Title
Evidence against the exon theory of genes derived from the triose-phosphate isomerase gene.

Journal
Proceedings of the National Academy of Sciences of the United States of America, 92(18)

ISSN
0027-8424

Authors
Kwiatowski, J
Krawczyk, M
Kornacki, M
et al.

Publication Date
1995-08-29

DOI
10.1073/pnas.92.18.8503

Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 8503–8506, August 1995
Evolution

Evidence against the exon theory of genes derived from the triose-phosphate isomerase gene

(intron evolution/origin of genes/exon shuffling/insect genes/phylogenetic inference)

JAN KWIATOWSKI*†, MICHAL KRAWCZYK*†, MACIEJ KORNACKI*, KEVIN BAILEY†, AND FRANCISCO J. AYALA†‡

*Institute of Botany, Warsaw University, Al. Ujazdowskie 4, 00-478 Warsaw, Poland; and †Department of Ecology and Evolutionary Biology, University of California, Irvine, CA 92717

Contributed by Francisco J. Ayala, June 1, 1995

ABSTRACT **The exon theory of genes proposes that the introns of protein-encoding nuclear genes are remnants of the DNA spacers between ancient minigenes. The discovery of an intron at a predicted position in the triose-phosphate isomerase (EC 5.3.1.1) gene of *Culex* mosquitoes has been hailed as an evidential pillar of the theory. We have found that that intron is also present in *Aedes* mosquitoes, which are closely related to *Culex*, but not in the phylogenetically more distant *Anopheles*, nor in the fly *Calliphora vicina*, nor in the moth *Spodoptera littoralis*. The presence of this intron in *Culex* and *Aedes* is parsimoniously explained as the result of an insertion in a recent common ancestor of these two species rather than as the remnant of an ancient intron. The absence of the intron in 19 species of very diverse organisms requires at least 10 independent evolutionary losses in order to be consistent with the exon theory.**

The exon theory of genes, also known as the "introns-early" theory, proposes that the exons of protein-coding genes are remnants of ancient minigenes and the introns derive from the spacers between them (1–3). There are two ways of testing the exon theory: functional and evolutionary. Functional tests investigate whether introns divide genes into segments that code for functional subunits of protein structure. Evolutionary tests ascertain whether the pattern of intron distribution across taxa conforms to the phylogenetic relationships among the taxa. Functional tests that yield results consistent with the theory may not be definitive, since constraints may exist at the protein or genome levels that restrict, or make more likely, the insertion of introns at certain sites.

A bulwark of the exon theory is the triose-phosphate isomerase (EC 5.3.1.1) gene (*Tpi*), a gene of ancient origin. The intron distribution of *Tpi* has been argued to be nonrandom (4–6) and, moreover, had elicited the prediction that an additional intron would be found at a site that would split the long exon III (5). The discovery of such an intron in the mosquito, *Culex tarsalis* (7), was, accordingly, hailed as strong support for the exon theory. We have investigated the presence of this intron in other dipterans and one lepidopteran and conclude that its taxonomic distribution is not consistent with the exon theory.§

MATERIALS AND METHODS

The four insects studied are, in increasing phylogenetic distance from *Culex*, the mosquitoes *Aedes* sp. and *Anopheles* sp., the fly *Calliphora vicina*, and the moth *Spodoptera littoralis*. The DNA of these species was isolated, amplified, and sequenced by described methods (8). The PCR products were cloned into the PCRII vector according to the manufacturer's (Invitrogen) protocol. Two primers, 5′-CGTKGGNGGNAACTGGAAGATGAAYGG-3′ (sense) and 5′-CGCTCYGAGTGBCCCAGGAYSACCCA-3′ (antisense) (K = G, T; S = G, C; Y = C, T; B = G, C, T; N = G, A, C, T), were derived from the conserved protein fragments VGGNWKMNG and WVILGHSER, respectively. The DNA sequences of other organisms were obtained from the GenBank data base maintained at IUBio archives at Indiana University.

The DNA sequences were aligned with the CLUSTAL V program (9). A maximum parsimony consensus tree (100 bootstrap resamplings) was obtained by consecutive execution of the SEQBOOT, PROTPARS, and CONSENSE programs of the PHYLIP 3.5 phylogenetic inference package (10). The branch lengths of the tree were based on accepted point mutation values (ref. 11; obtained with the PROTDIST program of PHYLIP) and were calculated with the FITCH program of PHYLIP.

RESULTS

Fig. 1 gives the alignment of 19 triose-phosphate isomerase protein sequences. The sequences for the mosquitoes *Aedes* and *Anopheles*, the fly *Calliphora*, and the moth *Spodoptera* are only partial, corresponding to DNA fragments that include the region where intron 5 is found in *Culex* mosquitoes. This intron is present in *Aedes*, a close relative of *Culex*, but not in the more distantly related *Anopheles* mosquitoes, nor in the flies *Drosophila* and *Calliphora*, nor in the moth *Spodoptera*. The absence of intron 5 from mosquitoes other than *Culex* and *Aedes* has been reported earlier (ref. 24; M. Tyshenko and V. K. Walker, personal communication).

Fig. 2 is a consensus tree of the 19 protein sequences, which corresponds well with phylogenies obtained by various methods, including ribosomal RNA gene sequences (25, 26). As noted earlier by Palmer and Logsdon (27), species in the early branches of the eukaryotic tree (*Giardia, Trypanosoma*, and *Leishmania*) have no *Tpi* introns. *Plasmodium* has one intron, the yeasts *Saccharomyces* and *Schizosaccharomyces* have none, but the mold *Aspergillus* has five. Introns are more abundant in animals and plants, with introns 3, 7, 8, 10, and 14 found in both, while introns 2, 13, and 15 are found in plants but not animals, and intron 12 is found only in animals. *Drosophila* has only one intron (at position 12). *Culex* also has only one intron, at position 5, shared with *Aedes* but not *Anopheles*, as noted.

DISCUSSION

The exon theory of genes proposes that modern protein-encoding genes were assembled from ancient minigenes, which correspond to modern exons, whereas the introns would be remnants of DNA spacers that separated the ancestral mini-

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.

Abbreviation: *Tpi*, triose-phosphate isomerase gene.

‡To whom reprint requests should be addressed.

§The *Tpi* fragment sequences of *Anopheles* sp., *Aedes* sp., *Calliphora vicina*, and *Spodoptera littoralis* have been deposited in the GenBank data base (accession nos. L38617, L42109, L38975, and L39011, respectively).

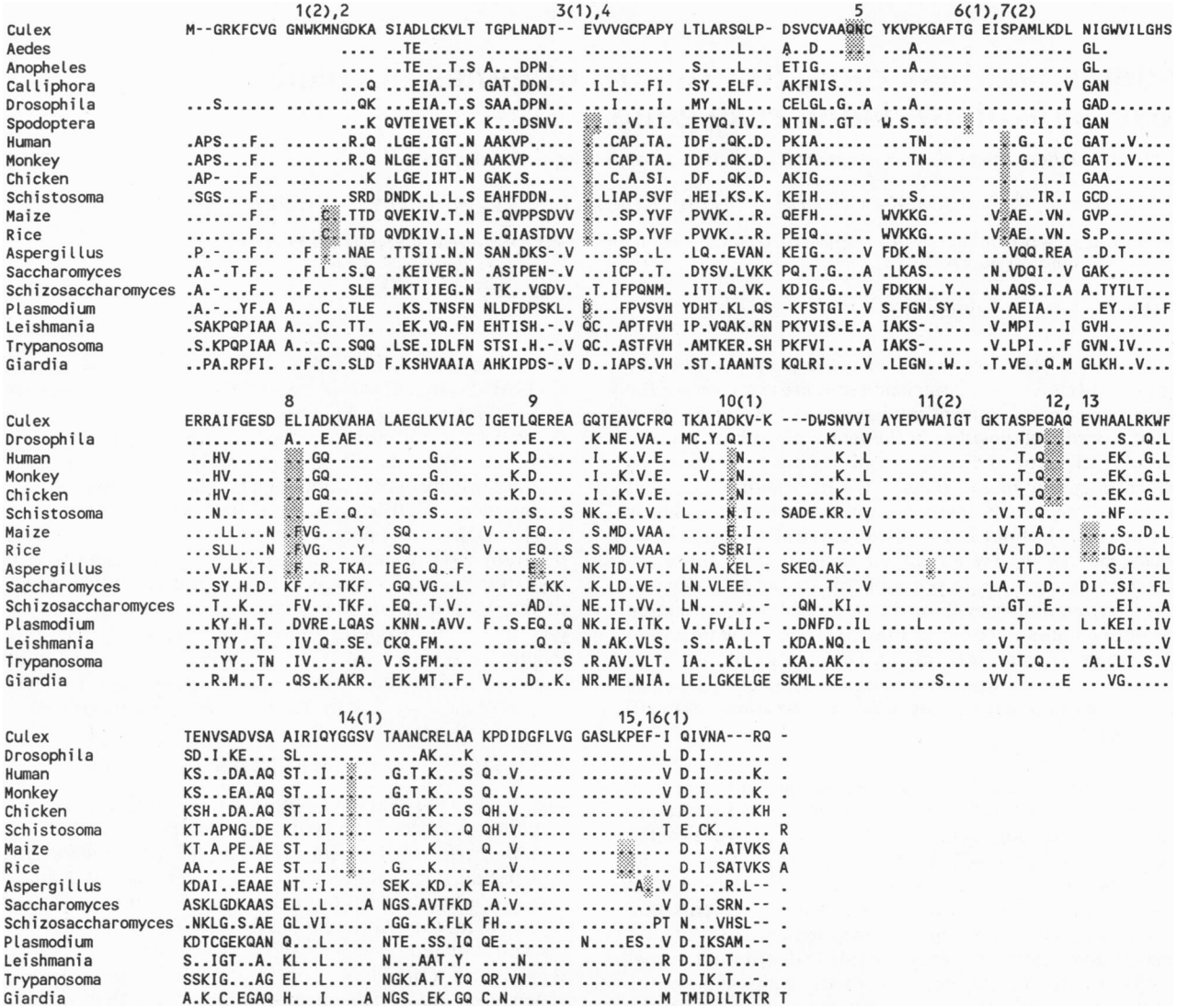

FIG. 1. Alignment of the amino acid sequence of triose-phosphate isomerase in 19 species. Positions identical to those of *Culex* are marked by dots. Intron positions are numbered and denoted by shading over two residues, when the site falls between two codons, or over one residue, when the codon is split, in which case the number in parentheses refers to the base in the triplet just before the intron. Proximal intron sites separated by 10 or fewer nucleotides are spaced by commas. Sources for published sequences are as follows: *Culex* (7), *Drosophila* (12), human (GenBank accession no. X69723), rhesus monkey (13), chicken (4), *Schistosoma* (14), rice (15), maize (16), *Aspergillus* (17), *Saccharomyces* (18), *Schizosaccharomyces* (19), *Plasmodium* (20), *Leishmania* (21), *Trypanosoma* (22), and *Giardia* (23).

genes. The absence of introns in eubacteria, archaebacteria, and several protist phyla would be due to their complete loss through evolutionary time in these groups of organisms. Many introns would also have been lost in genes from other protists and multicellular organisms. An alternative hypothesis (the "insertional" or "introns-late" theory) proposes that split genes have arisen from continuous genes by the insertion of introns. According to the introns-late theory, spliceosomal introns were never present in the ancestors of groups of organisms that now lack them; in other groups, introns have become inserted and occasionally deleted throughout their evolutionary history.

The claim that introns and exons derive from minigenes preexisting the divergence of bacteria and eukaryotes demands that the exon theory be tested in ancient genes, present in all sorts of organisms, such as *Tpi*—hence, the significance attached to the predicted discovery of *Tpi* intron 5 in *Culex* mosquitoes, splitting a DNA segment too long to correspond to only one ancient minigene.

Our results are not, however, consistent with the exon theory. The presence of intron 5 in *Culex* and *Aedes* but not in any of the other species represented in Fig. 2 is parsimoniously interpreted as the result of an evolutionary insertion at the position indicated by the arrow in Fig. 2. If this intron were present in the ancestor of all species shown in the figure, its absence from 17 of the 19 species would require a minimum of 10 independent evolutionary deletions of the intron.

More generally, the pattern of presence/absence of introns in the *Tpi* gene (Fig. 1) is more consistent with a dynamic process of occasional evolutionary insertions and deletions than with the exon theory.

The exon theory implies that introns at all 16 positions were present in the ancestral organisms, but each one was lost in numerous evolutionary events independently occurring in each of numerous lineages. The number of postulated events can be reduced by claiming that the pairs 1/2, 3/4, 12/13, and 15/16 represent each only one original intron that has slid in some lineages (17). The occurrence of intron slippage remains to be demonstrated and, in the case of *Tpi*, implies that it has occurred along as many as 7 or 9 nucleotides (for introns 15/16 and 12/13, respectively) that are as many as occur in some exons found in protein-coding genes (28).

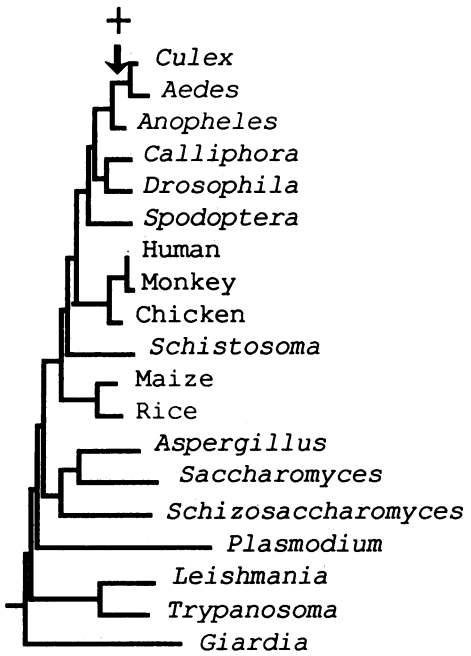

FIG. 2. Phylogenetic tree of 19 species derived from the *Tpi* protein sequences. The arrow indicates the postulated evolutionary insertion of intron 5. Branch lengths reflect genetic distances between sequences. The top six (insect) branches are proportionally based on only 77 amino acids. We first obtained a phylogeny that only included all complete sequences. The insect phylogeny, based on the 77 amino acids available for all six insect species, was separately obtained and then incorporated into the larger phylogeny.

The insertional theory accounts for introns as the result of a dynamic process of occasional insertions and deletions. A prediction made by this theory is that introns will be predominantly distributed in phylogenetic clusters, which allows for considerable elasticity and makes the theory all but untestable. Nevertheless, the number of independent evolutionary events required by this theory for *Tpi* is much smaller than the number required by the exon theory. In the case of intron 5, the insertional theory requires only one event (insertion in the *Culex/Aedes* ancestor) rather than at least 10 independent deletions. Other introns are shared by phylogenetically related organisms and can be explained by insertions in a common ancestor. Introns at sites 3, 7, 8, 10, and 14 are present in plants and animals and may have been inserted before the divergence of these two kingdoms. Intron 8 is present in the mold *Aspergillus* as well, which is consistent with its insertion before the divergence of multicellular organisms (although its absence from the two yeast species is also consistent with an independent insertion in the mold lineage). Introns 2, 13, and 15 are specific to plants, but only two species are represented in Fig. 1. Intron 12 may have been inserted early in the evolution of metazoans, since it is present in *Drosophila* as well as in the vertebrates. The mold *Aspergillus* has four introns in addition to the one at site 8, none of which is present in the two yeasts. Whether any or all of these represent early insertion in the fungi followed by loss in yeasts, or later insertion only in the mod lineage, is equally plausible on the basis of the data in Fig. 1.

Insects usually have fewer introns in protein-coding genes than vertebrates, which often share introns at specific sites, such as 3, 7, 8, 10, 12, and 14 for *Tpi*. This difference may be accounted for as a consequence of a distinctive high incidence of insertions early in the evolution of chordates. But it may also be, as often assumed, that many of these introns predate the divergence of insects and vertebrates but were mostly lost in insects (8, 28). The matter can only be settled after extensive data are collected for other animal phyla. In any case, *Tpi* conforms to the general pattern. Only one intron occurs in *Drosophila*, at site 12 where an intron is also present in vertebrates; and only one intron occurs in *Culex*, but at site 5, shared with *Aedes* but not with any other organisms. Intron insertions have rarely been detected in insects, and thus the *Tpi* gain of intron 5 in the *Aedes/Culex* lineage is particularly noteworthy. In dipterans, an intron insertion has been reported in a globin gene of the midge *Chironomus* (29), and we have uncovered in our laboratory an *Xdh* intron inserted in the *Drosophila willistoni* lineage, which is absent in other lineages of the subgenus *Sophophora* to which *D. willistoni* belongs, as well as in other *Drosophila* subgenera and related genera, such as *Chymomyza* and the medfly *Ceratitis capitata* (F. Rodriguez-Trelles and R. Tarrio, personal communication).

Logsdon *et al.* (30) have recently sequenced the *Tpi* gene in the insect *Heliothis virescens*, the nematode *Caenorhabditis elegans*, and the fungus *Coprinus cinereus* and have found introns at seven novel positions. They argue that their analysis showing the distribution of 21 intron sites in 20 different species is inconsistent not only with the exon theory of genes but also with the derivative notion that protein-coding genes are assembled by exon "shuffling." They argue that the large number of intron sites and their distribution conform to a random model of intron insertion. The exon-shuffling hypothesis maintains that introns are likely sites for intragenic recombination and thus contribute to the evolution of chimeric genes from preexisting sets of exons (31). This hypothesis is consistent with the exon theory of genes, but it does not specifically support it, since it is also consistent with the insertional theory. Exon shuffling may be a method of creating new genes, whether or not introns derive from the spacers between the minigenes of primeval organisms. Support for exon shuffling has been derived from statistical analysis of exon sequences (32) and from the tendency of intron positions to correspond to the recombinant junctions in some chimeric sequences, even though these may not be ancient (31, 33–35).

Our analysis showing 16 intron sites within a limited number of species, as well as the similar results of Logsdon *et al.* (30), does not particularly favor the existence at the genome or protein level of constraints that would allow for the insertion of introns at only a few specific sites within a gene. As Logsdon *et al.* (30) have argued, the finding of seven novel intron positions in the *Tpi* genes of just three, although diverse, organisms implies that more and more intron positions will be found as more *Tpi* genes are sequenced from remotely related eukaryotes. Stoltzfus *et al.* (31) have, moreover, concluded from the analysis of *Tpi* and three other ancient genes that there is no statistically significant correspondence between exons and units of protein structure, contrary to earlier claims (4–6).

We thank Elzbieta Wegner and Maciej Pszczolkowski for insect species; Joseph Felsenstein for making available the PHYLIP package; Jeffrey Palmer for making available the paper by J. M. Logsdon *et al.* (30) prior to publication; and Walter Fitch, Richard Hudson, Jeffrey Palmer, and Virginia Walker for comments about the manuscript. This work was supported by Grant 6P20303504 from the Committee for Scientific Research (Poland) to J.K. and GM42397 from the National Institutes of Health to F.J.A.

1. Darnell, J. E. J. (1978) *Science* **202,** 1257–1260.
2. Doolittle, W. F. (1978) *Nature (London)* **272,** 581–582.
3. Gilbert, W. (1987) *Cold Spring Harbor Symp. Quant. Biol.* **52,** 901–905.
4. Straus, D. & Gilbert, W. (1985) *Mol. Cell. Biol.* **5,** 3497–3506.
5. Gilbert, W., Marchionni, M. & McKnight, G. (1986) *Cell* **46,** 151–154.
6. Gilbert, W. & Glynias, M. (1993) *Gene* **135,** 137–144.
7. Tittiger, C., Whyard, S. & Walker, V. K. (1993) *Nature (London)* **361,** 470–472.

8. Kwiatowski, J., Skarecky, D. & Ayala, F. J. (1992) *Mol. Phylogenet. Evol.* **1,** 72–82.
9. Higgins, D. G., Bleasby, A. J. & Fuchs, R. (1992) *Comput. Appl. Biosci.* **8,** 189–191.
10. Felsenstein, J. (1989) *Cladistics,* **5,** 164–166.
11. Dayhoff, M. D. (1978) *Atlas of Protein Sequences and Structure* (Natl. Biomed. Res. Found., Washington, DC).
12. Shaw-Lee, R. L., Lissemore, J. L. & Sullivan, D. T. (1991) *Mol. Gen. Genet.* **230,** 225–229.
13. Old, S. E. & Mohrenweiser, H. W. (1988) *Nucleic Acids Res.* **16,** 9055.
14. Reis, M. G. I., Davis, R. E., Singh, H. & Shoemaker, C. B. (1993) *Mol. Biochem. Parasitol.* **59,** 235–242.
15. Xu, Y., Harris-Haller, L. W., McCollum, J. C., Hardin, S. H. & Hall, T. C. (1993) *Plant Physiol.* **102,** 697.
16. Marchionni, M. & Gilbert, W. (1986) *Cell* **46,** 133–141.
17. McKnight, G. L., O'Hara, P. J. & Parker, M. L. (1986) *Cell* **46,** 143–147.
18. Albert, T. & Kawasaki, G. (1982) *J. Mol. Appl. Genet.* **1,** 419–434.
19. Russell, P. R. (1985) *Gene* **40,** 125–130.
20. Ranie, J., Kumar, V. P. & Balaram, H. (1993) *Mol. Biochem. Parasitol.* **61,** 159–169.
21. Kohl, L., Callens, M., Wierenga, R. K., Oppeerdoes, F. R. & Michels, P. A. M. (1994) *Eur. J. Biochem.* **220,** 331–338.
22. Swinkels, B. W., Gibson, W. C., Osinga, K. A., Kramer, R., Veeneman, G. H., van Boom, J. H. & Borst, P. (1986) *EMBO J.* **5,** 1291–1298.
23. Mowatt, M. R., Weinbach, E. C., Howard, T. C. & Nash, T. E. (1994) *Exp. Parasitol.* **78,** 85–92.
24. Hurst, L. D. (1994) *Nature (London)* **371,** 381–382.
25. Smothers, J. F., von Dohlen, C. D., Smith, L. H. J. & Spall, R. D. (1994) *Science* **265,** 1719–1721.
26. Sogin, M. L. (1991) *Curr. Opin. Genet. Dev.* **1,** 457–463.
27. Palmer, J. D. & Logsdon, J. M. J. (1991) *Curr. Opin. Genet. Dev.* **1,** 470–477.
28. Hawkins, J. D. (1988) *Nucleic Acids Res.* **16,** 9893–9908.
29. Kao, W.-Y., Trewitt, P. M. & Bergtrom, G. (1994) *J. Mol. Evol.* **38,** 241–249.
30. 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.
31. Stoltzfus, A., Spencer, D. F., Zuker, M., Logsdon, J. M. J. & Doolittle, W. F. (1994) *Science* **265,** 202–207.
32. Dorit, R. L., Schoenbach, L. & Gilbert, W. (1990) *Science* **250,** 1377–1382.
33. Doolittle, R. F. (1985) *Trends Biochem. Sci.* **10,** 233–237.
34. Doolittle, R. F. (1991) *Science* **253,** 677–680.
35. Patthy, L. (1991) *BioEssays* **13,** 187–192.