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.

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