Available online at www.sciencedirect.com
ScienceDirect
Genomics 91 (2008) 315–325
GENOMICS
www.elsevier.com/locate/ygeno

Genomic, evolutionary, and expression analyses of *cee,* an ancient gene involved in normal growth and development

Jorge M.O. Fernandes [a,b,1], Daniel J. Macqueen [b,1], Hung-Tai Lee [b], Ian A. Johnston [b,*]

[a] *Department of Fisheries and Natural Sciences, Bodø University College, N-8049 Bodø, Norway*
[b] *School of Biology, University of St Andrews, St Andrews, Fife KY16 8LB, UK*

Received 21 August 2007; accepted 30 October 2007
Available online 31 January 2008

Abstract

The *cee* (conserved edge expressed protein) gene was recently identified in a genome-wide screen to discover genes associated with myotube formation in fast muscle of pufferfish. Comparative genomic analyses indicate that *cee* arose some 1.6–1.8 billion years ago and is found as a single-copy gene in most eukaryotic genomes examined. The complexity of its structure varies from an intronless gene in yeast and tunicates to nine exons and eight introns in vertebrates. *cee* is particularly conserved among vertebrates and is located in a syntenic region within tetrapods and between teleosts and invertebrates. Low *dN/dS* ratios in the *cee* coding region (0.02–0.09) indicate that the Cee protein is under strong purifying selection. In Atlantic salmon, *cee* is expressed in the superficial layers of developing organs and tissues. These data, together with functional screens in yeast and *Caenorhabditis elegans,* indicate that *cee* has a hitherto uncharacterized role in normal growth and development.
© 2008 Elsevier Inc. All rights reserved.

Keywords: Cee; Conserved edge-expressed protein; Myogenesis; Development; Purifying selection

The myotome of teleost fish contains anatomically discrete layers of slow and fast muscle fibers, predominantly involved in sustained and high-speed swimming activity, respectively [1–3]. Fast muscle fibers continue to be produced in adult fish until they reach around 44% of the maximum attainable body length [4,5]. Fiber recruitment involves myogenic progenitor cells fusing to form myotubes on the surface of existing muscle fibers, giving rise to a mosaic of fiber diameters in subsequent growth stages [6]. In contrast, the production of slow muscle fibers continues to occur in discrete zones until close to the final body size [7,8]. In the model pufferfish species *Takifugu rubripes*, we used subtracted cDNA libraries to identify a number of candidate myotube inhibitory genes that were specifically up-regulated in fast but not slow muscle, concomitant with the cessation of myotube production in fast muscle [9]. One of these genes, originally denoted *FRC386,* was particularly interesting because it corresponded to an uncharacterized protein that was conserved in a wide range of taxa. *FRC386* mRNA transcripts in pufferfish were up-regulated 15-fold in fast muscle following the end of fiber recruitment and were unchanged and present at concentrations more than 5-fold lower in a range of other tissues, including heart, liver, skin, and brain [9]. Large-scale RNAi screens in *Caenorhabditis elegans* revealed that disruption of function of the orthologue of *FRC386* resulted in a retardation of growth and development [10,11].

Since expression analysis in Atlantic salmon (*Salmo salar* L.) indicated that *FRC386* was localized on the surfaces of specific developing tissues and organs we renamed the gene *cee,* for conserved edge expressed protein. In the present study we have cloned the complete coding sequences of *cee* in four teleost species from various orders (Beloniformes, Cypriniformes, Salmoniformes, and Tetraodontiformes). These data, in conjunction with an additional 29 metazoan sequences retrieved by in silico data mining, were used to analyze the phylogeny, structure, and evolution of *cee* in multicellular animals.

* Corresponding author. Fax: +44 1334 463443.
E-mail address: iaj@st-and.ac.uk (I.A. Johnston).
[1] These authors contributed equally to this work.

0888-7543/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2007.10.017

Results and discussion

The origin of cee

We discovered *cee* in a previous study as a gene consistently up-regulated in the fast muscle of tiger pufferfish that had stopped producing new myotubes, compared to smaller fish in a growth phase of active muscle fiber recruitment [9]. The original clone containing *cee* was denoted *FRC386* (GenBank CK829928) and preliminary analyses revealed that it was both uncharacterized and highly conserved throughout evolution. This gene is named *cee* (conserved edge expressed protein) based on its developmental expression pattern in Atlantic salmon embryos (see below).

Exhaustive BLAST similarity searches were performed to identify *cee* in the available cDNA and genome databases. Details of the coverage of each genome assembly can be found in Supplementary Table S1. With three exceptions, a single *cee* gene was found to be present in all metazoan taxa examined, including insects (yellow fever and malaria mosquitoes, honey bee, fruit flies, and red flour beetle), nematodes (*Caenorhabditis* sp.), platyhelminthes (*Schistosoma* sp.), echinoderms (purple sea urchin), teleosts (tiger and green-spotted pufferfishes, medaka, salmon, stickleback, and zebrafish), amphibians (Western and African clawed frogs), birds (chicken), tunicates (ascidians) and mammals (human, chimp, rhesus monkey, mouse, rat, pig, guinea pig, shrews, cow, dog, cat, elephant, opossum, platypus, bushbaby, armadillo, hedgehogs, and microbat). Two *cee* sequences that share 95% identity at the nucleotide level are found in the African clawed frog *Xenopus laevis* (Supplementary Table S1). The yellow fever mosquito (*Aedes aegypti*) contains two paralogues that are present in distinct chromosomal regions (AAEL002521 in supercont1.59 and AAEL012936 in supercont1.765) but their coding sequences are 99.5% identical and indeed code for the same protein. *cee* could not be located in the rabbit genome, perhaps due to the low coverage (2X) and fragmentary nature of this preliminary assembly.

No apparent orthologue of *cee* could be identified in Archaea and eubacterial genomes, indicating that *cee* is specific to eukaryotes. To obtain an estimate for the origin of *cee* we screened all available protist genomes for *cee* orthologues. *cee* is absent in the amitochondriate eukaryote *Giardia lamblia,* which occupies a basal position in the phylogeny of protists and diverged from other eukaryotes circa 2.2 billion years ago (Gya) [12]. Similarly, *cee* orthologues are not present in euglenida (*Euglena gracilis*) and kinetoplastida (*Leishmania major, Trypanosoma brucei, T. vivax,* and *T. congolense*) euglenozoans. According to recent studies regarding the phylogeny of protists (reviewed by [13]), the most primitive eukaryotes in which *cee* orthologues can be identified are the Alveolata. This taxon (phylum Apicomplexa) comprises the malarial parasites *Plasmodium berghei* (GenBank XM_675171), *P. chabaudi* (XM_730628), *P. falciparum* (XM_001348503), *P. yoelii* (XM_721059), and tropical theileriosis parasite *Theileria annulata* (XM_950161). *cee* is also found in Amoebozoa (*D. discoideum,* XM_635525), fungi (*Saccharomyces cerevisiae,* YOR164C), and plants (*Arabidopsis thaliana,* AK176227). Taken together, our data place the origin of *cee* sometime after the most recent symbiotic event thought to have occurred approximately 1.8 Gya during the evolution of eukaryotic organisms [12] and prior to the divergence of animals/fungi and plants, which dates back to circa 1.6 Gya [14].

We have obtained experimental complete coding sequences for *cee* in four teleost fishes from the orders Beloniformes (medaka), Cypriniformes (zebrafish), Salmoniformes (Atlantic salmon), and Tetraodontiformes (tiger pufferfish). These nucleotide sequences were submitted to GenBank as *cee,* in conformity with the guidelines proposed by the zebrafish nomenclature committee, and their accession numbers are shown in Supplementary Table S1. For further characterization and evolutionary analysis of *cee* in metazoans, only complete coding sequences derived from high-quality predictions or with experimental support were used (Supplementary Table S1).

Cee has no known functional domains and is highly conserved in vertebrates

The putative proteins coded by *cee* range from 307 residues in chicken to 362 amino acids in *C. elegans.* Vertebrate Cee proteins have an acidic isoelectric point (5.2–5.6) and are particularly rich in leucine (~13%) and serine (~8–10%). There is a notable degree of conservation between Cee orthologues from different vertebrate taxa (Fig. 1), which share an overall identity of at least 80% when any two species are compared (Supplementary Table S2). In vertebrates, differences within Cee are generally distributed throughout the entire protein and many of the amino acid substitutions are isofunctional replacements, as shown in the multiple sequence alignment (Fig. 1). The regions of highest variability when all taxa are considered correspond to the amino- and carboxy-termini of Cee and to residues 91–102 and 179–186 in the zebrafish sequence (Fig. 1). The predicted chicken Cee protein, which is derived from an experimental sequence, has a deletion in a region (residues 151–166) that would otherwise be well conserved across vertebrates (Fig. 1). The primary structure of Cee from invertebrates is rather more diverse and shares approximately 30 to 40% identity at the protein level with its mammalian orthologues (Supplementary Fig. S1, Supplementary Table S2). The identity values between vertebrate Cee proteins and their orthologues in *P. chabaudi, Di. discoideum, Sac. cerevisiae,* and *C. elegans* are 19, 29, 26, and 23%, respectively. Despite the relatively low degree of similarity among invertebrate Cee orthologues, a conserved region corresponding to residues 39–52 can be identified within the invertebrate sequences (Supplementary Fig. S1). This domain is also highly conserved in all vertebrate species except the platypus, which has seven substitutions within this region (Supplementary Fig. S1). It is noteworthy that the motif YYEAHQ is also present in *Plasmodium* Cee (data not shown), suggesting that this might be an ancient functional domain.

Perhaps the most striking feature of Cee is the lack of known motifs or conserved domains to provide any insight as to what its cellular localization and molecular function might be. The limited information available regarding Cee is derived from high-throughput studies using *Sac. cerevisiae* and *C. elegans.* The yeast orthologue of Cee (YOR164C) is located in the

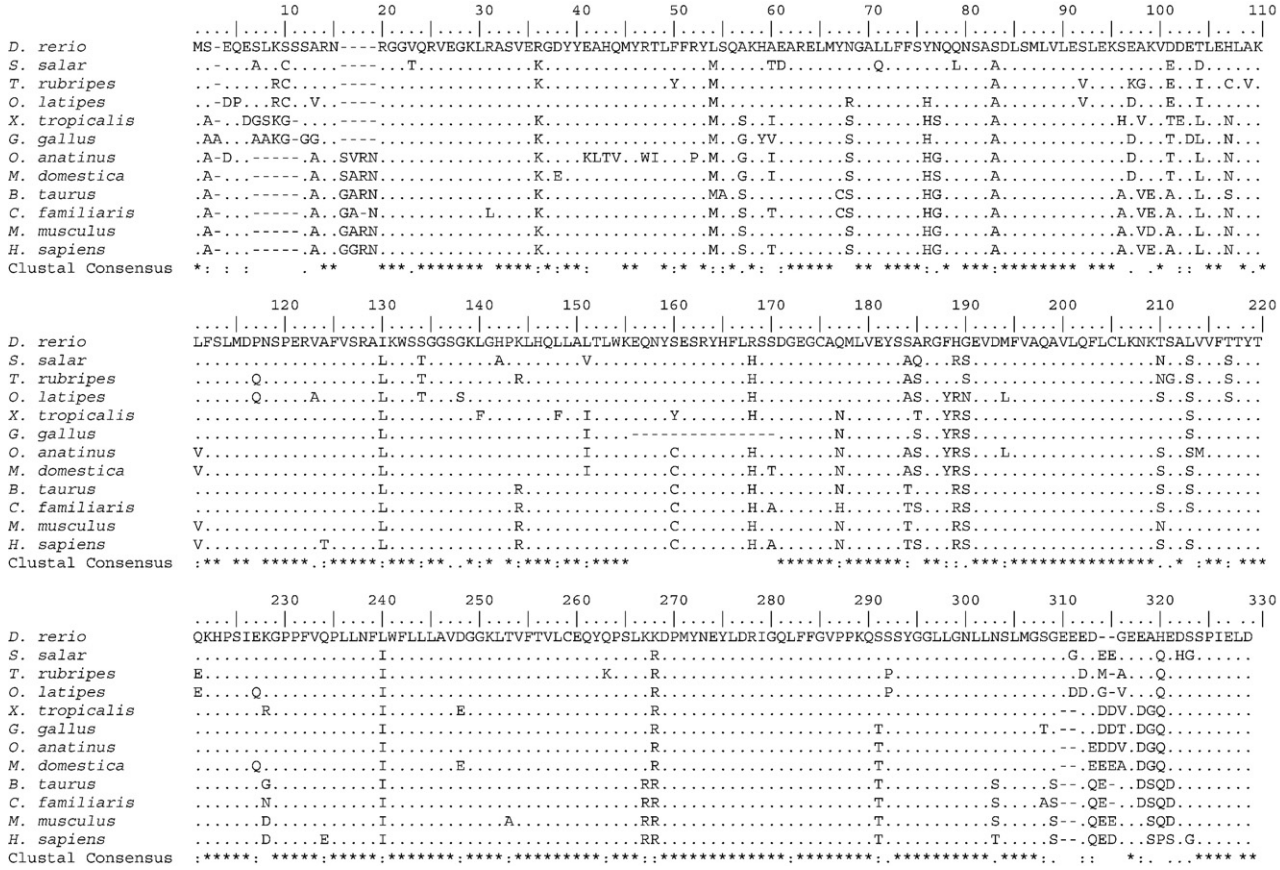

Fig. 1. Comparison of Cee polypeptide sequences from vertebrates. Putative protein sequences from cow (*B. taurus*), dog (*Can. familiaris*), zebrafish (*Da. rerio*), chicken (*Gal. gallus*), human (*H. sapiens*), opossum (*Mo. domestica*), mouse (*M. musculus*), platypus (*Orn. anatinus*), medaka (*Ory. latipes*), Atlantic salmon (*Sal. salar*), tiger pufferfish (*Ta. rubripes*), and Western clawed frog (*X. tropicalis*) were aligned with CLUSTALW. For a list of accession numbers please consult Supplementary Table S1. Amino acid residues identical to the zebrafish sequence are represented by a dot. In the consensus sequence, identical residues and conserved and semiconserved substitutions are indicated by asterisks, colons, and dots, respectively. Despite the extensive global similarity between vertebrate Cee proteins, no conserved domains of known function were identified.

cytoplasm [15]. Its function is unknown but affinity capture–mass spectrometry and two-hybrid experiments have revealed that yeast Cee interacts with Mdy2 [16] and Get3 [17]. Mdy2 has a ubiquitin-like domain, associates with ribosomes, and is required for efficient mating [18]. The ATPase Get3 is necessary for transporting proteins from the Golgi apparatus to the endoplasmic reticulum [19] and it is also involved in resistance to heat and metal stress [20]. The Cee yeast mutants, obtained by targeted deletion, were found to be viable but exhibited sensitivity at five generations when grown in the presence of the antifungal nystatin; however, it is not clear how *cee* contributes to yeast fitness [21]. When RNA interference was used to inhibit the function of *cee* in *C. elegans,* the mutants' development was retarded [10,11]. These data indicate that Cee may be a positive regulator of growth that is involved in protein binding, intracellular traffic, or translation. Since there are considerable sequence differences between the vertebrate and the invertebrate orthologues, it is plausible that Cee has additional molecular functions and is involved in other biological processes in vertebrates. It is noteworthy that some microarray experiments listed in the ArrayExpress database (http://www.ebi.ac.uk/microarray-as/aer) identify *cee* as the top differentially expressed gene in various human conditions, including Huntington disease and several types of malignant tumors.

Genomic architecture and synteny analysis of cee

The genomic structure of *cee* is identical among all the vertebrate orthologues examined and it comprises nine exons and eight introns that range from 3.6 to 43.6 kb in tiger pufferfish and zebrafish, respectively (Fig. 2). The lengths of all exons and the locations of their splice junctions are remarkably well conserved from fish to mammals. Despite some diversity of intron sizes within vertebrates, intron I–II is generally the largest, with the exception of zebrafish, in which intron VII–VIII spans approximately 30 kb (Fig. 2). The intron/exon structure of the *cee* gene is not conserved among invertebrates and its complexity varies from two to eight exons in the yellow fever mosquito and the purple sea urchin, respectively (Fig. 2). The nine-exon structure seems to have arisen during vertebrate evolution but it shares remarkable similarities with the eight-exon structure found in the purple sea urchin. Exons 1 to 6 and exon 8 of *cee* in the purple sea urchin have sizes similar to those of their vertebrate counterparts and the splice sites between

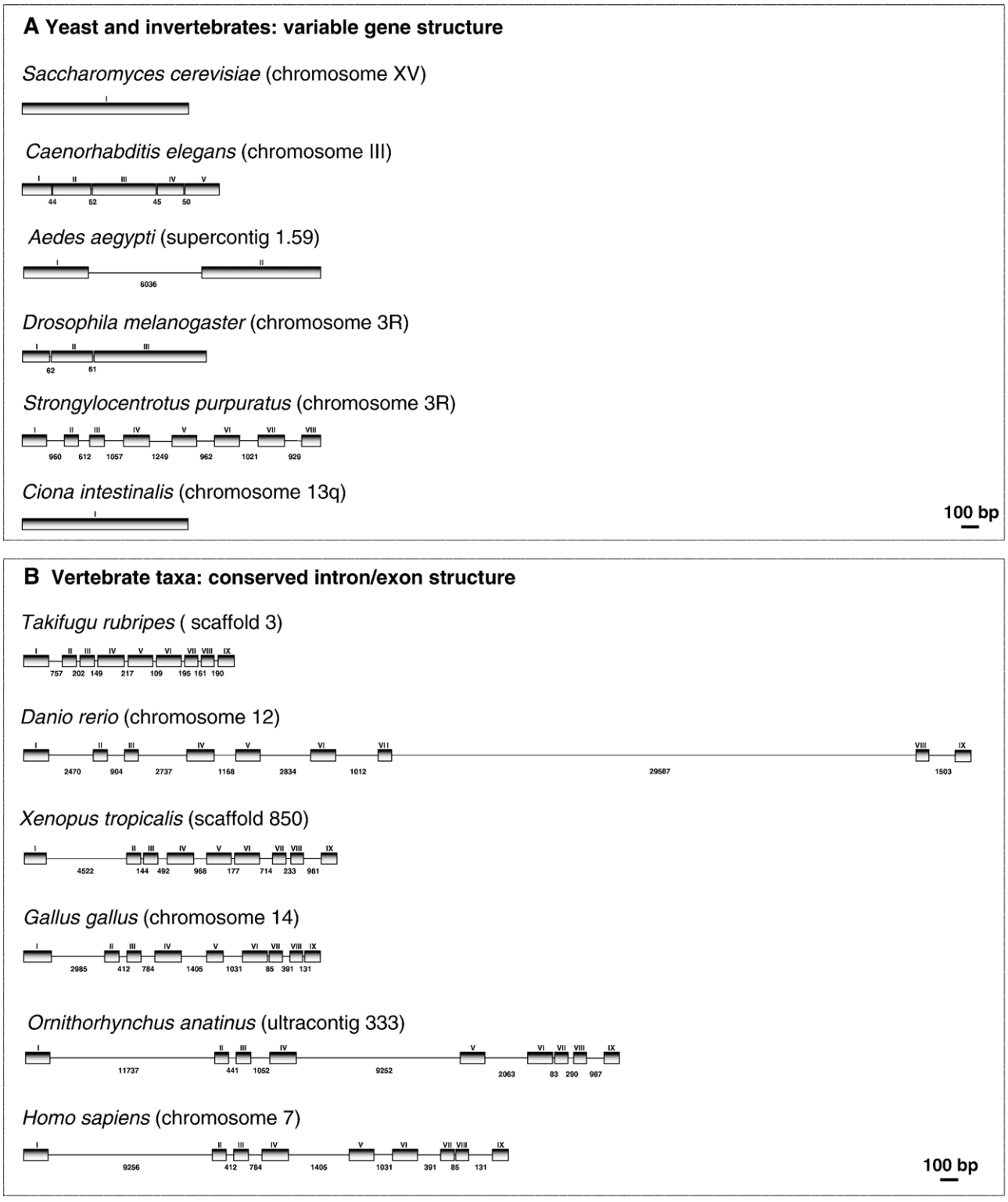

Fig. 2. Genomic structure of *cee* in several taxonomic groups. The structure of the *cee* gene (nine exons and eight introns) is highly conserved in vertebrates and different from that in other taxa, including lower chordates (tunicates). Exons are represented by bars labeled with Roman numerals. Introns are downscaled by 10 times but their real size (bp) is indicated. Only the gene regions corresponding to the open reading frame are shown in this diagram.

exons 1/2, 2/3, 4/5, 5/6, and 6/7 are conserved, as are the exon/intron borders between the last two exons. Moreover, the size of exon 7 in the purple sea urchin (149 bp) is equivalent to the combined sizes of exons 7 and 8 in vertebrates. The size, sequence similarity, and conservation of splice site junctions indicate that exons 7 and 8 found in the vertebrate *cee* orthologues were probably created by the introduction of a spliceosomal intron in an exon that is homologous to exon 7 in the purple sea urchin. In the sea squirt, *cee* has a genomic organization different from that of other vertebrate and invertebrate

taxa. In fact, it is an intronless gene similar to the baker's yeast orthologue (Fig. 2), which is rather surprising since tunicates are the closest extant relatives of vertebrates [22]. It is possible that the intronless version of *cee* found in the sea squirt has been created by reverse transcription of the processed mRNA followed by genome integration (retrotransposition), a common molecular mechanism of gene formation in eukaryotes [23]. However, we found no evidence supporting the existence of a putative parental *cee* gene in the ascidian's genome.

Phylogenetic footprinting between the tiger pufferfish, the zebrafish, the Western clawed frog, the chicken, and the human orthologues of *cee* failed to identify any evolutionarily conserved regions (i.e., 60% identity over 100 bp) within the intronic sequences. Similarly, no conserved and aligned transcription factor binding sites are present in the 10-kb sequence upstream of the putative translation start site of *cee.* Comparison of *cee* coding sequences reveals that exons 4, 5, and 6 are the best conserved across vertebrates (Supplementary Fig. S2).

The human orthologue of *cee* is a predicted Ensembl gene officially known as C7orf20 and it is found on the short arm of chromosome 7 (p22.3) at location 882,717–902,597. In the tiger pufferfish, *cee* is located on scaffold 3 and is surrounded by various genes, including the ATP-dependent DNA helicase 2 subunit 1, which plays a role in chromosome translocation (*xrcc6,* J04611); the transcription factor with antiapoptotic activity myocardin-like protein 1 (*mkl1,* AJ297257); the zinc finger DHHC domain-containing protein 16, involved in apoptosis (*zdhhc16,* AF176814); the DNA repair/transcription protein MMS19 (*mms19L,* AF357881); and a novel gene associated with esophageal cancer in humans (C16orf62) (Fig. 3A). As shown in Fig. 3A, the chromosomal region containing *cee* is syntenic in all teleost species examined and synteny is particularly

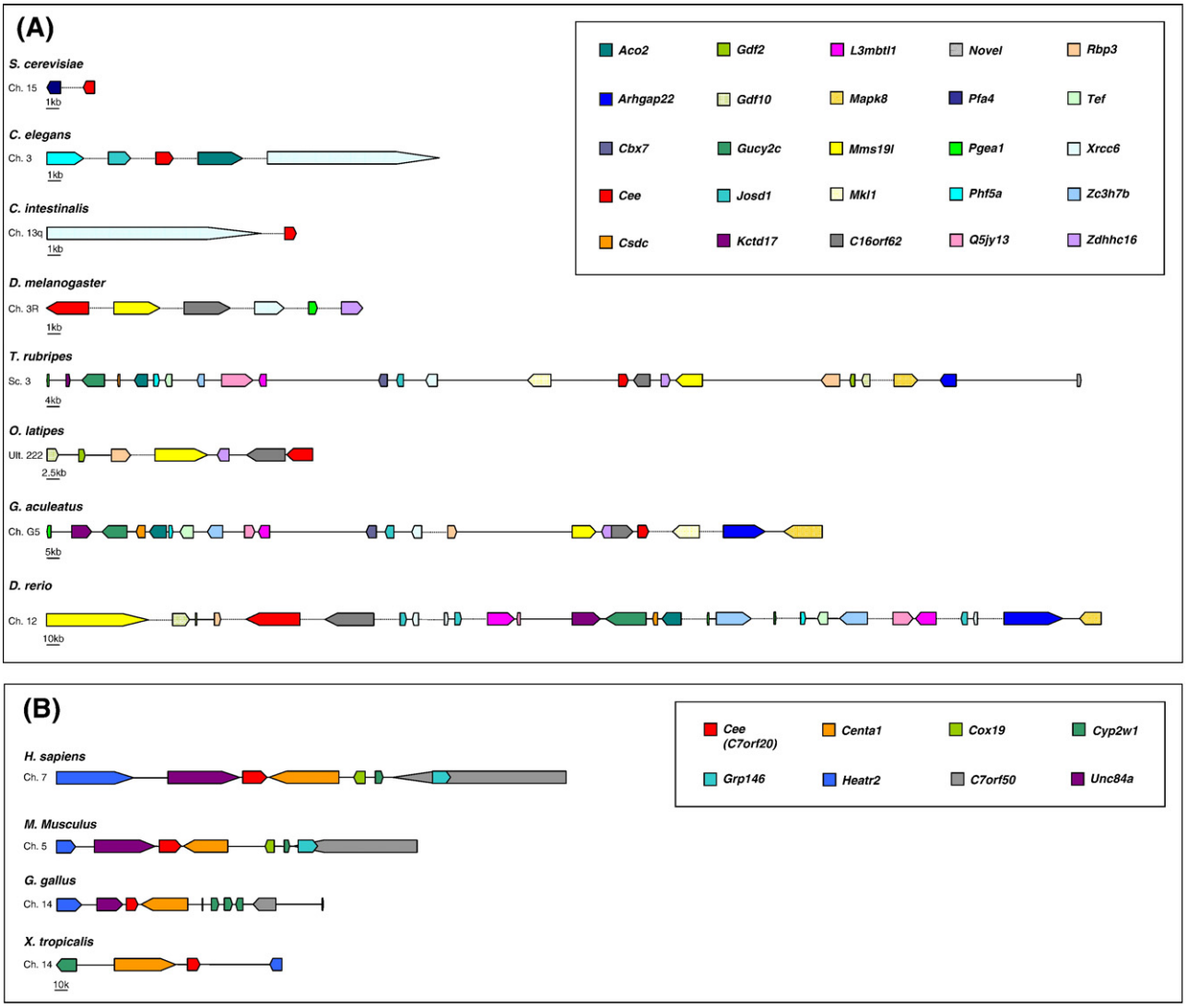

Fig. 3. Partial synteny map of *cee* and surrounding genes. (A) Genomic neighborhood of *cee* in *Ta. rubripes* and its orthologues in *C. elegans, Ci. intestinalis, D. melanogaster, Da. rerio, Gas. aculeatus, Ory. latipes,* and *Sac. cerevisiae.* Synteny is conserved between fish and invertebrates but disrupted between fish and tetrapods. (B) Diagram of the genes surrounding *cee* in *Gal. gallus, H. sapiens, M. musculus,* and *X. tropicalis,* illustrating the synteny conservation among these taxa. Genes are color coded and represented by block arrows that reflect their orientation in the genome. c*ee* is highlighted in red. Introns are not represented to scale.

well conserved between tiger pufferfish, zebrafish (chromosome 12), and stickleback (linkage group V), except for some gene inversions and chromosomal translocations. Despite the long evolutionary distance between these taxa, synteny is conserved in teleosts, tunicates, insects, and nematodes, albeit to a lesser extent (Fig. 3). In higher vertebrates the organization of the genomic region surrounding *cee* is very different from that observed in teleosts, but synteny is conserved among tetrapods (Fig. 3B).

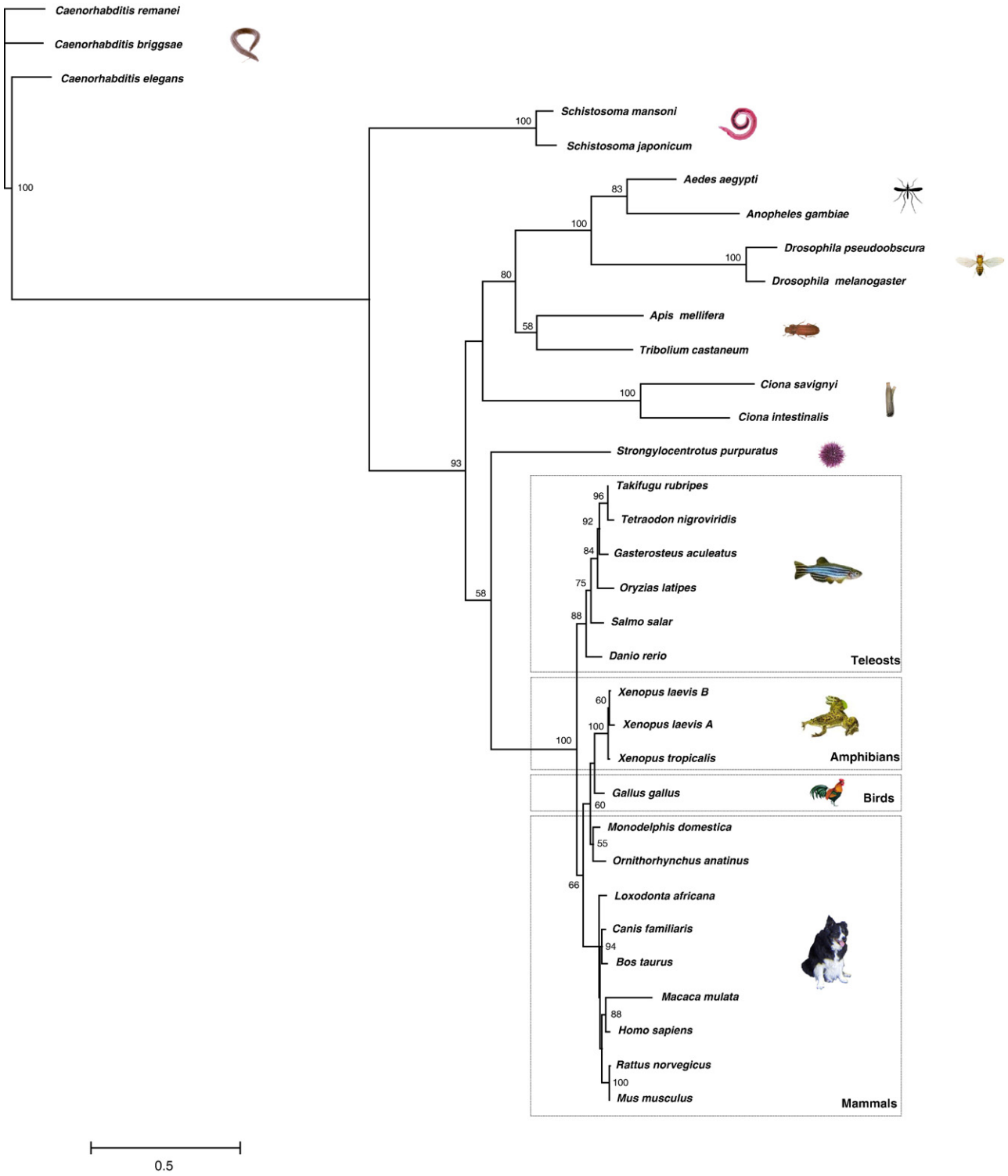

Fig. 4. Maximum-likelihood tree of Cee proteins from metazoans. This unrooted phylogram was obtained using a WAG model of amino acid evolution. Similar topologies were obtained using Bayesian-inference (Supplementary Fig. S3) and neighbor-joining (Supplementary Fig. S4) methods. Accession numbers for Cee sequences are listed in Supplementary Table S1.

Molecular evolution of cee

cee seems to be an orphan gene, in the sense that it is present in most organisms as a single copy and has no apparent homology to genes from any other family. This is rather unusual given that there is a tendency for genes found as single copies in protostomes or deuterostomes ancestral to vertebrates to be represented by up to four paralogues in most vertebrates (eight in ray-finned fish) due to whole-genome duplication events [24]. These paralogues effectively form families of closely related genes that are related to a single ancestral orthologue in subvertebrate groups. Several genes in the vicinity of *cee* have multiple paralogues (e.g., *mkl1* and *gdf2*), indicating that this paralogon was not lost after each round of whole genome duplication. The existence of a single *cee* gene in most species suggests that its expression needs to be tightly regulated (which would be difficult to achieve with two independent promoters) and is consistent with the rapid nonfunctionalization of duplicated paralogues by deleterious mutations in coding or regulatory regions, resulting in pseudogenes that would have degraded without evolutionary constraints to become unrecognizable within the genome of most vertebrates. There are two species in which *cee* is present as two paralogues. In the yellow fever mosquito (*Ae. aegypti*), two *cee* genes sharing 95% identity within their coding sequences are found in distinct chromosomal regions. The lack of divergence between these paralogues is consistent with a very recent species-specific gene duplication event. The tetraploid African clawed frog (*X. laevis*) has also retained two paralogues of *cee.* Since only one *cee* gene is present in the Western clawed frog (*X. tropicalis*), a close relative with a diploid genome, it is likely that the second copy of *cee* in *X. laevis* arose during the allotetraploidization event that occurred in the *Xenopus* lineage approximately 30 million years ago [25]. It is plausible that this recent *cee* paralogue will undergo subfunctionalization and perhaps degenerate into a pseudogene, since there is an overall relaxation of selective evolutionary pressure on duplicated gene pairs in *X. laevis* [26].

We have studied the phylogenetic relationships of *cee* orthologues and reconstructed trees using maximum-likelihood (Fig. 4), Bayesian-inference (Supplementary Fig. S3), maximum-parsimony (data not shown), and neighbor-joining (Supplementary Fig. S4) methods. All trees have overall similar topologies and most nodes have good bootstrap support. The branch lengths are relatively short, particularly within vertebrate clades, reflecting a high degree of conservation among these taxa. Cee from Coleopterans (beetles), Hymenopterans (bees), and Dipterans (flies and mosquitoes) forms a monophyletic group. Similarly, all vertebrate Cee proteins are part of the same clade, as expected. The topology of the teleost branch follows the currently accepted phylogenetic relationship between Cypriniformes, Salmoniformes, Beloniformes, and Tetraodontiformes [27]. Stickleback (Gasterosteiforme) Cee clusters closely with its Tetraodontiforme homologues. In three of the four phylogenetic trees, the two paralogues of Cee from *X. laevis* are grouped separately from their *X. tropicalis* orthologue (Fig. 4, Supplementary Fig. S3), supporting the hypothesis that they arose during the *Xenopus* allotetraploidization event [25], which did not occur in *X. tropicalis.* In *X. laevis,* the two *cee* genes have evolved asymmetrically and the paralogue herein designated B (GenBank BC074468) was found to be more similar to the ancestral gene (Fig. 4).

The ratio between nonsynonymous (amino acid-changing, *dN*) and synonymous (silent, *dS*) substitutions is an indicator of selective evolutionary constraints at the protein level and can be used to ascertain if a gene is under positive, neutral, or purifying selection [28]. We have determined the cumulative number of nonsynonymous and synonymous substitutions across *cee* coding regions and calculated their *dN/dS* ratios in the following taxa: insects (*Ae. aegypti, Anopheles gambiae, Apis mellifera, Drosophila melanogaster, D. pseudoobscura,* and *Tribolium castaneum*), platyhelminthes (*Schistosoma japonicum* and *Sch. mansoni*), nematodes (*C. briggsae, C. elegans,* and *C. remanei*), teleosts (*Da. rerio, Gasterosteus aculeatus, Oryzias latipes, Sal. salar,* and *Ta. rubripes*), amphibians (*X. laevis* and *X. tropicalis*), and mammals (*Bos taurus, Canis familiaris, Homo sapiens, Macaca mulatta, Monodelphis domestica, Mus musculus, Ornithorhynchus anatinus,* and *Rattus norvegicus*). In fish (Fig. 5A), mammals (Fig. 5B), and

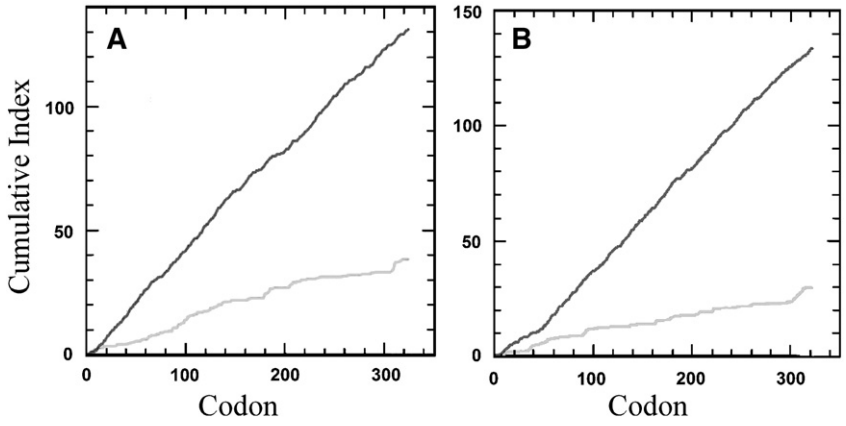

Fig. 5. Synonymous and nonsynonymous substitution rates in *cee.* Cumulative indices of average synonymous (black line) and nonsynonymous (gray line) substitutions are plotted against the aligned Cee protein sequences from (A) teleosts and (B) mammals. The number of synonymous substitutions per site is approximately constant throughout the coding sequence and higher than the number of nonsynonymous mutations, suggesting that *cee* is under strong purifying selection.

amphibians (data not shown) the number of synonymous substitutions per site is approximately constant in all exons and higher than the number of nonsynonymous mutations in teleosts and mammals, suggesting that *cee* is under strong purifying selection. The coding sequences of *cee* in nematodes, platyhelminthes, and insects have an overall higher number of synonymous substitutions compared to vertebrate taxa, which correspond to a lesser degree of *cee* conservation within these taxonomic groups (data not shown). The insect Cee proteins are particularly variable at their amino-and carboxy-termini (data not shown). The average *dN/dS* ratios of all pair-wise comparisons are 0.09, 0.06, 0.07, 0.04, 0.02, and 0.03 in insects, platyhelminthes, nematodes, teleosts, amphibians, and mammals, respectively. These low *dN/dS* values indicate that *cee* is

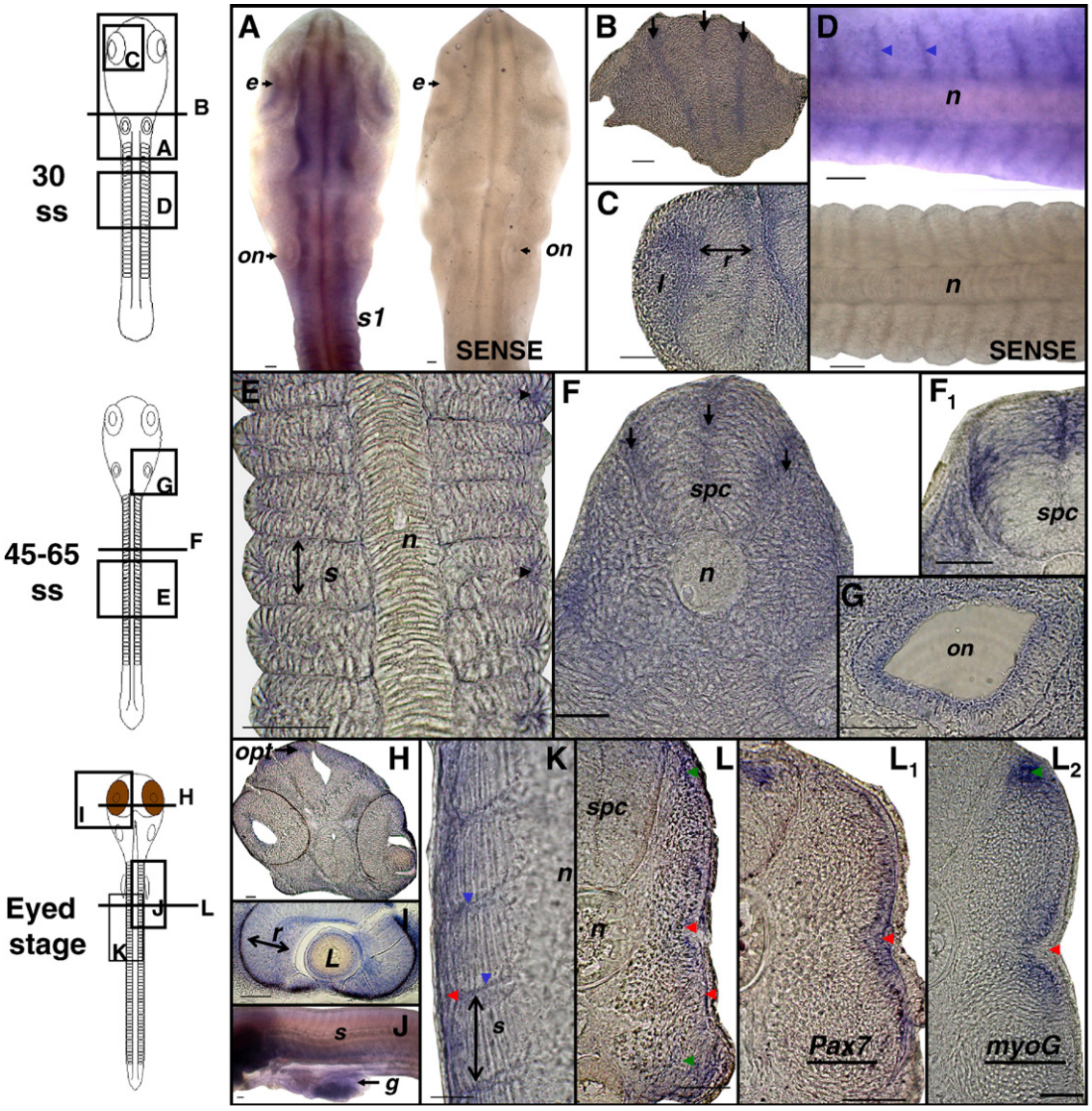

Fig. 6. Developmental expression pattern of cee in Atlantic salmon embryos. Schematic images of embryos to the left show the position of higher magnification flat mounts and sections (A–L). (A) Dorsal flat mounts showing the rostral region of a 30-ss embryo incubated with cee antisense or sense mRNA (marked SENSE). (B) Cross section through the midbrain (at the level of the optic tectum) at the 30-ss. Black arrows show cee expression at the midline and borders of the entire dorsal–ventral brain. (C) Cross section through the eye of a 30-ss embryo. *cee* was expressed at the lens–retina border throughout the segmentation period. (D) Dorsal flat mount of the somite region of 30-ss embryos showing *cee* staining at the somite borders. An equivalent sense control embryo (labeled SENSE) is shown for comparison. (E) Longitudinal section through the epithelial somites of a 45-ss embryo. *cee* mRNA clustered within cells on the lateral somite border and at the somite boundaries. (F) Somite cross section from an anterior somite at the 45-ss. Note the triple stripe of spinal cord expression marked by black arrows. Also shown as an inset is *pax7* staining at an equivalent stage (F_1). (G) Longitudinal section though the otolith nuclei; *cee* stained the internal edge. This image is representative of *cee* staining during segmentation and eyed stages. (H) Cross section through the midbrain of an eyed embryo at the level of the optic tectum. (I) Transverse section through the pigmented eye. *cee* expression continued on the boundaries of the lens and the retina beyond segmentation. (J) Lateral perspective flat mount of an eyed embryo showing the unrestricted expression of *cee* in the developing gut. (K) Longitudinal section along the somites of an eyed-stage embryo. *cee* was excluded from the medial myotome but was expressed at its lateral edge, in the external cell layer (red arrows), and along the somite borders (blue arrow). (L) Eyed-stage myotome cross section at the level just below the fin buds. *cee* was expressed in the external cell layer (red arrow) concomitant with *pax7* (L_1) and it was also present in the myotome in regions of new muscle production (green arrows), where muscle-specific markers like myogenin (L_2) were expressed. Abbreviations: e, eye; g, gut; l, lens; on, otolith nuclei; n, notochord; opt, optic tectum; r, retina; s, somites; and spc, spinal cord. Scale bars, 50 μm.

under selective pressure that favors silent substitutions in all taxa examined, particularly within vertebrates. A codon-based Fisher's exact test of positive selection [29] confirmed that $dN<dS$ ($p=1$) in all vertebrate groups, thus showing that the Cee protein is under strict purifying selection to resist amino acid changes and is very likely to have a conserved function across multiple vertebrate taxa. This strong evolutionary constraint might also explain why *cee* is found as a single gene in all vertebrate species examined, except in the tetraploid frog *X. laevis.*

Developmental expression of cee

We used in situ hybridization to analyze the mRNA distribution of *cee* in six developmental stages of Atlantic salmon embryogenesis, ranging from the onset of somitogenesis until postsegmentation. Specifically, these stages are (i) 1-somite stage (ss), (ii) 5-to 10-ss, (iii) 25-to 35-ss, (iv) 40-to 45-ss, (v) the end of segmentation (60- to 65-ss), and (vi) postsegmentation (designated the "eyed stage" due to the obvious dark eye pigmentation).

During segmentation, somites are added at the caudal end of the developing embryo until a maximum of 65 or 66 are formed. Sense controls were used for all stages and never produced specific staining. A signal for *cee* mRNA is absent at 0-to 10-ss but is consistently detected from ~25-ss in a pattern consistent with a role in the development of various organs and structures. *cee* is strongly expressed along the entire cranial–caudal axis of each embryo in three stripes marking the lateral edges and midline of the entire brain and neural tube/spinal cord (Fig. 6A, sense control included for comparison). Cross sections revealed that this staining extends through the entire dorsal–ventral axis of the brain and developing spinal cord (Figs. 6B and F). *pax7,* a transcription factor with important roles in the development of neural and muscular tissues [30], is also expressed at the midline and lateral edges of the dorsal spinal cord but is less restricted than *cee,* being present to a greater or lesser extent throughout the width of this region (Fig. 6F_1). By the eyed stage, *cee* transcripts are still present in the spinal cord but are no longer restricted to the edge and midline and colocalize with *pax7* across the spinal cord's dorsal width (not shown). At this stage, *cee* is expressed as a broad band surrounding the superficial edge of the entire cranial region (Fig. 6H) and at the borders of several structures within the developing brain (not shown). During somitogenesis *cee* is expressed at the boundaries of the somites as they develop from simple oval-shaped structures (Fig. 6D, marked by blue arrowheads, sense control included for comparison) to chevron-shaped structures with elongated muscle fibers at the eyed stage (Fig. 6K). *cee* transcripts are also detected diffusely throughout the myotome and ventral regions of somites during most of the segmentation period (Fig. 6F). A longitudinal section through the epithelial somites at the 45-ss revealed that staining is clustered mainly between cells at the superficial–lateral border of the somite (Fig. 6E, black arrows). *cee* is down-regulated in the medial myotome during late and postsegmentation stages when elongated muscle fibers are clearly present, but is expressed at the outer edge of the myotome, particularly in dorsal and lateral regions as well as the cell layer external to the myotome (Figs. 6K and L). In zebrafish embryos, *pax7*-expressing cells arise in the anterior compartment of epithelial somites and migrate laterally as the maturing somite rotates 90° from its starting state to form a cell layer external to the myotome [31,32]. The external cell layer contributes to myogenic precursors involved in larval and adult muscle growth and dermal cells of the skin [31]. In the regions where new muscle fibers are added, *myoD* family member genes such as *myoG* are also expressed [33]. Pax7 mRNA is present throughout the external cell layer at the end of the segmentation and eyed stage of Atlantic salmon development (Fig. 6L_1). However, *myoG* is expressed at the lateral edge of the myotome, particularly in dorsal and ventral regions and at the level of the horizontal myoseptum in zones of new muscle fiber production (Fig. 6L_2). The other myoD family genes (*myoD1a/1b/1c, myf5, myf6*) are also expressed in similar regions at the eyed stage of Atlantic salmon development [34]. Thus, *cee* is expressed concomitantly with *pax7* and myoD family genes in the external cell layer and myotomal compartment, respectively (Fig. 6L).

As the eye develops during the segmentation and postsegmentation stages, *cee* is expressed on the innermost and outermost surfaces of the retina bordering the lens and retinal epithelium, respectively (Figs. 6A, C, and I). *cee* transcripts can be found at the boundaries of several other structures throughout their embryonic development, such as the otolith nuclei (Fig. 6G), branchial arches, and fin buds (not shown). At the eyed stage, unrestricted *cee* staining is also present throughout several structures of the developing gut (Fig. 6J). The complex developmental expression pattern of *cee* raises the hypothesis that it may play an important role in multiple biological processes.

Materials and methods

In silico identification of cee orthologues

The putative translation product of *FRC386* from *Ta. rubripes* (GenBank Accession No. CK829928) was used as a probe in TBLASTN similarity searches [35] to identify all homologous metazoan sequences in the following databases: nonredundant sequence and expressed sequence tag databases at NCBI (http://www.ncbi.nlm.nih.gov/BLAST/), Uniprot (http://www.expasy.uniprot.org/), WormBase (http://www.wormbase.org/) and Ensembl (http://www.ensembl.org/). These predictions were further analyzed with the gene structure prediction software Genebuilder (http://l25.itba.mi.cnr.it/~webgene/genebuilder.html) and manually refined. For comparison purposes, *cee* orthologues were also retrieved in the yeast (*Sac. cerevisiae*), the social amoeba (*Di. discoideum*), and other protists (*L. major, Plasmodium* sp., and *Trypanosoma* sp.) from SGD (http://www.yeastgenome.org/; Accession Reference YOR164C), dictyBase (http://dictybase.org/; Accession Reference DDB0218329), and the Wellcome Trust Sanger Institute (http://www.sanger.ac.uk/), respectively. The *G. lamblia* (http://gmod.mbl.edu/), *E. gracilis* (http://tbestdb.bcm.umontreal.ca/), and microbial databases at NCBI were also screened for *cee* orthologues.

Animals and sample collection

A wild-caught tiger pufferfish of 1.4 kg was purchased from the local fish market in Maisaka (Shizuoka, Japan). Medaka (*Ory. latipes*) and zebrafish (*Da. rerio*) were bred in captivity at the Ocean Research Institute (The University of Tokyo, Japan) and at the Gatty Marine Laboratory (University of St Andrews, UK), respectively. Two adult Atlantic salmon (*Sal. salar*) were obtained from EWOS innovation (Lonningdal, Norway). All fish were humanely killed in conformity

Table 1
List of primer pairs used to amplify *cee* and corresponding amplicon sizes

Primer pair	Forward primer (5′ →3′)	Reverse primer (5′ →3′)	Size (bp)
Cee-CDS-Dr[a, b]	GTTCGGTTGGTCGGAGCAG	TTAATTATGCTCAATCACACCTC	1715
Cee-CDS-Ol[a]	GCGGAGAAGGATCGACCATGTC	CGGCTGTTGGGTCAGTCCAG	1000
Cee-CDS-Ss[a]	ATGTCGGAGCAGGAGGCTCTG	TCAGTCCAGCTCAATGGGGC	969
Cee-CDS-Tr[a]	GCAACGATGTCGGAACAAGAATC	TTTATCTTTGTCCTGAGGTGGG	1001
Cee-ISH-Ss[c]	GAACGGATGCTCAGCTTTATAGC	TCAGTCCAGCTCAATGGGGC	1158

[a] Sequences of the primer pairs used to amplify the complete coding sequence of *cee* orthologues in zebrafish (Cee-CDS-Dr), medaka (Cee-CDS-Ol), Atlantic salmon (Cee-CDS-Ss), and tiger pufferfish (Cee-CDS-Tr).
[b] The primers for zebrafish *cee* also amplify the full 3′ untranslated region, hence the larger product size in relation to the other fish species.
[c] Primer sets used to clone *cee* for in situ hybridization of salmon embryos.

with the British Home Office guidelines. Fast muscle samples were carefully dissected and stored in RNAlater (Ambion) for subsequent RNA isolation. Atlantic salmon embryos reared at Akvaforsk (Sunndalsora, Norway) at 10 °C were fixed in 4% (m/v) paraformaldehyde (Sigma) in phosphate-buffered saline (Sigma), dehydrated in a graded methanol (Fisher) series, and then stored in 100% methanol at −70 °C.

Cloning and sequencing of cee cDNAs

Total RNA from fast muscle was extracted, quantified, and used for cDNA synthesis as previously described [36]. Controls lacking reverse transcriptase were also included. To amplify the full coding sequences of *cee,* 1 μl of cDNA from Atlantic salmon, medaka, tiger pufferfish, or zebrafish was used in standard PCR containing the primer pairs Cee-CDS-Ss, Cee-CDS-Ol, Cee-CDS-Tr, and Cee-CDS-Dr, respectively (Table 1).<T1>PCR products were then ligated onto a pCR4-TOPO plasmid vector (Invitrogen) before chemical transformation into TOP10 *Escherichia coli* cells (Invitrogen). Plasmids were extracted, purified, and sent for fluorescent dideoxynucleotide sequencing at the University of Dundee (UK).

Sequence analyses

Experimental sequencing data were assembled into contigs using the SeqMan software from the DNASTAR package (USA). Nucleotide sequences were translated with DNAMAN (Lynnon Biosoft, Canada) and the putative proteins aligned by CLUSTALW at the Kyoto Bioinformatics server (http://align.genome.jp/) using a BLOSSUM matrix with the default parameters. Pair-wise protein sequence comparisons were performed with BioEdit [37]. The ScanProsite software (http://www.expasy.ch/prosite/) was used to identify structural and functional motifs in the Cee protein sequences. The intron/exon structures of *cee* in various species were determined with Spidey (http://www.ncbi.nlm.nih.gov/spidey/). Analyses of synteny between *cee* from *Ta. rubripes* and its orthologues in *C. elegans, Ciona intestinalis, D. melanogaster, Da. rerio, Gas. aculeatus, Gallus gallus, H. sapiens, M. musculus, Ory. latipes, Sac. cerevisiae,* and *X. tropicalis* were performed with the data mining tool BioMart (http://www.ensembl.org/biomart/martview/). The presence of evolutionarily conserved regions and transcription factor binding sites in the 10,000 bp upstream of the translation start site was investigated using Mulan and multiTF [38].

Phylogenetic inference and tests of selection

Bayesian inference of phylogeny was performed with MrBayes [39] with an average mixed amino acid model of protein evolution. Markov chain Monte Carlo runs of 1,000,000 generations were used and Bayesian posterior probabilities were estimated on the final 9000 trees. A maximum likelihood analysis was performed with PhyML [40] using a WAG model of amino acid evolution and assuming a gamma distribution of substitution rates. The reliability of this tree was tested using a bootstrap test with 500 pseudoreplicates. Neighbor-joining (amino acid model with Poisson correction) and maximum-parsimony trees were reconstructed with MEGA [41]. A bootstrap test with 10,000 replicates was used to test the inferred phylogenies. The coding sequences corresponding to Cee in various species were retrieved from the appropriate databases (Supplementary Table S1) and grouped by taxon, as follows: nematodes, platyhelminthes, insects, teleosts, amphibians, tunicates, and mammals. PAL2NAL [42] was used to align the coding sequences within each group according to the respective protein sequence alignment. The average numbers of synonymous (*dS*) and nonsynonymous (*dN*) substitutions, insertions, and deletions in the codon alignments were determined with SNAP [43], which is based on the method developed by Nei and Gojobori [44]. A codon-based Fisher's exact test of positive selection based on the *dN/dS* ratios between sequences was performed with MEGA [41].

RNA probe preparation and whole-mount in situ hybridization

A 1157-bp amplicon containing 189 bp of the 5′ untranslated region and the full coding sequence of salmon *cee* was amplified by PCR with T3/T7 primers from a pCR4-TOPO plasmid containing the appropriate insert. This PCR product was used to synthesize sense and antisense DIG-labeled *cee* RNA probes by in vitro transcription with T3 or T7 RNA polymerases (Roche), according to the manufacturer's instructions. Whole-mount in situ hybridization was performed following a standard procedure [45] using six embryos per developmental stage. Bound DIG-labeled probes were detected with alkaline phosphatase conjugated to anti-DIG Fab fragments (Roche) using the chromogenic substrates 5-bromo-4-chloro-3′-indoly phosphate *p*-toluidine (Roche) and nitroblue tetrazolium (Roche). Cryosections were prepared by flash freezing embryos mounted in Cryomatrix (Thermo Electron Corp., UK) in isopentane cooled to near freezing (−159 °C) over liquid nitrogen before 18-μm sections were cut on a CM1850 cryostat (Leica Microsystems). Whole-mount embryos and sections were photographed using a Leica DMRB compound or Leica MZ7.5 binocular microscope and a Nikon Cool-Pix camera.

Acknowledgments

This work was funded by the Integrated Project SEAFOOD-plus, EU (Contract 506359). D.J.M. was supported by a studentship (NER/S/A/2004/12435) from the Natural Environment Research Council, UK.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2007.10.017.

References

[1] J.D. Altringham, I.A. Johnston, Modelling muscle power output in a swimming fish, J. Exp. Biol. 148 (1990) 395–402.
[2] Q. Bone, J. Kicenuik, D.R. Jones, On the roles of the different fibre types in fish myotomes at intermediate speeds, Fish Bull. 76 (1978) 691–699.
[3] I.A. Johnston, W. Davison, G. Goldspink, Energy metabolism of carp swimming muscles, J. Comp. Physiol. 114 (1977) 203–216.

[4] I.A. Johnston, et al., Plasticity of muscle fibre number in seawater stages of Atlantic salmon in response to photoperiod manipulation, J. Exp. Biol. 206 (2003) 3425–3435.
[5] A.H. Weatherley, H.S. Gill, A.F. Lobo, Recruitment and maximal diameter of axial muscle fibres in teleosts and their relationship to somatic growth and ultimate size, J. Fish Biol. 33 (1988) 851–859.
[6] A. Rowlerson, A. Veggetti, Cellular mechanisms of post-embryonic muscle growth in aquaculture species, in: I.A. Johnston (Ed.), Muscle Development and Growth, Academic Press, San Diego, 2000, pp. 103–140.
[7] I.A. Johnston, et al., Rapid evolution of muscle fibre number in post-glacial populations of Arctic charr *Salvelinus alpinus*, J. Exp. Biol. 207 (2004) 4343–4360.
[8] W. van Raamsdonk, W. Mos, M.J. Smit-Onel, W.J. van der Laarse, R. Fehres, The development of the spinal motor column in relation to the myotomal muscle fibers in the zebrafish (*Brachydanio rerio*). I. Posthatching development, Anat. Embryol. (Berlin) 167 (1983) 125–139.
[9] J.M. Fernandes, et al., A genomic approach to reveal novel genes associated with myotube formation in the model teleost, *Takifugu rubripes*, Physiol. Genomics 22 (2005) 327–338.
[10] R.S. Kamath, et al., Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi, Nature 421 (2003) 231–237.
[11] F. Simmer, et al., Genome-wide RNAi of C. elegans using the hypersensitive rrf-3 strain reveals novel gene functions, PLoS Biol. 1 (2003) E12.
[12] S.B. Hedges, et al., A genomic timescale for the origin of eukaryotes, BMC Evol. Biol. 1 (2001) 4.
[13] S.B. Hedges, The origin and evolution of model organisms, Nat. Rev., Genet. 3 (2002) 838–849.
[14] J.E. Blair, P. Shah, S.B. Hedges, Evolutionary sequence analysis of complete eukaryote genomes, BMC Bioinformatics 6 (2005) 53.
[15] W.K. Huh, et al., Global analysis of protein localization in budding yeast, Nature 425 (2003) 686–691.
[16] T.C. Fleischer, C.M. Weaver, K.J. McAfee, J.L. Jennings, A.J. Link, Systematic identification and functional screens of uncharacterized proteins associated with eukaryotic ribosomal complexes, Genes Dev. 20 (2006) 1294–1307.
[17] T. Ito, et al., A comprehensive two-hybrid analysis to explore the yeast protein interactome, Proc. Natl. Acad. Sci. U. S. A. 98 (2001) 4569–4574.
[18] Z. Hu, B. Potthoff, C.P. Hollenberg, M. Ramezani-Rad, Mdy2, a ubiquitin-like (UBL)-domain protein, is required for efficient mating in *Saccharomyces cerevisiae*, J. Cell Sci. 119 (2006) 326–338.
[19] M. Schuldiner, et al., Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile, Cell 123 (2005) 507–519.
[20] J. Shen, C.M. Hsu, B.K. Kang, B.P. Rosen, H. Bhattacharjee, The *Saccharomyces cerevisiae* Arr4p is involved in metal and heat tolerance, Biometals 16 (2003) 369–378.
[21] G. Giaever, et al., Functional profiling of the *Saccharomyces cerevisiae* genome, Nature 418 (2002) 387–391.
[22] F. Delsuc, H. Brinkmann, D. Chourrout, H. Philippe, Tunicates and not cephalochordates are the closest living relatives of vertebrates, Nature 439 (2006) 965–968.
[23] D.V. Babushok, E.M. Ostertag, H.H. Kazazian Jr., Current topics in genome evolution: molecular mechanisms of new gene formation, Cell. Mol. Life Sci. 64 (2007) 542–554.
[24] A. Meyer, M. Schartl, Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions, Curr. Opin. Cell Biol. 11 (1999) 699–704.
[25] B.J. Evans, D.B. Kelley, R.C. Tinsley, D.J. Melnick, D.C. Cannatella, A mitochondrial DNA phylogeny of African clawed frogs: phylogeography and implications for polyploid evolution, Mol. Phylogenet. Evol. 33 (2004) 197–213.
[26] R.D. Morin, et al., Sequencing and analysis of 10,967 full-length cDNA clones from *Xenopus laevis* and *Xenopus tropicalis* reveals post-tetraploidization transcriptome remodeling, Genome Res. 16 (2006) 796–803.
[27] D. Steinke, W. Salzburger, A. Meyer, Novel relationships among ten fish model species revealed based on a phylogenomic analysis using ESTs, J. Mol. Evol. 62 (2006) 772–784.
[28] Z. Yang, R. Nielsen, Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models, Mol. Biol. Evol. 17 (2000) 32–43.
[29] J. Zhang, S. Kumar, M. Nei, Small-sample tests of episodic adaptive evolution: a case study of primate lysozymes, Mol. Biol. Evol. 14 (1997) 1335–1338.
[30] D. Lang, S.K. Powell, R.S. Plummer, K.P. Young, B.A. Ruggeri, PAX genes: roles in development, pathophysiology, and cancer, Biochem. Pharmacol. 73 (2007) 1–14.
[31] G.E. Hollway, et al., Whole-somite rotation generates muscle progenitor cell compartments in the developing zebrafish embryo, Dev. Cell 12 (2007) 207–219.
[32] F. Stellabotte, B. Dobbs-McAuliffe, D.A. Fernandez, X. Feng, S.H. Devoto, Dynamic somite cell rearrangements lead to distinct waves of myotome growth, Development 134 (2007) 1253–1257.
[33] D.J. Macqueen, D. Robb, I.A. Johnston, Temperature influences the coordinated expression of myogenic regulatory factors during embryonic myogenesis in Atlantic salmon (*Salmo salar* L.), J. Exp. Biol. 210 (2007) 2781–2794.
[34] S.H. Devoto, et al., Generality of vertebrate developmental patterns: evidence for a dermomyotome in fish, Evol. Dev. 8 (2006) 101–110.
[35] E.M. Gertz, Y.K. Yu, R. Agarwala, A.A. Schaffer, S.F. Altschul, Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST, BMC Biol. 4 (2006) 41.
[36] J.M. Fernandes, J.R. Kinghorn, I.A. Johnston, Differential regulation of multiple alternatively spliced transcripts of MyoD, Gene 391 (2007) 178–185.
[37] T.A. Hall, BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT, Nucleic Acids Symp. Ser. 41 (1999) 95–98.
[38] I. Ovcharenko, et al., Mulan: multiple-sequence local alignment and visualization for studying function and evolution, Genome Res. 15 (2005) 184–194.
[39] F. Ronquist, J.P. Huelsenbeck, MrBayes 3: Bayesian phylogenetic inference under mixed models, Bioinformatics 19 (2003) 1572–1574.
[40] S. Guindon, F. Lethiec, P. Duroux, O. Gascuel, PHYML online—a Web server for fast maximum likelihood-based phylogenetic inference, Nucleic Acids Res. 33 (2005) W557–W559.
[41] S. Kumar, K. Tamura, I.B. Jakobsen, M. Nei, MEGA2: molecular evolutionary genetics analysis software, Bioinformatics 17 (2001) 1244–1245.
[42] M. Suyama, D. Torrents, P. Bork, PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments, Nucleic Acids Res. 34 (2006) W609–W612.
[43] B. Korber, HIV Sequence Signatures and Similarities, in: A.G. Rodrigo, G.H. Learn (Eds.), Computational and Evolutionary Analysis of HIV Molecular Sequences, Kluwer Academic, Dordrecht, 2000, pp. 55–72.
[44] M. Nei, T. Gojobori, Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions, Mol. Biol. Evol. 3 (1986) 418–426.
[45] T. Jowett, Double in situ hybridization techniques in zebrafish, Methods 23 (2001) 345–358.