutational deletion/insertion bias or by a highly biased repair mechanism (see discussion) associated with recombination rates.

Size distribution of polymorphic indels: Figure 3

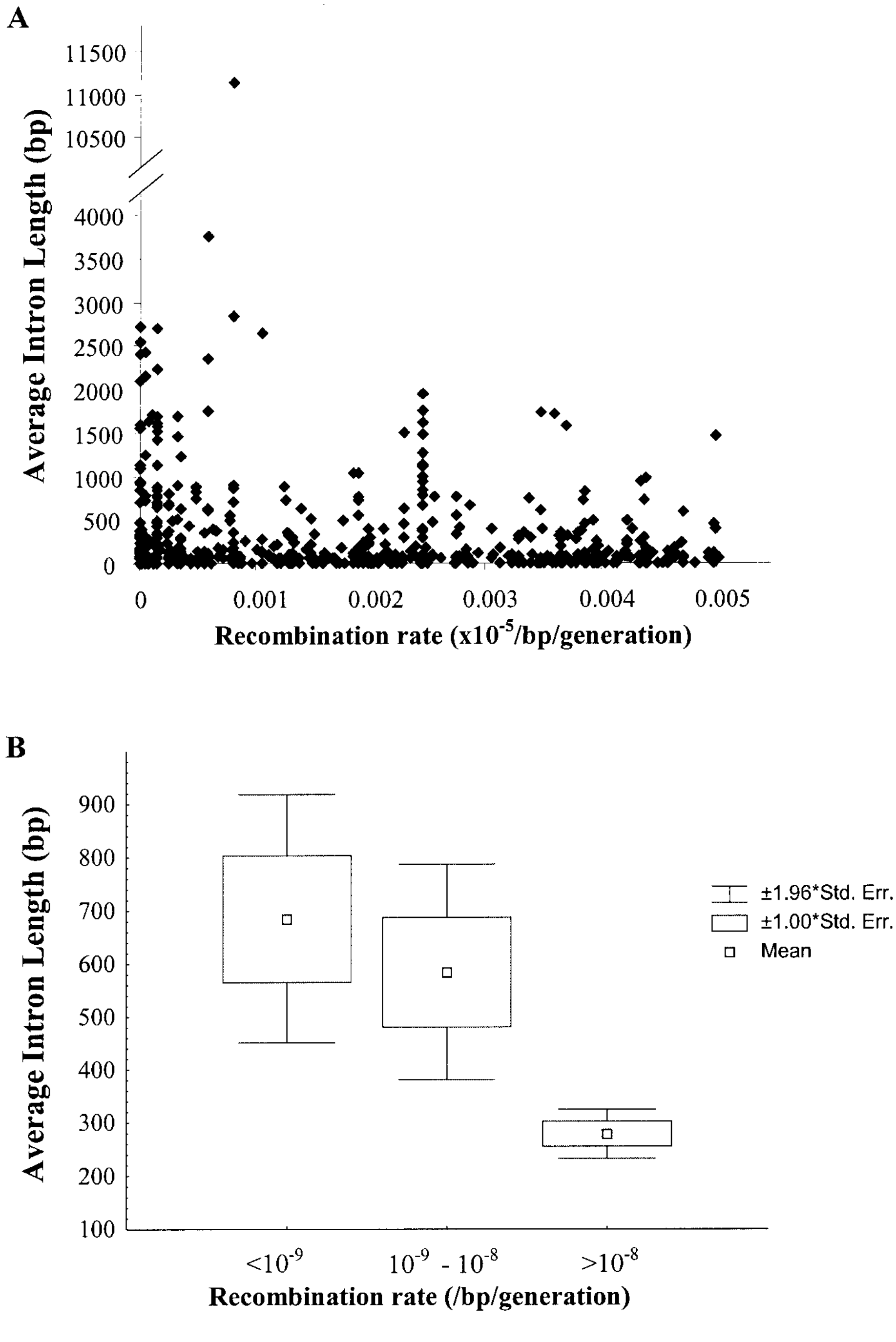

FIGURE 2.—(A) Relationship between the average intron length and recombination rates in *D. melanogaster* based on 447 genes with 1 or more introns and a total of 1345 introns; Spearman's rank correlation, $R = -0.2048$; $P = 1.3 \times 10^{-5}$. (B) Average and standard errors of the intron length when recombination rates are divided into three groups ($<1 \times 10^{-9}$, 1×10^{-9}–1×10^{-8}, and $>1 \times 10^{-8}$/bp/generation, with 55, 124, and 268 genes, respectively). A nonparametric Kruskal-Wallis ANOVA test of three recombination groups reveals that the average intron lengths in genes located in regions with low recombination are longer than those in regions of high recombination ($H = 21.12$; $P < 1 \times 10^{-5}$). Equivalent results are obtained when the gene EG:25E8.4 (FBgn0023526) with a single intron of 11.1 kb and an estimated recombination rate of 0.8×10^{-8}/bp/generation is not taken into account [Spearman's rank correlation $R = -0.2039$ ($P = 1.4 \times 10^{-5}$) and Kruskal-Wallis ANOVA test $H = 20.59$ ($P < 1 \times 10^{-5}$); $n = 446$].

shows the frequency distribution of indel sizes obtained from the polymorphism analyses in noncoding regions. Analyzing all indels shorter than 1000 bp, most length polymorphisms are 1–2 bp long (43.0 and 43.7% for insertion and deletions, respectively) and the great majority are within the range 1–10 bp (84.0 and 77.0% for insertion and deletions, respectively). The equivalent analyses only for indels located in introns and intergenic regions also show that the great majority of them are within the range 1–10 bp (89.2 and 75.4% for insertions and deletions, respectively, in introns, and 75.6 and 78.8% for insertions and deletions, respectively, in intergenic regions).

Even though long indels (longer than 100 bp) are a small fraction of all length changes, they nevertheless affect the estimation of average lengths. When only indels shorter than 100 bp are taken into account, the average length for insertions and deletions is 5.13 (±0.65 SE) and 7.01 (±0.76) bp, respectively, for all regions, and 5.04 (±0.84) and 7.52 (±1.20) bp for introns. The analyses of all indels shorter than 1000 bp, however, yield average insertion and deletion lengths of 40.3 (±15.1) and 16.3 (±4.7) bp, respectively, and 10.3 (±5.3) and 19.6 (±8.8) for insertions and deletions, respectively, in introns. Moreover, the only three polymorphic indels longer than 1000 bp detected in introns are insertions in genes located in regions of high recombination. This last observation, together with

TABLE 1

Polymorphic insertion and deletion events in *D. melanogaster*

					Insertion events (IE)				Deletion events (DE)			
Genomic region	Cytological position	Rec.[a]	n^b	m^c	Total	Introns	Intergenic	Coding	Total	Introns	Intergenic	Coding
Adh	35D1	2.01	20	8s, 13y	7	4	3	0	7	3	4	0
ase	1B4	0.00	6	1s, 1y	0	0	0	0	0	0	0	0
boss	96F11	3.86	5	5s	0	0	0	0	0	0	0	0
cec	99E4–99E5	2.77	10	1s, 1y	13	0	13	0	25	1	24	0
Ci	101F	0.00	10	9s, 1y	0	0	0	0	0	0	0	0
dpp	22F1	3.60	18		3	3	0	0	8	6	2	0
a-Est4	84D4	0.36	12	1s, 1y	0	0	0	0	3	0	3	0
Est-6	69A1	3.32	28	4s	0	0	0	0	0	0	0	0
hsp83	63C1	4.07	15	1s	0	0	0	0	1	0	0	1
Lcp-Psi	44D1–8	0.53	10	1s	1	0	1	0	1	0	1	0
MlcI	98B1–8	3.50	19	8s, 1y	28	22	5	1	17	16	1	0
Mst26Aa/Ab	26A1–5	4.40	92	1s, 1y	0	0	0	0	5	0	4	1
per	3B2	1.59	9	6s, 1y	0	0	0	0	2	1	1	0
Pgd	2D6	0.49	13 + 142*	1s	7	4	3	0	8	5	3	0
Pgi	44A–46D	0.75	21	14s, 13y	0	0	0	0	1	0	1	0
prune	2F1	0.81	5	3s	0	0	0	0	1	1	0	0
Ref(2)p	37E2–F1	0.56	10	1s	3	3	0	0	5	3	0	2
Rh3	92D1	3.47	5	5s	0	0	0	0	0	0	0	0
run	19E2	1.61	11	11s	0	0	0	0	1	1	0	0
Sod	68A8–9	3.85	25	1s	1	1	0	0	6	5	1	0
su(f)	20E1–2	0.39	50*		2	0	2	0	3	0	3	0
su(s)	1B10–C1	0.00	50**		1	0	1	0	2	1	1	0
su(wa)	1E1–2	0.07	50**		3	1	2	0	3	3	0	0
Tpi	99E	2.77	25	9s, 1y	2	2	0	0	3	3	0	0
tra	73A9	1.32	11	1s	1	0	1	0	0	0	0	0
Ver	10A1	2.55	71	70s	2	1	1	0	6	1	5	0
white	3C2	2.45	15	1s	12	11	1	0	15	14	1	0
y-ac-sc	1B	0.00	287*		13	5	7	1	18	2	16	0
yp2	9A2–5	3.54	6	6s	0	0	0	0	1	1	0	0
zeste	3A3	1.29	6	6s	1	0	0	1	0	0	0	0
Zw(G6pd)	18E1–5	2.48	33 + 64*	12s, 1y	9	7	2	0	5	5	0	0
Subtotal high recombination[d]					79	51	26	2	102	57	43	2
Subtotal low recombination[d]					30	13	16	1	45	15	28	2
Total					109	64	42	3	147	72	71	4

All studies were done on the basis of sequence comparisons except those labeled * and **, indicating fine restriction map length polymorphism, and SSCP and sequencing, respectively.

[a] Recombination rate × 10^{-8}/bp/generation (see MATERIALS AND METHODS).

[b] *n*, the number of *D. melanogaster* sequences under analysis.

[c] *m*, the number of *D. simulans* (s) and *D. yakuba* (y) sequences used in the analyses.

[d] Regions of high and low recombination are defined as those with recombination rates $>1 \times 10^{-8}$ and $<1 \times 10^{-8}$/bp/generation, respectively.

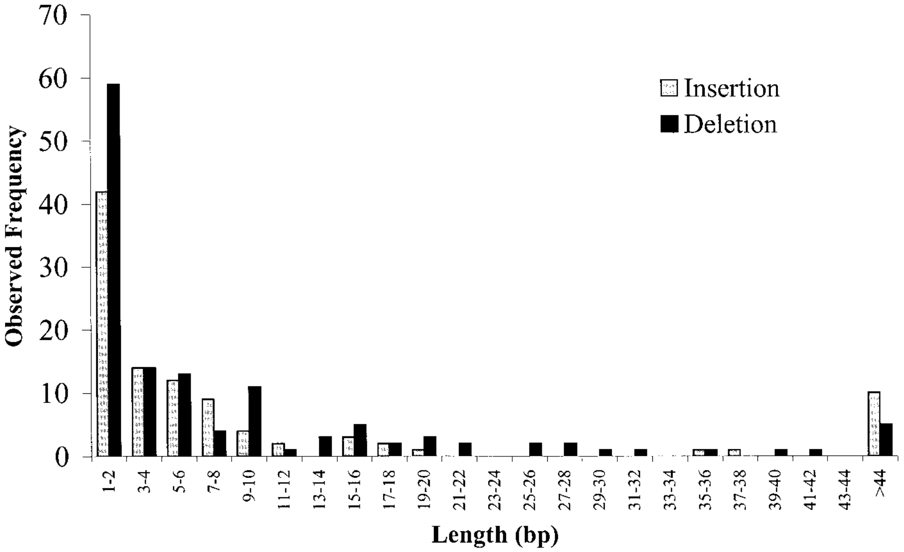

Figure 3.—Frequency distribution of indel sizes of 256 indel polymorphisms in *D. melanogaster* (see text for details).

the previously indicated DE/IE ratio, confirms that there are no identifiable differences in the mutational deletion bias between regions of high and low recombination that could generate longer introns in regions of low recombination.

Shorter introns in *D. simulans* than in *D. melanogaster*: Consistently larger effective population sizes (N_e) have been reported for *D. simulans* than for *D. melanogaster*, on the basis of analyses of nucleotide polymorphism levels (Aquadro *et al.* 1988; Moriyama and Powell 1996), synonymous codon usage, and the rate of synonymous and replacement evolution (Akashi 1995, 1996; Eanes *et al.* 1996). Moriyama and Powell's (1996) review of nucleotide polymorphism studies indicates an average difference of two- to threefold in N_e between *D. melanogaster* and *D. simulans*. *D. simulans* also has a higher rate of recombination than *D. melanogaster*, most conspicuous in regions where the recombination rate is low in *D. melanogaster* (True *et al.* 1996). If the presence of shorter introns in regions of higher recombination in *D. melanogaster* is a consequence of weak selection in *D. melanogaster*, then a species with greater recombination and larger N_e, *i.e.*, *D. simulans*, will be expected to exhibit smaller intron lengths. Akashi (1996) detected the presence of significantly smaller coding regions in *D. simulans* compared with *D. melanogaster*, suggesting a possible difference in effectiveness of selection as a cause (see discussion), but the comparison of total intron length in 22 genes (26 intron length differences) did not show any significant trend.

We compared a total of 211 fixed length differences between homologous sequences of the two species (Table 2). To reduce the probability of classifying polymorphic indels in one species as fixed differences, we used only those regions where multiple sequences are available in *D. melanogaster* and, when possible, in *D. simulans* (see Table 1). There is a significant excess of cases where the *D. melanogaster* noncoding sequence is longer than the homologous region in *D. simulans* (130 *vs.* 81, $P = 0.0009$; two-tailed sign test). This trend is highly significant for length differences in introns (66 *vs.* 27, $P = 0.00007$) but not in flanking regions (55 *vs.* 50, $P = 0.70$).

We also investigated the distribution of fixed indel differences along the two lineages leading to the common ancestor of these two species by comparing the homologous sequences in *D. yakuba*. Both the lineage and the direction of the indel mutation could be assigned without ambiguity for only 14 fixed indel mutations in introns (located in the genes *Adh*, *Ci*, *per*, and *Pgi*). The *D. melanogaster* lineage shows three IE and three DE, while the *D. simulans* lineage shows two IE and six DE. While suggestive of a possible difference, these numbers are too small to conclude that the observed trend is real.

DISCUSSION

The deletion to insertion mutational bias in Drosophila: Results from our analysis of 256 length polymorphisms suggest an overall deletion to insertion bias of 1.35; *i.e.*, an excess of deletions compared to insertions ($P = 0.017$). A 95% confidence interval (two-tailed *G*-test for goodness of fit) for deletion bias compatible with the estimated PDB is 1.02–1.78. In mammals, the average deletion bias obtained from analyzing 156 processed pseudogenes is 2.74 (Ophir and Graur 1997) and in primates it is 2.15 (Saitou and Ueda 1994); both values are close to the PDB observed in the our study.

This deletion bias in *D. melanogaster* is in the same direction although smaller ($G = 178.1$; $P < 1 \times 10^{-6}$) than the value of 8.7 estimated from analysis of the retroposon *Helena* in both *D. melanogaster* and *D. virilis* groups (Petrov *et al.* 1996; Petrov and Hartl 1998). The discrepancy between our results obtained from the analysis of polymorphic indels and those obtained from

TABLE 2

Fixed length differences between *D. melanogaster* and *D. simulans*

	Longer in *D. melanogaster*				Longer in *D. simulans*			
Genomic region	Total	Introns	Intergenic	Coding	Total	Introns	Intergenic	Coding
Adh	3	1	2	0	2	1	1	0
cec	19	0	19	0	9	0	9	0
Ci	0	0	0	0	1	1	0	0
α-*Est4*	9	0	9	0	16	0	16	0
Est-6	1	0	0	1	0	0	0	0
hsp83	2	0	2	0	2	0	2	0
Lcp-Psi	4	0	4	0	5	0	5	0
Mst26Aa/Ab	6	0	2	4	2	0	1	1
per	4	2	0	1	2	0	2	0
Pgd	18	6	12	0	14	4	10	0
Pgi	4	4	0	0	1	1	0	0
prune	1	0	1	0	2	1	1	0
Ref(2)p	6	5	1	0	2	2	0	0
run	2	2	0	0	1	1	0	0
Sod	4	2	2	0	3	3	0	0
tra	2	1	0	1	0	0	0	0
Ver	1	1	0	1	0	0	0	0
white	42	41	1	0	16	13	3	0
yp2	1	1	0	0	0	0	0	0
zeste	1	0	0	1	3	0	0	3
Total	130	66	55	9	81	27	50	4

Genomic regions *ase*, *boss*, *Tpi*, *MlcI*, *rh3*, and *Zw* do not show any fixed difference between *D. melanogaster* and *D. simulans*. See Table 1 for the number of alleles studied in each region and MATERIALS AND METHODS for details.

comparison of paralogous *Helena* elements indicates the presence of different mutation or selection mechanisms in retroelements *vs.* actively transcribed genes or their flanking regions.

Charlesworth (1996) indicated that because *Helena* retroposon length differences were between elements at different chromosomal locations and not allelic polymorphisms, it is not possible to reject an adaptive role in the fixation of retroposon indels. Empirical observations in Drosophila suggest that the elimination of transposable element sequences is favored by moderate negative selection (Langley *et al.* 1988; Charlesworth *et al.* 1992a,b). The preferential fixation of deletions over insertions by selection in *Helena* sequences might explain why longer deletions on average are found in *Helena* comparisons than in our noncoding and intron polymorphism data (≈50% of deletions longer than 10 bp in *Helena* sequences compared to 23% of deletions longer than 10 bp in our data). The lack of correspondence between data sets and approaches indicates the need for additional studies on the mutational deletion bias, the putative differences between heterochromatic and euchromatic (*i.e.*, actively transcribed) regions, and the correlation with recombination rates. Comparison of polymorphism data and fixed differences, we propose, will be a valuable tool.

Potential effect of selection on *D. melanogaster* indel polymorphism: We have investigated and can reject the possibility that our estimate of the ratio of DEs:IEs based on polymorphism data is strongly biased (*i.e.*, reduced) by the action of selection under a strong deletion bias. Strong selection ($\sigma \gg 1$; $\sigma = 4N_es$) acting on small noncoding indels can be eliminated from consideration because, if it were the predominant form of selection, each species would be essentially at mutation-selection balance for indels, and each species would be monomorphic, which is not the case. Weak selection ($\sigma \approx 1$) is also unlikely under a strong deletion bias because it would inescapably produce a difference in the PDB (see Figure 4) in regions of high and low recombination. With a 10- to 20-fold difference in the effective population sizes of high and low recombination regions of the *D. melanogaster* genome (Berry *et al.* 1991; Begun and Aquadro 1992), weakly selected mutations will be governed by selection in regions of high recombination and by genetic drift in regions of low recombination, leading to a difference in the PDB in the two regions of the genome. No difference in the PDB is detected in our data between regions of high and low recombination (Table 1). That this difference is not observed allows us to conclude that our estimated PDB = 1.35 based on segregating indels cannot be a strongly biased estimate of the mutational deletion bias (MDB). On the basis of this argument we can conclude that the

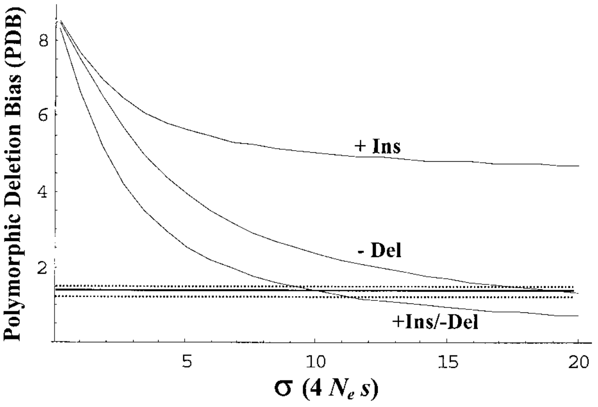

Figure 4.—Expected deletion bias at polymorphic level (PDB) under the infinitely many sites model with selection, free recombination, and semidominance (Crow and Kimura 1970), based on an MDB of 8.7 deletion events for each insertion event. The number of detected polymorphic insertion and deletion events was used to parameterize the expected number of segregating sites in a population and we obtained PDB as a joint function of σ ($4N_es$), the MDB, and the sample size. Under the model, the effect of selection and tight linkage would affect the actual number of indels but not their ratio. Moreover, the number of alleles under analysis has only a very small effect on PDB and does not change the expectation of PDB for very strong selection. Three different selective scenarios reducing the expected PDB are depicted: (1) positive selection on small insertions (+Ins), (2) negative selection on small deletions (−Del), and (3) positive and negative selection on small insertions and deletions, respectively (+Ins/−Del). The horizontal line at PDB = 1.35 indicates the observed polymorphic deletion bias based on the 256 indel events, while the dotted lines above and below indicate the polymorphic deletion bias in regions of low (PDB = 1.50) and high (PDB = 1.29) recombination, respectively. A sample size of 10 alleles has been assumed.

mutational process generates only a modest excess of deletions over insertions.

Intron presence and recombination rates in *D. melanogaster*: We report here a weak, but significant, negative correlation between recombination rates and the number of exons ($R = -0.085$, $P = 0.034$; $n = 620$). This observation is quite unexpected because there has been an extensive gene rearrangement within chromosomal arms in Drosophila evolution, meaning that genes located in regions of low recombination in *D. melanogaster* have been in the present recombinational environment for a relatively short period of time. Therefore, the observation of a tendency for genes located in regions of low recombination to exhibit a greater number of introns may be an indication that the tempo of intron gain/loss might be faster than usually believed.

Longer coding regions in the region of low recombination in *D. melanogaster*: We also noted a trend toward longer coding regions in regions of low recombination than in regions of high recombination in *D. melanogaster* ($R = 0.1047$, $P = 0.0091$; $n = 620$), a relationship previously undetected (Comeron *et al.* 1999). A possible explanation for this observation was proposed by Akashi (1996) to account for longer coding regions in *D. melanogaster* than in *D. simulans*. He argued that amino acid insertions are likely to be less deleterious than amino acid deletions and that a fraction of amino acid insertions will be weakly deleterious. Because selection is predicted to be less efficient in regions of low recombination, slightly deleterious amino acid insertions would be expected to differentially accumulate in these regions. This explanation is also in agreement with a higher rate of replacement substitutions (K_a) between *D. melanogaster* and *D. simulans* in regions of low recombination than in regions of high recombination (Takano 1998). Indeed, in the interspecific analysis between *D. melanogaster* and *D. simulans*, we found that both the K_a and the K_a/K_s values are significantly higher in those genes located in regions of low recombination than in those genes located in regions of high recombination (see materials and methods for details).

Natural selection acting on minimum intron length: It is generally accepted that there are structural constraints limiting minimum intron length and that natural selection must govern this and other requirements for proper splicing (Upholt and Sandell 1986; Tsurushita and Korn 1987; Mount *et al.* 1992). Additional constraints acting on intron sequences include the presence of gene regulatory controls in some introns and the formation of stem-loop structures that contribute to pre-mRNA stability (Schaeffer and Miller 1993; Stephan and Kirby 1993; Kirby *et al.* 1995; Leicht *et al.* 1995).

Carvalho and Clark (1999) indicated that both short and long introns tend to be overrepresented in regions of low recombination. They suggested that the lower effectiveness of selection in regions of low recombination allowed the differential accumulation of large and very short introns in genes located in these regions. As commented above, a minimum intron length is required for correct splicing, and introns that are shorter than this length are expected to be deleterious. We question, however, whether such a lower limit in intron length is governed by weak selection and therefore subject to the mutation-selection-drift balance. Based on our data, the length of "short" introns (defined as shorter than 60 bp when the modal intron length is 58 bp) is not correlated with recombination rates ($R = -0.0276$, $P = 0.656$; $n = 262$), whereas "long" introns (>60 bp or >80 bp) show the overall negative correlation between their length and recombination rates ($R = -0.2276$, $n = 1083$; and $R = -0.2194$, $n = 656$, respectively; $P < 1 \times 10^{-6}$ in both cases). Furthermore, short introns (<60 bp) are more abundant in region of *high* recombination than in regions of *low* recombination, contrary to weak selection predictions (21.1 and 16.1% for regions with recombination rates $>1 \times 10^{-8}$ and $<1 \times 10^{-9}$/bp/generation, respectively; an equivalent trend is observed when short introns are defined as those shorter than 80 bp). On the basis of these results, we suggest that minimum intron length is mainly subject to strong selection in Drosophila.

Natural selection acting on average intron length: Two lines of evidence suggest that natural selection is acting on intron length. First, as indicated in RESULTS, a mutational bias alone is unlikely to be causing the observed correlation between intron and recombination rate. Polymorphism data based both on the ratio of indel events and on the total size differences in introns comparing regions with high and low rates of recombination rule out this possibility. Thus there is a disparity between polymorphic data and the observed average intron length in *D. melanogaster* in relationship with recombination rates. Second, there are a higher number of fixed insertion differences in introns in *D. melanogaster* than in *D. simulans* that are unlikely to be caused by distinct mutational mechanisms between these two closely related species.

Among selective explanations for the general negative relationship between intron length and recombination, only weak selection acting on indels in introns would allow the evolution of different intron lengths without evident differences at the polymorphism level. Restriction fragment length polymorphism and direct sequencing studies show that small length polymorphisms are common in introns, indicating that many of them are not subject to strong selection. We propose, therefore, that the observed difference in intron length in regions of high and low recombination is molded by weak selection, and specifically that intron length in *D. melanogaster* is the result of a mutation-selection-drift balance.

Weak selection has considerably less influence on the presence of polymorphisms and their frequency in populations than it does on the probability of fixation. Figure 5, A and B, shows the predicted effect of weak selection on the deletion bias with respect to both polymorphism (PDB) and fixation probabilities [fixed deletion bias (FDB)] under the infinitely many sites model with free recombination and semidominance, and assuming a MDB ≈ 1.50. As the figure shows, selection coefficients small enough to be hardly detected in polymorphism data on the basis of small sample sizes can lead to conspicuous differences in intron length evolution (fixed differences) due to small differences in the expected segregating frequency. In accord with this prediction, and as commented above, there is no significant difference between the PDB observed in the region of high and low recombination. Also, we do not detect a significant difference in the average frequency of the 256 polymorphic insertions and deletions we analyzed (0.295 and 0.183, respectively; $P = 0.32$, applying Mann-Whitney *U*-test). We do find, however, significantly elevated frequency of insertions in regions of high recombination (0.3696 *vs.* 0.1732, respectively; Mann-Whitney *U*-test, $P = 0.025$). A similar trend is not detected in regions of low recombination (0.178 and 0.197 for inser-

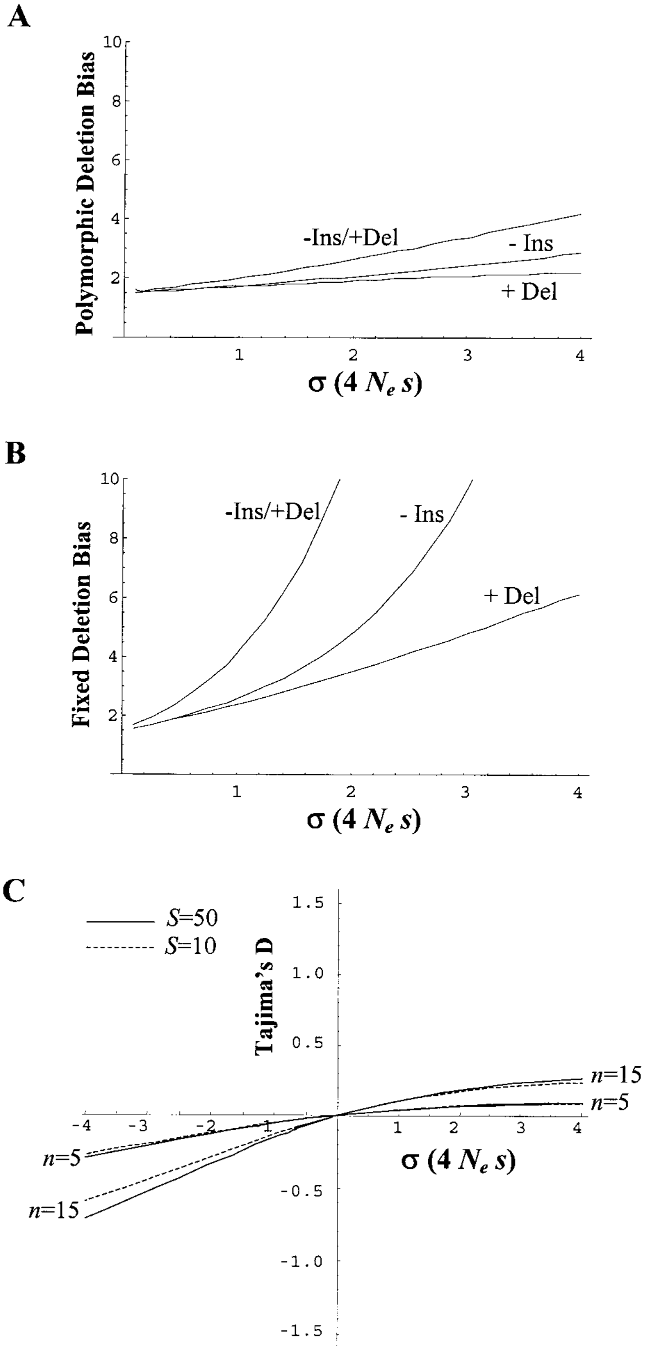

FIGURE 5.—Expected deletion bias at polymorphic (A) and at fixed difference (B) levels based on an MDB of 1.5 under the infinitely many sites model for weak selection ($\sigma < 4$). Three different selective possibilities are depicted as +Del, −Ins, and −Ins/+Del, indicating positive selection on small deletions, negative selection on small insertions, and negative and positive selection on small insertions and deletions, respectively. (C) Expected TAJIMA's (1989) *D* statistic for positive and negative weak selection for conditions of 5 and 15 (*n*) alleles and an expected number of segregating sites (*S*) under neutrality of 10 and 50, using both the expected number of segregating sites and their frequencies in the populations based on the infinitely many sites model (CROW and KIMURA 1970). In all cases, free recombination and semidominance were assumed.

tions and deletions, respectively). Using the 14 genomic regions with at least two IEs and at least two DEs, we also calculated Tajima's D statistic (Tajima 1989) with the aim of detecting possible directional frequency differences between insertions and deletions. No significant trend was detected: 8 regions showed Tajima's D greater (more positive) for insertions and 6 regions showed the opposite trend ($P > 0.8$). Nonsignificant results were also obtained for D statistic comparisons of indels in regions of high and low recombination. Tajima's D statistic, however, is insensitive to very weak selection (Figure 5C), especially when sample sizes are small (Akashi 1999). Thus we cannot reject the possibility of weak selection operating on intron indels.

In yeast, the repair of mispaired loops in heteroduplex DNA shows distinct biases, efficiencies, and repair pathways, for short and long indels, for nicked and continuous DNA, as well as for different genomic regions (Bishop *et al.* 1989; Kramer *et al.* 1989; Nag *et al.* 1989; Vincent and Petes 1989; Detloff *et al.* 1991; Kirkpatrick and Petes 1997; Lamb 1998; Corrette-Bennett *et al.* 1999). A generally biased repair toward deletions associated with recombination, which could cause shorter sequences in regions of high recombination, is, however, unlikely to be a contributing factor in indel dynamics in Drosophila based on our study of indel polymorphisms in *D. melanogaster*. A strong prediction of such a model is that deletions will segregate at a higher frequency than insertions, which is opposite to the observation (see above), and/or deletions will be more frequently present as polymorphisms than insertions (*i.e.*, a higher PDB) in regions of high recombination than in regions of low recombination, which again is not detected in the data. In addition, a model with DNA repair favoring deletions would not account for lengths longer than expected in regions of low recombination given a mutational excess of deletions compared to insertions. Nonetheless, biased mispair repair due to heteroduplex formation between heterozygotes is predicted to have effects similar to weak selection under appropriate (and very restrictive) rates of gene conversion, repair biases, and levels of heterozygosity, and the possible influence on indel evolution in specific eukaryotic genomes warrants further investigation.

Models of weak selection: What kind of selection might be operating to generate a correlation between intron length and recombination rate? Two nonmutually exclusive selective hypotheses are proposed and below we discuss their possible roles in shaping intron length in Drosophila: (i) selection favoring shorter transcripts to reduce transcriptional time and energy, and (ii) the selective advantage of long introns as reducers of interference among selected mutations.

Transcriptional efficiency: Biologically, longer transcripts must be inherently more costly to synthesize than short transcripts, owing to the increment in both transcription time and energetic costs of DNA and RNA synthesis. Assume for the purpose of this discussion that the absolute selective cost of a small insertion (or conversely the selective advantage of a small deletion) is small, *i.e.*, $\sigma \approx 1$. Then, given the 10- to 20-fold difference in the effective population size between high and low recombination regions of the *D. melanogaster* genome, it follows that the efficacy of selection acting on these indels will differ. Genes located in regions with higher recombination rates in *D. melanogaster* would be expected, under this scenario, to have shorter introns; these same introns will also be expected to be longer than their homologues in *D. simulans*, a species with a somewhat larger effective population size.

There are problems, however, with this hypothesis. Imagine that selection is nearly neutral in regions of low recombination ($\sigma_{LR} \approx 1$). Then, for regions of high recombination selection will be relatively strong, $\sigma_{HR} > 10$, and this will generate an obvious difference in the PDB, one that is not detected between regions (see above). As an alternative, imagine that $\sigma_{LR} \ll 1$ and $\sigma_{HR} = 1$. In this case, evolution of insertions in regions of low recombination will be determined by the mutational process (and biases) and not by selection. Since there is a significant deletion bias detected at the polymorphism level, mutational processes would push intron length toward uniformly short lengths, again something that is not observed in the data.

We used a forward computer simulation to explore a more complex mutation-selection-drift scenario (see materials and methods for details; Figure 6). In the presence of a mutational deletion bias, even a relatively weak one, and no selection, average intron length will evolve to be close to the minimum permissible length. Our simulations show that only when the deletion bias is weak and the occurrence of long insertions is frequent, is it possible for average intron length to be far enough from the minimum length to be compatible with the observed average size of introns in low recombination regions. Under these conditions, it could be possible for selection to act in regions of high recombination, where the effective population size is larger, to produce smaller average intron lengths. This particular scenario is unlikely, however, for the simple reason that sufficiently long insertions must occur at a vastly greater rate than is indicated from population genetic and evolutionary data to explain the difference between the average intron length in regions of high and low recombination (a difference of ∼400 bp). Thus, we do not believe that the energetic cost hypothesis and a mutation-selection-drift balance is a tenable *unique* explanation for the observed correlation between intron length and recombination rate in *D. melanogaster*.

An additional indication that selection favoring shorter introns to increase transcriptional efficiency is unlikely to be the general cause of longer introns in regions of low recombination comes from the study of human introns. Isochore GC richness is known to be

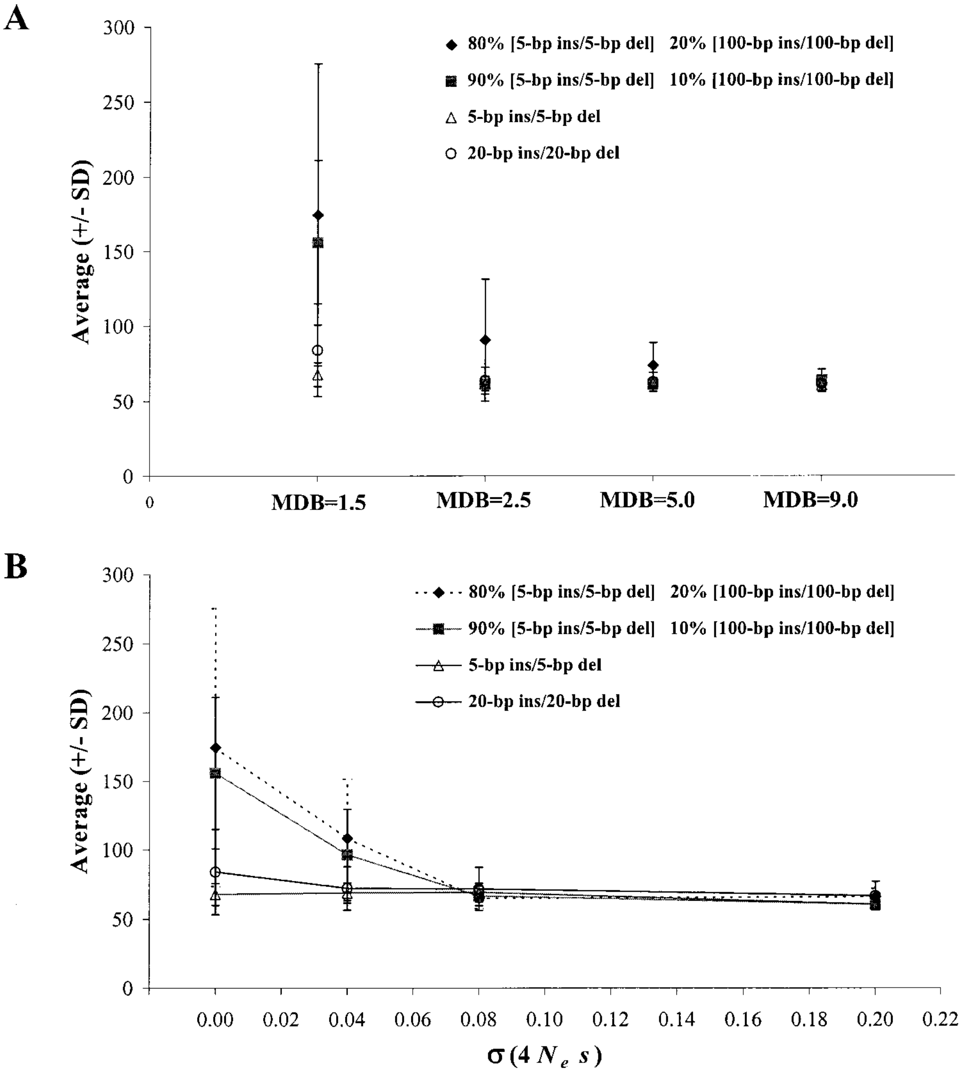

Figure 6.—Intron length (bp) average and standard deviation (SD) obtained from forward computer simulations under a mutation-selection-drift model (see materials and methods for details). The model imposes a minimum intron length ($L_{\min}$), with mutations that decrease the length below $L_{\min}$ being strongly selected against. Four different scenarios with diverse combinations of insertion (ins) and deletion (del) lengths were studied. The average and modal (data not shown) lengths were estimated each $4N$ generations and the results depicted are based on between 500 and 2000Ngenerations. (A) Intron length under neutrality for four different ratios of deletion-to-insertion events (MDB). (B) Intron length when the deleterious scaled selection coefficients for insertions, $\sigma(4Ns)$, per base pair are 0, 0.04, 0.08, and 0.2. In all cases $L_{\min}$ = 60 bp. As expected, scenarios with deletions longer than insertions are equivalent to a higher MDB (data not shown).

correlated positively with recombination rates (Eyre-Walker 1993) and negatively with intron length (Duret *et al.* 1995), thus suggesting that long introns are preferentially located in regions of low recombination. We have confirmed the overall negative relationship between intron length and recombination in humans, where there is also a mutational tendency toward DNA loss (Saitou and Ueda 1994; Ophir and Graur 1997). The analysis of 507 introns reveals a significantly negative relationship ($R = -0.1868$, $P = 2.3 \times 10^{-5}$) between intron length and recombination rates (a more detailed analysis will be presented elsewhere). As with the Drosophila data this trend is also detected when introns are divided into three recombination rate groups (Kruskal-Wallis ANOVA test; $H = 25.22$; $P < 0.0001$). Humans have, at least, one order of magnitude smaller long-term effective population size (N_e) than Drosophila (Li and Sadler 1991; Moriyama and Powell 1996), and weakly selected mutations in Drosophila ($\sigma \approx 1$) are expected to behave as effectively neutral in humans ($\sigma \ll 1$). Therefore, the observed negative relationship between intron length and recombination in both Drosophila and humans is hardly explained *only* by weak selection acting on transcriptional costs in both species, unless selection coefficients against longer transcripts in humans were much higher than in Drosophila for unknown reasons. Efficacy of selection arguments alone cannot explain the large differences in intron length observed across recombination rates.

Introns as modifiers of recombination in Drosophila: Given a mutational bias toward DNA loss, even if it is moderate, why do most eukaryotic genes have introns, and why, indeed, are some of them very long? The possible insertion of repetitive elements as a cause of long introns does not seem to have a significant role in *D. melanogaster* based on database search (see materials and methods; also see Moriyama *et al.* 1998).

As an explanation for intron persistence and long lengths, we propose that there might be situations in which a longer intron length is selectively advantageous. Otto and Barton (1997) have showed that mutations that increment recombination even slightly can increase the probability of fixation of advantageous mutations (see also Hey 1998). Length polymorphisms can be viewed as modifiers of recombination. As modifiers of recombination, deletions are expected to behave as dominant mutations and insertions as recessive ones. Insertions will effectively increase the distance and the expected recombination rate between bordering sites only in the homozygous state; *i.e.*, paired homologous chromosomes with a heterozygous insertion will create an insertion loop, not increasing the genetic distance between the sites on either side. Conversely, deletions will decrease the effective recombination rate when either heterozygous (creating a deletion loop) or homozygous. Interestingly, the expected FDB based on deleterious dominant deletions and favorable recessive insertions is equivalent to that expected for semidominant deleterious deletions and advantageous insertions (in both cases, FDB = MDB × $e^{-\sigma}$).

Longer introns reduce the Hill-Robertson effect compared to the same intron with a deletion and will also increase the hitchhiking effect between the favorable mutations and the longest variant. One selective advantage of introns, then, is to increase the recombination rate between selected mutations in different exons. This selective advantage of longer introns is expected to be greatest when the recombination rate is low and to diminish with increasing recombination (Barton 1995; Otto and Barton 1997), unless the number of beneficial mutations is large. Thus, the recombination modifier hypothesis (with variable selection coefficients related to recombination rates) can in principle explain why introns in regions of lower recombination are longer than those in regions of high recombination. We propose that the intron length-recombination rate correlation is an indication that selective interference across introns is sufficiently strong to allow selection to act against deletions and/or to favor insertions in introns in regions of low recombination. This scenario is congruent with the data: no clear pattern at the polymorphism level but manifest length differences both between regions of high and low recombination in *D. melanogaster* and between *D. melanogaster* and *D. simulans.* Without interference selection favoring insertions in introns, we cannot explain why intron length should be inversely correlated with recombination rate.

We further speculate that interference among many relatively weakly selected mutations will be the major cause of selection to reduce interference. This class of mutations is both abundant as polymorphisms and segregates at high frequencies in populations. Moreover, interference is maximized when selection coefficients are of equivalent magnitude (Hill and Robertson 1966). All of these conditions will increase the magnitude of interference when *summed over all interactions.* Mutations under strong positive selection, on the other hand, move through the population too fast to be affected by very small changes of recombination. Strongly deleterious mutations are also unlikely to be affected by indels since they are rare as polymorphisms and segregate at very low frequencies.

In Drosophila, synonymous mutations are the canonical example of weakly selected mutations, commonly observed as polymorphisms in coding regions (Akashi 1995; Moriyama and Powell 1996; Akashi and Schaeffer 1997; Kliman 1999; Llopart and Aguadé 2000). On the basis of analysis of 620 genes, we estimate that the average σ for synonymous mutations in *D. melanogaster* is 1.46 (95% confidence intervals of 0.26–2.9 and a mode of 1.31; see materials and methods for details). In addition, we have previously shown that selection for codon bias is even detectable in regions of very low recombination (Comeron *et al.* 1999). Interference among synonymous mutations may provide one of the selective benefits to the presence of introns (and to insertions making them longer). In support of this hypothesis, we recently showed that interference (*i.e.*, the Hill-Robertson effect) among synonymous mutations should be occurring in Drosophila, on the basis of simulation results of weakly selected (synonymous) mutations under a model with realistic parameters of recombination, mutation, effective population size, and selection (Comeron *et al.* 1999).

Codon bias is not the only form of weak selection that can promote selection for interference reduction. In Drosophila, both amino acid replacement changes (Ohta 1993; Akashi 1996; Moriyama and Powell 1996; Takano 1998; Schmid *et al.* 1999) and mutations in regulatory regions (Ludwig and Kreitman 1995; Ludwig *et al.* 2000) are likely to be weakly selected and this may be true in humans as well (Ohta 1995; Cargill *et al.* 1999). In mammals, selection for biased codon usage remains equivocal (Mouchiroud *et al.* 1995), but Eyre-Walker has recently shown that selection favors G/C ending codons over A/T ending codons in human genes and that this is likely to be driven by weak selection (Eyre-Walker 1999). A question that needs to be addressed for mammalian species is whether nucleotide polymorphism, even if it is all subject to selection, is abundant enough to promote selection for interference reduction.

A scenario can be envisaged in which a mutational bias favoring deletions and a selection bias favoring long introns in regions of low recombination will establish a dynamic equilibrium that will change according to the recombinational environment within a given genome and/or gene. The fact that intron length and recombination are negatively correlated in both Drosophila and humans suggests that selection acting on intron length might be a general characteristic in eukaryotic genomes. The outcome of this complex equilibrium with mutational biases toward shorter introns, weak selection (which would be species specific), and strong selection controlling endpoints for proper splicing, might explain different modal intron lengths for different organisms.

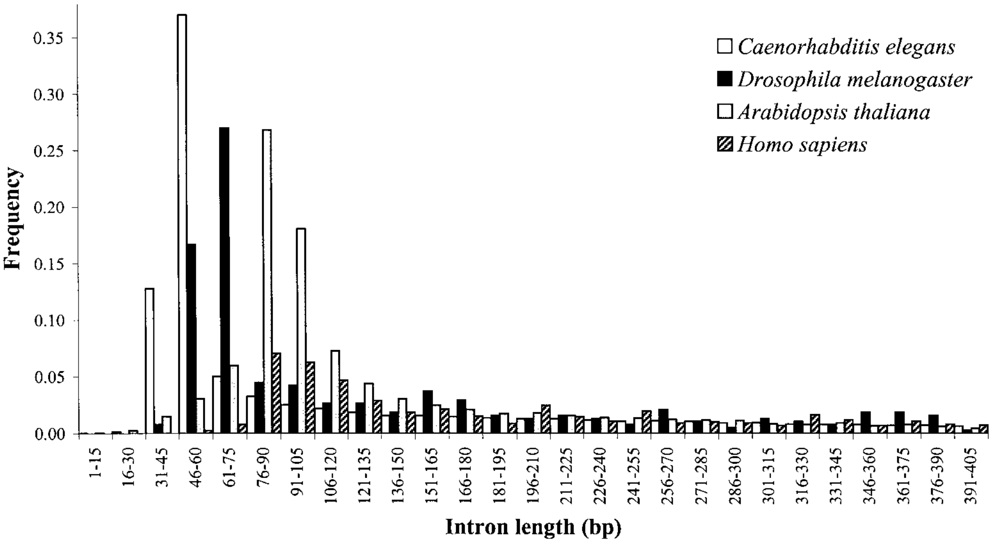

Figure 7.—Intron length distribution of four model organisms. The modal and average lengths (in parentheses) are 47 (277.0), 58 (402.5), 86 (183.6), and 130 bp (975.0) for *C. elegans* (n = 35,171 introns), *D. melanogaster* (n = 1345 introns), *A. thaliana* (n = 17,471 introns), and *H. sapiens* (n = 3240 introns), respectively. The percentages of introns >405 bp are 18.4, 25.7, 8.8, and 46.6% for *C. elegans*, *D. melanogaster*, *A. thaliana*, and *H. sapiens*, respectively.

For instance, comparison of model organisms (Figure 7) shows modal lengths of 47, 58, 86, and 130 bp for *Caenorhabditis elegans*, *D. melanogaster*, *Arabidopsis thaliana*, and *Homo sapiens*, respectively. Small differences in the deletion bias alone will not explain these differences.

Lengyel and Penman (1975) showed in an innovative study that the size of hnRNA but not mature mRNA increases with genome size in dipterans, indicative of a positive relationship between genome size and total intron length. A positive relationship between intron and genome size has now been documented for many other eukaryotic organisms (Hughes and Hughes 1995; Moriyama *et al.* 1998; Deutsch and Long 1999; Vinogradov 1999). A general prediction of our model is an expected negative relationship between recombination rates per physical unit and the distance between any cluster of weakly selected sites (*e.g.*, genes or exons). Accurate studies of recombination rates, polymorphism levels, and rates of genome evolution in natural populations of model organisms other than Drosophila and humans may allow us to investigate whether interference selection is required to account not only for intron lengths but for genome sizes as well. If so, this will shed entirely new light on the age-old "*C* value paradox."

We thank C. Bergman for alerting us to ways in which DNA repair mechanisms may influence in/del bias. We also thank P. Andolfatto, C. Bergman, A. Llopart, M. Long, T. Nagylaki, S. Orzack, E. Stahl, K. Thornton, and C. Toomajian for helpful discussions and J. Hey and three anonymous reviewers for their constructive comments. J.M.C. is supported by a Postdoctoral Fellowship from Ministerio de Educación y Ciencia, Spain. This research was supported by National Institutes of Health grant GM-39355 to M.K., and a Postdoctoral Fellowship from Ministerio de Educación y Ciencia, Spain, to J.M.C. J.M.C. dedicates this work to Maria Teresa Graupera Vila.

LITERATURE CITED

Aguadé, M., and C. H. Langley, 1994 Polymorphism and divergence in regions of low recombination in Drosophila, pp. 67–76 in *Non-Neutral Evolution: Theories and Molecular Data*, edited by G. B. Golding. Chapman & Hall, New York.

Akashi, H., 1995 Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. Genetics **139:** 1067–1076.

Akashi, H., 1996 Molecular evolution between *Drosophila melanogaster* and *D. simulans*: reduced codon bias, faster rates of amino acid substitution, and larger proteins in *D. melanogaster.* Genetics **144:** 1297–1307.

Akashi, H., 1999 Inferring the fitness effects of DNA mutations from polymorphism and divergence data: statistical power to detect directional selection under stationarity and free recombination. Genetics **151:** 221–238.

Akashi, H., and S. W. Schaeffer, 1997 Natural selection and the frequency distributions of "silent" DNA polymorphism in Drosophila. Genetics **146:** 295–307.

Aquadro, C. F., K. M. Lado and W. A. Noon, 1988 The rosy region of *Drosophila melanogaster* and *Drosophila simulans.* I. Contrasting levels of naturally occurring DNA restriction map variation and divergence. Genetics **119:** 875–888.

Aquadro, C. F., D. J. Begun and E. C. Kindahl, 1994 Selection, recombination and the levels of DNA polymorphism in Drosophila, pp. 45–56 in *Non-Neutral Evolution: Theories and Molecular Data*, edited by G. B. Golding. Chapman & Hall, New York.

Barton, N. H., 1995 Linkage and the limits of natural selection. Genetics **140:** 821–841.

Begun, D. J., and C. F. Aquadro, 1992 Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster.* Nature **356:** 519–520.

Berget, S. M., C. Moore and P. A. Sharp, 1977 Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. Proc. Natl. Acad. Sci. USA **74:** 3171–3175.

Berry, A. J., J. W. Ajioka and M. Kreitman, 1991 Lack of polymorphism on the Drosophila fourth chromosome resulting from selection. Genetics **129:** 1111–1117.

Bishop, D. K., J. Andersen and R. D. Kolodner, 1989 Specificity of mismatch repair following transformation of *Saccharomyces cerevisiae* with heteroduplex plasmid DNA. Proc. Natl. Acad. Sci. USA **86:** 3713–3717.

Cargill, M., D. Altshuler, J. Ireland, P. Sklar, K. Ardlie *et al.*, 1999 Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat. Genet. **22:** 231–238.

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

Charlesworth, B., 1994 The effect of background selection against deleterious mutations on weakly selected, linked variants. Genet. Res. **63:** 213–227.

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

Charlesworth, B., A. Lapid and D. Canada, 1992a The distribution of transposable elements within and between chromosomes in a population of *Drosophila melanogaster.* I. Element frequencies and distribution. Genet. Res. **60:** 103–114.

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

Charlesworth, B., M. Morgan and D. Charlesworth, 1993 The effect of deleterious mutations on neutral molecular variation. Genetics **134:** 1289–1303.

Chow, L. T., R. E. Gelinas, T. R. Broker and R. J. Roberts, 1977 An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. Cell **12:** 1–8.

Comeron, J. M., 1999 K-estimator: calculation of the number of nucleotide substitutions per site and the confidence intervals. Bioinformatics **15:** 763–764.

Comeron, J. M., M. Kreitman and M. Aguadé, 1999 Natural selection on synonymous sites is correlated with gene length and recombination in Drosophila. Genetics **151:** 239–249.

Corrette-Bennett, S. E., B. O. Parker, N. L. Mohlman and R. S. Lahue, 1999 Correction of large mispaired DNA loops by extracts of Saccharomyces cerevisiae. J. Biol. Chem. **274:** 17605–17611.

Crow, J. F., and M. Kimura, 1970 *An Introduction to Population Genetics Theory.* Alpha Editions, Edina, MN.

Deloukas, P., G. D. Schuler, G. Gyapay, E. M. Beasley, C. Soderlund *et al.*, 1998 A physical map of 30,000 human genes. Science **282:** 744–746.

de Souza, S. J., M. Long and W. Gilbert, 1996 Introns and gene evolution. Genes Cells **1:** 493–505.

Detloff, P., J. Sieber and T. D. Petes, 1991 Repair of specific base pair mismatches formed during meiotic recombination in the yeast *Saccharomyces cerevisiae.* Mol. Cell. Biol. **11:** 737–745.

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

Duret, L., D. Mouchiroud and C. Gautier, 1995 Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. J. Mol. Evol. **40:** 308–317.

Eanes, W. F., M. Kirchner, J. Yoon, C. H. Biermann, I.-N. Wang *et al.*, 1996 Historical selection, amino acid polymorphism and lineage-specific divergence at the *G6pd* locus in *Drosophila melanogaster* and *D. simulans.* Genetics **144:** 1027–1041.

Eyre-Walker, A., 1993 Recombination and mammalian genome evolution. Proc. R. Soc. Lond. Ser. B Biol. Sci. **252:** 237–243.

Eyre-Walker, A., 1999 Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. Genetics **152:** 675–683.

Felsenstein, J., 1974 The evolutionary advantage of recombination. Genetics **78:** 737–756.

FlyBase, 1998 FlyBase—A Drosophila database. Nucleic Acids Res. **26:** 85–88.

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.

Hawkins, J. D., 1988 A survey on intron and exon lengths. Nucleic Acids Res. **16:** 9893–9905.

Hey, J., 1998 Selfish genes, pleiotropy and the origin of recombination. Genetics **149:** 2089–2097.

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

Hudson, R. R., 1994 How can the low levels of DNA sequence variation in regions of the Drosophila genome with low recombination rates be explained? Proc. Natl. Acad. Sci. USA **19:** 6815–6818.

Hudson, R. R., and N. L. Kaplan, 1995 Deleterious background selection with recombination. Genetics **141:** 1605–1617.

Hughes, A. L., and M. K. Hughes, 1995 Small genomes for better flyers. Nature **377:** 391.

Hughes, A. L., and M. Yeager, 1997 Comparative evolutionary rates of introns and exons in murine rodents. J. Mol. Evol. **45:** 125–130.

Irvine, K. D., S. L. Helfand and D. S. Hogness, 1991 The large upstream control region of the Drosophila homeotic gene Ultrabithorax. Development **111:** 407–424.

Izban, M. G., and D. S. Luse, 1992 Factor-stimulated RNA Polymerase-II transcribes at physiological elongation rates on naked DNA but very poorly on chromatin templates. J. Biol. Chem. **267:** 13647–13655.

Kaplan, N. L., R. R. Hudson and C. H. Langley, 1989 The "hitchhiking effect" revisited. Genetics **123:** 887–899.

Kersanach, R., H. Brinkmann, M. F. Liaud, D. X. Zhang, W. Martin *et al.*, 1994 Five identical intron positions in ancient duplicated genes of eubacterial origin. Nature **367:** 387–389.

Kirby, D. A., S. V. Muse and W. Stephan, 1995 Maintenance of pre-mRNA secondary structure by epistatic selection. Proc. Natl. Acad. Sci. USA **92:** 9047–9051.

Kirkpatrick, D. T., and T. D. Petes, 1997 Repair of DNA loops involves DNA-mismatch and nucleotide-excision repair proteins. Nature **387:** 929–931.

Kliman, R. M., 1999 Recent selection on synonymous codon usage in Drosophila. J. Mol. Evol. **49:** 343–351.

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

Kramer, B., W. Kramer, M. S. Williamson and S. Fogel, 1989 Heteroduplex DNA correction in Saccharomyces cerevisiae is mismatch specific and requires functional PMS genes. Mol. Cell. Biol. **9:** 4432–4440.

Kreitman, M., and M. Aguadé, 1986 Genetic uniformity in two populations of Drosophila melanogaster as revealed by filter hybridization of four-nucleotide-recognizing restriction enzyme digests. Proc. Natl. Acad. Sci. USA **83:** 3562–3566.

Lamb, B. C., 1998 Gene conversion disparity in yeast: its extent, multiple origins, and effects on allele frequencies. Heredity **80:** 538–552.

Langley, C. H., E. Montgomery, R. Hudson, N. Kaplan and B. Charlesworth, 1988 On the role of unequal exchange in the containment of transposable element copy number. Genet. Res. **52:** 223–235.

Leicht, B. G., S. V. Muse, M. Hanczyc and A. G. Clark, 1995 Constraints on intron evolution in the gene encoding the myosin alkali light chain in Drosophila. Genetics **139:** 299–308.

Lengyel, J., and S. Penman, 1975 hnRNA size and processing as related to different DNA content in two dipterans: Drosophila and Aedes. Cell **5:** 281–290.

Li, W.-H., 1987 Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. J. Mol. Evol. **24:** 337–345.

Li, W.-H., and D. Graur, 1991 *Fundamentals of Molecular Evolution.* Sinauer, Sunderland, MA.

Li, W.-H., and L. A. Sadler, 1991 Low nucleotide diversity in man. Genetics **129:** 513–523.

Llopart, A., and M. Aguadé, 2000 Nucleotide polymorphism at the RpII215 gene in *Drosophila subobscura*: weak selection on synonymous mutations. Genetics **155:** 1245–1252.

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., C. Bergman, N. H. Patel and M. Kreitman, 2000 Evidence for stabilizing selection in a eukaryotic enhancer element. Nature **403:** 564–567.

Marchionni, M., and W. Gilbert, 1986 The triosephosphate isomerase gene from maize: introns antedate the plant-animal divergence. Cell **46:** 133–141.

Maynard Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favorable gene. Genet. Res. **23:** 23–35.

Moriyama, E. N., and D. L. Hartl, 1993 Codon usage bias and base composition of nuclear genes in Drosophila. Genetics **134:** 847–858.

Moriyama, E. N., and J. R. Powell, 1996 Intraspecific nuclear DNA variation in Drosophila. Mol. Biol. Evol. **13:** 261–277.

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

Mouchiroud, D., C. Gautier and G. Bernardi, 1995 Frequencies of synonymous substitutions in mammals are gene-specific and correlated with frequencies of nonsynonymous substitutions. J. Mol. Evol. **40:** 107–113.

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.

Nag, D. K., M. A. White and T. D. Petes, 1989 Palindromic sequences in heteroduplex DNA inhibit mismatch repair in yeast. Nature **340:** 318–320.

Ogata, H., W. Fujibuchi and M. Kanehisa, 1996 The size differences among mammalian introns are due to the accumulation of small deletions. FEBS Lett. **390:** 99–103.

Ohta, T., 1993 Amino acid substitution at the Adh locus of Drosophila is facilitated by small population size. Proc. Natl. Acad. Sci. USA **90:** 4548–4551.

Ohta, T., 1995 Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. J. Mol. Evol. **40:** 56–63.

Ophir, R., and D. Graur, 1997 Patterns and rates of indel evolution in processed pseudogenes from humans and murids. Gene **205:** 191–202.

Otto, S. P., and N. H. Barton, 1997 The evolution of recombination: removing the limits to natural selection. Genetics **147:** 879–906.

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.

Petrov, D. A., and D. L. Hartl, 1999 Patterns of nucleotide substitution in Drosophila and mammalian genomes. Proc. Natl. Acad. Sci. USA **96:** 1475–1479.

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

Powell, J. R., and E. N. Moriyama, 1997 Evolution of codon usage bias in Drosophila. Proc. Natl. Acad. Sci. USA **94:** 7784–7790.

Saitou, N., and S. Ueda, 1994 Evolutionary rates of insertion and deletion in noncoding nucleotide sequences of primates. Mol. Biol. Evol. **11:** 504–512.

Sambrook, J., 1977 Adenovirus amazes at Cold Spring Harbor. Nature **268:** 101–104.

Schaeffer, S. W., and E. L. Miller, 1993 Estimates of linkage disequilibrium and the recombination parameter determined from segregating nucleotide sites in the alcohol dehydrogenase region of *Drosophila pseudoobscura*. Genetics **135:** 541–552.

Schmid, K. J., L. Nigro, C. F. Aquadro and D. Tautz, 1999 Large number of replacement polymorphisms in rapidly evolving genes of Drosophila. Implications for genome-wide surveys of DNA polymorphism. Genetics **153:** 1717–1729.

Shah, D. M., R. C. Hightower and R. B. Meagher, 1983 Genes encoding actin in higher plants: intron positions are highly conserved but the coding sequences are not. J. Mol. Appl. Genet. **2:** 111–126.

Sharp, P. M., and W.-H. Li, 1989 On the rate of DNA sequence evolution in Drosophila. J. Mol. Evol. **28:** 398–402.

Shields, D. C., P. M. Sharp, D. G. Higgins and F. Wright, 1988 “Silent” sites in Drosophila genes are not neutral: evidence of selection among synonymous codons. Mol. Biol. Evol. **5:** 704–716.

Sorsa, V., 1988 *Chromosome Maps of Drosophila*. CRC Press, Inc., Boca Raton, FL.

Statistica for Windows 5.1, 1997 StatSoft, Tulsa, OK.

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

Stephan, W., T. H. Wiehe and M. W. Lenz, 1992 The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. Theor. Popul. Biol. **41:** 237–254.

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.

Tachida, H., 2000 Molecular evolution in a multisite nearly neutral mutation model. J. Mol. Evol. **50:** 69–81.

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

Takano, T. S., 1998 Rate variation of DNA sequence evolution in the Drosophila lineages. Genetics **149:** 959–970.

Tatusova, T. A., and T. L. Madden, 1999 Blast 2 sequences—a new tool for comparing protein and nucleotide sequences. FEMS Microbiol. Lett. **174:** 247–250.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin and D. G. Higgins, 1997 The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **25:** 4876–4882.

True, J. R., J. M. Mercer and C. C. Laurie, 1996 Differences in crossover frequency and distribution among three sibling species of Drosophila. Genetics **142:** 507–523.

Tsurushita, N., and L. J. Korn, 1987 Effects of intron length on differential processing of mouse mu heavy-chain mRNA. Mol. Cell. Biol. **7:** 2602–2605.

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.

Vincent A., and T. D. Petes, 1989 Mitotic and meiotic gene conversion of Ty elements and other insertions in *Saccharomyces cerevisiae*. Genetics **122:** 759–772.

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

Zeng, L.-W., J. M. Comeron, B. Chen and M. Kreitman, 1998 The molecular clock revisited: the rate of synonymous vs. replacement change in Drosophila. Genetica **102/103:** 369–382.

Communicating editor: J. Hey