*Proc. Natl. Acad. Sci. USA*
Vol. 95, pp. 1658–1662, February 1998
Evolution

# New *Drosophila* introns originate by duplication

ROSA TARRÍO, FRANCISCO RODRÍGUEZ-TRELLES, AND FRANCISCO J. AYALA*

Department of Ecology and Evolutionary Biology, University of California, Irvine, CA 92697-2525

*Contributed by Francisco J. Ayala, December 17, 1997*

**ABSTRACT** **We have analyzed the phylogenetic distribution of introns in the gene coding for xanthine dehydrogenase in 37 species, including 31 dipterans sequenced by us. We have discovered three narrowly distributed novel introns, one in the medfly *Ceratitis capitata*, the second in the *willistoni* and *saltans* groups of *Drosophila*, and the third in two sibling species of the *willistoni* group. The phylogenetic distribution of these introns favors the "introns-late" theory of the origin of genes. Analysis of the nucleotide sequences indicates that all three introns have arisen by duplication of a preexisting intron, which is pervasive in *Drosophila* and other dipterans (and has a homologous position as an intron found in humans and other diverse organisms).**

The "exon theory of genes" (ref. 1; and the related "introns-early" theory, refs. 2–3) proposes that the first genes were made of short DNA sequences that coded for small polypeptide chains and correspond to modern exons. According to this theory, introns were in existence at the very beginning of evolution and were used to assemble the early genes, but many were lost during evolution (4). The "introns-late" theory proposes that introns arose late in evolution, perhaps around the origin of multicellularity, and may have been instrumental in creating, during the Cambrian explosion, a profusion of new genes by exon shuffling (5–9). The two theories make different predictions with regard to the evolutionary rates of intron insertion and deletion after diversification of the progenote. According to the exon theory, later divergent eukaryotes have lost many of their introns, and the ones that remain are a subset of the total number that existed in earliest times. In contrast, the introns-late theory claims that introns were inserted sometime after archeozoan divergence and have proliferated in various lineages. Recent evidence manifests that the patterns of intron distribution and insertion across lineages typically show a restrictive phylogenetic distribution indicative of a recent origin (9–13). However, the process of spliceosomal intron gain remains to be elucidated.

We have studied the phylogenetic distribution of introns in a gene coding for xanthine dehydrogenase (*Xdh*; EC 1.2.1.37) across a broad phylogenetic spectrum that includes 33 dipterans, 1 lepidopteran, 2 vertebrates, and 1 fungus. We have uncovered in the dipterans the presence of three narrowly distributed introns, which cannot parsimoniously be reconciled with the exon theory. The evidence indicates that the most likely explanation for the origin of these introns is a recent duplication from preexisting but not very proximal introns.

## MATERIALS AND METHODS

A 1,605-bp-long fragment of the *Xdh* gene (corresponding to positions 1,078 to 2,682 in the *Drosophila melanogaster* sequence; ref. 14) was amplified, cloned, and sequenced in 31 species of dipterans. The region comprises about half (975 bp) of exon 2, intron 2 (between positions 2,052 and 2,335), and most of exon 3 (348 bp).

Genomic DNA was prepared according to Kawasaki (15); the region was amplified by PCR with Ampli-tak DNA polymerase (Perkin–Elmer) by using high-fidelity conditions (16). PCR primers were designed by identifying conserved regions in homologous *Xdh* sequences available in GenBank for distantly related insects, vertebrates, and fungi. PCR primer sequences were sense, 5′-GCTCCTGGAGGCATGATHGCCTATCGTCG-3′; antisense, 5′-ATGAATTCCARTGTGAANAGKCCRTAGCCYTGCAT-3′ (H = A,C,T; K = G,T; N = G,A,C,T; R = A,G; Y = C,T). PCR conditions: template denaturation for 1 min at 95°C, primer annealing for 1 min at 56°C, and primer extension for 2 min at 72°C (30 sec added for each successive cycle), for a total of 32 cycles. PCR products were purified with the Wizard PCR Preps DNA purification kit (Promega), and the amplified *Xdh* region was ligated to the PCR II vector from the TA-cloning kit (Invitrogen) and cloned into *Escherichia coli* INV $\alpha$ F′ competent cells. Plasmid DNA was prepared for sequencing by using the QIA prep kit (Qiagen) and sequenced manually by the dideoxy termination method (17) with Sequenase 228 V3.0 DNA sequencing kit (Amersham). Both strands of the *Xdh* region were completely sequenced with 14 primers that included, in addition to the standard M13/pUC sequencing oligonucleotides, the two amplifying primers and the following sequencing primers: *XL1*, 5′-GATGACAWCCMCGMATGGA-3′; *XL2*, 5′-GTGATTGTGACCATHGAGSAGGC-3′; *XL3*, 5′-TCCACTCAGCATCCGTCNKAGGTRCA-3′; *XL4*, 5′-ACAGCCTGYGABATTGARTGCTA-3′; *XL5*, 5′-GGAATCGCATTCGGNGYNATGCA-3′; *XL6*, 5′-GATCCGATCTVAAYGGMATGGC-3′; *XR1*, 5′-GCCTTGTTGATCCAYTCYTKCCA-3′; *XR2*, 5′-CCAGCCTGGTTNARRTGCAT-3′; *XR3*, 5′-AAGGACARATCCATNGACCA-3′; *XR4*, 5′-TGCACCTCNGANGGATGCTG-3′; *XR5*, 5′-TAGGAGTTGTGCTYDATRGCCT-3′; *XR6*, 5′-CCAGATAGAGYTCDCCDTCCAT-3′ (B = C,G,T; D = A,G,T; M = A,C; R = A,G; S = C,G; V = A,C,G; W = A,T). The primer XL3 failed in *Drosophila sturtevanti*, for which it was replaced by *XL3s*, 5′-TACTTAAATAGGAACGGATG-3′.

Sequences were aligned with CLUSTAL W (18), with minor adjustments by eye. Intron positions were inferred by reference to the coding and the canonical spliceosomal intron splice sites.

## RESULTS

Fig. 1 displays the 443 aa corresponding to the *Xdh* region sequenced in this study. The seven species listed include three *Drosophila* (out of the 28 species we have analyzed), the medfly *Ceratitis capitata*, the moth *Caliphora vicina*, the mold *Aspergillus nidulans*, and *Homo sapiens*. The region includes 11 of the

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.



Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. AF039603–AF039649).

*To whom reprint requests should be addressed at: Department of Ecology and Evolutionary Biology, 321 Steinhaus Hall, University of California, Irvine, CA 92697-2525. e-mail: fjayala@uci.edu.

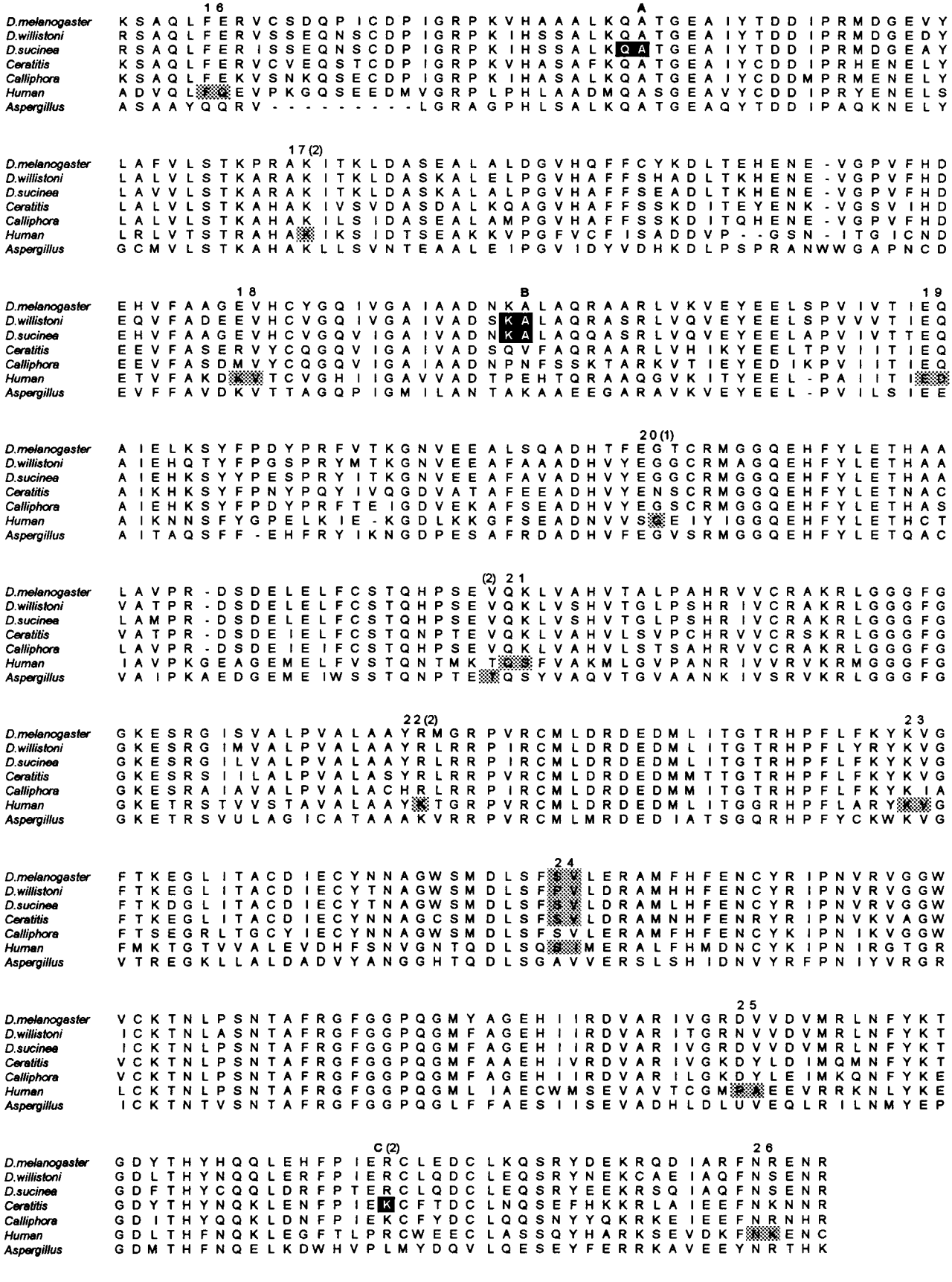


FIG. 1. Alignment of 443 residues of the amino acid sequence of xanthine dehydrogenase in seven species. Previously known intron positions are numbered (as in humans) above the sequences and denoted by shading: over two residues when the site falls between two codons, or over one residue when the codon is split, in which case the number in parentheses refers to the base in the triplet just before the intron. Black shading denotes novel inserted introns, labeled A, B, and C. Sources for published sequences are as follows: *Drosophila melanogaster* (14, 19), *Calliphora vicina* (20), *Homo sapiens* (21), and *Aspergillus nidulans* (22). Human intron 24 is at the same position as *Drosophila* intron 2.

35 introns reported for the human sequence (numbered 16–26 on top of the sequences), which are precisely the same also present in the mouse (not shown). The moth *Caliphora* does not have any intron in this region. The mold has only one intron, located 4 bp upstream from human intron 21. Human intron 24 is present in all other species (with the noted exception of the moth and the fungus) and is the only one present in this region of *D. melanogaster* (where it is known as intron 2). The figure shows the position of three novel introns (designated A, B, and C), previously unknown in *Drosophila*. Intron C is found only in the medfly. Intron B is present in *D. willistoni* and *D. sucinea*, as well as in other species of the *willistoni* species group (not shown). Intron A is present only in *D. sucinea* and its sibling *D. capricorni* (not shown).

The three new introns fall into positions that are clearly distinct from any previously known intron position in this region. The physical distances between these introns and any of those previously described is 48 bp (between intron B and human intron 18), 67 bp (between intron C and human intron 26), and 75 bp (between intron A and human intron 16). In the *Drosophila Xdh* coding region, intron A is located 276 bp (92 codons) upstream from intron B, which is located 606 bp (202 codons) upstream from intron 2 (human intron 24). Intron C is located 269 bp (89.7 codons) downstream from intron 2. The newly discovered *Drosophila* introns have sizes ranging from 52 to 66 bp, which are typical for *Drosophila* introns (except for the first one for each gene, which is typically longer). Introns A and B are in phase 0 and intron C is in phase 2. All three introns fit to the "GT-AG" consensus splicing signal.

The phylogeny depicted in Fig. 2 indicates the distribution and insertion of the three newly discovered introns. Each of the three introns exhibits a clustered distribution. Intron A is present in the two closely related species *D. capricorni* and *D. succinea*, but not in the other species of the *willistoni* group even though these are all close relatives, nor in any other

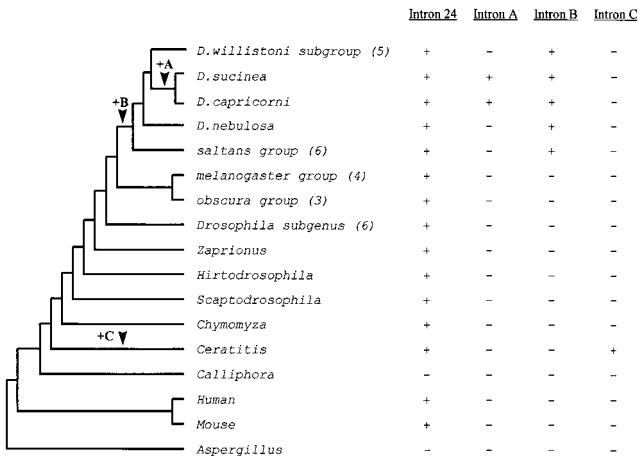


FIG. 2. Phylogenetic distribution of intron 24 and of the three novel introns, A, B, and C, detected in the *Xdh* studied region. Presence of an intron is represented by a plus sign, and absence is represented by a minus. The putative phylogenetic insertion of the novel introns is indicated by arrows. The consensus cladogram is based on data from refs. 22–24. The number of species included within taxonomic categories is shown in parentheses. The species are: for the *willistoni* subgroup, *D. equinoxialis*, *D. paulistorum*, *D. willistoni*, *D. tropicalis*, and *D. insularis*; for the *saltans* group, *D. saltans*, *D. prosaltans*, *D. emarginata*, *D. neocordata*, *D. sturtevanti*, and *D. subsaltans*; for the *melanogaster* group, *D. melanogaster*, *D. simulans*, *D. teissieri*, and *D. erecta*; for the *obscura* group, *D. subobscura, D. pseudoobscura*, and *D. persimilis*; for the *Drosophila* subgenus, *D. gymnobasis, D. robusta*, *D. virilis*, *D. hydei*, *D. funebris*, and *D. busckii*; *Zaprionus tuberculatus*; for *Hirtodrosophila, D. pictiventris*; *Scaptodrosophila lebanonensis*; *Chymomyza amoena*; and *Ceratitis capitata*. The nucleotide sequences of the introns in the *Drosophilid* species have GenBank accession numbers AF039603–AF039649.

*Drosophila* species. Intron B is found in all the representatives of the *saltans* (six sequenced species) and *willistoni* (five sequenced species) groups of *Drosophila*, but is not present in other species of the Sophophora subgenus, such as the *melanogaster* and *obscura* groups, nor in any other *Drosophila* species. Intron C is exclusively present in the mediterranean fruit fly *C. capitata* (but we have not sequenced any other species of the family Tephritidae, to which *Ceratitis* belongs).

Fig. 3 (*Upper*) shows an alignment of introns A, B, and 2 in *D. sucinea* and *D. capricorni*. Intron A is identical in both species, and there is only one nucleotide difference between the species in intron B and one in intron 2 (indicated by arrows). These similarities reflect the very recent divergence of these two species.

Unexpectedly, for seemingly nonhomologous noncoding sequences, introns A, B, and 2 exhibit striking sequence similarity. Ignoring the 5′ and 3′ consensus splicing sites, there are several identical runs of consecutive nucleotides in regions 1 and 2 (highlighted in Fig. 3). Introns A and B show two strings of 6 and 4 consecutive, identical nucleotides in region 1, and a string of 6 identical nucleotides in region 2. Introns B and 2 show two strings of 4 consecutive identical nucleotides in region 1 (ATCT and TTAA) and two strings of 6 and 5 in region 2 (separated by a single-nucleotide putative indel). Similarity between introns 2 and A is much less apparent; their alignment would be difficult if considered separately from intron B.

To ascertain the probability of observing this much similarity by chance only, we have carried out a simple, random permutation test as follows. For each pairwise comparison, the first 6 and the last 2 nt of each sequence, corresponding to the intron consensus splicing sites (5′-GTRAGT and AG-3′, respectively) were excluded. The sequence of intron B, which is the largest, is placed as the reference against which two sets of 100,000 randomly permuted sequences of equal length are compared. As nucleotide frequencies for each set we use those from the second intron involved in the comparison (A or 2, depending on the set). In this respect, the test is conservative because it disregards observed differences that could easily be accounted by transitional events. In none of the two sets of 100,000 random permutations did we observe a single sequence carrying a combination of strings as large as, or larger than, the observed ones. This means that the estimated chance probability of the observed identity is $P \ll 0.001$ for either A to B, or A to 2, or B to 2 intron comparisons.

Fig. 3 (*Lower*) displays the alignment between introns C and 2 of *Ceratitis*. Ignoring the 5′ and 3′ consensus splicing sites, we can see four strings of 6, 4, 6, and 4 identical consecutive nucleotides. The probability that this much similarity might have come about by chance is also $P < 0.001$ (only one case out of 100,000 random permutations).

## DISCUSSION

The introns-early theory proposes the presence of many introns in the common ancestor of all organisms alive today. In the early primitive organisms, introns would have been critically important for assembling short, primordial minigenes (encoding 15–20 aa) into more complex genes through exon shuffling. In contrast, the introns-late theory assumes that present-day exon/intron structures arose through evolution by random insertion of introns into continuous genes after the emergence of the eukaryotic cell or the emergence of the mitochondria. According to the introns-late theory, spliceosomal introns were never present in the remote ancestors of groups of organisms that now lack them, such as the eubacteria, archaeobacteria, and archaezoan protists. One source of evidence for or against the two theories has been the pattern of intron distribution across taxa.

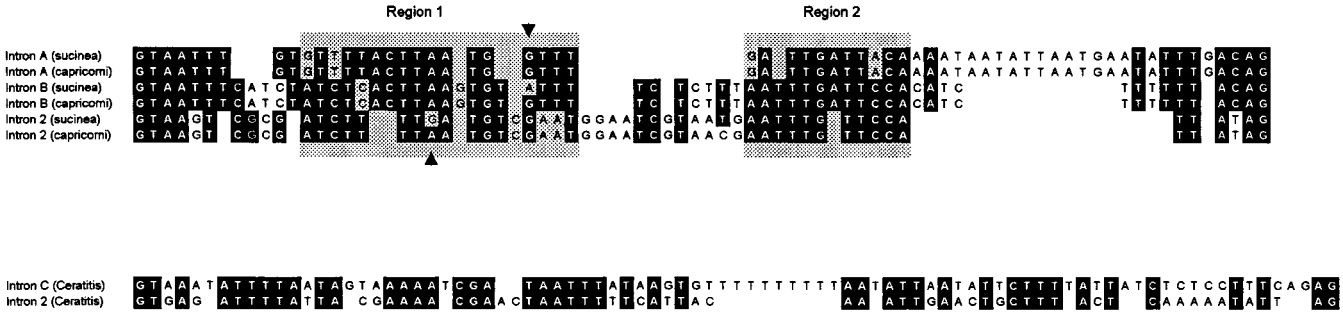


FIG. 3. (*Upper*) Alignment of the nucleotide sequences of introns A, B, and 2 for *D. sucinea* and *D. capricorni.* Black shading over sites denotes identical nucleotides shared by at least two nonhomologous introns. Regions 1 and 2, denoted by light-shading overlay, indicate two regions of extensive site identity between the *Drosophila* introns. The two arrowheads indicate nucleotide substitutions between *D. sucinea* and *D. capricorni*. (*Lower*) Alignment of the nucleotide sequences of introns C and 2 in *C. capitata*.

Parsimony arguments generally have favored the introns-late hypothesis. To accommodate the observed distribution of introns in the tree of life, proponents of the early-intron view have to postulate massive parallel intron losses from many different protistan lineages, including complete intron extinction from the several earliest lineages, as well as repeated, independent losses in numerous lineages of all animal and plant lineages extensively studied (7, 9–13).

The clustered phylogenetic distribution of the three novel introns reported in this study suggests that they are recently derived evolutionary features. The presence of intron A exclusively in *D. sucinea* and *D. capricorni* (Fig. 2) is parsimoniously explained by a single gain in the recent common ancestor of both species with no subsequent losses. If this intron would have been present in the ancestor of all species shown in Fig. 2, it would be necessary to invoke a minimum of 13 independent losses to account for the observed distribution. Similarly, the presence of intron B in all the representatives of the *willistoni* and *saltans* groups, but not in the other species, can be parsimoniously interpreted as the result of a single gain in their shared ancestral lineage after divergence from other Sophophora subgenus groups (*melanogaster* and *obscura*). Alternatively, 10 independent losses would be required if this intron would have been present in the ancestor of all species in Fig. 2. Parsimony also favors an insertional origin of intron C in the *Ceratitis* lineage. Otherwise, at least four independent losses must be invoked, rather than a single gain.

Intron sliding from one to another position within a gene is one possible way to account for the appearance of introns in new positions, while claiming that the introns are ancient. Intron sliding has been invoked to account for otherwise putative intron insertions in several genes (e.g., ref. 25; review in ref. 9). Evidence for putative intron sliding is scarce (review in ref. 26). In any case, this claim has been vigorously challenged, because the cases of intron slippage required by the early-introns theory seem to be too numerous to be consistent with the reported correlation between the boundaries of the ancient protein modules and the ends of ancient exons (27). The intron-sliding hypothesis has recently been evaluated on the basis of its implications with respect to the spatial and phylogenetic distribution of intron positions. Examination of 205 distinct intron positions from five sets of genes shows no sign of the excess closeness or clustering expected from sliding, leading to the conclusion that intron sliding has been a negligible phenomenon (9). In the present case, intron sliding could hardly account for the observed phylogenetic distribution of introns. The best candidate new intron position to be involved in sliding is 48 bp from the nearest intron (intron B with regard to human intron 18). This distance falls well outside the supposed 12 bp limit for intron displacement (28). The observation of three putative new intron insertions within the short nucleotide sequences sampled in this study, two of them within a set of closely related species, suggests that new spliceosomal introns can be gained, at least in some regions, more easily than it is often assumed.

A proposal of the introns-late theory is that spliceosomal introns are originally derived from group II introns that invaded the nuclear genome from organelles. Given that group II introns appear to be absent from animal mitochondria and that animals do not have chloroplasts, this hypothesis leaves unexplained how recent introns, such as the ones uncovered in this and other studies, may be acquired in animals (e.g., refs. 7 and 9–13). At least two hypothetical mechanisms can account for recent intron gains (reviewed in ref. 26): (*i*) the insertion of a transposable element that can be removed from the transcript utilizing splice sites within the transposon or in the flanking host sequences, and (*ii*) duplication of a preexisting intron. There is evidence that some transposable elements in maize can be spliced out like introns. However, transposable elements do not seem to be good candidates as sources for the putative intron gains in the *Xdh* region because intron insertion by this mechanism will delete or insert a few amino acids at the site of integration (29). The results obtained in our study suggest that intron duplication might be a powerful mechanism at generating new intron positions in the nuclear genomes, evidence of which, to date, has been lacking.

We have shown that the sequence similarity between intron 2 and introns A, B, and C cannot be explained by chance. Two processes that could account for the similarities are functional constraints (natural selection) and duplication. Apart from the 5′ and 3′ spliceosomal signals, the most obvious candidate region to be conserved in nonhomologous short introns is the branch point. This signal appears to be involved in the recognition of the 3′ splice site, being critical for the correct splicing of the intron. Based on the distribution of branch points in *Drosophila* introns, Mount *et al*. (30) have suggested that the minimum distances between branch points and the farthest 5′ and 3′ splice sites are 38 and 15 nt, respectively. This allows us to discard the possibility of the branch point to be located in region 1 of either intron A or B (Fig. 3), because it would be too close to the 5′ splice site. Similarly, region 2 of intron A is fully encompassed by the first 37 nt downstream from the 5′ splice site, and most of region 2 in intron 2 (9 out of 11 nt) is included within the last 15 nt from 3′. Hence, it seems unlikely that branch points can account for the similar sequences in these introns. This conclusion is supported by analysis of the distribution of nucleotide bases within the conserved region 2 (Fig. 3) in relation to the *Drosophila* branch-point consensus "CTAAT." The branched nucleotide (underlined) is predominantly A. Introns 2, A, and B share a single A within region 2, but this is preceded by a G, which is very rarely observed at this position in *Drosophila* introns (30). Thus, even if the branch point were located in region 2, it would be unlikely that it would be located in the same central position in all the three

introns and thereby account for this region's sequence similarity among the three introns.

The hypothesis that the observed similarity among the introns is due to causes other than natural selection is reinforced by examination of the intron sequences in related species. We were unsuccessful in aligning introns B and 2 (intron A is present only in *D. capricorni* and *D. sucinea*) across all the species of the *willistoni* and *saltans* groups. Inability to obtain a reliable alignment indicates that the sequence similarities across regions 1 and 2 are not conserved outside the lineages of *D. sucinea* and *D. capricorni*. Also, a BLAST (31) computer search over the sequences currently stored in GenBank and other molecular databases did not yield any similar motifs in other *Drosophila* regions. We conclude that natural selection by itself cannot account for the sequence similarity among the introns. Other factors, such as gene conversion, also seem unlikely because the coding regions surrounding the introns are quite dissimilar from the intron sequences. If the intron sequences proceed from unrelated origins, they would likely be too different to undergo gene conversion. We propose that intron duplication is the most likely explanation for the new introns A and B in the *Xdh* region of *Drosophila* (and by extension of the argument, we suggest that intron C of *Ceratitis* has also been acquired by duplication, although the evidence is more limited).

Intron duplication could have occurred via reverse splicing of some preexisting nuclear intron in the new intron positions of a pre-mRNA, followed by reverse transcription and homologous recombination (32). Reverse splicing has recently been proposed for the addition of spliceosomal introns to the U6 small nuclear RNA genes in yeast (33–35). The phylogenetic distribution of intron A indicates that this intron was gained more recently than intron B, and that these two introns are much more recent than intron 2, which is present across a broad range of taxa from mammals to *Drosophila*. The sequence of events leading to the duplication of introns A and B could be as follows: at some time point during the evolution of the common ancestor of the *willistoni* and *saltans* groups, and after it had diverged from the ancestor of the *melanogaster* and *obscura* lineages (some 30–50 million years ago; see ref. 24), intron 2 spliced out of the mRNA and inserted into the location currently occupied by intron B according to the mechanism described above. Subsequently, during the evolution of the *willistoni* group, a similar event would have taken place in the common ancestor of *D. sucinea* and *D. capricorni*, after it had separated from the main *willistoni* lineage, originating intron A. This intron could originate either from intron B or intron 2, because it is about equally different from both (7 indels plus 11.5 substitutions or 10 indels plus 8.4 substitutions, respectively). Because *D. capricorni* and *D. sucinea* are very similar at the molecular level (which contrasts with their substantial degree of morphological differentiation), the insertion of intron A must have been a recent event.

In conclusion, we have provided evidence of (*i*) recent acquisition of new introns in Diptera, two (A and B) in *Drosophila* and one (C) in *Ceratitis*; and (*ii*) duplication of a preexisting intron present in the same gene but at a remote location. Two intriguing questions arise. Why is it that intron 24 but not other introns within this *Xdh* region has persisted from a remote ancestor into mammals and flies? What is distinctive of intron 24 that has made it a source of duplication? This second question may, of course, have a trivial answer; namely, that it is the only intron that has persisted in the Diptera and thus the only one available for repeated duplication in *Drosophila* and in *Ceratitis*. Similarly, the first question has a possible trivial answer; namely, that the persistence of intron 24 is a matter of chance, through the one billion years (half a billion in each lineage) of separate evolution of mammals and flies.

We are indebted to Walter Fitch, Alberto Garcia-Saez, Richard R. Hudson, Carlos Machado, Harry Mangalam, and Andrey Tatarenkov for helpful discussions. This work was supported by National Institutes of Health Grant GM42397 to F.J.A. and COR-0077/94 from the Comunidad Autónoma de Madrid to Antonio Gómez Sal.

1. Gilbert, W., Marchionni, M. & McKnight, G. (1986) *Cell* **46,** 151–154.
2. Darnell, J. E. J. (1978) *Science* **202,** 1257–1260.
3. Doolittle, W. F. (1978) *Nature (London)* **272,** 581–582.
4. Gilbert, W. (1987) *Cold Spring Harbor Symp. Quant. Biol.* **52,** 901–905.
5. Cavalier-Smith, T. (1978) *J. Cell Sci.* **34,** 247–278.
6. Cavalier-Smith, T. (1985) *Nature (London)* **315,** 283–284.
7. Palmer, J. D. & Logsdon, J. M., Jr. (1991) *Curr. Opin. Genet. Dev.* **1,** 470–477.
8. Stoltzfus, A., Spencer, D. F., Zuker, M., Logsdon, J. M., Jr., & Doolittle, W. F. (1994) *Science* **265,** 202–207.
9. Stoltzfus, A., Logsdon, J. M., Jr., Palmer, J. D. & Doolittle, W. F. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 10739–10744.
10. Logsdon, J. M., Jr., & Palmer, J. D. (1994) *Nature (London)* **369,** 526.
11. Logsdon, J. M., Jr., Tyshenko, M. G., Dixon, C., D.-Jafari, J., Walker, V. K. & Palmer, J. D. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 8507–8511.
12. Kwiatowski, J., Skarecky, D. & Ayala, F. J. (1992) *Mol. Phylogenet. Evol.* **1,** 72–82.
13. Kwiatowski, J., Krawczyk, M., Kornacki, M., Bailey, K. & Ayala, F. J. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 8503–8506.
14. Keith, T. P., Riley, M. A., Kreitman, M., Lewontin, R. C., Curtis, D. & Chambers, G. (1987) *Genetics* **116,** 67–73.
15. Kawasaki, E. S. (1990) in *PCR Protocols: A Guide to Methods and Applications*, eds. Innis, M. A., Gelfand, D. H., Sninsky, J. J. & White, T. J. (Academic, San Diego), pp. 146–152.
16. Kwiatowski, J., Skarecky, D., Hernandez, S., Pham, D., Quijas, F. & Ayala, F. J. (1991) *Mol. Biol. Evol.* **8,** 884–887.
17. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5463–5467.
18. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994) *Nucleic Acids Res.* **22,** 4673–4680.
19. Lee, C. S., Curtis, D., McCarron, M., Love, C., Gray, M., Bender, W. & Chovnick, A. (1987) *Genetics* **116,** 55–66.
20. Houde, M., Tiveron, M. & Bregegere, F. (1989) *Gene* **85,** 391–402.
21. Ichida, K., Amaya, Y., Noda, K., Minoshima, S., Hosoya, T., Sakai, O., Shimizu, N. & Nishino, T. (1993) *Gene* **133,** 279–284.
22. Glatigny, A. & Scazzocchio, C. (1995) *J. Biol. Chem.* **270,** 3534–3550.
23. Grimaldi, D. A. (1990) *Bull. Am. Mus. Nat. Hist.* **197,** 1–139.
24. Powell, J. R. & DeSalle, R. (1995) *Evol. Biol.* **28,** 87–138.
25. Gilbert, W., de Souza, S. J. & Long, M. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 7698–7703.
26. Li, W-H. (1997) *Molecular Evolution* (Sinauer, Sunderland, MA).
27. Rzhetsky, A., Ayala, F. J., Hsu, L. C., Chang, C. & Yoshida, A. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 6820–6825.
28. Cerff, R. (1995) in *Tracing Biological Evolution in Protein and Gene Structures*, eds. Go, M. & Schimmel, P. (Elsevier, New York), pp. 205–227.
29. Patthy, L. (1995) *Protein Evolution by Exon-Shuffling* (R. G. Landes Company, Austin, TX).
30. Mount, M. S., Burks, C., Hertz, G., Stormo, G. D., White, O. & Fields, C. (1992) *Nucleic Acids Res.* **20,** 4255–4262.
31. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. T. (1990) *J. Mol. Biol.* **215,** 403–410.
32. Fink, G. R. (1987) *Cell* **49,** 5–6.
33. Takahashi, Y., Urushiyama, S., Tani, T. & Oshima, Y. (1993) *Mol. Cell. Biol* **13,** 5613–5619.
34. Tani, T. & Oshima, Y. (1989) *Nature (London)* **337,** 87–90.
35. Tani, T. & Oshima, Y. (1991) *Genes Dev.* **5,** 1022–1031.