J Mol Evol (2004) 59:448–457
DOI: 10.1007/s00239-004-2636-5

JOURNAL OF MOLECULAR EVOLUTION



# Hypervariable and Highly Divergent Intron–Exon Organizations in the Chordate *Oikopleura dioica*

**Rolf B. Edvardsen,[1] Emmanuelle Lerat,[2] Anne Dorthea Maeland,[1] Mette Flåt,[1] Rita Tewari,[1] Marit F. Jensen,[1] Hans Lehrach,[3] Richard Reinhardt,[3] Hee-Chan Seo,[1] Daniel Chourrout[1]**

[1] Sars Centre for Marine Molecular Biology, Bergen High Technology Centre, Thormoehlensgt. 55, 5020 Bergen, Norway
[2] Laboratoire de Biométrie et Biologie Evolutive, Université Lyon 1, 16 rue Dubois, 69622 Villeurbanne Cedex, France
[3] Max-Planck Institute for Molecular Genetics, Ihnestr. 73, 14195 Berlin, Germany

Received: 2 October 2003 / Accepted: 5 February 2004

**Abstract.** *Oikopleura dioica* is a pelagic tunicate with a very small genome and a very short life cycle. In order to investigate the intron–exon organizations in *Oikopleura*, we have isolated and characterized ribosomal protein EF-1α, Hox, and α-tubulin genes. Their intron positions have been compared with those of the same genes from various invertebrates and vertebrates, including four species with entirely sequenced genomes. *Oikopleura* genes, like *Caenorhabditis genes*, have introns at a large number of nonconserved positions, which must originate from late insertions or intron sliding of ancient insertions. Both species exhibit hypervariable intron–exon organization within their α-tubulin gene family. This is due to localization of most nonconserved intron positions in single members of this gene family. The hypervariability and divergence of intron positions in *Oikopleura* and *Caenorhabditis* may be related to the predominance of short introns, the processing of which is not very dependent upon the exonic environment compared to large introns. Also, both species have an undermethylated genome, and the control of methylation-induced point mutations imposes a control on exon size, at least in vertebrate genes. That introns placed at such variable positions in *Oikopleura* or *C. elegans* may serve a specific purpose is not easy to infer from our current knowledge and hypotheses on intron functions. We propose that new introns are retained in species with very short life cycles, because illegitimate exchanges including gene conversion are repressed. We also speculate that introns placed at gene-specific positions may contribute to suppressing these exchanges and thereby favor their own persistence.

**Key words:** Intron — Conversion — α-Tubulin — Hox — Ribosomal protein — Urochordate — *Oikopleura*

## Introduction

The debate on intron evolution has strongly focused on whether introns first appeared in prokaryotes or in eukaryotes (Gilbert et al. 1986, 1997; Cavalier-Smith 1991; Logsdon et al. 1998). The hypothesis of an eukaryotic origin is increasingly favored because introns have not been found in sequenced bacterial genomes but have been discovered in a number of basal eukaryotic taxons (Archibald et al. 2002; Nixon et al. 2002). The phylogeny of intron sequences is not easily soluble due to their rapid evolution. The positions of introns are, in contrast, fairly stable, and genes of distantly related phyla indeed often have introns at conserved positions. Conversely, introns occupying nonconserved positions probably have been gained relatively late (Logsdon et al. 1998; Robertson 2000),

*Correspondence to:* Daniel Chourrout; *email:* Daniel.Chourrout@sars.uib.no

or they may have moved from ancient intron positions through so-called intron sliding (Rogozin et al. 2000). Comparisons of genes between related taxons also reveal intron losses (Robertson 1998, 2000). Several mechanisms have been proposed to explain intron gain, sliding, and loss. These include precise "transposition-like" insertions or excisions, conversion between genes having different intron contents, and conversion between genes and related transcripts (Lynch and Richardson 2002).

The turnover of introns is generally slow, and intron–exon organizations can be almost invariant in large taxons, such as all vertebrates (Venkatesh et al. 1999). When modelized as a stochastic population-genetic process, it is found to be dependent upon effective population size and upon weak selective pressures (Lynch 2002). Introns that have efficiently colonized genes can be recruited into essential processes, with alternative splicing as the most speaking example (Hanke et al. 1999). A phenomenon receiving increasing attention is the control of abnormal transcripts through nonsense mediated decay (NMD) (Hentze and Kulozik 1999; Lykke-Andersen 2001). Ancient introns, especially those separating distinct protein-coding domains, may have contributed to the assembly of novel genes through exon shuffling (de Souza et al. 2001).

Recent studies have revealed significant differences in intron–exon organization between vertebrates and invertebrate deuterostomes (Wada et al. 2002). Genome sequencing of invertebrate deuterostomes, initiated with the ascidian *Ciona intestinalis* (Urochordata [Dehal et al. 2002]), will reveal the extent of intron–exon remodelling that took place before and after the emergence of chordates. We have undertaken genome sequencing in *Oikopleura dioica*, which belongs to larvaceans, another class of urochordates. *Oikopleura* has a very short life cycle (4 days) and its genome is remarkably small and compact (70 megabases; Mb), with very short intergenic distances and a majority of introns shorter than 50 bp (Seo et al. 2001). In this initial work, we observed that only a minority of intron positions were conserved between four genes of *Oikopleura* and their probable human orthologs. In order to test whether these differences are part of critical reorganizations after the divergence of both lineages, we extended our comparisons between *Oikopleura* and other animals to a variety of genes and gene families, namely, ribosomal protein genes, the elongation factor EF-1α gene, Hox genes, and α-tubulin genes. These genes have been characterized in a large number of animal phyla and they are sufficiently conserved to allow unambiguous comparisons of intron positions.

Here we report a considerable divergence and variability of intron positions in *Oikopleura dioica*. The nematode *Caenorhabditis elegans* is the only other species displaying such derived intron–exon organization features. Our observations are discussed using existing and novel hypotheses, in relation to the strong genome compaction and the short life cycle of these two species.

## Materials and Methods

### Oikopleura *Culture*

Methods for permanent culture of the species have been reported elsewhere (Spada et al. 2001).

### *Isolation of α-Tubulin Genes*

The first α-tubulin gene of *Oikopleura* (gene B) was revealed by random sequencing of a cDNA library prepared at the adult stage, and its genomic sequence was obtained through screening of a sperm genomic library. Cloning of all other α-tubulin genes was started by aligning this sequence with a data set of genomic sequences produced by whole-genome shotgun sequencing (Seo et al. 2001). This data set consisted of 44,797 contigs representing a total of 4098 Mb of nonredundant sequences. Alignments of 800 nonredundant expressed sequence tags (EST) with this shotgun data set showed an average coverage of 65%, with 83% of ESTs matched on more than one-quarter of their length and 48% covered on at least three-quarters of it. Each BLAST hit sequence obtained with the tubulin sequence was retrieved and allowed primer design for PCR amplification from genomic DNA and/or vector-anchored PCR on a plasmid cDNA library. Almost full-length genes (amino acid 16 to amino acid 412) were obtained from other species (hagfish *Myxine glutinosa*, amphioxus *Branchiostoma floridae*, sea squirt *Ciona intestinalis*, sea urchin *Strongylocentrotus droebechiensis*, lobster *Homarus gammarus*) through PCR amplification of genomic DNA using degenerate primers (AlphaDeg-1fw 5′-CAYGTBGGBCAGGCTGG WGTYC AGAT-3′, AlphaDeg-2rv 5′-GCCTCRGAGAAYTCNC CYCCT CCAT-3′, where Y = C or T, R = G or A, B = G or T or C, W = A or T, N = any).

### *Genomic Localization of* Oikopleura *α-Tubulin Genes*

A BAC library containing 9216 clones (135-kb average insert size) was screened with probes generated from the sequences of genes B, D, F, H, and I. Gene-specific primers were designed for tentative PCR amplification of every known α-tubulin gene from positive BAC clones (see Table S1, supplemental information).

### *Retrieval of* Oikopleura *EF-1α, Hox, and Ribosomal Protein Gene Sequences*

The *Oikopleura* shotgun data set was also used to retrieve sequences showing strong similarity to vertebrate and invertebrate EF-1α, Hox, and ribosomal protein genes (see Table S2, supplemental information, for accession numbers).

### *Sequence Analysis*

The initial sequence alignments were performed using ClustalX version 1.81 (Jeanmougin et al. 1998) and visually inspected for misalignments. Alignments of *Oikopleura* genomic and cDNA sequences, when available, easily revealed the presence of introns interrupting the coding regions. The translated DNA sequences

**Table 1.** Intron positions summed up for 12 ribosomal protein genes in five species

| *H. sapiens* | *C. intestinalis* | *O. dioica* | *D. melanogaster* | *C. elegans* | |
|---|---|---|---|---|---|
| 56 positions<br>41 shared | 37 | 2 | 12 | 13 | *H. sapiens* |
| | 48 positions<br>38 shared | 1 | 9 | 13 | *C. intestinalis* |
| | | 30 positions<br>3 shared | 1 | 0 | *O. dioica* |
| | | | 17 positions<br>14 shared | 4 | *D. melanogaster* |
| | | | | 28 positions<br>15 shared | *C. elegans* |

*Note*. The 12 genes analyzed were 60S ribosomal proteins L5, L11, L19, L21, L23, L24, L29, L30, and P0 and 40S ribosomal protein genes S2, S3, and S6. The total number of positions of each species and the total number of positions shared with four other species are indicated on the diagonal. Each cell contains the number of positions shared between two species. For accession numbers see supplemental information, Table S2. The *Ciona intestinalis* ribosomal protein gene sequences were retrieved from http://genome.jgi-psf.org/ciona4/ciona4.home.html.

were first aligned using ClustalW 1.74 (Thompson et al. 1994). After visual corrections using the sequence editor Seaview (Galtier et al. 1996), the protein alignment was used as a reference in order to keep the codon information when aligning the corresponding DNA coding sequences. Trees were constructed using the neighbor-joining method implemented in the Phylo_win program (Galtier et al. 1996) on the nonsynonymous rate ($K_a$)-based distances, codon distances computed according to Li (1993), using the pairwise gap removal option and with 500 bootstrap replicates. The analysis of codon usage was performed according to Lerat et al. (2002), using the relative frequency of the 59 degenerated codons. The Factorial Correspondence Analysis (FCA), a multivariate analysis, using the ADE-4 software package (Thioulouse et al. 1997), calculates the position of each sequence in a multidimensional space according to codon usage, allowing the detection of differences between sequences and identification of the codons involved. Potential conversion events between distinct α-tubulin genes were searched using Sawyer's (1989) method and his program GENECONV (www.math.wustl.edu/~sawyer/geneconv). The basis of the method is the identification of silent sites (synonymous codon substitutions) at which two DNA sequences agree (but differ from other sequences) and the segmentation of the sequences according to contiguous stretches of these sites. Gene conversion increases the lengths of these stretches. The significance of these lengths is then estimated by comparison with values obtained from 10,000 artificial data sets constructed by randomly permuting the silent polymorphic sites.

## Results

### *Ribosomal Protein Genes*

Ribosomal protein genes are highly conserved across distantly related phyla. Twelve distinct ribosomal protein genes of *Oikopleura dioica* were retrieved from the shotgun sequence dataset, each matching one or several *Oikopleura* ESTs. Intron–exon organization of these genes was determined by gene–cDNA alignment and compared with organization of their orthologs in human, the sea squirt *Ciona intestinalis*, the fruitfly *Drosophila melanogaster*, and the nematode *Caenorhabditis elegans*. The total number of introns of these 12 genes varied strongly between species, with the highest number in human and the lowest number in the fruitfly (Table 1). Most intron positions of the sea squirt were also found in human genes, and most intron positions of the fruitfly were found in human and/or the sea squirt. This suggests that for these 12 genes of human, sea squirt, and fly, most introns have an ancient origin.

In contrast, about half of the *C. elegans* intron positions were not found in the respective gene orthologs of other species. As the fly and the *C. elegans* lineages started to diverge from the chordate lineage at the same time, either the fly genes have lost many introns which are still found in *C. elegans* or most *C. elegans* introns result from lineage-specific insertions. A more extreme picture was observed for *Oikopleura* genes, which displayed 27 specific intron positions, for a total number of 30. Intron–exon organizations of *Oikopleura* ribosomal protein genes differed far more from those of human and *Ciona* genes than *Drosophila* and even *C. elegans* genes did. This indicates a "crisis" of gene organization in the *Oikopleura* lineage, which probably involved numerous intron gains and/or sliding, as well as numerous intron losses. At this point we cannot rule out that some *Oikopleura*-specific introns are ancient and were lost in other animal lineages.

### *Elongation Factor EF-1α Gene*

Partial sequences of two distinct EF-1α genes were also collected in the *Oikopleura dioica* genome data set. As for the single EF-1α gene identified in another *Oikopleura* species (Wada et al. 2002), their intron–exon organizations markedly differed from those of other deuterostome EF-1α genes, including those from ascidians (Fig. 1). The *Oikopleura* genes had introns in at least eight positions, six were *Oikopleura*-specific and one was found in urochordates only. The

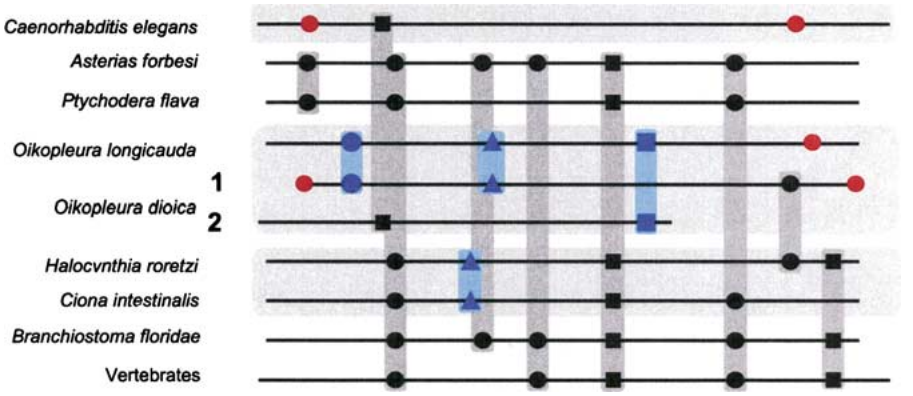


**Fig. 1.** Intron–exon organizations of animal elongation factor 1α genes. Intron positions in genes of each species are identified in the protein sequence (amino acids 1 to 462). Black circles (phase 0), squares (phase 1), and triangles (phase 2) represent intron positions shared by at least two large animal groups. Red symbols represent intron positions recorded in a single species and a single gene. Blue symbols represent intron positions recorded in several species of one class. See accession numbers for *O. dioica* and *C. elegans* in the supplemental information (Table S2), and Wada et al. (2002).

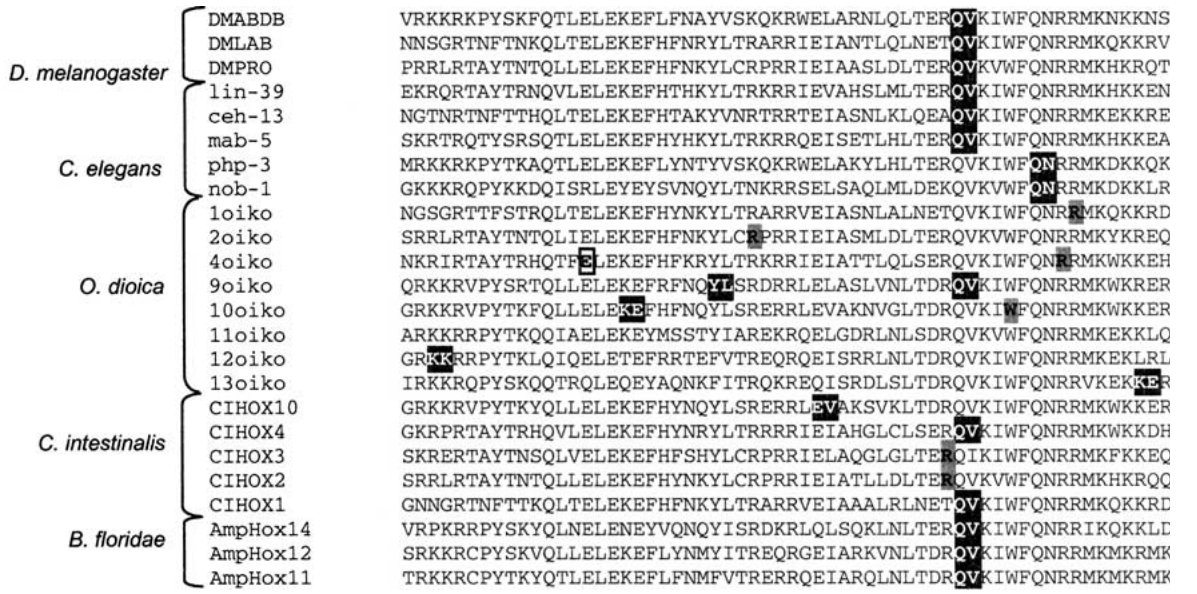


**Fig. 2.** Intron positions in the homeodomain of Hox proteins. Black boxes around two amino acids indicate an intron in phase 0, white boxes around single amino acids indicate an intron in phase 1, and black boxes around single amino acids indicate an intron in phase 2. The following Hox genes have no introns in the homeobox and are not included in the figure: *D. melanogaster* DMSCR, DMANTP, DMABDA, DMUBX, and DMDFD, *C. elegans* egl-5; *C.* intestinalis Hox5, Hox6, Hox12, and Hox13; *H. erythrogramma* Hox1, Hox6, Hox7, Hox9, and Hox10; and *B. floridae* Hox1, Hox2, Hox3, Hox4, Hox5, Hox6, Hox7, Hox8, Hox9, Hox10 and Hox13. None of the *H. sapiens* genes have introns in their homeobox. For accession numbers see supplemental information, Table S2.

intron distributions of the three *Oikopleura* genes were markedly different from each other, suggesting that gene organizations were remodeled until relatively late. Apart from *Oikopleura*, *C. elegans* was the only species displaying species-specific intron positions.

## *Hox Genes*

Hox proteins bind DNA through a helix–turn–helix (HTH) motif within the homeodomain, which is encoded by a conserved region of the gene, the homeobox. We cloned eight *Oikopleura* Hox genes and their cDNAs and localized their introns by gene–cDNA alignments (unpublished). Hox genes identified in several vertebrates and invertebrates, including human, the amphioxus *Branchiostoma floridae*, the sea squirt *Ciona intestinalis*, the echinoderm *Heliocidaris erythrogramma*, *Drosophila melanogaster*, and *Caenorhabditis elegans* were also retrieved from public databases. The homeodomains of all genes were aligned and intron positions in the homeobox compared.

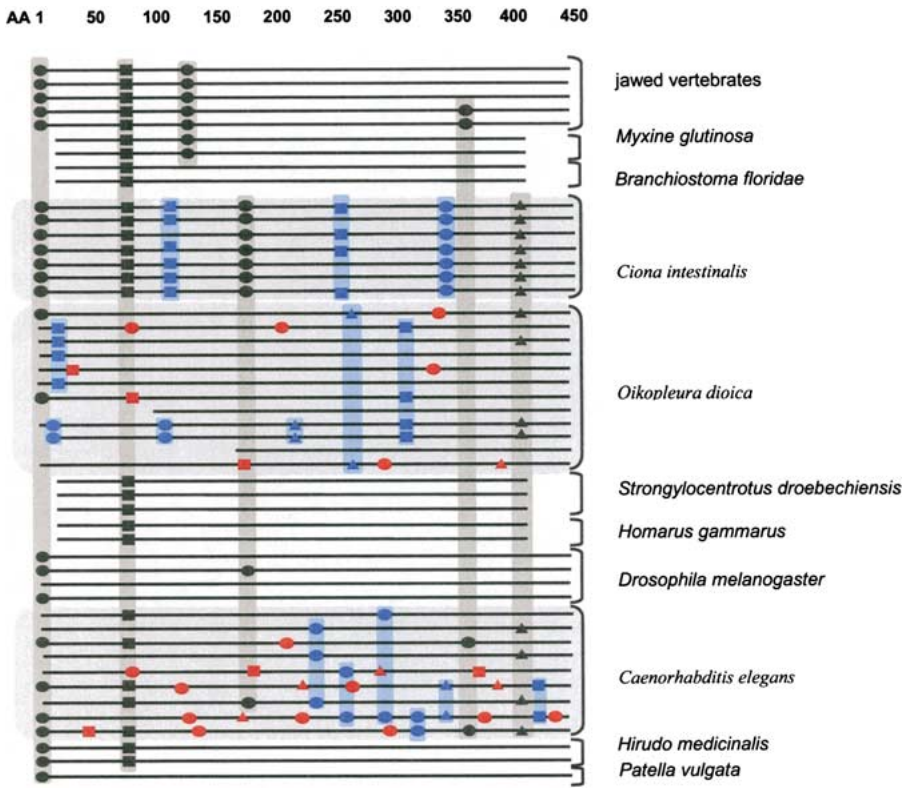


**Fig. 3.** Intron–exon organizations of animal α-tubulin genes. Intron positions in genes of each species are identified in the protein sequence by circles (phase 0 introns), squares (phase 1), and triangles (phase 2). Black symbols represent intron positions shared by at least two large animal groups. Red symbols represent intron positions recorded in a single species and a single gene. Blue symbols represent intron positions recorded in a single species but in several genes. Intron positions common to distinct genes are linked by vertical bars. Genes of hagfish, amphioxus, *Oikopleura dioica*, sea urchin, and lobster have been cloned and characterized in this work. For accession numbers see supplemental information, Table S2.

No intron was found in the homeobox of any human Hox gene. We also did not detect introns in the homeoboxes of five *Heliocidaris* Hox genes. Introns were identified in Hox genes from the five other species illustrated in Fig. 2. One intron position separated codons 44 and 45 of the homeobox in one or several of their Hox genes. Twelve other intron positions were species-specific. One of them was found in two genes of *Caenorhabditis elegans*, between codon 50 and codon 51. Two positions were found in *Ciona intestinalis*, in Hox10 (between codon 33 and codon 34), and in Hox2 and Hox3 (in codon 43). Strikingly, 9 of the 12 species-specific intron positions were found in *Oikopleura dioica*, with 2 in Hox10 and 2 in Hox4. The other five were found in five distinct Hox genes. Species-specific positions were well dispersed all over the homeobox sequence. Only one of the *Oikopleura* genes (Hox11) had no intron in the homeobox.

## α-*Tubulin Genes*

Twelve α-tubulin genes were cloned using specific sequences of the *Oikopleura* genome data set and were characterized in detail. Ten cDNAs matching these genes were also isolated for determination of intron positions and generation of RNA probes for *in situ* hybridizations during larval development. A variety of expression patterns, some fairly ubiquitous and others clearly tissue-specific, was observed (not shown). The 12 *Oikopleura* genes were compared with those of other invertebrates and vertebrates for their coding sequence and for their intron content. Most genes from those species were retrieved from public databases, but we cloned several of them from other invertebrate deuterostomes and from the lobster. In total we examined 45 genes from 10 invertebrate groups, as well as a large number of vertebrate α-tubulin genes including all mouse genes (Fig. 3).

Like in other gene families, distinct α-tubulin genes of vertebrates shared similar intron–exon organization: in each vertebrate species, all genes had three introns, at positions 1/2, 76, and 125/126, with some genes having an additional intron at position 352/353. Each of these four positions was also found in some invertebrate genes, suggesting an ancient origin. Two other positions (176/177, 407) were also shared between distantly related invertebrates. Altogether,

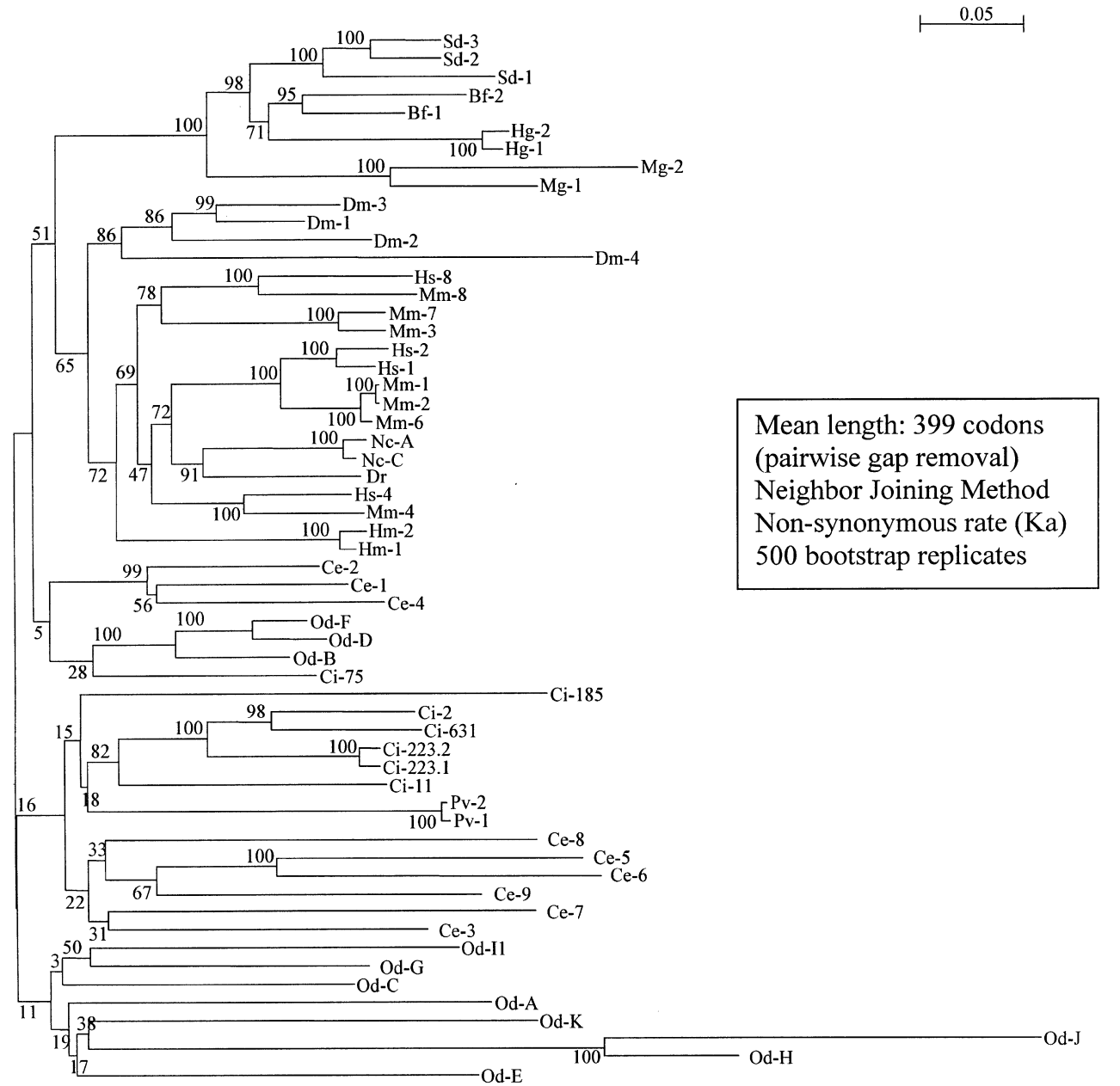


**Fig. 4.** Tree obtained from the alignment of the nucleic coding sequences of α-tubulin by the neighbor-joining method based on the $K_a$ distance. Numbers given along the branches are the bootstrap values after 500 replicates. Bf, *Branchiostoma floridae*; Ce, *Caenorhabditis elegans*; Ci, *Ciona intestinalis*; Dm, *Drosophila melanogaster*, Dr;- *Danio rerio*; Hg, *Homarus gammarus*; Hs, *Homo sapiens*; Mg, *Myxine glutinosa*; Mm, *Mus musculus*; Nc, *Notothenia coriiceps*; Pv, *Patella vulgata*; Od, *Oikopleura dioica* Sd, *Strongylocentrotus droebechiensis*. The *Ciona intestinalis* sequences are annotated using their genome scaffold number.

these six conserved positions accounted for the entire intron content of 8 of the 11 animal groups studied here.

In the three other species we identified no fewer than 41 species-specific positions (3 in *Ciona intestinalis*, 15 in *Oikopleura dioica*, and 23 in *Caenorhabditis elegans*). The three specific positions of *Ciona* were shared by four, six, and seven of its seven α-tubulin genes. In contrast, most specific positions of *Oikopleura* (9/15) and *Caenorhabditis* (17/23) were found in only 1 of their 12 and 9 genes, respectively. The majority of the remaining species-specific intron positions were found in only two genes.

Intron positions shared by several α-tubulin genes of a given species may have been multiplied along with gene duplication events. Alternatively, introns can be transferred from one gene to another, for example, by gene conversion (Mange and Prudhomme 1999). Phylogenetic analyses were carried out to reveal the evolutionary relationship between the α-tubulin coding sequences. Trees generated on the basis of amino acid sequences were poorly resolved

**Table 2.** Results of gene conversion analysis using α-tubulin coding sequences and the GENECONV program

| Species | No. genes | GS | Genes involved | SimP | Begin | End | Length |
|---|---|---|---|---|---|---|---|
| *Mus musculus* | 7 | 0 | Mm-3; Mm-8 | 0.0071 | 592 | 644 | 53 |
| | | | Mm-4; Mm-8 | 0.0071 | 670 | 701 | 32 |
| | | | Mm-7; Mm-8 | 0.0097 | 592 | 644 | 53 |
| | | | Mm-1; Mm-6 | 0.0125 | 865 | 1310 | 446 |
| | | | Mm-6; Mm-2 | 0.0403 | 1 | 392 | 392 |
| | | 2 | Mm-3; Mm-8 | 0.0066 | 592 | 644 | 53 |
| | | | Mm-4; Mm-8 | 0.0066 | 670 | 701 | 32 |
| | | | Mm-6; Mm-2 | 0.0088 | 1 | 861 | 861 |
| | | | Mm-7; Mm-8 | 0.0092 | 592 | 644 | 53 |
| | | | Mm-1; Mm-6 | 0.0294 | 865 | 1310 | 446 |
| | | 1 | Mm-3; Mm-8 | 0.0131 | 592 | 644 | 53 |
| | | | Mm-4; Mm-8 | 0.0131 | 670 | 701 | 32 |
| | | | Mm-3; Mm-6 | 0.0148 | 70 | 134 | 65 |
| | | | Mm-7; Mm-8 | 0.0170 | 592 | 644 | 53 |
| | | | Mm-3; Mm-1 | 0.0205 | 70 | 134 | 65 |
| | | | Mm-3; Mm-2 | 0.0351 | 70 | 134 | 65 |
| *Drosophila melanogaster* | 4 | 0 | Dm-1; Dm-3 | 0.0152 | 418 | 543 | 126 |
| | | 2 | Dm-1; Dm-3 | 0.0065 | 388 | 543 | 156 |
| | | 1 | Dm-1; Dm-3 | 0.0122 | 77 | 543 | 467 |
| *Strongylocentrotus droebechiensis* | 3 | 0 | Sp-2; Sp-3 | 0.0487 | 397 | 476 | 80 |
| | | 2 | Sp-2; Sp-3 | 0.0388 | 262 | 476 | 215 |
| | | 1 | Sp-2; Sp-3 | 0.0076 | 262 | 476 | 215 |
| | | | Sp-2; Sp-3 | 0.0076 | 598 | 785 | 188 |
| *Ciona intestinalis* | 7 | 0 | 223-2; 631 | 0.0000 | 823 | 996 | 174 |
| | | | 223-1; 631 | 0.0000 | 823 | 996 | 174 |
| | | | 223-2; 631 | 0.0000 | 46 | 176 | 131 |
| | | | 223-1; 631 | 0.0000 | 46 | 176 | 131 |
| | | | 223-2; 631 | 0.0015 | 1048 | 1136 | 89 |
| | | | 223-1; 631 | 0.0018 | 1048 | 1136 | 89 |
| | | | 223-1; 223-2 | 0.0258 | 724 | 1223 | 500 |
| | | 2 | 223-1; 631 | 0.0000 | 823 | 1304 | 482 |
| | | | 223-2; 631 | 0.0000 | 823 | 1223 | 401 |
| | | | 223-1; 631 | 0.0000 | 46 | 224 | 179 |
| | | | 223-2; 631 | 0.0000 | 46 | 176 | 131 |
| | | | 223-1; 2 | 0.0000 | 55 | 197 | 143 |
| | | | 223-2; 2 | 0.0000 | 55 | 188 | 134 |
| | | 1 | 223-1; 631 | 0.0000 | 773 | 1304 | 532 |
| | | | 223-2; 631 | 0.0000 | 773 | 1304 | 532 |
| | | | 223-1; 631 | 0.0000 | 46 | 224 | 179 |
| | | | 223-2; 631 | 0.0000 | 46 | 224 | 179 |
| | | | 223-1; 2 | 0.0000 | 55 | 206 | 152 |
| | | | 223-2; 2 | 0.0000 | 55 | 206 | 152 |
| | | | 223-1; 2 | 0.0002 | 949 | 1280 | 332 |
| | | | 223-2; 2 | 0.0007 | 949 | 1280 | 332 |
| *Oikopleura dioica* | 9 | 0 | Od-B; Od-C | 0.0245 | 151 | 197 | 47 |
| | | 2 | Od-B; Od-C | 0.0390 | 151 | 197 | 47 |
| | | 1 | Od-B; Od-C | 0.0079 | 25 | 296 | 272 |
| *Caenorhabditis elegans* | 9 | 0,1,2 | None detected | | | | |
| *Caenorhabditis briggsae* | 7 | 0,1,2 | None detected | | | | |

*Note*. Gene conversion events as detected by the GENECONV program (method of Sawyer 1989). *Oikopleura* genes H, J, and I2 were omitted due to partial 5′ sequences. No. genes, number of genes used for each species; GS, g scale value, where 0 allows no mismatch, 2 allows some mismatch, and 1 allows more mismatches in the conversion tracts; SimP, simulated $p$ values based on 10,000 permutations. The converted region was considered significant at $p < 0.05$. Begin, first nucleotide of converted region; End, last nucleotide in converted region; Length, length of converted region.

due to the high level of conservation of α-tubulins. Trees constructed using coding nucleotide sequences were far more robust and showed clustering of genes by species, with no recognition of clear orthologs (not shown). Codon usage differences between species were examined in detail and proved not to be sufficient to explain the observed clustering by species (supplementary material, Fig. S1). Trees based on nonsynonymous nucleotide substitutions also showed the clustering for most species (Fig. 4): rather robust clusters were found for all vertebrate, hagfish, amphioxus, sea urchin, lobster, fly, and *Patella* gene

complements. Another cluster contained five of the seven *Ciona* genes. One of the two other genes of *Ciona* was very weakly linked to this cluster. The way in which these species-specific clusters were distributed relative to each other in the tree is unlikely to reveal the true phylogeny, since the fly and the vertebrate clusters fell into the same major branch of the tree, and the lobster cluster was in another branch together with three deuterostome clusters. The distribution of *Oikopleura* and *Caenorhabditis* α-tubulin gene complements was somewhat different from those of the other species: each of them was divided into two clusters, with very low bootstrap values. This can be explained by the larger diversity of *Caenorhabditis* and *Oikopleura* genes, as shown by their long branches in the tree (see Table S3, supplemental information, for identity matrixes).

The fact that α-tubulin genes of most species cluster together in phylogenetic analyses can indicate that each gene complement has been generated through an independent multiplication after the speciation events. Alternatively, α-tubulin genes may have converged secondarily in each lineage through concerted evolution. To test the latter possibility, the coding sequences of six species were analyzed with the GENECONV program and putative conversion tracts were detected in all species except *C. elegans* (Table 2). Only one conversion tract was proposed for *Oikopleura*, between two of its nine full-length genes. The frequency of gene conversion is negatively correlated with intergenic distance (at least in the yeast; see Discussion). Therefore we started to physically map the *Oikopleura* α-tubulin genes by screening a BAC library with five gene-specific probes. Approximately 200 positive clones were found, indicating cross-hybridizations between distinct genes. PCR amplification on 41 of these positive clones with 11 pairs of gene-specific primers allowed characterization of their gene content. Ten genes were found at eight distinct genome locations (not shown). In one location, gene B colocalized with gene I2 (10-kb distance), whereas gene H colocalized with gene I1 in another locus (intergenic distance not known). BAC walking from both sides of a clone harboring genes B and I2 revealed no common positive clones with BAC walking from an H- and I1-containg clone. Interestingly, genes I1 and I2, which were unlinked, were almost identical, including within their intronic sequences, and were fairly divergent from genes H and B, to which they were linked, respectively.

## Discussion

This study shows that *Oikopleura* intron–exon organizations have diverged greatly from those of other animals, including other chordates and other urochordates, in particular. Most *Oikopleura* introns occupy nonconserved positions and probably originate from numerous late intron gains or sliding of ancient introns. *Oikopleura* genes also have few conserved intron positions, as if new introns have replaced the old ones. Intron positions also have evolved rapidly in the lineage of *Caenorhabditis elegans*. The genomewide organizations of *C. elegans* and *C. briggsae* genes have been compared to estimate the rate of intron turnover since the species diverged (Kent and Zahler 2000; Coghlan and Wolfe 2002), revealing that most intron positions are common to both species. We made the same observation with our sample of genes (not shown) and found that the few differences are easier to explain by intron loss/gain than with local intron sliding.

The literature offers several lines of interpretations for the strong divergence of intron positions in *O. dioica* and *C. elegans*. First, *O. dioica* and *C. elegans* have, compared to others species studied here, very compact genomes and a majority of very short introns ($< 50$ bp in *Oikopleura*). The splicing of short introns begins with pairing of intron ends (intron definition model), whereas the splicing of large introns is thought to involve pairing of splice sites across exons (exon definition model) (Berget 1995). Short introns from a variety of species seem to possess all information required for their recognition and splicing (Lim and Burge 2001), whereas the excision of larger introns depends upon additional signals from flanking exons (Blencowe 2000). Such an ''informational autonomy'' of short introns could give them more positional freedom within the coding sequence. Second, a constraint on intron positioning was proposed for vertebrate genes, as part of a mechanism to control illegitimate exchanges (Kricker et al. 1992). Interlocus recombination creates chromosomal instability, and gene conversion from pseudogenes to active genes can transfer undesirable mutations. Illegitimate exchanges are reduced by sequence divergence, which is accelerated by the high mutation rate at methylated CpG dinucleotides. These mutations were found to preferentially affect repeats, intronless pseudogenes, and large exons, but not exons smaller than 300 base pairs (bp). This constraint of exon size for mutational control should not prevail in most invertebrates, since their genomes are largely undermethylated (Tweedie et al. 1997). Consistent with this, we found equal frequencies of CG and GC dinucleotides in *Oikopleura* genes (not shown). In summary, both the small intron size and the undermethylation would relax some constraints on the exon size and, consequently, on the positioning of introns.

The fact that *Oikopleura* and *C. elegans* genes contain many more species-specific intron positions than other species could indicate that a large number

of new introns have invaded the genomes of their ancestors and displaced the old introns and/or that new introns installed at divergent positions have been advantaged because they serve some unusual purpose. This question is part of a more global and incompletely resolved issue of whether and how introns confer selective advantages that outweigh their extra cost during gene replication and transcription (Duret 2001; Lynch and Richardson 2002). The functions of introns that depend upon their position are naturally our main focus here. Introns are instrumental for generating a gene product diversity through alternative splicing (Hanke et al. 1999), but most, if not all, genes examined here are not known to be subjected to alternative splicing. Introns can harbor enhancers, alternative promoters, or even entire genes (Maxwell and Fournier 1995; Duret and Bucher 1997). How important the positioning of introns containing such elements is remains to be clarified in general. In the present case, we do not see why this function would require more variable intron positions in *Oikopleura* and *C. elegans*. In fact, we expect fewer of the introns in *Oikopleura* and *C. elegans* to harbor such elements than in other species, since they are generally very small. Introns also play essential roles for the elimination of aberrant transcripts through nonsense-mediated decay in providing spatial landmarks with respect to termination codons (Hentze and Kulozik 1999; Lykke-Andersen 2001). This function may have helped the proliferation of introns and, in addition, may have constrained the intron–exon organization in several fashions (Lynch and Richardson 2002). Among predictions based upon the function of introns in NMD, the exon size should be more uniform than in a model of random insertion, and the number of introns should increase with the gene size. Indeed, these predictions have been successfully tested in several genomes (Lynch and Kewalramani 2003). It will be interesting to learn whether or not NMD also occurs in *Oikopleura* and, when the genome sequence becomes available, how NMD may have influenced the intron–exon organizations. At present, we have no evidence for such an influence, since in all genes studied here both the exon sizes and the density of introns are actually more variable in *Oikopleura* than in other species. Finally, introns can affect the frequency of recombination and, via differential recombination rates, play on the selection of optimal combination of genes (Duret 2001). However, the influence of introns on recombination is essentially viewed through the variation of their length and not from the variation of their positions.

Though the literature provides possible explanations for a relaxation of intron positions in *Oikopleura* genes, there is as yet no clear indication on how divergent and variable intron–exon organizations may be beneficial. New introns have been able to replace most ancient introns but have remained in single members a gene family (e.g., α-tubulin genes). This could indicate that it is the variable configuration of gene organization rather than some particular positions that has been selected and/or that the transfer of introns between well-related genes has been severely limited. We provide indications for a concerted evolution of α-tubulin genes, which had less impact on *C. elegans* and *Oikopleura* gene complements than on those of other animals. Since gene conversion is a good candidate mechanism for the transfer of introns between genes (Mange and Prudhomme 1999), a suppression of conversion would explain both the greater diversity of α-tubulin genes in *Oikopleura* and in *C. elegans* and the heterogeneity of their intron–exon organizations. A genomewide study has indeed concluded that *C. elegans* has been little affected by gene conversion (Semple and Wolfe 1999). This might appear to be a paradox, since *C. elegans* has a very short generation time and, consequently, highly frequent meiotic cycles, unless specific mechanisms counteract gene conversion in short-lived species. Such mechanisms could also be operating in *Oikopleura*. How conversion can be negatively controlled is unclear. An examination of the *C. elegans* genome also shows that a minority of gene duplicates are physically linked (Semple and Wolfe 1999), as if they were quickly separated through fast genome rearrangements (Coghlan and Wolfe 2002). Since gene conversion rates, at least in yeast, correlate negatively with the physical distance between genes (Drouin 2002), a fast separation of related genes is an attractive candidate mechanism for conversion suppression. As an example, α-tubulin genes are often found in clusters of almost-identical genes (Table S4, supplemental information), but all *C. elegans* α-tubulin genes are dispersed. Those of *Oikopleura* are found in distinct genome locations as well, except two pairs of genes, each of which is composed of two fairly divergent genes. Finally, it is tempting to speculate that introns placed at variable positions in members of a gene family could themselves and directly hinder illegitimate exchanges. One may argue that variation of intron length, and not position, can produce the same effect. However, the flexibility of intron length may be compromised in *Oikopleura* and *Caenorhabditis* due to considerable pressure for genome compaction, so that heterologies must instead be obtained through a diversification of intron positions.

*Acknowledgments.* André Adoutte has contributed important advice throughout the course of this work. We also thank David Liberles for suggestions on the phylogenetic analysis. We thank the personnel of the UoB/Sars sequencing facility and the Sars Centre *Oikopleura* culture facility for their continued assistance.

## References


Archibald JM, O'Kelly CJ, Doolittle WF (2002) The chaperonin genes of jakobid and jakobid-like flagellates: Implications for eukaryotic evolution. Mol Biol Evol 19:422–431

Berget SM (1995) Exon recognition in vertebrate splicing. J Biol Chem 270:2411–2414

Blencowe BJ (2000) Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. Trends Biochem Sci 25:106–110

Cavalier-Smith T (1991) Intron phylogeny: A new hypothesis. Trends Genet 7:145–148

Coghlan A, Wolfe KH (2002) Fourfold faster rate of genome rearrangement in nematodes than in Drosophila. Genome Res 12:857–867

Dehal P, Satou Y, Campbell RK, Chapman J, Degnan B, De Tomaso A, Davidson B, Di Gregorio A, Gelpke M, Goodstein DM, Harafuji N, Hastings KE, Ho I, Hotta K, Huang W, Kawashima T, Lemaire P, Martinez D, Meinertzhagen IA, Necula S, Nonaka M, Putnam N, Rash S, Saiga H, Satake M, Terry A, Yamada L, Wang HG, Awazu S, Azumi K, Boore J, Branno M, Chin-Bow S, DeSantis R, Doyle S, Francino P, Keys DN, Haga S, Hayashi H, Hino K, Imai KS, Inaba K, Kano S, Kobayashi K, Kobayashi M, Lee BI, Makabe KW, Manohar C, Matassi G, Medina M, Mochizuki Y, Mount S, Morishita T, Miura S, Nakayama A, Nishizaka S, Nomoto H, Ohta F, Oishi K, Rigoutsos I, Sano M, Sasaki A, Sasakura Y, Shoguchi E, Shin-i T, Spagnuolo A, Stainier D, Suzuki MM, Tassy O, Takatori N, Tokuoka M, Yagi K, Yoshizaki F, Wada S, Zhang C, Hyatt PD, Larimer F, Detter C, Doggett N, Glavina T, Hawkins T, Richardson P, Lucas S, Kohara Y, Levine M, Satoh N, Rokhsar DS (2002) The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. Science 298:2157–2167

de Souza SJ, Long M, Schoenbach L, Roy SW, Gilbert W (1996) Intron positions correlate with module boundaries in ancient proteins. Proc Natl Acad Sci USA 93:14632–14636

Drouin G (2002) Characterization of the gene conversions between the multigene family members of the yeast genome. J Mol Evol 55:14–23

Duret L (2001) Why do genes have introns? Recombination might add a new piece to the puzzle. Trends Genet 17:172–175

Duret L, Bucher P (1997) Searching for regulatory elements in human noncoding sequences. Curr Opin Struct Biol 7:399–406

Galtier N, Gouy M, Gautier C (1996) SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput Appl Biosci 12:543–548

Gilbert W, de Souza SJ, Long M (1997) Origin of genes. Proc Natl Acad Sci USA 94:7698–7703

Gilbert W, Marchionni M, McKnight G (1986) On the antiquity of introns. Cell 46:151–153

Hanke J, Brett D, Zastrow I, Aydin A, Delbruck S, Lehmann G, Luft F, Reich J, Bork P (1999) Alternative splicing of human genes: more the rule than the exception? Trends Genet 15:389–390

Hentze MW, Kulozik AE (1999) A perfect message: RNA surveillance and nonsense-mediated decay. Cell 96:307–310

Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ (1998) Multiple sequence alignment with Clustal X. Trends Biochem Sci 23:403–405

Kent WJ, Zahler AM (2000) Conservation, regulation, synteny, and introns in a large-scale *C. briggsae–C. elegans* genomic alignment. Genome Res 10:1115–1125

Kricker MC, Drake JW, Radman M (1992) Duplication-targeted DNA methylation and mutagenesis in the evolution of eukaryotic chromosomes. Proc Natl Acad Sci USA 89:1075–1079

Lerat E, Capy P, Biemont C (2002) Codon usage by transposable elements and their host genes in five species. J Mol Evol 54:625–637

Li WH (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol 36:96–99

Lim LP, Burge CB (2001) A computational analysis of sequence features involved in recognition of short introns. Proc Natl Acad Sci USA 98:11193–11198

Logsdon JM Jr (1998) The recent origins of spliceosomal introns revisited. Curr Opin Genet Dev 8:637–648

Lykke-Andersen J (2001) mRNA quality control: Marking the message for life or death. Curr Biol 11:88–91

Lynch M (2002) Intron evolution as a population-genetic process. Proc Natl Acad Sci USA 99:6118–6123

Lynch M, Kewalramani A (2003) Messenger RNA surveillance and the evolutionary proliferation of introns. Mol Biol Evol 20:563–571

Lynch M, Richardson AO (2002) The evolution of spliceosomal introns. Curr Opin Genet Dev 12:701–710

Mange A, Prudhomme JC (1999) Comparison of *Bombyx mori* and *Helicoverpa armigera* cytoplasmic actin genes provides clues to the evolution of actin genes in insects. Mol Biol Evol 16:165–172

Maxwell ES, Fourier MJ (1995) The small nucleolar RNAs. Annu Rev Biochem 64:897–934

Nixon JE, Wang A, Morrison HG, McArthur AG, Sogin ML, Loftus BJ, Samuelson J (2002) A spliceosomal intron in *Giardia lamblia*. Proc Natl Acad Sci USA 99:3701–3705

Robertson HM (1998) Two large families of chemoreceptor genes in the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae* reveal extensive gene duplication, diversification, movement, and intron loss. Genome Res 8:449–463

Robertson HM (2000) The large srh family of chemoreceptor genes in *Caenorhabditis* nematodes reveals processes of genome evolution involving large duplications and deletions and intron gains and losses. Genome Res 10:192–203

Rogozin IB, Lyons-Weiler J, Koonin EV (2000) Intron sliding in conserved gene families. Trends Genet 16:430–432

Sawyer S (1989) Statistical tests for detecting gene conversion. Mol Biol Evol 6:526–538

Semple C, Wolfe KH (1999) Gene duplication and gene conversion in the *Caenorhabditis elegans* genome. J Mol Evol 48:555–564

Seo HC, Kube M, Edvardsen RB, Jensen MF, Beck A, Spriet E, Gorsky G, Thompson EM, Lehrach H, Reinhardt R, Chourrout D (2001) Miniature genome in the marine chordate *Oikopleura dioica*. Science 294:2506

Spada F, Steen H, Troedsson C, Kallesoe T, Spriet E, Mann M, Thompson EM (2001) Molecular patterning of the oikoplastic epithelium of the larvacean tunicate *Oikopleura dioica*. J Biol Chem 276:20624–20632

Thioulouse J, Chessel D, Dolédec S, Olivier JM (1997) ADE-4: A multivariate analysis and graphical display software. Stat Comput 7:75–83

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Tweedie S, Charlton J, Clark V, Bird A (1997) Methylation of genomes and genes at the invertebrate-vertebrate boundary. Mol Cell Biol 17:1469–1475

Venkatesh B, Ning Y, Brenner S (1999) Late changes in spliceosomal introns define clades in vertebrate evolution. Proc Natl Acad Sci USA 96:10267–10271

Wada H, Kobayashi M, Sato R, Satoh N, Miyasaka H, Shirayama Y (2002) Dynamic, insertion-deletion of introns in deuterostome EF-1 alpha genes. J Mol Evol 54:118–128