Selective Constraints on Intron Evolution in Drosophila

John Parsch[1]

Department of Biology II, Section of Evolutionary Biology, University of Munich (LMU), Munich 80333, Germany

Manuscript received May 5, 2003
Accepted for publication August 11, 2003

ABSTRACT

Intron sizes show an asymmetrical distribution in a number of organisms, with a large number of "short" introns clustered around a minimal intron length and a much broader distribution of longer introns. In *Drosophila melanogaster*, the short intron class is centered around 61 bp. The narrow length distribution suggests that natural selection may play a role in maintaining intron size. A comparison of 15 orthologous introns among species of the *D. melanogaster* subgroup indicates that, in general, short introns are not under greater DNA sequence or length constraints than long introns. There is a bias toward deletions in all introns (deletion/insertion ratio is 1.66), and the vast majority of indels are of short length (<10 bp). Indels occurring on the internal branches of the phylogenetic tree are significantly longer than those occurring on the terminal branches. These results are consistent with a compensatory model of intron length evolution in which slightly deleterious short deletions are frequently fixed within species by genetic drift, and relatively rare larger insertions that restore intron length are fixed by positive selection. A comparison of paralogous introns shared among duplicated genes suggests that length constraints differ between introns within the same gene. The *janusA*, *janusB*, and *ocnus* genes share two short introns derived from a common ancestor. The first of these introns shows significantly fewer indels than the second intron, although the two introns show a comparable number of substitutions. This indicates that intron-specific selective constraints have been maintained following gene duplication, which preceded the divergence of the *D. melanogaster* species subgroup.

INTRONIC sequences, which interrupt exons and are removed through splicing, are nearly universal in eukaryotes (Nixon *et al.* 2002; Simpson *et al.* 2002). However, the general functional and evolutionary importance of introns remains unclear. Large-scale comparisons of intron sequences within genomes indicate that only a small fraction of their sequence contains information necessary for proper splicing (Mount *et al.* 1992). Aside from GT and AG dinucleotides at the 5′ and 3′ splice sites, respectively, and an A nucleotide required for branchpoint formation, there are no intronic sequences under absolute constraint. Preferred consensus sequences providing information for splice site and branchpoint selection are limited to a few nucleotides surrounding those positions and show a relatively high level of variation among introns (Mount *et al.* 1992; Long and Deutsch 1999). In addition, interspecific comparisons of orthologous introns indicate that there is little constraint on nucleotide sequence, as introns undergo nucleotide substitutions at rates comparable to pseudogenes and fourfold degenerate codon positions (Graur and Li 2000). This suggests that introns evolve neutrally (or nearly so) at the level of DNA sequence. Despite this apparent lack of primary sequence constraint, several observations suggest that intron size is subject to natural selection. For example, the distribution of intron lengths in *Drosophila melanogaster* and several other organisms with well-characterized genomes is asymmetrical, with many introns falling into a narrow distribution around a "minimal" intron length and the remaining introns showing a much broader distribution of lengths ranging from hundreds to thousands of base pairs (Mount *et al.* 1992; Deutsch and Long 1999; Yu *et al.* 2002). In *D. melanogaster*, minimal introns have lengths centered around 61 ± 10 bp (Yu *et al.* 2002), although the boundary separating introns into the "short" and "long" classes is not discrete (Comeron and Kreitman 2000). The relatively narrow length distribution of short introns suggests that natural selection may be involved in the maintenance of intron size.

Over evolutionary time, transitions from the short to the long intron size class appear to be rare events (Stephan *et al.* 1994; Moriyama *et al.* 1998). Stephan *et al.* (1994) compared 17 intron sequences available from at least two species of the *D. melanogaster* species subgroup and observed no changes in length class. In comparisons between more distantly related species (*i.e.*, *D. melanogaster vs. D. pseudoobscura* or *D. melanogaster vs. D. virilis*) transitions between size classes were observed, although these transitions were typically accompanied by an increase in polypyrimidine content just upstream of the 3′ splice site in the longer intron (Stephan *et al.* 1994). This observation is consistent with the proposal

[1] *Address for correspondence:* Department of Biology II, Section of Evolutionary Biology, University of Munich (LMU), Luisenstrasse 14, Munich 80333, Germany.
E-mail: parsch@zi.biologie.uni-muenchen.de

that different splicing mechanisms are used for short and long introns (Mount *et al.* 1992) and suggests that multiple compensatory mutations may be necessary for a size transition to occur.

Further evidence for natural selection acting on intron size comes from the relationship between intron length and recombination rate. Carvalho and Clark (1999) reported a significant negative correlation between intron length and recombination rate in *D. melanogaster*. This observation can be explained by natural selection, which is expected to be stronger in regions of high recombination, favoring shorter introns. In addition, introns in the size range of 60–80 bp occur on average more in regions of higher recombination than do introns shorter than 60 bp or introns longer than 80 bp, suggesting weak natural selection for both minimal and maximal intron length (Carvalho and Clark 1999). Comeron and Kreitman (2000) found a similar negative correlation between intron length and recombination in *D. melanogaster*, although they did not find evidence for weak natural selection against very short introns (<60 bp). These authors proposed that introns act as modifiers of recombination. Longer introns increase the probability of recombination between weakly selected sites in adjacent exons and thus reduce interference selection. Since interference between selected sites is expected to be greater in regions of low recombination, this model also predicts a negative correlation between intron length and recombination rate.

Finally, there is growing evidence for a functional link between intron length and gene expression. Castillo-Davis *et al.* (2002) reported a strong negative correlation between intron length and expression level in genomic surveys of both *Caenorhabditis elegans* and *Homo sapiens*. This can be explained by a negative fitness cost associated with the transcription of long introns. Since the transcription of apparently unnecessary intronic sequences costs the organism both time and energy (in the form of ATP), natural selection is expected to minimize intron length in genes that are transcribed at high levels. Further evidence suggests that introns of minimal length may be selectively maintained in genes due to a synergistic relationship between RNA processing and RNA export from the nucleus (Yu *et al.* 2002). A number of experimental studies in yeast, mice, and Drosophila have indicated that the presence of a short intron leads to higher levels of gene expression relative to an intronless gene (Choi *et al.* 1991; Palmiter *et al.* 1991; Holstege *et al.* 1998; Llopart *et al.* 2002). However, selection may not always favor the presence of short introns that increase gene expression. In the case of the *jingwei* gene, which shows an intron presence-absence polymorphism within *D. teissieri*, population genetics data suggest that the intronless form is favored by selection (Llopart *et al.* 2002).

In this article, patterns of nucleotide substitution, insertion, and deletion are analyzed for 15 introns from nine different genes across species of the *D. melanogaster* species subgroup. The advantage of comparing introns from within this species group is that they are divergent enough (at least 10 million years) for many changes to have occurred, yet similar enough to allow for reliable alignment. Because the phylogenetic relationship of these species is known, it is possible to classify indels as either insertions or deletions in most cases. In addition, the observed sequence changes are those that have been fixed between species and thus are changes that are positively selected, neutral, or only very slightly deleterious. The results indicate that, in general, short introns are not under greater sequence or length constraints than long introns. There is an overall indel bias toward short deletions. However, intron length is relatively well conserved across species, suggesting the selective fixation of less-frequent, longer insertions. Finally, a comparison of paralogous introns shared among duplicated genes suggests that length constraints may be intron-specific and can differ between introns within the same gene.

MATERIALS AND METHODS

Intron-containing sequences that were available from at least seven of the eight species of the *D. melanogaster* species subgroup (*D. melanogaster*, *D. simulans*, *D. sechellia*, *D. mauritiana*, *D. yakuba*, *D. teissieri*, *D. erecta*, or *D. orena*) were downloaded from GenBank. A recently described member of the species group, *D. santomea* (Lachaise *et al.* 2000), was not included in this study due to the paucity of available sequences. The final data set consisted of 15 introns from nine different genes: *Alcohol dehydrogenase* (*Adh*), *Amylase-related* (*Amyrel*), *Andropin* (*Anp*), *Cecropin C* (*CecC*), *janusA* (*janA*), *janusB* (*janB*), *ocnus* (*ocn*), *roughex* (*rux*), and *Superoxide dismutase* (*Sod*). The GenBank accession numbers for each gene are as follows: *Adh* (M17827, M36582, X04672, M19264, X54120, X54118, X54116, Z00032), *Amyrel* (U69607, U96159, AF039558, U96157, AF280878, AF280879, AF039562, U96158), *Anp* (X56726, AB047040–AB047045), *CecC* (Z11167, AB047056–AB047062), *janA* (M27033, AY013339–AY013344), *janB* (M27033, AY013345–AY013351), *ocn* (AF231190, AY013352–AY013358), *rux* (AE003436, AF327884–AF327890), and *Sod* (X13780, X15685, AF127155–AF127160).

To construct a phylogenetic tree of the *D. melanogaster* species subgroup, protein-encoding sequences from a subset of the above genes for which orthologous sequences were available from the outgroup species, *D. pseudoobscura*, were used. The accession numbers for the *D. pseudoobscura* sequences are X64489 (*Adh*), U82556 (*Amyrel*), S77099 (*janA* and *janB*), and U47871 (*Sod*). A 50% majority-rule consensus parsimony tree based on the concatenated protein-encoding sequences was generated using PAUP* (Swofford 2000). All nodes of this tree were supported by bootstrap values of at least 68%, with the exception of those connecting species of the *D. simulans* clade (*D. simulans*, *D. sechellia*, and *D. mauritiana*), which could not be resolved with >50% support. In this case, the three species were assumed to be equally related to each other, descending from a common polytomic node.

Intron sequences were aligned using a hierarchical approach. That is, the sequences were first aligned within three subsets on the basis of their phylogenetic relationships. The subsets were: (1) *D. melanogaster*, *D. simulans*, *D. sechellia*, and

TABLE 1

Intron lengths (in base pairs) in species of the *D. melanogaster* subgroup

Gene	Intron	Align[a]	mel	sim	sec	mau	yak	tei	ere	ore	CV[b]
Adh	1	758	654	654	653	665	617	666	614	619	0.035
	2	75	65	67	67	65	63	64	65	65	0.021
	3	89	70	66	66	68	64	67	76	61	0.066
Amyrel	1	62	56	58	57	58	57	57	57	60	0.021
Anp	1	62	62	62	62	62	62	62	—	56	0.037
CecC	1	70	69	70	70	61	70	70	70	70	0.046
janA	1	58	58	58	—	58	58	58	58	58	0.000
	2	128	105	106	—	106	103	78	102	103	0.010
janB	1	73	58	64	64	64	64	64	65	69	0.047
	2	57	57	57	57	57	57	57	57	57	0.000
	3	68	61	61	61	61	59	59	66	63	0.037
ocn	1	54	54	54	54	54	52	52	52	50	0.028
	2	69	55	57	48	66	52	59	55	56	0.094
rux	1	116	90	90	90	90	94	104	105	95	0.067
Sod	1	822	725	730	731	739	708	783	709	782	0.040

mel, *D. melanogaster*; sim, *D. simulans*; sec, *D. sechellia*; mau, *D. mauritiania*; yak, *D. yakuba*; tei, *D. tessieri*; ere, *D. erecta*; ore, *D. orena*.

[a] Total base pairs (including gaps) in the sequence alignment.

[b] Coefficient of variation for intron length among species.

D. mauritiana; (2) *D. yakuba* and *D. teissieri*; and (3) *D. erecta* and *D. orena*. Initial alignments were performed using ClustalX (Thompson *et al.* 1997) with a gap opening penalty of 15 and a gap extension penalty of 5. A complete alignment of all species was then generated by aligning the subsets using the gap penalties given above and without resetting gaps. For some of the introns, the computer-generated alignments were adjusted by eye. In these cases, the general strategy was to favor mismatches to minimize the number of gaps, while ensuring that the 5′ (GT) and 3′ (AG) splice signals and other conserved sequence blocks remained aligned. The complete alignments are presented in supplemental Figure 1 available at http://www.genetics.org/supplemental/. The numbers of substitutions, insertions, and deletions that have occurred in each intron were inferred by parsimony, assuming the phylogenetic relationship indicated by the protein-encoding sequences. In the case of the *D. simulans* species complex, for which the phylogenetic relationship was unclear, a conservative approach was used. That is, a substitution or indel shared by any two of the three species was assumed to have a single origin. In the case of ambiguous indels (those that could not be classified as insertions or deletions due to the lack of an appropriate outgroup sequence), the indel was assigned the minimum length possible under parsimony. A complete list of all indels and their lengths is provided in supplemental Figure 2 available at http://www.genetics.org/supplemental/.

RESULTS

Intron length variation in the *D. melanogaster* species subgroup: The data set consists of 15 introns from nine different genes (Table 1). Of the 15 introns, 13 fall into the short-size class (average length range is 53–100 bp), and 2 fall into the long-size class (average lengths are 643 and 738 bp). Consistent with previous reports (Stephan *et al.* 1994), there are no changes from the short to the long intron class within the *D. melanogaster* species subgroup. For each intron, sequences were available from all eight species of the subgroup, with the exception of the two *janA* introns (which were unavailable from *D. sechellia*) and the *Anp* intron (which was unavailable from *D. erecta*). The total length of the aligned intron sequences was 2561 bp. This includes 981 bp from short introns and 1580 bp from long introns. A summary of the intron lengths is given in Table 1. The two long introns show greater length changes among species in terms of numbers of base pairs, but there is not greater variance in intron length in long introns after correcting for intron size. The average coefficient of variation (CV) for short intron length is 4.3% and the average CV for long intron length is 3.9%. Thus there is no evidence for greater length constraints on short introns. If anything, the short introns show greater length variation, although this is not significant, given the limited sample size of long introns.

Comparison of nucleotide substitutions and indels: A consensus parsimony tree of the *D. melanogaster* species subgroup based on the concatenated protein-encoding sequences of the *Adh*, *Amyrel*, *janA*, *janB*, and *Sod* genes is shown in Figure 1. These genes were chosen due to the availability of an orthologous sequence in *D. pseudoobscura*, which was used as an outgroup. The same general topology was produced using the concatenated intron sequences of all nine genes used in this study (not shown), although an outgroup sequence could not be used for the introns due to either the lack of an available sequence or ambiguity of alignment. There is some uncertainty as to the relationship of the species of the *D. simulans* complex (*D. simulans*, *D. sechellia*, and *D. mauritiana*). This uncertainty is likely due to shared ancestral alleles persisting in the three extant species fol-

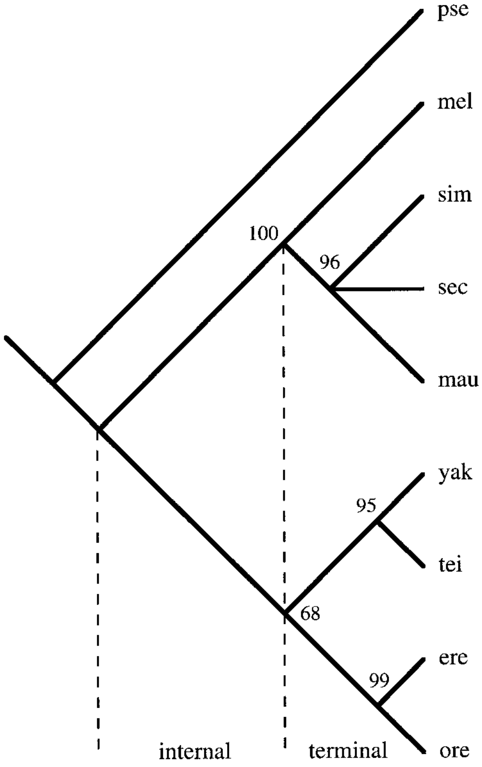

Figure 1.—Bootstrap 50% majority-rule consensus cladogram of the *D. melanogaster* species subgroup. The tree is based on concatenated protein-encoding sequences of the *Adh*, *Amyrel*, *janA*, *janB*, and *Sod* genes. mel, *D. melanogaster*; sim, *D. simulans*; sec, *D. sechellia*; mau, *D. mauritiana*; yak, *D. yakuba*; tei, *D. tessieri*; ere, *D. erecta*; ore, *D. orena*. *D. pseudoobscura* (pse) was used as an outgroup to root the tree. Bootstrap values (1000 replicates) are given at each node. This topology was used to infer numbers of substitutions and indels occurring within introns. Branches connecting the two major clades within the species subgroup were considered internal, while those within each clade were considered terminal.

lowing speciation (Kliman *et al.* 2000; Ting *et al.* 2000). To be conservative, a tree in which these three species coalesce at a common, polytomic ancestral node was assumed for this article (Figure 1).

The "two-clade" structure of the *D. melanogaster* species subgroup presented in Figure 1 differs slightly from the traditionally assumed phylogeny for this group, which places *D. yakuba* and *D. teissieri* in a clade with *D. melanogaster* and the *D. simulans* complex species (Ashburner 1989; Powell 1997). It should be noted, however, that this traditional phylogeny was based on nonmolecular data or on DNA sequence from a single gene, *Adh.* The phylogenetic relationship presented here is based on DNA sequences from *Adh,* plus four other genes. The same topology is generated using maximum-likelihood and distance methods, which support the *D. yakuba, D. teissieri, D. erecta,* and *D. orena* clade with bootstrap values of 66 and 97%, respectively. This clade is further supported by a recently developed Bayesian method, which samples the posterior probability of trees generated by maximum likelihood (Huelsenbeck and Ronquist 2001). Using this method, the posterior probability of the above clade is 64%. None of the above methods support the traditional phylogeny with probabilities >15%. If each gene is considered separately (instead of using a concatenated sequence), only *Adh* provides consistent support for the traditional phylogeny. The *janA* and *janB* genes each support the phylogeny shown in Figure 1. *Sod* supports a third tree that places *D. erecta* and *D. orena* in a clade with *D. melanogaster* and the *D. simulans* complex. The *Amyrel* sequence does not support any of the above trees with bootstrap values >50%. A recent phylogenetic study of the *D. melanogaster* subgroup using DNA sequences of the *Adh, Adhr, Gld,* and *ry* genes and more closely related outgroup species also strongly supports the tree shown in Figure 1 (Ko *et al.* 2003). On the basis of these results, the relationship depicted in Figure 1 was used to infer the numbers of base substitutions and indels occurring in the intron sequences by parsimony (see materials and methods).

For the entire intron data set, 972 nucleotide substitutions and 176 indels were inferred. The 13 short introns had 486 substitutions and 74 indels, while the 2 long introns had 486 substitutions and 102 indels. The difference in the substitution/indel ratio between short and long introns is significant ($\chi^2 = 3.8$; $P = 0.05$). This difference could be due to either an increased rate of indels or a decreased rate of substitutions in long introns relative to short introns. The latter explanation is better supported by the data. Indel rates (corrected for intron length) are very similar between the short and long introns, with short introns showing 0.08 indels/bp and long introns showing 0.06 indels/bp. However, substitution rates differ significantly between the two intron classes, with 0.50 substitutions/bp in short introns and 0.31 substitutions/bp in long introns ($\chi^2 = 39.7$; $P < 0.001$). It should be noted that the above comparison of substitution rates is conservative, due to the fact that three of the short intron sequences were available from only seven of the eight species compared in this study. The total number of substitutions inferred by parsimony from an alignment of seven sequences will necessarily be less than (or equal to) that inferred from an alignment of eight sequences. This result suggests greater selective constraint on the DNA sequence of long introns, perhaps because they contain additional regulatory sequences that are subject to purifying selection. However, this interpretation is inconsistent with the observation that conserved intronic regions with presumed regulatory function experience far fewer indels than substitutions in comparisons between *D. melanogaster* and *D. virilis* (Bergman and Kreitman 2001). More sequences of long introns from across the *D. melanogaster* species subgroup are needed to confirm the substitution rate difference between short and long introns.

Indel size distribution: Of the 176 indels inferred from the intron alignments, 93 (53%) could be classified as deletions and 56 (32%) could be classified as insertions. The remaining 27 (15%) of the indels were ambiguous. This is due mainly to cases where the indels differed between the two clades within the species subgroup (Figure 1). That is, *D. melanogaster*, *D. simulans*, *D. sechellia*, and *D. mauritiana* all shared an indel not present in *D. yakuba*, *D. teissieri*, *D. erecta*, or *D. orena*. For the entire data set, there is a significant excess of deletions relative to insertions (χ^2 = 9.2; P = 0.002), with a deletion/insertion ratio of 1.66. This pattern holds for both the short and long intron classes. For the short introns, the deletion/insertion ratio is 1.71 (χ^2 = 4.5; P = 0.035); for the long introns, it is 1.63 (χ^2 = 4.8; P = 0.029). The above estimate is in reasonable agreement with the 1.35 deletion/insertion ratio reported for indel polymorphisms within *D. melanogaster* introns (Comeron and Kreitman 2000).

The indel size distribution is also in good agreement with that observed by Comeron and Kreitman (2000), with 57% of the deletions and 48% of the insertions being either 1 or 2 bp in length (Figure 2). Ninety percent of the deletions and 94% of the insertions were <10 bp. In general, deletions tended to be slightly longer than insertions, with average lengths of 4.59 and 3.50 bp, respectively, although this difference is not significant (Mann-Whitney test, P = 0.70). For the short introns, deletions and insertions averaged 3.54 and 3.63 bp, respectively (Mann-Whitney test, P = 0.28); for long introns, deletions and insertions averaged 5.42 and 3.41 bp, respectively (Mann-Whitney test, P = 0.67).

Lengths of indels occurring along internal and terminal branches: As mentioned above, 15% of the indels were classified as "ambiguous," because they could not be polarized as either insertions or deletions. It is likely, however, that many of these events represent insertions, because the total intron length is well conserved among species (Table 1) and deletions are predominant among the indels that could be classified (Table 2). In general, the ambiguous indels are longer than those that could be classified as insertions or deletions (Figure 2). The average length of the ambiguous indels is 7.22 bp, while the average length of all other indels (insertions and deletions combined) is 4.18 bp. The length difference between the two classes is highly significant (Mann-Whitney test, P = 0.008). This pattern holds for both the short and long introns: 7.11 bp for ambiguous *vs.* 3.57 bp for all other indels within the short introns and 7.28 bp for ambiguous *vs.* 4.65 bp for all other indels within the long introns. The length difference is marginally significant within both the short (Mann-Whitney test, P = 0.066) and long (Mann-Whitney test, P = 0.062) intron classes.

The *D. melanogaster* species subgroup is composed of two clades of closely related species separated by relatively long internal branches. Most of the ambiguous indels occur on these internal branches and cannot be classified as either insertions or deletions due to the lack of an appropriate outgroup sequence. However, some indels are classified as ambiguous if they overlap with other indels occurring within a particular clade. Of the 27 ambiguous indels, 24 fall into the first category (average length is 7.88 bp) and 3 fall into the second category (average length is 2.00 bp). When the indels are classified as either internal branch or terminal branch (Figure 1), there is a highly significant length difference with internal branch indels averaging 7.88 bp and the terminal branch indels averaging 4.14 bp (Mann-Whitney test, P = 0.0017). The length difference between internal branch and terminal branch indels is

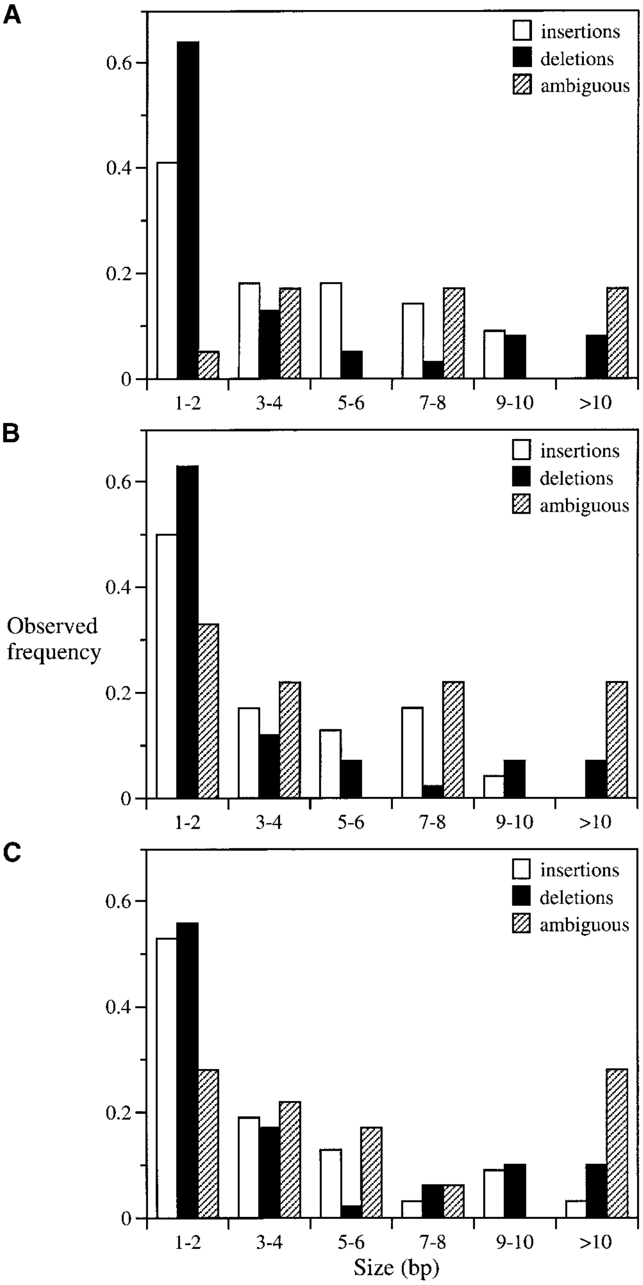

Figure 2.—Size distribution of insertions, deletions, and ambiguous indels in (A) all introns, (B) short introns, and (C) long introns.

TABLE 2

Numbers of substitutions and indels in introns

Gene	Intron	Sub	Sub/bp[a]	Indels	Indels/bp[a]	Indels/sub	Del	Ins	Del/ins
Adh	1	169	0.22	56	0.07	0.33	24	25	0.96
	2	34	0.45	10	0.13	0.29	5	5	1.00
	3	45	0.51	10	0.11	0.22	3	6	0.50
Amyrel	1	41	0.66	5	0.08	0.12	2	3	0.67
Anp	1	30	0.48	2	0.03	0.07	2	0	—
CecC	1	36	0.51	2	0.03	0.06	2	0	—
janA	1	40	0.69	0	0.00	0.00	0	0	—
	2	60	0.47	12	0.09	0.20	7	2	3.50
janB	1	35	0.48	6	0.08	0.17	4	1	4.00
	2	28	0.49	0	0.00	0.00	0	0	—
	3	25	0.37	3	0.04	0.12	1	1	1.00
ocn	1	28	0.52	3	0.06	0.11	3	0	—
	2	34	0.49	10	0.14	0.29	6	4	1.50
rux	1	50	0.43	11	0.09	0.22	6	2	3.00
Sod	1	317	0.39	46	0.06	0.15	28	7	4.00
All introns		972	0.38	176	0.07	0.18	93	56	1.66
Short introns		486	0.50	74	0.08	0.15	41	24	1.71
Long introns		486	0.31	102	0.06	0.21	52	32	1.63

Sub, substitutions; del, deletions; ins, insertions.

[a] Total base pairs (including gaps) in the sequence alignment.

significant for both the short and long introns. For short introns, internal branch indels average 7.88 bp and terminal branch indels average 3.53 bp (Mann-Whitney test, $P = 0.019$). For long introns, internal branch indels average 7.88 bp and terminal branch indels average 4.60 bp (Mann-Whitney test, $P = 0.033$).

Length constraints on paralogous introns: The *janA*, *janB*, and *ocn* genes arose through two separate gene duplication events, both of which predate the divergence of the *D. melanogaster* species subgroup (Yanicostas *et al.* 1995; Parsch *et al.* 2001b). The three genes share two paralogous introns derived from a common ancestral gene (Figure 3). Although these introns are too divergent among genes to be aligned by DNA sequence, their paralogy is supported by their conserved location within the aligned protein-encoding regions and by the phase with which they interrupt codons. In all three genes the first intron is located between a first and a second codon position, while the second intron is located between a third and a first codon position. The *janB* gene has an additional 5′ intron that is not present in *janA* or *ocn* (Figure 3). In comparisons among species of the *D. melanogaster* species subgroup, the two parlogous introns show comparable numbers of base substitutions, but differ markedly in numbers of indels. For the three genes combined, the first paralogous intron has 96 substitutions and 3 indels, while the second has 119 substitutions and 25 indels. This difference in indel/substitution ratios is highly significant ($\chi^2 = 11.8$; $P < 0.001$), indicating different rates of indel accumulation in the two introns. The difference is unlikely to be explained by indel-specific mutational differences, because the introns are only 125 bp apart within each gene and the three genes lie in tandem within a 2.5-kb region of chromosome arm 3R. Thus it appears that selective constraints with regard to indels may differ among short introns within the same gene. In the case of *janA*, *janB*, and *ocn*, the first paralogous intron appears to be under much stronger selective constraints to maintain length than the second.

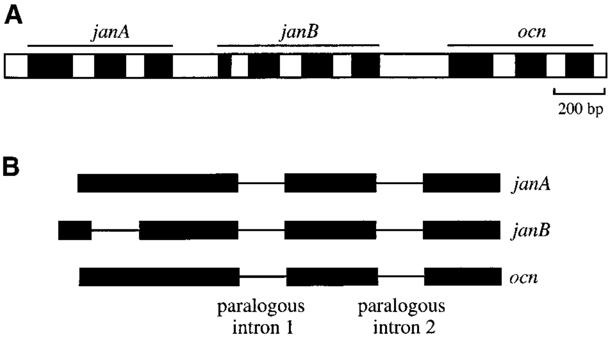

Figure 3.—(A) Genomic organization of the *janA*, *janB*, and *ocn* genes. In *D. melanogaster*, the three genes lie in tandem in a 2.5-kb region of chromosome arm 3R. (B) Schematic alignment of the three paralogous genes. Protein-encoding regions are shown as solid boxes.

DISCUSSION

A comparison of 15 orthologous intron sequences from eight species of the *D. melanogaster* species subgroup revealed a total of 176 indels that have occurred since the divergence of the species subgroup ∼10 million years ago. Of the indels that could be classified as either insertions or deletions, there was a significant excess of deletions (deletion/insertion ratio is 1.66). Furthermore, the vast majority of the indels were <10 bp in length (90% for deletions, 94% for insertions). These results are comparable to those reported by Comeron and Kreitman (2000) for indel polymorphisms occuring within introns of *D. melanogaster*. Those authors reported a deletion/insertion ratio of 1.35, with 77% of the deletions and 84% of the insertions <10 bp. This suggests that the intronic indels segregating within species closely reflect those that become fixed between species. In the more distantly related *D. pseudoobscura*, a slightly different pattern of indel polymorphism has been observed. Schaeffer (2002) surveyed polymorphism in the *Adh* and *Adhr* genes and found a slight excess of insertions (deletion/insertion ratio is 0.83), with 77% of the deletions and 94% of the insertions <10 bp. Although this survey was based on a small number of introns, it suggests that there may be mutational and/or selective differences between *D. melanogaster* and *D. pseudoobscura* that may contribute to the genome and intron size differences between these two species (Moriyama *et al.* 1998).

A bias toward deletions has been observed in studies of "dead-on-arrival" non-LTR retrotransposons in the *D. melanogaster* and *D. virilis* species groups (Petrov *et al.* 1996; Petrov and Hartl 1998) and in a survey of five different transposable elements in the complete *D. melanogaster* genome (Blumenstiel *et al.* 2002). These results suggest that there is a relatively high rate of spontaneous DNA loss within these species, with deletion/insertion ratios ranging from ∼4 to 8. The same qualitative pattern is also seen for the introns examined in this study (Table 2), although the deletion bias is not as extreme. This is likely due to the fact that introns are under constraints for proper splicing and that indel mutations that disrupt splicing and alter the protein sequence encoded by a gene will quickly be eliminated from the population by purifying selection (Ptak and Petrov 2002).

There is an overall bias toward deletions relative to insertions in introns (Table 2), but there is not a significant difference between deletion and insertion lengths. This suggests that, in general, introns should evolve toward shorter lengths. However, it is clear that introns maintain relatively constant lengths over evolutionary time (Table 1; Stephan *et al.* 1994; Moriyama *et al.* 1998). How can this be explained? A possible explanation based on compensatory evolution is as follows. Assuming that natural selection maintains a minimal length for short introns, as is indicated by the tight distribution of short intron lengths observed in many genomes (Mount *et al.* 1992; Deutsch and Long 1999; Yu *et al.* 2002), deletions that bring intron length below the minimum will be disfavored by natural selection. However, since the vast majority of deletions are of very short length (Figure 2), they may be only very slightly deleterious and can become fixed in a species through genetic drift. A general mutational bias toward small deletions and their successive fixation by drift may result in a "ratchet" effect in which intron length decreases by small steps. Because the length change is small at each step, the effect on relative fitness may be negligible. Eventually, a rare, large insertion may occur. Since this insertion is longer than the previous deletions that have gone to fixation in the species, it may have a larger effect on fitness, and if it restores the minimal intron length, it will be driven to fixation by positive selection.

The above model is supported by the observation that internal branch indels are significantly longer than terminal branch indels. The former are indels that occur on the branches separating the two major clades of the *D. melanogaster* species subgroup (Figure 1) and cannot be classified as either insertions or deletions due to the lack of an appropriate outgroup sequence. However, the observation that intron length is well conserved between the two clades (Table 1) and is generally well conserved between more distantly related species (Stephan *et al.* 1994; Moriyama *et al.* 1998) suggests that many of these indels represent insertions. Otherwise, the observed deletion bias would lead to a persistent decrease in intron length over time. Thus, the data are consistent with the relatively frequent occurrence and fixation of small deletions (within each of the two major clades) and with the less-frequent occurrence and fixation of larger insertions (between clades). Since the same pattern is observed in the two large introns, a similar process may also occur in introns of this size class. In this case, the fixation of large insertions may be selectively favored not to maintain a minimal intron length for efficient splicing, but to reduce interference between selected sites in adjacent exons (Comeron and Kreitman 2000). More orthologous sequences from long introns are needed to investigate this possibility.

The process described above should be continuous and not limited to only the internal branches of the phylogeny. However, it may be difficult to detect such an effect from the terminal branch indels, especially with a limited sample size of introns. This is because the ratchet model requires the successive fixation of multiple small deletions before a large insertion is favored by selection. The terminal branch species used in the current analysis typically differ by 5% or less in noncoding DNA sequence. Since indel rates are ∼15–20% of substitution rates (Table 2), only one indel is likely to occur along a particular terminal branch in a short intron. Thus there is little opportunity for the ratchet process to function over relatively short time scales. It should also be noted that the model does not require that all deletions be deleterious and all insertions beneficial. Selection for (or against) indels occurs only after intron length falls below a minimum required for efficient splicing. As can be seen from Figure 2, large deletions (>10 bp) do become fixed within the short intron class. However, it is noteworthy that the three large deletions detected within this sample occur within three of the larger introns of this size class (23 bp in *janA* intron 2, 11 bp in *janB* intron 1, and 11 bp in *rux*).

Indels were partitioned into three categories (insertion, deletion, and ambiguous) using parsimony and assuming the relationship shown in Figure 1. This tree is strongly supported by several methods of phylogenetic reconstruction used in this article (see RESULTS) and by other recent molecular analyses (Ko *et al.* 2003), but differs slightly from the relationship traditionally assumed for the *D. melanogaster* species subgroup (Ashburner 1989; Powell 1997). Assuming the traditional relationship, however, does not alter the major results reported here. For example, there is still a significant bias toward deletions (deletion/insertion ratio is 1.96) and no significant difference between deletion and insertion sizes (average lengths of 4.11 and 3.58 bp, respectively). Because the traditional tree allows the *D. erecta/D. orena* clade to be used as an outgroup to all other species, there are fewer ambiguous indels under this assumption. However, the ambiguous indels do not differ significantly in size from classified indels (average lengths of 3.83 and 3.93 bp, respectively). Thus, assuming the traditional phylogeny also predicts that intron length should consistently decrease over time, but does not suggest a process by which length can be restored and maintained relatively constantly over evolutionary timescales.

Comparison of indel rates in the paralogous introns of the *janA*, *janB*, and *ocn* genes indicates that the level of selective constraint on intron length may vary between introns within the same gene. Of the two paralogous introns shared among these three genes, the first shows significantly fewer length changes than the second when compared among species of the *D. melanogaster* species subgroup. Several observations indicate that this difference cannot be explained by different mutational processes in the two introns. First, the introns are only 125 bp apart within each gene and all three genes lie in tandem within a genomic region of 2.5 kb. It is extremely unlikely that mutation rates could vary so extensively over a very small portion of the genome. Second, the two paralogous introns show similar numbers of nucleotide substitutions among species (Table 2), suggesting equal mutation rates with respect to single base changes. Finally, a comparison of intraspecific polymorphism (which is expected to be less sensitive to weak selection than interspecific divergence) in these introns suggests equal mutation rates (Parsch *et al.* 2001a; C. Meiklejohn, personal communication). A survey of polymorphism in the *janA*, *janB*, and *ocn* genes in 36 alleles of *D. simulans* and in 8 alleles of *D. melanogaster* revealed a total of 26 single nucleotide polymorphisms in the first paralogous intron and 30 in the second. The number of indels observed within species was too low to be informative, with one indel in the first intron and two in the second.

Comparison of the lengths of the two introns among the three paralogous genes suggests that the difference in selective constraint most likely predates the divergence of the *D. melanogaster* species subgroup. Among the three genes, the first intron shows relatively little length variation, ranging from 50 bp (*ocn* in *D. orena*) to 58 bp (*janA* in all species). The second intron shows much greater length differences among the paralogs, ranging from 48 bp (*ocn* in *D. sechellia*) to 106 bp (*janA* in *D. simulans* and *D. mauritiana*). The conservation of intron length across the paralogs is surprising, given that the selective constraints on protein-encoding sequences appear to differ among the three genes. The *janA*, *janB*, and *ocn* genes show significant differences from each other in their nonsynonymous/synonymous substitution rates, indicating that they have likely undergone functional divergence following duplication (Parsch *et al.* 2001b).

The observation that two short introns within the same gene are under different length constraints is difficult to explain. Could it be that intron order plays a role? Perhaps the first intron of a gene is under stronger length constraints than are subsequent introns. This possibility is not supported by the limited data that are available. Aside from *janA*, *janB*, and *ocn*, only one of the other genes surveyed, *Adh*, contains multiple short introns (considering the two short introns of the adult transcriptional unit). In *Adh*, the first short intron shows 10 indels and 34 substitutions, while the second short intron shows 10 indels and 45 substitutions. The difference in the indel/substitution ratio is not significant ($\chi^2 = 0.31$; $P = 0.58$). Furthermore, the *janB* gene contains a 5′ intron that is not present in *janA* or *ocn* (Figure 3). This intron does not appear to be under stronger length constraints than the two subsequent *janB* introns. The length of the first *janB* intron ranges from 58 bp in *D. melanogaster* to 69 bp in *D. orena.* This intron shows an indel/substitution ratio of 0.17, which is comparable to that of the third intron (0.12 indels/substitution), but much greater than that of the second intron, which is invariant in length across the entire species subgroup. Additional interspecific comparisons of paralogous and other genes containing multiple introns are needed to determine if the pattern seen in the *janA*, *janB*, and *ocn* genes is common. If so, it would indicate that intron-length evolution cannot be accurately modeled as a general process in which all introns within a particular size or recombination class are under the same selective constraints, but rather that unique constraints applying to individual introns must also be taken into account. Further studies of substitution and indel rates in long introns are needed to elucidate differences in selective constraint between introns of the two size classes.

I thank Tina Hambuch, David de Lorenzo, Wolfgang Stephan, Colin Meiklejohn, and Justin Blumenstiel for constructive comments on the manuscript. This work was supported by funds from the University of Munich (LMU).

LITERATURE CITED

Ashburner, M., 1989 *Drosophila: A Laboratory Handbook.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

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

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

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

Castillo-Davis, C. I., S. L. Mekhedov, D. L. Hartl, E. V. Koonin and F. A. Kondrashov, 2002 Selection for short introns in highly expressed genes. Nat. Genet. **31:** 415–418.

Choi, T., M. Huang, C. Gorman and R. Jaenisch, 1991 A generic intron increases gene expression in transgenic mice. Mol. Cell. Biol. **11:** 3070–3074.

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.

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

Graur, D., and W.-H. Li, 2000 *Fundamentals of Molecular Evolution*, Ed 2. Sinauer Associates, Sunderland, MA.

Holstege, F. C., E. G. Jennings, J. J. Wyrick, T. I. Lee, C. J. Hentgartner *et al.*, 1998 Dissecting the regulatory circuitry of a eukaryotic genome. Cell **95:** 717–728.

Huelsenbeck, J. P., and F. Ronquist, 2001 MRBAYES: Bayesian inference of phylogeny. Bioinformatics **17:** 754–755.

Kliman, R. M., P. Andolfatto, J. A. Coyne, F. Depaulis, M. Kreitman *et al.*, 2000 The population genetics of the origin and divergence of the *Drosophila simulans* complex species. Genetics **156:** 1913–1931.

Ko, W.-Y., R. M. David and H. Akashi, 2003 Molecular phylogeny of the *Drosophila melanogaster* species subgroup. J. Mol. Evol. **57:** 562–573.

Lachaise, D., M. Harry, M. Solignac, F. Lemeunier, V. Benassi *et al.*, 2000 Evolutionary novelties in islands: *Drosophila santomea*, a new *melanogaster* sister species from Sao Tome. Proc. R. Soc. Lond. B Biol. Sci. **267:** 1487–1495.

Llopart, A., J. M. Comeron, F. G. Brunet, D. Lachaise and M. Long, 2002 Intron presence-absence polymorphism in *Drosophila* driven by positive Darwinian selection. Proc. Natl. Acad. Sci. USA **99:** 8121–8126.

Long, M., and M. Deutsch, 1999 Association of intron phase with conservation at splice site sequences and evolution of spliceosomal introns. Mol. Biol. Evol. **16:** 1528–1534.

Moriyama, 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.

Nixon, J. E., A. Wang, H. G. Morrison, A. G. McArthur, M. L. Sogin *et al.*, 2002 A spliceosomal intron in *Giardia lamblia*. Proc. Natl. Acad. Sci. USA **99:** 3701–3705.

Palmiter, R. D., E. P. Sandgren, M. R. Avarbock, D. D. Allen and R. L. Brinster, 1991 Heterologous introns can enhance expression of transgenes in mice. Proc. Natl. Acad. Sci. USA **88:** 478–482.

Parsch, J., C. D. Meiklejohn and D. L. Hartl, 2001a Patterns of DNA sequence variation suggest the recent action of positive selection in the *janus-ocnus* region of *Drosophila simulans*. Genetics **159:** 647–657.

Parsch, J., C. D. Meiklejohn, E. Hauschteck-Jungen, P. Hunziker and D. L. Hartl, 2001b Molecular evolution of the *ocnus* and *janus* genes in the *Drosophila melanogaster* species subgroup. Mol. Biol. Evol. **18:** 801–811.

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

Powell, J. R., 1997 *Progress and Prospects in Evolutionary Biology: The Drosophila Model.* Oxford University Press, New York.

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

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

Simpson, A. G., E. K. MacQuarrie and A. J. Roger, 2002 Eukaryotic evolution: early origin of canonical introns. Nature **419:** 270.

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.

Swofford, D. L., 2000 *PAUP**: Phylogenetic Analysis Using Parsimony (*and Other Methods)*, Version 4. Sinauer Associates, Sunderland, MA.

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.

Ting, C. T., S. C. Tsaur and C.-I Wu, 2000 The phylogeny of closely related species as revealed by the genealogy of a speciation gene, *Odysseus*. Proc. Natl. Acad. Sci. USA **97:** 5313–5316.

Yanicostas, C., P. Ferrer, A. Vincent and J.-A. Lepesant, 1995 Separate *cis*-regulatory sequences control expression of *serendipity* β and *janus A*, two immediately adjacent *Drosophila* genes. Mol. Gen. Genet. **246:** 549–560.

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

Communicating editor: S. Schaeffer