The structure, splicing, synteny and expression of lamprey COE genes and the evolution of the COE gene family in chordates

Ricardo Lara-Ramírez[1,2], Guillaume Poncelet[1], Cédric Patthey[1,3], Sebastian M. Shimeld[1]*
[1]Department of Zoology, University of Oxford, South Parks Road, Oxford, OX1 3PS, United Kingdom
[2] Present address: Centro de Investigación en Ciencias Biológicas Aplicadas. Instituto Literario No. 100 Colonia Centro. CP 50000, Mexico
[3] Present address: Umeå Center for Molecular Medicine, Umeå University, Umeå, Sweden

*Corresponding author: sebastian.shimeld@zoo.ox.ac.uk

Abstract
COE genes encode transcription factors that have been found in all metazoans examined to date. They possess a distinctive domain structure that includes a DNA-binding domain (DBD), an IPT/TIG domain, and a helix-loop-helix (HLH) domain. An intriguing feature of the *COE* HLH domain is that in jawed vertebrates it is composed of three helices, compared to two in invertebrates. We report the isolation and expression of two *COE* genes from the brook lamprey *Lampetra planeri,* and compare these to *COE* genes from the lampreys *Lethenteron japonicum* and *Petromyzon marinus*. Molecular phylogenetic analyses do not resolve the relationship of lamprey *COE* genes to jawed vertebrate paralogues, though synteny mapping shows they all derive from duplication of a common ancestral genomic region. All lamprey genes encode conserved DBD, IPT/TIG and HLH domains, however the HLH domain of lamprey *COE-A* genes encodes only two helices while *COE-B* encodes three helices. We also identified *COE-B* splice variants encoding either two or three helices in the HLH domain, along with other *COE-A* and *COE-B* splice variants affecting the DBD and C-terminal transactivation regions. In situ hybridisation revealed expression in the lamprey nervous system including the brain, spinal cord and cranial sensory ganglia. We also detected expression of both genes in mesenchyme in the pharyngeal arches, and underlying the notochord. This allows us to establish the primitive vertebrate expression pattern for *COE* genes, and compare this to that of invertebrate chordates and other animals to develop a model for *COE* gene evolution in the chordates.

Keywords: COE, Ebf, neurogenesis, lamprey, cranial ganglia, pharyngeal arch, brain.

Introduction

Lampreys are jawless vertebrates. Together with hagfishes they form the cyclostomes, a lineage that separated early in vertebrate evolution from the lineage that gave rise to the jawed vertebrates (gnathostomes). While lampreys have core vertebrate features such as a dorsal tubular nervous system, neural crest cells, placode-derived sensory ganglia, and a cranial and axial skeleton, they lack gnathostome characters such as hinged jaws and paired appendages. As such, they have been an important model system for understanding the early morphological evolution of vertebrates (Shimeld and Donoghue 2012). The expression of genes involved in developmental processes has been a critical piece of evidence in such studies, allowing insight into the evolution of new characters.

The placement of lampreys relative to the two rounds of genome duplication (2R) thought to have occurred early in vertebrate evolution (Putnam et al. 2008) is also important for understanding vertebrate character evolution. Jawed vertebrate genomes are characterised by large paralagous blocks of genes deriving from duplications, each traceable to a small number of ancestral linkage groups in a reconstructed pre-duplication ancestor (Nakatani et al. 2007). Consequently, many gene families comprise multiple paralagous genes in jawed vertebrates but a single copy gene in the vertebrates' nearest relatives, amphioxus and urochordates (Putnam et al. 2008). However, it is currently unclear whether lampreys share the 2R duplications. While lamprey genomes have clearly undergone genome duplication, such that multiple paralogues of many gene families are found and leading some authors to suggest both genome duplications are shared (Smith et al. 2013), in molecular phylogenetic analyses lamprey genes rarely group robustly with jawed vertebrate paralogue groups (Kuraku et al. 2009). This raises the possibility that lampreys and jawed vertebrates might have undergone parallel genome duplication, or, as a recent study based on an improved genome map suggests, there may have been only one genome duplication coupled with a number of segmental duplications (Smith and Keinath 2015). Furthermore, some or all of these duplications may have occurred at a similar time to lineage separation, such that gene relationships are obscured and/or the return from tetraploidy to diploidy happened independently in the two lineages (Furlong and Holland 2002). These uncertainties mean evolutionary comparisons involving lamprey genes need to be made at the level of gene families rather than individual orthologues.

The *COE* genes (also known as *Ebf* genes) are a family of HLH transcription factors that are involved in many developmental aspects of the vertebrate nervous system. Their restricted expression reflects important aspects of the structure of the nervous system in different vertebrate species. For example, in mice, *COE* genes highlight the regionalisation of the brain and spinal cord, particularly marking post-mitotic neurons. In the spinal cord they are expressed ventrolaterally in a region

corresponding to motor neurons (Garel et al. 1997). In the PNS, mouse *COE* genes are expressed in the olfactory epithelium, vomeronasal organ, trigeminal and glossopharyngeal cranial ganglia, inner ear, dorsal root ganglia (DRG) and peripheral glia (Corradi et al. 2003; Malgaretti et al. 1997; Wang et al. 1997). Also, in keeping with their expression sites in the nervous system, COE proteins have been shown to induce neuronal differentiation and associated cell cycle exit (Garcia-Dominguez et al. 2003). Altogether, the expression of *COE* genes aids both the identification of areas of neuronal differentiation and the assessment of neuronal cell types in the nervous system of species with less well described anatomy and developmental processes.

Apart from the nervous system, vertebrate *COE* genes are also expressed in mesodermal structures during development and in mesodermal derivatives at late developmental stages and in adulthood (Hesslein et al. 2009; Jimenez et al. 2007; Kieslinger et al. 2005). *COE* genes are expressed in somites (El-Magd et al. 2014a; El-Magd et al. 2014b; Green and Vetter 2011) and have an important role in the commitment and differentiation of muscle cells (Green and Vetter 2011; Jin et al. 2014). *COE* genes are also expressed in the mesodermal and neural crest components of pharyngeal arches at embryonic stages (Dubois et al. 1998; El-Magd et al. 2014b; Jin et al. 2014; Kieslinger et al. 2005; Pozzoli et al. 2001). In the lymphocyte lineage, *Ebf1*/*COE1* participates in the specification of B-cell lymphocytes (Hagman et al. 1995; Lin and Grosschedl 1995; Treiber et al. 2010). Also, *COE* genes seem to be involved in the regulation of adipocyte and osteoblast commitment and differentiation (Akerblad et al. 2002; Akerblad et al. 2005; Hesslein et al. 2009; Kieslinger et al. 2005) and in the specification of brown versus white adipocyte identity (Rajakumari et al. 2013).

While *COE* genes encode an HLH domain with a typical Helix1(H1)-Linker(L)-Helix2(H2) structure, they lack the DNA-binding basic region typical of most HLH proteins. Instead, they possess a large DNA-binding domain (DBD) which includes an atypical Zn-finger (Fields et al. 2008; Hagman et al. 1995). Between the DBD and HLH domains, an IPT/TIG (Immunoglobulin-like, Plexins, Transcription-factors/Transcription factor Immunoglobulin) domain is also present and has been suggested to be involved in protein-protein interactions, dimerization, and even DNA binding (Siponen et al. 2010; Treiber et al. 2010). In addition, COE proteins have a transactivation (TA) domain at the carboxy terminus (Hagman et al. 1995). *COE* genes have been found in a wide variety of metazoans (Crozatier and Vincent 1999; Demilly et al. 2011; Jackson et al. 2010; Mazet et al. 2004; Pang et al. 2004), with presence in ctenophores and sponges showing they date from at least the last common ancestor of animals (Daburon et al. 2008; Jackson et al. 2010; Pang et al. 2004). Invertebrates, including cephalochordates and tunicates, generally have a single *COE* gene (Dubois and Vincent 2001; Jackson et al. 2010; Mazet et al. 2004; Pang et al. 2004).

Invertebrate and jawed vertebrate *COE* genes have another key difference: all invertebrate *COE* genes analysed to date have a typical H1-L-H2 organisation in the HLH domain, while all jawed vertebrate *COE* genes encode a duplicated H2 (H2d) such that the organisation of their HLH domain is H1-L-H2d-H2a (Crozatier et al. 1996; Dubois and Vincent 2001; Mazet et al. 2004; Pang et al. 2004). Daburon and colleagues characterised two *COE* genes, which they named *COE-A* and *COE-B*, from the lamprey *Petromyzon marinus* (Daburon et al. 2008). *COE-A* appeared to lack the duplicated H2 (H2d), possessing only one H2 (H2a), though this was not definitive due to poor quality genome data in this region. *COE-B*, however, had the duplicated H2. Furthermore, Expressed Sequence Tag (EST) data from a second lamprey species, *Lampetra fluviatilis*, showed *COE-B* transcripts are alternately spliced, such that transcripts could have the structure H1-L-H2d-H2a (as in jawed vertebrates), or H1-L-H2a (as in invertebrates). These studies present an intriguing picture of *COE* family evolution in chordates, but leave unanswered questions; for example it is unclear when the H2 duplication occurred, relative to the timing of genome duplications, and how this relates to gene orthology.

Even though significant progress has been made in the understanding of *COE* gene expression and function in jawed vertebrates, particularly in mice, and new information of *COE* gene expression and function in several metazoans is emerging (Jackson et al. 2010; Pang et al. 2004), there is still a gap in understanding of *COE* gene expression and function at the invertebrate-vertebrate transition. *COE* gene expression in the cephalochordate *Branchiostoma floridae* has been described in detail, and shown to be expressed in scattered cells throughout the brain and nerve cord, as well as in the ventral part of the anterior somites and in scattered ectodermal cells presumed by these authors and others to be peripheral epidermal sensory neurons (Mazet et al. 2004; Schubert et al. 2004). In urochordates, *COE* expression in the ascidian *Ciona intestinalis* is first detected at the gastrula stage in the A9.32 cell pair (Imai et al. 2004), which give rise to central nervous system cells. By the neurula and tailbud stages *COE* expression is more widespread in the central nervous system, as well as in palp cells that may be sensory neurons (Mazet et al. 2005). *C. intestinalis COE* is also expressed in mesodermal cells, specifically in the pharyngeal muscle lineage (Razy-Krajka et al. 2014). When compared to studies of other invertebrate taxa, these data support a primitive role for *COE* genes in neural differentiation, and a possible ancient role for *COE* genes in mesodermal cells (Jackson et al. 2010; Pang et al. 2004). However, little is known about *COE* gene expression and function in cyclostomes, which can bridge the gap between what is known in invertebrates and jawed vertebrates, and clarify important aspects of expression patterns and gene evolution within the *COE* family.

To gain more insight into *COE* gene evolution and expression at the invertebrate-vertebrate transition, we identified and studied two lamprey *COE* genes in the brook lamprey *Lampetra planeri*. We address the presence/absence of the duplicated H2 in lamprey *COE* genes (Daburon et al., 2008), and their relation to vertebrate paralogy groups. To accomplish this, we used genome data from *P. marinus* and *Lethenteron japonicum* (also known as *Lethenteron camtschaticum*) (Mehta et al. 2013; Smith et al. 2013), as well as transcript data from *L. planeri*, and show that lamprey *COE-B* has the duplicated H2 in all these species, while *COE-A* lacks a duplicated H2. We further explored alternative splicing of both *COE-A* and *COE-B* in *L. planeri* by RT-PCR and transcriptomics, showing alternate splicing of both genes. We also expand previous molecular phylogenetic analyses and couple this with synteny comparisons to develop evolutionary models of *COE* loci in chordates. These analyses show all four gnathosotome *COE* loci and both lamprey *COE* loci share syntenic characters with each other and with the amphioxus *COE* locus. In addition, we describe the expression patterns of both *L. planeri COE* genes during embryogenesis by *in situ* hybridisation, and show widespread expression in central and peripheral nervous systems, as well as in pharyngeal mesenchyme and other mesodermal populations. Our data suggest conservation of *COE* gene function in neuronal differentiation and pharyngeal development in all vertebrate lineages, but highlight differences within the vertebrate lineage such as the absence of expression in dorsal root ganglia and muscle derivatives during lamprey embryonic development with respect to jawed vertebrates.

Materials and Methods

Embryo collection, fixation, in situ hybridisation and gene cloning

We extracted total RNA from stage 24-26 *L. planeri* embryos. For PCR we based primer design on *Petromyzon marinus* genomic information from the Ensembl website. For *LpCOE-A* we used sense primer 5'-CTAGCGCGGGCCCACTTCGAGAA-3' and antisense primer 5'-TGGGAGGCCTCGGACACGCTGAT-3'. For *LpCOE-B* we used sense primer 5'-GAGGGCACACTTTGAGAAGCAGCCA-3' and antisense primer 5'-GCCGGGCTCCGAAACGCTCAC-3'. Cloned sequences have been deposited in Genbank, accession numbers MF539934 (COE-A) and MF539935 (COE-B). *L. planeri* embryos were collected from a shallow river in the New Forest National Park, United Kingdom, with permission from the Forestry Commission. Embryos were brought to the laboratory and placed in Petri dishes with filtered river water from the same river where they were caught. They were kept at 13-15 °C and fixed at different stages of development following the staging system of Tahara (1988). All experiments were performed under local ethical approval. When necessary, embryo chorions were removed with fine forceps before fixation.

Embryos were fixed in 4% phosphate-buffered saline (PBS)-buffered paraformaldehyde (PFA) pH 7.5. PFA was cooled on ice before use. Embryos were fixed in an approximately 10X excess volume of 4% PFA-PBS with respect to river water at 4 °C overnight or longer. After fixation, embryos were washed twice in diethyl pyrocarbonate (DEPC)-treated 1X PBS for 10 minutes each, and then dehydrated through a graded series of PBS:methanol (25%, 50%, and 75% of methanol in 1X PBS) once for 10 minutes each. Finally, they were washed twice in 100% methanol for 10 minutes each and stored in fresh methanol at -20 °C. In situ hybridisation experiments and histology were carried out as previously described (Lara-Ramirez et al. 2015).

Molecular phylogenetic analysis

Sequence analysis and manipulation were performed using BioEdit (Hall, 1999). Accession numbers for sequences used for phylogenetic analysis are shown in Figure 1. We selected sequences from representatives of major vertebrate groups (mammal: *Homo sapiens*. Lepidosaur: *Anolis carolinensis, Gallus gallus*. Amphibian: *Xenopus tropicalis*. Sarcopteryigian fish: *Latimeria chalumnae*. Spotted gar: *Lepisosteus oculatus*. Teleost: *Danio rerio*. Elasmobranch: *Callorhincus milii, Raja eglantaria*) plus selected invertebrate outgroups. Multiple sequence alignments were carried out using MAFFT software version 6.864b for (Katoh et al. 2002; Katoh and Toh 2008). The parameter for strategy for MAFFT alignments was set as “auto”. All other parameters were as the defaults. This alignment was trimmed using GBLOCKS to regions included in all sequences, and a third alignment which was further adjusted by eye, removing both Helix 2’s from all sequences because of uncertainty over alignment in this region. For phylogenetic tree construction, the Maximum Likelihood (ML) and Bayesian algorithms were used and tree construction was conducted using MEGA version 5.2 (Tamura et al. 2011) for ML, and MrBayes version 3.2 (Ronquist and Huelsenbeck 2003) for Bayesian analyses. For ML we used the Whelan and Goldman (WAG) amino acid substitution matrix (Whelan and Goldman 2001) and 100 bootstrap replicates to obtain support values at each node. Bayesian inference was performed using the Markov chain Monte Carlo method. Two independent Markov chains were run, each with 1 million iterations with default heating parameters. The first 25% of the trees were discarded as burn-in before compiling consensus trees and summary statistics. Posterior probabilities at each internal branch were taken as a measure of statistical support. Both methods produced the same tree topology at key nodes, though with differing support values. Initial analyses included the COE sequence from *C. intestinalis*, and this was consistently placed outside the vertebrate genes as previously reported (Daburon et al. 2008). However in our analyses inclusion of this sequence reduced resolution within the vertebrate genes, so we removed it for subsequent analyses.

Intron-exon organisation and synteny analysis

COE genomic loci were identified in the genomes of *Branchiostoma floridae* (amphioxus), *L. japonicum*, *P. marinus*, *Callorhincus milii* (elephant shark) and *Homo sapiens* by BLAST. In addition we compared *COE* loci in H. sapiens and *L. oculatus* using Genomicus v89.01 (Louis et al. 2015). We chose *H. sapiens* as an extensively annotated vertebrate genome, and *C. milii* as a member of the earliest diverging jawed vertebrate lineage; together they encompass extant jawed vertebrate diversity. Intron-exon structures were extracted from gene models or determined by comparing transcripts to the genomic sequence. To map synteny and paralogy relationships, genes adjacent to the *COE* loci in each species were searched by TBLASTN of their predicted proteins across the other genomes. The top chromosomal or scaffold hits was recorded to predict orthology. Human paralogues were extracted from Ensembl predictions of paralogy.

Transcriptome analysis and RT-PCR

Total embryo RNA from *L. planeri* embryos stages 25, 26 and 27 (Tahara 1988) was pooled and sequenced by Illumina Hi-Seq. Raw data have been deposited in the SRA (Bioproject: PRJNA371458). Reads were assembled using Trinity (Grabherr et al. 2011) to persevere candidate splice variants, and putative *COE* transcripts extracted from the assembly. We also remapped the raw transcriptome data back onto the *COE* gene models to examine read distribution across the splice variants; briefly, all reads mapping to each gene were extracted using BLASTN with default parameters. A kmer search strategy was used to identify individual fragments (derived from the paired-end reads) that crossed exon boundaries, taking into account single nucleotide variations. All fragments matching to the 12 last or 12 first nucleotides of each exon were examined to see whether they corresponded to the canonical splice form, or to potential alternate splice forms. For each putative exon-exon junction, fragments containing at minimum a predicted sequence of 7 nucleotides of the upstream exon followed by 20 nucleotides of the downstream exon, or the converse, were counted as occurrences of the splice junction.

Splice variants predicted from the transcriptome data and represented by >1 fragment were verified by RT-PCR on *L. planeri* RNA deriving from staged lamprey embryos. Primers were designed to span introns to exclude the possibility of genomic DNA contamination confounding the results. Bands were cloned and sequenced, confirming all the splice variants. Primer sequences used in the RT-PCR are: COEA TIG 5'; AATAACTCCAAGCACGGGCG. COEA TIG 3'; CTGATGGCTTTGATGCACGG. COEB ZNF 5'; AGAATCCGGAGATGTGTCGG. COEB ZNF 3'; CGATGGGGTCTCGTTTCTGT. COEB TIG 5'; GACAACTTCTTCGACGGGCT. COEB TIG 3'; AGGGTGACCTCCACCACTC. COEB TA1 5'; GTGAGCGTTTCGGAGCCTG. COEB TA1 3'; GGGACACGCTGCTCGTATT. COEB Helix1 5'; TCTGAACGAGCCCACCATTGACTAC. COEB Helix2a 3'; TGGTGGCGGGGCATGCTGTACAGAGCT.

Results

Molecular phylogenetic and synteny analysis of lamprey *COE-A* and *COE-B*

cDNA fragments encoding *LpCOE-A* and *LpCOE-B* were initially isolated from stage 24-26 *L. planeri* embryo RNA using primers based on *COE* sequences identified in the *P. marinus* genome. To extend these sequences we generated and searched an *L. planeri* transcriptome assembly, allowing us to deduce the whole open reading frame for each gene. Alignment of their predicted amino acid sequences with mouse COE proteins COE1-4 revealed high sequence similarity in the DBD, IPT/TIG and HLH domains (Supplementary file 1). This process also identified splice variants, which are described further below.

We also analysed *COE* sequences from other cyclostomes (Supplementary file 1): specifically *P. marinus COE-A* (Daburon et al. 2008), sequences from another lamprey species (*L. japonicum*: (Mehta et al. 2013)) and sequences from a transcriptome from the hagfish *Eptatretus burgheri* (kindly provided by Juan Pascal Anaya). Molecular phylogenetic analysis of these cyclostome sequences, including jawed vertebrate COE sequences plus invertebrate deuterostome COE sequences as outgroups, placed the cyclostome genes with *COE* genes from jawed vertebrates, though failed to clarify relationships to jawed vertebrate *COE1-4* (Figure 1A). The cyclostome sequences fell into two clades with reasonable support (Figure 1A), which we call *COE-A* and *COE-B* following previous nomenclature (Daburon et al. 2008).

To further examine the relationships between lamprey and other chordate *COE* loci, we examined the synteny surrounding *COE* loci of human, spotted gar, elephant shark, lamprey and amphioxus. Jawed vertebrate *COE1-4* lie in paralagous regions of the genome (Figure 1B, S3), with sufficient similarity in neighbouring genes to conclude the loci evolved by block duplication. For example, human *COE2* and *COE4* are linked to *GNRH* paralogues, while *COE1* and *COE3* are linked to *FoxI* paralogues. Moreover, *COE1-4* were predicted to form a family of whole genome duplication paralogs in a pre-computed whole genome assessment based on synteny (Singh et al. 2015). We found weak evidence for syntenic organisation of these regions with the amphioxus *COE* locus. Genes linked to amphioxus *COE* on scaffold 381 had orthologues on the same chromosome as human *COE* loci, though the genomic distance was relatively large (Figure 1B). Both lamprey *COE* loci also showed evidence of similarity in organisation to jawed vertebrate *COE* loci. For example, both are linked to *GNRH* paralogues, lamprey *COE-B* and elephant shark *COE1* are linked to *CLNT1A*, and lamprey *COE-A* and human *COE4* are linked to *NOP56*. However, there is no clear one-to-one relationship between the lamprey loci and the four jawed vertebrate loci such that orthology can be deduced (Figure 1B). We hence conclude that both jawed vertebrate and lamprey *COE* regions evolved by block duplications from a single ancestral locus

as seen in amphioxus, but we cannot determine whether these are shared duplications, or occurred in parallel.

Alternate splicing of lamprey *COE* genes

Daburon and colleagues (Daburon et al. 2008) previously provided evidence that the *P. marinus COE-B* locus included a duplicated helix 2 as found in all jawed vertebrate *COE* genes. They also showed *COE-B* to be alternately spliced in *L. fluviatilis*, yielding transcripts encoding one or both H2 copies, but failed to identify similar splicing for *COE-A*. Since the presence or absence of the duplicated H2 appears to be a major structural difference between jawed vertebrate and invertebrate *COE* genes, we sought to clarify the structure and splicing of both lamprey *COE* genes.

Schematic intron-exon maps of the two lamprey *COE* loci, inferred by mapping *L. planeri* transcripts to the *L. japonicum* genome, are show in Figure 2A. Exons are numbered sequentially based on *COE-B*, with numbering preserved in the other genes to illustrate exon homology. The intron-exon structure and intron phase class of chordate *COE* genes has been previously shown to be relatively conserved (Daburon et al. 2008), and our analysis confirmed this, with *COE-A* and *COE-B* very similar to each other and to jawed vertebrate *COE* genes, with the exception of the duplicated H2, and some variation in C terminal exon structure (Figure 2A). This confirms the previous report (Daburon et al. 2008) of a duplicated H2 in *COE-B*, and RT-PCR of *L. planeri* demonstrated alternate splicing here in this species (Figure 2, S2). We found no evidence for a similar exon duplication in *COE-A*; a second H2 was not present in either the *L. japonicum* or *P. marinus* genome assemblies, and neither RT-PCR across this region (not shown) or transcriptome analysis (see below) identified alternate splicing in this region for *LpCOE-A*. Neither *COE-A* or *COE-B* hagfish transcripts contained a duplicated H2, though in the absence of a hagfish genome we cannot determine if this is due to splicing or if hagfish lack the duplicated exon.

Amongst the RT-PCR clones generated for *LpCOE-B* were two that encoded truncated ORFs. One contained a small insertion of 14bp in the ZNF of the DBD. This altered the reading frame of the ORF, leading to premature truncation of the predicted protein within the DBD. Comparison of this sequence to the *L. japonicum* genome indicates it derived from splicing at a site located 14bp upstream of the canonical exon 6 splice acceptor site, within the intron between exons 5 and 6 (Figure 2A, S1). The other contained a deletion in the IPT/TIG domain, also altering the ORF and truncating the predicted protein within this domain. This derived from alternate splicing between exons 9 and 10, removing 23bp from exon 10 (Figure S1). To examine whether these were biologically meaningful, and to test for other splice variants, we extracted predicted *COE* transcripts from an *L. planeri* embryo transcriptome assembly (Supplementary file 2). In addition we mined individual fragments (derived from the paired-end reads) from the sequence data used to construct this assembly, identifying and

counting fragments that bridged between exons (Table 1). Both these analyses confirmed the H2 splice variants of *LpCOE-B*. Neither showed evidence of *LpCOE-B* splicing in either the ZNF or IPT/TIG domains, with 18 and 24 fragments respectively reflecting the canonical splice and none reflecting the putative alternate splices discussed above (Table 1). However these analyses did identify other splice products for both *LpCOE-A* and *LpCOE-B* (Table 1, Figure 2A-E). *LpCOE-A* showed a single alternate product, at the boundary between the DBD and IPT/TIG domains. This resulted from use of an alternate splice site in exon 8 which maintained the reading frame, with the two forms represented by 19 and 10 fragments respectively (Table 1; Figure 2A, E). *LpCOE-B* showed multiple alternate products in the C terminal TA domain, all supported by multiple fragments (Table 1; Figure 2A, C, D). One, which we named TA1, results from loss of 12bp at the junction between exons 13 and 14. In addition to this, four alternate variants (jointly named TA2) were detected from splicing of exons 14, 15, 16 and 17, resulting in the loss of varying amounts of sequence from the C-terminal region of the protein, with the 5' end of exon 15 also represented by fragments mapping to two closely spaced acceptor sites (Figure 2D; Table 1). Two additional variants were reflected by single fragments in the raw transcriptome data (Table 1) but not found in the assembly; these were not considered further.

Finally, we sought to test alternate splicing experimentally (with the exception of the variants represented only by single transcriptome fragments, which we did not address further). We designed primers for each and amplified them from staged *L. planeri* embryo RNA, confirming band identity by sequencing (Figure S2). This confirmed all the variants predicted by the transcriptome analysis were genuine, with all amplified from embryo and larval RNA. However, as with the transcriptome, it failed to confirm the variants detected in the ZNF and IPT/TIG domains *of LpCOE-B*, instead only amplifying a single band encoding the canonical ORF from each region (Figure S2). Since these variants were originally identified as clones from an amplification-cloning experiment, they may reflect rare events difficult to detect by RT-PCR or transcriptomics, or PCR amplification artefacts. As their biological relevance is questionable, we have not sought to investigate this further.

In summary, analysis of splice variants and intron-exon structures in multiple lamprey species indicate that the two lamprey *COE* loci are structurally distinct. *COE-B* has a duplicated H2, and alternately splices this to produce jawed vertebrate-like and invertebrate-like *COE* transcripts. *COE-A* encodes only one H2, and has no alternate splicing in this region. In *L. planeri* both genes are spliced at other points, either by skipping exons, or by use of closely-spaced alternate splice sites near intron-exon boundaries resulting in the loss/gain of small numbers of amino acids.

Expression of *LpCOE-A* in *L. planeri* embryos

Expression of *LpCOE-A* was analysed from stages 21 to 28. At stage 21, expression is seen in the developing mandibular arch as well as in a domain under the notochord (Figure 3A). At stage 22,

expression appears in the nascent second and third pharyngeal arches (Figure 3B). At stage 23, transcripts are seen in the spinal cord and ventral regions of the diencephalon, midbrain and hindbrain, in the trigeminal, geniculate and posterior lateral line placode/ganglia, as well as in the developing fourth pharyngeal arch and in an extending domain below the notochord (Figure 3C, D). At this stage, faint expression starts to be seen in the nasohypophyseal plate (NHP) (Figure 3D, arrowhead). At stage 24, *LpCOE-A* expression is seen as two parallel lateral stripes from a dorsal view. This expression is found in the mantle layer of the neural tube and it extends from the diencephalon posteriorly all along the growing spinal cord (Figure 3E). Expression progresses dorsally in the diencephalon, midbrain and hindbrain with respect to stage 23, leaving a gap of expression at the level of the midbrain-hindbrain boundary (MHB) (Figure 3F, G). Expression also increases in the NHP and ophthalmic, maxillomandibular and posterior lateral line placodes/ganglia, and also appears in the developing fifth pharyngeal arch (Figure 3F, G).

At stage 25, expression expands in most of the CNS, running the entire length of the spinal cord (Figure 4A). In the brain, expression covers the hindbrain, midbrain and most of the forebrain, except its ventral-most region which possibly corresponds to the hypothalamus (Figure 4D). At this stage, transcripts are now present in the telencephalon (Figure 4D). Faint expression is also seen in the epiphysis. In the hindbrain, a big patch of expression is seen at the level of r3-r5 dorsally (Figure 4D, asterisk). Notably, the MHB remains unstained. Outside the brain, expression increases in the forming olfactory epithelium and trigeminal, geniculate and petrosal ganglia (Figure 4D). At this stage, transcripts are also seen in the eight pharyngeal arches and in the upper lip (Figure 4D). From a ventral view, expression is distinguished in the mesoderm –and possibly neural crest– of each pharyngeal arch, but is clearly excluded from the endoderm and epidermis (Figure 4F). From a dorsal view, staining is seen as two lateral stripes along the neural tube as in stage 24 embryos (Figure 4E). Expression under the notochord also progresses posteriorly as the embryo elongates (Figure 4A, I, arrows). Cross sections of stage 25 embryos reveal strong staining on the lateral sides of the neural tube and a thin, unstained domain in the middle demarcating the mantle layer and the ventricular zone, respectively (Figure 4B, C, G, H). At the level of the otic vesicle, there is a gap of expression in the middle of the mantle layer with respect to the dorso-ventral axis (Figure 4C, white asterisks). Interestingly, there is also a relatively wide unstained area in the dorsal-most part of the brain (Figure 4B, C, black asterisks), whereas in the forming spinal cord it is mostly seen in the ventral-most part (Figure 4H, black asterisk). Outside the nervous system, expression is seen in mesenchyme of the pharyngeal arches (Figure 4B, F). Expression is also seen surrounding roughly the ventral half of otic vesicles (Figure 4C, arrows). Medial expression in this domain possibly corresponds to the acoustic ganglion. In the body, transcripts are found in mesenchyme ventro-lateral to the notochord and in a more lateral domain

just above the vitellum (Figure 4G, black and white arrows, respectively). This expression is not observed more posteriorly (Figure 4H).

At stage 26, *LpCOE-A* expression diminishes both in the head and spinal cord (Figure 4J). In the head, reduction of expression is seen in the ventral midbrain, dorsal diencephalon and olfactory epithelium (Figure 4K). *LpCOE-A* is no longer observed in the epiphysis. A marked patch of expression is maintained in the dorsal hindbrain at the level of the otic vesicle (Figure 4K, white asterisk). Expression in pharyngeal arches remains relatively strong and staining is now seen in a forming ninth pharyngeal arch depicting eight pharyngeal pouches (Figure 4K). From a dorsal view, expression is restricted to the lateral side of the brain and spinal cord delimiting the extent of the ventricular zone in the middle (Figure 4L). In the brain, the ventricular zone expands at the level of the epiphysis, the MHB, and posterior hindbrain (Figure 4L, white asterisks). Expression is still seen in trigeminal and geniculate ganglia. In the spinal cord expression is considerably downregulated with respect to previous stages and expression under the notochord disappears (Figure 4M, arrows). At stage 27, expression diminishes even more with respect to stage 26 becoming more restricted, although it maintains the same general expression pattern (data not shown). At stage 28, expression is more reduced and confined to the head (Figure 4N). In the brain the same pattern is maintained overall, with transcripts still observed in the telencephalon, very faintly in the dorsal diencephalon, dorsal midbrain, and restricted regions of the hindbrain (Figure 4O). The trigeminal, geniculate and petrosal ganglia are still stained, and in the olfactory epithelium transcripts are observed in the ventral half of the epithelium (Figure 4O). Expression in all pharyngeal arches is maintained relatively strong along their dorso-ventral axes (Figure 4O).

Expression of *LpCOE-B* in *L. planeri* embryos

Expression of *LpCOE-B* was analysed from stages 21 to 28. At stage 21, weak expression is observed in the anterior spinal cord as distinct, widely-spaced dots forming two lateral stripes as seen from a dorsal view (Figure 5A, B, arrows). Later at stage 22, this expression pattern increases in intensity and extends posteriorly along the growing spinal cord (Figure 5C). Expression is still seen as widely-spaced dots from a dorsal view forming two parallel stripes of staining (Figure 5D). At this stage expression is also observed in the trigeminal (maxillomandibular) placodes and presumptive NHP (Figure 5C, black and white arrowheads, respectively). Expression in the presumptive NHP is observed as two relatively large lateral dots in its posterior facet, which weakly extend towards the anterior until they progressively meet in the middle (Figure 5C, inset). At stage 23, expression is maintained in the trigeminal placode and it now appears in the developing geniculate and petrosal placodes (Figure 5E). At this stage expression persists in the spinal cord. At stage 24, *LpCOE-B* appears in the hindbrain, midbrain and diencephalon (Figure 5F). Expression intensifies in the forming maxillomandibular,

geniculate and petrosal ganglia, as well as in the NHP (Figure 5F, H). Expression in the spinal cord is considerably upregulated with respect to earlier stages and is observed roughly in the ventral half (Figure 5G), preserving its position on the lateral sides of the neural tube as seen from a dorsal view (Figure 5H). Outside the nervous system, strong expression is seen in mesenchyme of the first pharyngeal arch and in an extending domain above the pharynx (Figure 5F, G).

At stage 25, *LpCOE-B* is strongly expressed in restricted regions of the diencephalon, midbrain and hindbrain, as well as in trigeminal, geniculate and petrosal ganglia and olfactory epithelium (Figure 6A). In the hindbrain, transcripts are mostly observed dorsally in an anterior and a strong posterior patch (Figure 6A, asterisks), with diffuse staining in between. Faint expression is observed in the epiphysis (Figure 6A, black arrowhead). Notably, *LpCOE-B* is not expressed in the telencephalon at this stage. Expression in the spinal cord becomes very restricted to the dorsal side and is continuous with the posterior patch of expression in the hindbrain (Figure 6B). *LpCOE-B* is not observed under the notochord in the trunk region as with *LpCOE-A* (Figure 6B, arrows; compare with Figure 4I, arrows). At this stage, transcripts are seen in the five anterior-most pharyngeal arches with strongest staining in the first arch (Figure 6A). Like *LpCOE-A*, an expression domain between the notochord and pharyngeal arches is also present, although it does not extend as far to the posterior (Figure 6A, arrow). From a dorsal view, staining is clearly restricted to the lateral sides of the neural tube similar to *LpCOE-A*, delineating the ventricular zone in the middle (Figure 6C).

At stage 26, expression dramatically increases in the head and expression in the spinal cord is maintained dorsally (Figure 6D). In the brain expression is more refined, and transcripts are now detected in the telencephalon (Figure 6F, white arrow). In the diencephalon, mild staining is observed in the epiphysis (Figure 6F, black arrowhead) and in a small region next to the telencephalon and two more dorsal domains next to the midbrain (Figure 6F, black asterisks). In the midbrain, two large expression domains, one dorsal and one ventral, are separated by an unstained region (Figure 6F, white open arrowheads). In the hindbrain, two big patches of staining are seen on the dorsal side, one at the level of rhombomeres 2-4 and another one at the transition with the spinal cord (Figure 6F, white asterisks). Expression in the olfactory epithelium remains strong and signals are now detected in the upper and lower lips (Figure 6F). Eight pharyngeal arches are now stained all along their dorsoventral axes (Figure 6F). The spinal cord maintains its dorsal expression domain although it is observed slightly weaker, and no expression is observed below the notochord in the spinal cord region (Figure 6G). From a dorsal view, staining is observed at the lateral sides of the brain and spinal cord demarcating the ventricular zone in the middle (Figure 6E, H). Similar to *LpCOE-A*, in the brain the unstained medial region follows expansions of the ventricular zone at the level of the epiphysis, MHB and posterior hindbrain (Figure 6E, white asterisks). Note, however, that at stage 25 these expansions

are not so evident (Figure 6C). Cross sections at stage 26 confirm the lateral expression in the neural tube, generally stronger dorsally (Figure 6I-K, white arrows). Cross sections also reveal expression in maxillomandibular ganglia (Figure 6I) as well as in mesenchyme of the pharyngeal arches and developing vellum (Figure 6I, arrowheads and asterisks, respectively). A wide expression domain is observed at the dorsal edge of each pharyngeal arch (Figure 6J, asterisks), which is contiguous with a line of stained cells located between the neural tube and myotomes, passing lateral to the notochord (Figure 6J, black arrows). More posteriorly, expression is only observed in the spinal cord (Figure 6K). At stage 27, expression clearly diminishes in the head and in the spinal cord it is virtually absent (Figure 6L, M). Expression in the diencephalon disappears almost completely (Figure 6L). The two big patches of staining in the dorsal hindbrain persist although at a much lower level, and a third small patch is distinguished between them (Figure 6L, asterisks). Expression remains strong in the olfactory epithelium, upper lip, and mandibular arch, in dorsal domains of each pharyngeal arch and in the trigeminal, geniculate and petrosal ganglia. Also, expression in pharyngeal arches starts to fade, mostly ventrally, and expression disappears in the eighth pharyngeal arch (Figure 6L). At stage 28, expression diminishes even more in the head but the same expression sites are maintained, with strongest expression in the olfactory epithelium, upper lip, cranial ganglia and dorsal and ventral domains of each pharyngeal arch (Figure 6N).

Discussion

The evolution of *COE* gene structure, splicing and duplication in vertebrates

Gene relationships between lampreys and jawed vertebrates have often proven difficult to decipher. While lamprey genome sequences provide evidence for at least one and possibly two genome duplication (Mehta et al. 2013; Smith and Keinath 2015; Smith et al. 2013), analysis of individual gene families often fails to clarify whether these are shared with the genome duplications of jawed vertebrates (Kuraku et al. 2009). Daburon and colleagues original analysis of *COE* H2d evolution showed at least one lamprey *COE* gene had this duplicated exon, but poor genome quality prevented them from determining the status of the second *COE* gene (Daburon et al. 2008).

Combining molecular phylogenetics, synteny, intron-exon maps and splice form analysis helped us further explore this. First, molecular phylogenetics suggest cyclostomes have only two COE genes: our data includes three lamprey species, plus one hagfish, and all these sequences clearly fall into two orthologue groups. These genes do not appear as orthologues to specific jawed vertebrate *COE* genes in this analysis, and while synteny shows all these copies have derived by duplications of large gene blocks consistent with genome duplication, they too do not determine whether these are shared or evolved in parallel. However, the duplicated H2d is shared between lamprey *COE-B* and

jawed vertebrate *COE* genes. Unless we consider this change has evolved in parallel, the exon duplication that formed H2d must have occurred prior to the separation of the two lineages, and prior to the duplications that formed *COE1-4* and *COE-A* and *COE-B*. This model, detailed in Figure 7, implies that lamprey *COE-A* has lost the duplicated H2 exon and reverted to an invertebrate-like state. Intron phasing across these exons (Daburon et al. 2008), plus the presence in *COE-B* genes of an alternate splice product lacking H2d, indicate this loss could occur relatively easily while still preserving a functional ORF. The functional implications of having two versus three helices are not understood, though since the HLH region is involved in dimerization, COE protein heterodimers have been reported (Wang et al. 2002), and dimerization between two and three helix versions of COE proteins appears feasible (Daburon et al. 2008), we can speculate that *COE-B* H2d splicing in lampreys allows for a wider array of dimers.

We also observed other splice variants for both *LpCOE-A* and *LpCOE-B*. Splicing in the C-terminal TA domain of *LpCOE-B* resulted primarily from use of different exons, including what appears to be a new exon (exon 16 in Figure 2A) which lacks an equivalent in either lamprey *COE-A* genes or jawed vertebrate *COE* genes. The sequence encoded by this exon was also absent from the hagfish *COE-B* sequence, though without a genome sequence this is not conclusive. C-terminal splicing of jawed vertebrate *COE* genes has been little-studied, having only been experimentally verified for mouse *COE-4* (Wang et al. 2002). Other experimentally-validated lamprey splice products resulted from use of alternative splice donor or acceptor sites, resulting in the addition or loss of small blocks of sequence. These splice products have no described counterparts in jawed vertebrates, so likely represent lamprey innovations, as depicted in Figure 7.

***COE* expression in the lamprey CNS and lineage specific expression domains**

The expression of *LpCOE-A* and *LpCOE-B* in the lamprey nervous system is in agreement with a conserved role of *COE* genes in neuronal differentiation. From the early stages of development, when expression of *LpCOE-A* and *LpCOE-B* appear in the rhombospinal region, both transcripts are observed at the lateral margins of the neural tube. This is similar to the expression of *COE* genes in postmitotic neurons in the spinal cord of mouse, chicken, *Xenopus* and zebrafish (Bally-Cuif et al. 1998; Dubois et al. 1998; Garel et al. 1997). Expression of both *LpCOE-A* and *LpCOE-B* in the rhombospinal region appears complementary to that of *LpNgnA* (Lara-Ramirez et al. 2015), which is restricted to the ventricular zone of the neural tube.

In mouse and chicken, expression of *COE* genes is observed in the motor column at the ventrolateral margin of the developing spinal cord, and in a thinner domain that extends dorsally at the level of the subventricular zone (El-Magd et al. 2014b; Garel et al. 1997). In the lamprey, however, expression of *LpCOE-A* is proportionately more extended, covering nearly all the mantle layer, and no subventricular

zone-like domain of expression is observed with either lamprey *COE* gene. This expression could reflect an anatomical difference between lamprey and gnathostome spinal cords, or a gene regulation difference between lamprey and gnathostome *COE* genes. For example, given that *COE* genes are strongly expressed in the ventrolateral motor column of gnathostome spinal cords, it is possible that most of the lamprey mantle layer marked by *COE* genes corresponds to populations of motor neurons, which are proportionately more extended in the lamprey spinal cord than in jawed vertebrates. Alternatively, it is possible that the lamprey spinal cord also possesses a subventricular zone as in gnathostomes, but the expression of *COE* genes is activated in a bigger cell repertoire that also includes different interneuron subpopulations. In either case, a more detailed characterisation of neuronal populations of the developing lamprey spinal cord is needed to clarify the particular cell types where lamprey *COE* genes are expressed.

In the hindbrain, both lamprey *COE* genes may mark branchiomotor nuclei, which are morphologically discerned from stage 26 (Murakami et al. 2004). In the mouse, *COE1-3* are strongly expressed in facial branchiomotor neurons of r4, although at different time points (Garel et al. 2000). We also observed a clear distinction in hindbrain expression between lampreys and jawed vertebrates. In zebrafish and mouse, *COE* genes are first expressed in r2 and r4, leaving gaps of expression in r1, r3 and r5-7, although later in development they are activated in these remaining rhombomeres. In the lamprey, however, no such r2/r4 initial expression was observed. This is in keeping with the spatial distribution of branchiomotor nuclei with respect to rhombomeric boundaries, since, in lampreys, branchiomotor neurons are not in register with rhombomeres as in gnathostomes (Murakami et al. 2004). We also did not observe expression of either lamprey *COE* gene in rhombomeric boundaries as with mouse *COE* genes (Garel et al. 1997). These comparisons support the view that lampreys, while having a hindbrain that is fundamentally similar to gnathostomes in terms of broad rhombomeric organisation, differ with respect to the precise organisation of the cells that differentiate in each (Parker et al. 2016).

According to previous interpretations of the regionalisation of the lamprey diencephalon at stage 26, a time at which the embryonic lamprey brain acquires a well-defined compartmentalisation (Lara-Ramirez et al. 2015; Murakami et al. 2002; Murakami et al. 2001), both lamprey genes seem to be expressed in the alar (pretectum) and basal plates of prosomere (P) 1, but only *LpCOE-A* is expressed in P2 in the dorsal thalamus, except for a very small patch of *LpCOE-B* expression in the epiphysis. The presence or absence of a P3 territory in the embryonic lamprey brain is not resolved by either lamprey *COE* gene. We also note other differences in expression: (i) The region of the embryonic lamprey hypothalamus has been clearly defined previously by expression of *TTF-1/Nkx2.1* (Osorio et al. 2005), and both lamprey *COE* genes seem to be absent from the hypothalamus, which is in sharp contrast

with mouse *COE1-3* which are strongly expressed in this region (Garel et al. 1997). (ii) We did not detect specific expression of either lamprey *COE* gene in the midbrain-hindbrain boundary (MHB) as with zebrafish *COE2* and *COE3* (Li et al. 2010). (iii) We did not observe lamprey *COE* expression in the ventricular zone as occurs with mouse *COE* genes in the anterior hindbrain (Garel et al. 1997). (iv) We did not observe expression of either *LpCOE-A* or *LpCOE-B* in the lamprey retina, while retinal expression is seen with *Xenopus COE3* and mouse *COE1-4* (Garel et al. 1997; Pozzoli et al. 2001; Wang et al. 1997).

In the telencephalon, tetrapod *COE1* genes specifically mark the striatum of the lateral ganglionic eminence (LGE) and a cell corridor passing through the medial ganglionic eminence (MGE) to the ventral thalamus (Bielle et al. 2011; Lopez-Bendito et al. 2006). In the lamprey, we observe a thin domain of expression running continuously from the telencephalon to the ventral diencephalon close to the hypothalamus. The lamprey telencephalon has been shown to be divided into pallium and subpallium, and the lamprey subpallium has been proposed to be equivalent to the LGE only, with the MGE being a gnathostome innovation (Sugahara et al. 2011). However a recent reanalysis of this question including study of the hagfish embryonic brain came to the alternative view, that the MGE is primitive and present in hagfishes and lampreys (Sugahara et al. 2016). Our data are consistent with this later interpretation of the lamprey telencephalon, and suggest a conserved thalamo-striatal connection marked by *COE* expression in the vertebrate MGE.

COE expression in the vertebrate PNS

In the peripheral nervous system, both lamprey *COE* genes were observed in the olfactory placode/epithelium as well as in the placode-derived cranial ganglia. Jawed vertebrate olfactory epithelia and cranial ganglia also express *COE* genes, though there is variation in paralogue group(s) depending on the species examined (Bally-Cuif et al. 1998; Dubois et al. 1998; El-Magd et al. 2014b; Pozzoli et al. 2001; Wang and Reed 1993; Wang et al. 2002; Wang et al. 1997). The only exception is in lateral line placodes/ganglia where *LpCOE- A* was expressed, while to our knowledge no other vertebrate *COE* gene has been reported to have lateral line expression. However lateral line ganglia have not been well-studied as they are absent from some model species, so this probably just reflects missing data. Overall this likely reflects subfunctionalisation of *COE* expression, and suggests cranial placode/ganglia expression preceded gene duplication. *COE* genes also show PNS expression in amphioxus and *Ciona* in cells postulated to be placode homologues (Mazet et al. 2005; Mazet et al. 2004), showing this is a chordate-wide character.

No expression of either gene was observed in lamprey DRG, a prominent expression site for mouse (Davis and Reed 1996; Wang et al. 1997) and chicken *COE* genes (El-Magd et al. 2014b). Lamprey embryonic DRG are visible at the stages examined, and have been shown to express another bHLH

gene, *LpNgnA*, at these stages (Lara-Ramirez et al. 2015). Since *COE* genes elsewhere mark differentiating neurons, this may reflect a delay in the differentiation of DRG neurons in lampreys compared to other vertebrates.

***COE* genes in mesodermal tissues**

We observed lamprey *COE* gene expression in the mesenchyme of the pharyngeal arches and the mesenchyme dorsal to the pharynx that runs posteriorly, ventral to the notochord. Expression in pharyngeal arches has been observed with *Xenopus COE2* and *COE3* (Dubois et al. 1998; Pozzoli et al. 2001), chicken *COE1-3* (El-Magd et al. 2014b) and mouse *COE2* and *COE3* (Kieslinger et al. 2005) during embryonic development. Chicken *COE1* and *COE3* are mainly expressed in the neural crest component of pharyngeal arches, though weak expression in the mesodermal component was also observed (El-Magd et al. 2014b). In contrast, chicken *COE2* is mainly expressed in the mesodermal component of pharyngeal arches. We did not observe such a distinction with either lamprey *COE* gene. In addition, there is a time difference in pharyngeal expression of *COE* genes among tetrapods, as mouse *COE2* and *COE3* appear in pharyngeal arches before neural expression, whereas *Xenopus COE2* and *COE3* and chicken *COE1-3* are expressed after the onset of neural expression. In this sense, lamprey *COE* activation is more similar to *Xenopus* and chicken than to mouse. Zebrafish *COE2* expression has been reported in migrating cranial neural crest cells (Bally-Cuif et al. 1998), although its expression has not been followed at further developmental stages to determine if zebrafish *COE2* is activated in pharyngeal arches. Interestingly, in *Ciona COE* is also expressed in pharyngeal mesoderm and has an important role in its specification (Razy-Krajka et al. 2014), but pharyngeal expression has not been reported from amphioxus. This suggests stepwise acquisition of *COE* pharyngeal mesenchyme expression in chordates: from no expression primitively, to pharyngeal mesodermal expression in the vertebrate-tunicate ancestor, and acquisition of neural crest expression in vertebrates.

We also observed expression of both lamprey *COE* genes above the pharynx and running posteriorly ventral to the notochord, although their expression domains are not identical. Comparison with a recent anatomical description of lamprey mesoderm (Tulenko et al. 2013) reveals that *LpCOE-A* is expressed in nephric tissues. Tunicates lack a homologue of the nephric system, but in amphioxus *COE* expression in the mesoderm is restricted to the ventral part of the left anterior somite (Mazet et al. 2004), where a structure known as Hatschek's nephridium will form. Hatschek's nephridium is proposed to be homologous to the vertebrate nephric system, and expresses nephric marker genes (Kozmik et al. 1999). Expression of *COE* genes in the lamprey intermediate mesoderm suggests nephric expression may be ancestral for chordates. In agreement with this, mouse *COE4* is expressed in kidney (Wang et al. 2002).

No expression of either lamprey *COE* gene was observed in pre-somitic mesoderm, somites or their derivatives. Somitic expression of *COE* genes has been observed with mouse *COE2* (Kieslinger et al. 2005) and chicken *COE1-3* in the forming sclerotome, excluded from the dermomyotome (El-Magd et al. 2014b). Chicken *COE1* and *COE3* expression in these tissues is related to cartilage blastemas of the dorsal neural arches and proximal ribs, and to mesenchyme lateral to hypaxial and epaxial muscle precursors (El-Magd et al. 2014a). In mice, *COE* genes are expressed in osteoblasts and have been shown to have an important role in the differentiation of osteoclasts (Hesslein et al. 2009; Kieslinger et al. 2005). Thus, at least in mice and chicken, *COE* genes seem to have a role in the development of bone tissues. Absence of lamprey *COE* expression in somites might thus be explained by the lack of a mineralised skeleton.

In gnathostomes, *COE* genes are also expressed in the muscle lineage. In *Xenopus*, *COE2* and *COE3* are expressed in somites and in ventrally-migrating hypaxial muscle, and direct expression of a number of genes that are involved in the commitment, migration and differentiation of muscle cells (Green and Vetter 2011). Also, mouse *COE1* is expressed in skeletal muscle and *COE3* is expressed in the diaphragm, where they have an important role in regulating muscle cell relaxation (Jin et al. 2014). However, we did not observe expression of either lamprey *COE* gene in any muscle derivatives. We also did not observe expression of lamprey *COE* genes associated to the heart as with chicken *COE3* (El-Magd et al. 2014b). Thus, our results suggest that all muscle-related expression of *COE* genes as observed in mouse and *Xenopus* is completely absent in lampreys and, to a broader extent, all somitic *COE* expression is absent in lampreys. Amphioxus *COE* is only expressed in the anterior-most somite, where the nephric system develops (Mazet et al. 2004) and so is also not muscle related. However in the tunicate *Ciona COE* has a role in specifying the heart and pharyngeal muscle lineages (Razy-Krajka et al. 2014). It is therefore unclear whether muscle expression is an ancient character either lost or undetected in lampreys and amphioxus, or has evolved independently in *Ciona* and gnathostomes. The role of the *Drosophila COE* orthologue, *Collier*, in muscle development (Crozatier and Vincent 1999) would suggest the former.

Summary and Conclusions

Our analysis of lamprey *COE* gene identifies both similarities and differences in gene structure and expression between *COE* genes in vertebrates, other chordates and other animals. *COE* gene structure is generally conserved in vertebrates, though lamprey *COE-B* has a new exon towards its 3' end. Intron phasing through this region is the same in both lamprey *COE* genes and in vertebrate *COE* genes (Figure 2A; see (Daburon et al. 2008) for other vertebrate *COE* genes). This means alternate splicing through this region is potentially viable for all these genes, though it has only been

identified in lamprey *COE-B* and mouse *COE4*. Unless it has remained undetected for other *COE* genes, this would hence appear to have evolved in parallel. Splicing of the duplicated H2 also appears to be unique to lamprey *COE-B* genes. The model for the evolutionary origin of the duplicated H2 proposed by Daburon and colleagues (Daburon et al. 2008), in which duplication of the H2-encoding exon is followed by loss of the 3' end of the most 5' duplicate, is consistent with our data. However we find no evidence for this in *COE-A* genes. This suggests an evolutionary scenario in which this exon has been lost by *COE-A* following genome duplication (Figure 7).

We are also now able to map *COE* gene expression onto this model for the major chordate lineages: amphioxus, tunicates, cyclostomes and gnathostomes (further detailed in Figure 7). Both CNS and PNS expression are ancestral for chordates, and in vertebrates PNS expression has become incorporated into the peripheral ganglia (although on current data, DRG expression is in gnathostomes only). Some aspects of mesodermal expression also appear ancestral: nephric expression for all chordates, and pharyngeal expression for the tunicate-vertebrate ancestor. General expression in the nervous system and some mesodermal derivatives seems to be ancestral for all metazoans (Jackson et al. 2010), though descriptions are insufficient to determine if the later corresponds to nephric and/or pharyngeal-type cells. In gnathostomes COE expression appears in other mesodermal cell types, and gnathostome COE expression in the brain differs to that of lampreys (discussed above). While the latter could map onto the evolution of patterns of neurogenesis in the vertebrate brain, mesodermal expression may reflect diversification of cell types early in vertebrate evolution.

Acknowledgements

We thank the Forestry Commission of England for permission to collect lamprey embryos and Dr. Jo Begbie for access to histology facilities. RL-R was supported by the Mexican National Council for Science and Technology (CONACYT). CP was supported by a Royal Society Newton International Fellowship from the Royal Society and by an EMBO Long Term Fellowship.

References

Akerblad P, Lind U, Liberg D, Bamberg K, Sigvardsson M (2002) Early B-cell factor (O/E-1) is a promoter of adipogenesis and involved in control of genes important for terminal adipocyte differentiation Mol Cell Biol 22:8015-8025

Akerblad P et al. (2005) Gene expression analysis suggests that EBF-1 and PPARgamma2 induce adipogenesis of NIH-3T3 cells with similar efficiency and kinetics Physiol Genomics 23:206-216 doi:10.1152/physiolgenomics.00015.2005

Bally-Cuif L, Dubois L, Vincent A (1998) Molecular cloning of Zcoe2, the zebrafish homolog of Xenopus Xcoe2 and mouse EBF-2, and its expression during primary neurogenesis Mech Dev 77:85-90 doi:S0925-4773(98)00144-0 [pii]

Bielle F et al. (2011) Slit2 activity in the migration of guidepost neurons shapes thalamic projections during development and evolution Neuron 69:1085-1098 doi:10.1016/j.neuron.2011.02.026
Corradi A et al. (2003) Hypogonadotropic hypogonadism and peripheral neuropathy in Ebf2-null mice Development 130:401-410
Crozatier M, Valle D, Dubois L, Ibnsouda S, Vincent A (1996) Collier, a novel regulator of Drosophila head development, is expressed in a single mitotic domain Curr Biol 6:707-718
Crozatier M, Vincent A (1999) Requirement for the Drosophila COE transcription factor Collier in formation of an embryonic muscle: transcriptional response to notch signalling Development 126:1495-1504
Daburon V, Mella S, Plouhinec JL, Mazan S, Crozatier M, Vincent A (2008) The metazoan history of the COE transcription factors. Selection of a variant HLH motif by mandatory inclusion of a duplicated exon in vertebrates BMC Evol Biol 8:131 doi:10.1186/1471-2148-8-131
Davis JA, Reed RR (1996) Role of Olf-1 and Pax-6 transcription factors in neurodevelopment J Neurosci 16:5082-5094
Demilly A, Simionato E, Ohayon D, Kerner P, Garces A, Vervoort M (2011) Coe genes are expressed in differentiating neurons in the central nervous system of protostomes PLoS One 6:e21213 doi:10.1371/journal.pone.0021213
Dubois L, Bally-Cuif L, Crozatier M, Moreau J, Paquereau L, Vincent A (1998) XCoe2, a transcription factor of the Col/Olf-1/EBF family involved in the specification of primary neurons in Xenopus Curr Biol 8:199-209 doi:S0960-9822(98)70084-3 [pii]
Dubois L, Vincent A (2001) The COE--Collier/Olf1/EBF--transcription factors: structural conservation and diversity of developmental functions Mech Dev 108:3-12 doi:S0925477301004865 [pii]
El-Magd MA, Saleh AA, El-Aziz RM, Salama MF (2014a) The effect of RA on the chick Ebf1-3 genes expression in somites and pharyngeal arches Dev Genes Evol 224:245-253 doi:10.1007/s00427-014-0483-y
El-Magd MA, Saleh AA, Farrag F, Abd El-Aziz RM, Ali HA, Salama MF (2014b) Regulation of chick Ebf1-3 gene expression in the pharyngeal arches, cranial sensory ganglia and placodes Cells Tissues Organs 199:278-293 doi:10.1159/000369880
Fields S, Ternyak K, Gao H, Ostraat R, Akerlund J, Hagman J (2008) The 'zinc knuckle' motif of Early B cell Factor is required for transcriptional activation of B cell-specific genes Mol Immunol 45:3786-3796 doi:10.1016/j.molimm.2008.05.018
Furlong RF, Holland PW (2002) Were vertebrates octoploid? Philos Trans R Soc Lond B Biol Sci 357:531-544 doi:10.1098/rstb.2001.1035
Garcia-Dominguez M, Poquet C, Garel S, Charnay P (2003) Ebf gene function is required for coupling neuronal differentiation and cell cycle exit Development 130:6013-6025 doi:10.1242/dev.00840
Garel S, Garcia-Dominguez M, Charnay P (2000) Control of the migratory pathway of facial branchiomotor neurones Development 127:5297-5307
Garel S, Marin F, Mattei MG, Vesque C, Vincent A, Charnay P (1997) Family of Ebf/Olf-1-related genes potentially involved in neuronal differentiation and regional specification in the central nervous system Dev Dyn 210:191-205 doi:10.1002/(SICI)1097-0177(199711)210:3<191::AID-AJA1>3.0.CO;2-B [pii]
Grabherr MG et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome Nat Biotechnol 29:644-U130 doi:10.1038/nbt.1883
Green YS, Vetter ML (2011) EBF proteins participate in transcriptional regulation of Xenopus muscle development Dev Biol 358:240-250 doi:10.1016/j.ydbio.2011.07.034
Hagman J, Gutch MJ, Lin H, Grosschedl R (1995) EBF contains a novel zinc coordination motif and multiple dimerization and transcriptional activation domains EMBO J 14:2907-2916
Hesslein DG et al. (2009) Ebf1-dependent control of the osteoblast and adipocyte lineages Bone 44:537-546 doi:10.1016/j.bone.2008.11.021

Imai KS, Hino K, Yagi K, Satoh N, Satou Y (2004) Gene expression profiles of transcription factors and signaling molecules in the ascidian embryo: towards a comprehensive understanding of gene networks Development 131:4047-4058 doi:10.1242/dev.01270

Jackson DJ et al. (2010) Developmental expression of COE across the Metazoa supports a conserved role in neuronal cell-type specification and mesodermal development Dev Genes Evol 220:221-234 doi:10.1007/s00427-010-0343-3

Jimenez MA, Akerblad P, Sigvardsson M, Rosen ED (2007) Critical role for Ebf1 and Ebf2 in the adipogenic transcriptional cascade Mol Cell Biol 27:743-757 doi:10.1128/MCB.01557-06

Jin S et al. (2014) Ebf factors and MyoD cooperate to regulate muscle relaxation via Atp2a1 Nat Commun 5:3793 doi:10.1038/ncomms4793

Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform Nucleic Acids Res 30:3059-3066

Katoh K, Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program Brief Bioinform 9:286-298 doi:10.1093/bib/bbn013

Kieslinger M et al. (2005) EBF2 regulates osteoblast-dependent differentiation of osteoclasts Dev Cell 9:757-767 doi:10.1016/j.devcel.2005.10.009

Kozmik Z, Holland ND, Kalousova A, Paces J, Schubert M, Holland LZ (1999) Characterization of an amphioxus paired box gene, AmphiPax2/5/8: developmental expression patterns in optic support cells, nephridium, thyroid-like structures and pharyngeal gill slits, but not in the midbrain-hindbrain boundary region Development 126:1295-1304

Kuraku S, Meyer A, Kuratani S (2009) Timing of genome duplications relative to the origin of the vertebrates: did cyclostomes diverge before or after? Mol Biol Evol 26:47-59 doi:10.1093/molbev/msn222

Lara-Ramirez R, Patthey C, Shimeld SM (2015) Characterization of two neurogenin genes from the brook lamprey lampetra planeri and their expression in the lamprey nervous system Dev Dyn doi:10.1002/dvdy.24273

Li S, Yin M, Liu S, Chen Y, Yin Y, Liu T, Zhou J (2010) Expression of ventral diencephalon-enriched genes in zebrafish Dev Dyn 239:3368-3379 doi:10.1002/dvdy.22467

Lin H, Grosschedl R (1995) Failure of B-cell differentiation in mice lacking the transcription factor EBF Nature 376:263-267 doi:10.1038/376263a0

Lopez-Bendito G et al. (2006) Tangential neuronal migration controls axon guidance: a role for neuregulin-1 in thalamocortical axon navigation Cell 125:127-142 doi:10.1016/j.cell.2006.01.042

Louis A, Nguyen NT, Muffato M, Roest Crollius H (2015) Genomicus update 2015: KaryoView and MatrixView provide a genome-wide perspective to multispecies comparative genomics Nucleic Acids Res 43:D682-689 doi:10.1093/nar/gku1112

Malgaretti N et al. (1997) Mmot1, a new helix-loop-helix transcription factor gene displaying a sharp expression boundary in the embryonic mouse brain J Biol Chem 272:17632-17639

Mazet F, Hutt JA, Milloz J, Millard J, Graham A, Shimeld SM (2005) Molecular evidence from Ciona intestinalis for the evolutionary origin of vertebrate sensory placodes Dev Biol 282:494-508 doi:10.1016/j.ydbio.2005.02.021

Mazet F, Masood S, Luke GN, Holland ND, Shimeld SM (2004) Expression of AmphiCoe, an amphioxus COE/EBF gene, in the developing central nervous system and epidermal sensory neurons Genesis 38:58-65 doi:10.1002/gene.20006

Mehta TK et al. (2013) Evidence for at least six Hox clusters in the Japanese lamprey (Lethenteron japonicum) Proc Natl Acad Sci U S A 110:16044-16049 doi:10.1073/pnas.1315760110

Murakami Y, Ogasawara M, Satoh N, Sugahara F, Myojin M, Hirano S, Kuratani S (2002) Compartments in the lamprey embryonic brain as revealed by regulatory gene expression and the distribution of reticulospinal neurons Brain Res Bull 57:271-275

Murakami Y, Ogasawara M, Sugahara F, Hirano S, Satoh N, Kuratani S (2001) Identification and expression of the lamprey Pax6 gene: evolutionary origin of the segmented brain of vertebrates Development 128:3521-3531
Murakami Y, Pasqualetti M, Takio Y, Hirano S, Rijli FM, Kuratani S (2004) Segmental development of reticulospinal and branchiomotor neurons in lamprey: insights into the evolution of the vertebrate hindbrain Development 131:983-995 doi:10.1242/dev.00986
Nakatani Y, Takeda H, Kohara Y, Morishita S (2007) Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates Genome Res 17:1254-1265 doi:10.1101/gr.6316407
Osorio J, Mazan S, Retaux S (2005) Organisation of the lamprey (Lampetra fluviatilis) embryonic brain: insights from LIM-homeodomain, Pax and hedgehog genes Dev Biol 288:100-112 doi:10.1016/j.ydbio.2005.08.042
Pang K, Matus DQ, Martindale MQ (2004) The ancestral role of COE genes may have been in chemoreception: evidence from the development of the sea anemone, Nematostella vectensis (Phylum Cnidaria; Class Anthozoa) Dev Genes Evol 214:134-138 doi:10.1007/s00427-004-0383-7
Parker HJ, Bronner ME, Krumlauf R (2016) The vertebrate Hox gene regulatory network for hindbrain segmentation: Evolution and diversification: Coupling of a Hox gene regulatory network to hindbrain segmentation is an ancient trait originating at the base of vertebrates Bioessays 38:526-538 doi:10.1002/bies.201600010
Pozzoli O, Bosetti A, Croci L, Consalez GG, Vetter ML (2001) Xebf3 is a regulator of neuronal differentiation during primary neurogenesis in Xenopus Dev Biol 233:495-512 doi:10.1006/dbio.2001.0230
Putnam NH et al. (2008) The amphioxus genome and the evolution of the chordate karyotype Nature 453:1064-1071 doi:10.1038/nature06967
Rajakumari S et al. (2013) EBF2 determines and maintains brown adipocyte identity Cell Metab 17:562-574 doi:10.1016/j.cmet.2013.01.015
Razy-Krajka F, Lam K, Wang W, Stolfi A, Joly M, Bonneau R, Christiaen L (2014) Collier/OLF/EBF-dependent transcriptional dynamics control pharyngeal muscle specification from primed cardiopharyngeal progenitors Dev Cell 29:263-276 doi:10.1016/j.devcel.2014.04.001
Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models Bioinformatics 19:1572-1574
Schubert M, Holland ND, Escriva H, Holland LZ, Laudet V (2004) Retinoic acid influences anteroposterior positioning of epidermal sensory neurons and their gene expression in a developing chordate (amphioxus) Proc Natl Acad Sci U S A 101:10320-10325 doi:10.1073/pnas.0403216101
Shimeld SM, Donoghue PC (2012) Evolutionary crossroads in developmental biology: cyclostomes (lamprey and hagfish) Development 139:2091-2099 doi:10.1242/dev.074716
Singh PP, Arora J, Isambert H (2015) Identification of Ohnolog Genes Originating from Whole Genome Duplication in Early Vertebrates, Based on Synteny Comparison across Multiple Genomes PLoS Comput Biol 11:e1004394 doi:10.1371/journal.pcbi.1004394
Siponen MI et al. (2010) Structural determination of functional domains in early B-cell factor (EBF) family of transcription factors reveals similarities to Rel DNA-binding proteins and a novel dimerization motif J Biol Chem 285:25875-25879 doi:10.1074/jbc.C110.150482
Smith JJ, Keinath MC (2015) The sea lamprey meiotic map improves resolution of ancient vertebrate genome duplications Genome Res 25:1081-1090 doi:10.1101/gr.184135.114
Smith JJ et al. (2013) Sequencing of the sea lamprey (Petromyzon marinus) genome provides insights into vertebrate evolution Nat Genet 45:415-421, 421e411-412 doi:10.1038/ng.2568
Sugahara F, Aota S, Kuraku S, Murakami Y, Takio-Ogawa Y, Hirano S, Kuratani S (2011) Involvement of Hedgehog and FGF signalling in the lamprey telencephalon: evolution of regionalization

and dorsoventral patterning of the vertebrate forebrain Development 138:1217-1226 doi:10.1242/dev.059360
Sugahara F et al. (2016) Evidence from cyclostomes for complex regionalization of the ancestral vertebrate brain Nature 531:97-100 doi:10.1038/nature16518
Tahara Y (1988) Normal Stages of Development in the Lamprey, Lampetra-Reissneri (Dybowski) Zoological Science 5:109-118
Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods Molecular Biology and Evolution 28:2731-2739 doi:DOI 10.1093/molbev/msr121
Treiber N, Treiber T, Zocher G, Grosschedl R (2010) Structure of an Ebf1:DNA complex reveals unusual DNA recognition and structural homology with Rel proteins Genes Dev 24:2270-2275 doi:10.1101/gad.1976610
Tulenko FJ et al. (2013) Body wall development in lamprey and a new perspective on the origin of vertebrate paired fins Proc Natl Acad Sci U S A 110:11899-11904 doi:10.1073/pnas.1304210110
Wang MM, Reed RR (1993) Molecular cloning of the olfactory neuronal transcription factor Olf-1 by genetic selection in yeast Nature 364:121-126 doi:10.1038/364121a0
Wang SS, Betz AG, Reed RR (2002) Cloning of a novel Olf-1/EBF-like gene, O/E-4, by degenerate oligo-based direct selection Mol Cell Neurosci 20:404-414 doi:S1044743102911383 [pii]
Wang SS, Tsai RY, Reed RR (1997) The characterization of the Olf-1/EBF-like HLH transcription factor family: implications in olfactory gene regulation and neuronal development J Neurosci 17:4149-4158
Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach Mol Biol Evol 18:691-699

Figure legends

Figure 1

(A) Molecular phylogenetic analysis of deuterostome *COE* sequences. The tree shown was constructed by Bayesian inference. Shown above key nodes are posterior probabilities from this analysis, while numbers below these nodes are percentage bootstrap support for the same node derived from Maximum Likelihood analysis. Gnathostome and cyclostome *COE* orthology groups are boxed. Support for orthology of *HsCOE4* with *CmCOE4* and *LoCOE4* is weak, however *CmCOE4* and *LoCOE4* are well supported as orthologues, and synteny supports orthology of *HsCOE4* and *LoCOE4* (Figure S3). Species abbreviations: Ac, *Anolis carolinensis*; Bf, *Branchiostoma floridae*; Cm, *Callorhincus milii*; Dr, *Danio rerio*; Eb, *Eptatretus burgeri*; Gg, *Gallus gallus*; Hs, *Homo sapiens*; Lc, *Latimeria chalumnae*; Lj, *Lethenteron japonicum*; Lo, *Lepisosteus oculatus.* Lp, *Lampetra planeri*; Pm, *Petromyzon marinus*; Re, *Raja eglanteria*; Sk, *Saccoglossus kowalevski*; Sp, *Strongylocentrotus purpuratus*; Xt, *Xenopus tropicalis*.

(B) Schematic maps of *COE* locus paralogy and synteny in jawed vertebrate, lamprey and amphioxus genomes. *COE* genes are in red, colour coding of other genes is as follows: Grey; genes with no orthologues or paralogues identified in the analysed regions. White; genes with syntenic orthologues

and/or paralogues in human and shark. Black; genes linked to *COE* in amphioxus and their orthologue positions in other species. Other colours; genes linked to lamprey *COE* genes and their orthologues and/or paralogues in other species. Discontinuities shown as angled bars indicate where genes map to the same chromosome arm or scaffold, but are separated by multiple intervening genes which are not shown. While the current amphioxus (*B. floridae*) genome assembly has tandem *COE* gene models, sequence comparisons (not shown) indicate these derive from a single gene as depicted. Additional synteny data for *L. oculatus* are in Figure S3.

Figure 2

Alternate splicing of lamprey *COE* genes.

(A) Schematic intron-exon maps of *COE-A* and *COE-B*, generated by mapping *L. planeri* sequence to the *L. japonicum* genome. The various protein domains are colour coded, and alternate splicing verified by transcriptome and RT-PCR is indicated. These are shown in further detail in panels D-E below. Intron phase class is also shown, and the structure of Mouse *COE1*, as previously described (Daburon et al. 2008), is shown for comparison. Exons are numbered sequentially in *COE-B*, with numbering preserved in *COE-A* and mouse *COE1* to indicate exon homology. An expansion of this with accompanying RT-PCR data is in Figure S2.

(B) Splicing site in the HLH domain. *COE-B* shows alternate splicing of the H2d exon. Black arrows at the top indicate intron sites in *COE-B*, grey arrows at the bottom indicate intron sites in COE-A.

(C) Splicing within the TA1 region of *LpCOE-B.* Variant 2 removes 4 amino acids from the predicted ORF adjacent to the exon 13-exon 14 boundary.

(D) Splicing within the TA2 region of *LpCOE-B*. Four splice variants were identified. Comparison to intron positions illustrates this occurs by skipping of exon 15, exon 16, or both. *LpCOE-A* is shown for comparison and has a relatively different sequence and a different exon structure in this region.

(E) Splicing at the junction between exons 8 and 9 of *LpCOE-A*. This results in the insertion of 10 amino acids, as compared to *LpCOE-B*.

Figure 3

LpCOE-A expression during *L. planeri* development at stages 21-24. (A-D, F and G) are lateral views, (E) dorsal view. In all images anterior is to the left. (A) At stage 21, expression is seen in the mandibular arch (pa1) and below the notochord (n: arrow). (B) At stage 22, expression appears in the second and third pharyngeal arches (pa2, pa3). (C, D) At stage 23, expression appears in the spinal cord (sc), ventral diencephalon (di: white asterisk), midbrain (mb: behind the ophthalmic placode (opV)) and hindbrain (hb), in the maxillomandibular (mmV), geniculate (g) and posterior lateral line (llp) placodes, the

olfactory/neurohypophyseal plate (o/NHP: black arrowhead), as well as in the fourth pharyngeal arch (pa4). Expression below the notochord extends posteriorly (arrow). (E) At stage 24, from a dorsal view, expression is seen as two lateral stripes (arrows) all along the neural tube except in the telencephalon. (F, G) In the brain, expression progresses dorsally in the diencephalon (white asterisk), midbrain and hindbrain. Expression increases in the o/NHP (arrowheads) and in the opV, mmV and llp placodes/ganglia. At this stage, transcripts are also observed in the nascent fifth pharyngeal arch (pa5) and below the notochord (G, arrow). Additional abbreviation: MHB, midbrain-hindbrain boundary.

Figure 4

LpCOE-A expression during *L. planeri* development at stages 25-28. (A, D, I-K, M-O) Lateral views. (F) Ventral view of the head of the embryo shown in (A). (E, L) Dorsal views of the anterior trunk and head, respectively. Anterior is to the left in all images except in (E) where anterior is to the top, and (B, C, G, H) are cross-sections of the embryo shown in (A) and (D). Levels of cross-sections are indicated in (A) and (D) and dorsal side is to the top. (I) and (M) are lateral views of the trunk at the levels marked by lines in (A) and (J), respectively. (A) At stage 25, expression expands in most of the CNS. (D) In the head, expression is seen in the telencephalon (tc), diencephalon, midbrain and hindbrain as well as in the epiphysis (ep). Strong expression in the dorsal hindbrain is seen at the level of rhombomeres 3-5 (asterisk). Transcripts are also observed in the forming olfactory epithelium and trigeminal, geniculate and petrosal ganglia. At this stage, transcripts are seen in the eight pharyngeal arches and in the upper lip (ul). (F) From a ventral view of the head, expression is distinguished in mesenchyme of each pharyngeal arch that possibly corresponds to the mesoderm and neural crest components of each arch. (E) From a dorsal view, staining is seen as two lateral stripes along the neural tube. Expression under the notochord progresses posteriorly as the embryo elongates (A, I, arrows). (B, C, G, H) Cross sections reveal strong staining on the sides of the neural tube corresponding to the mantle layer (ml). (B) Expression is seen in trigeminal ganglia and mesenchyme of pharyngeal arches. (C) Transcripts are also seen surrounding the otic vesicle (ov) and possibly in the acoustic ganglia (white arrows). (G) In the body, transcripts are found in mesenchyme ventro-lateral to the notochord (black arrows) and in an extended domain (white arrows) just above the vitellum (vi). (H) This expression is not observed more posteriorly. (J, K, M) At stage 26, *LpCOE-A* expression is downregulated both in the head and spinal cord maintaining the same general pattern, but expression under the notochord disappears (M, arrows). All nine pharyngeal arches are stained at this stage (K). (L) From a dorsal view, expression is restricted to the lateral sides of the neural tube delimiting the extent of the ventricular zone (vz) in the middle, with expansions of the ventricular zone at the level of the epiphysis, the MHB, and posterior hindbrain (white asterisks). (N, O) At stage 28, expression is even more reduced and confined to the head. Additional abbreviations as previous figures plus: hy, hypothalamus; tc, telencephalon.

Figure 5

LpCOE-B expression during *L. planeri* development at stages 21-24. (A, C, E, F and G) are lateral views. (B, D and H) are dorsal views. In all images anterior is to the left. (A, B) At stage 21, *LpCOE-B* is observed in the anterior spinal cord as two lateral stripes (arrows). (C, D) At stage 22, expression extends posteriorly along the growing spinal cord (C, arrows), which is seen as two parallel stripes from a dorsal view (D). Transcripts are also observed in the trigeminal placodes (black arrowheads) and the o/NHP anlage (white arrowheads). Inset, frontal view of the embryo shown in (C). (E) At stage 23, expression is maintained in the spinal cord (arrows) and trigeminal placodes, and it now appears in the developing geniculate and petrosal placodes. (F-H) At stage 24, *LpCOE-B* appears in the hindbrain, midbrain and in a small region of the diencephalon (white asterisk) (F). Expression considerably intensifies in the trigeminal, geniculate and petrosal ganglia, as well as in the NHP (white arrowheads) and spinal cord (G). Outside the nervous system, strong expression is seen in mesenchyme of the first pharyngeal arch (pa1) and in an extending domain between the pharynx and notochord (black arrows). Expression in the neural tube is localised to the lateral sides as in previous stages (H, white arrows). Abbreviations as in previous figures.

Figure 6

LpCOE-B expression during *L. planeri* development at stages 25-28. (A, B, D, F, G, L-N) are lateral views; (B) and (G) are lateral views of the trunk region at the stages indicated. (C, E, H) are dorsal views of the head and trunk region. (I-K) are cross-sections of a stage 26 embryo, dorsal to the top. In all lateral and dorsal views anterior is to the left. (A-C) At stage 25, *LpCOE-B* is strongly expressed in specific regions of the diencephalon, midbrain and hindbrain, as well as in the maxillomandibular, geniculate and petrosal ganglia and olfactory epithelium (A, white arrowhead). In the hindbrain, two patches of staining are observed dorsally (A, asterisks), with diffuse staining in between. Faint expression is also observed in the epiphysis (A, black arrowhead). *LpCOE-B* is not expressed in the telencephalon at this stage. Expression in the spinal cord becomes very restricted to the dorsal side, and transcripts are not observed under the notochord as with *LpCOE-A* (B, arrows). At this stage, transcripts are seen in the five anterior-most pharyngeal arches and in an extending domain just above the pharynx (A, black arrow). From a dorsal view, staining is still restricted to the lateral sides of the neural tube as with *LpCOE-A*, delineating the ventricular zone in the middle (C). (D-K) At stage 26, expression increases even more in the head. Staining is highly increased in the trigeminal (t), geniculate and petrosal ganglia as well as in the first pharyngeal arch (D, F). In the spinal cord transcripts are restricted dorsally, and are not observed under the notochord (G, arrows). Expression is localised to the lateral sides of the brain and spinal cord as seen in dorsal views (E, H). In the brain, expansions of the ventricular zone are observed at the level of the epiphysis, MHB and posterior hindbrain (E, asterisks). (I-K) Cross-sections

reveal staining on the lateral sides of the neural tube corresponding to the mantle layer, with stronger expression dorsally (white arrows). Expression is also observed in pharyngeal arch mesenchyme possibly corresponding to both mesoderm and neural crest (arrowheads). Transcripts are also observed in the velum (asterisks) and in a stream of cells ventrolateral to the neural tube that reach pharyngeal expression (black arrows). Transcripts are also observed in maxillomandibular ganglia (I). (L, M) At stage 27, expression is clearly downregulated in the head, although maintaining the same pattern as stage 26 embryos, except in the diencephalon (L) and in the spinal cord (M) where transcripts are virtually absent. (N) At stage 28, expression is reduced even more but the same pattern as in previous stages is still observed in the head, with strongest expression seen in olfactory epithelium, upper lip, cranial ganglia and dorsal and ventral domains of each pharyngeal arch. Additional abbreviations as previous figures plus: ll, lower lip; T, tegmentum; te, optic tectum.

Figure 7

A model for the evolution of the *COE* genes in chordates. Duplication of H2 is shown before the duplication and divergence of the vertebrate *COE* paralogues; we have shown vertebrate *COE* paralogue relationships as unresolved, as while synteny data shows it happened in both lineages by block duplication probably tied to genome duplication, lamprey-gnathostome orthology is not resolvable by either molecular phylogenetics or synteny. Lamprey *COE-A* then reverts to a 2 helix state. A summary of expression is shown next to each lineage; for lampreys and gnathostomes, this summarises the combined expression of all paralogues. Expression of *COE* genes in CNS and PNS is ancestral. Some mesodermal expression is also likely ancestral. We found no evidence for endodermal expression. Neural and mesodermal expression may both be older, as both are found in many other animals, though the type of mesoderm expressing *COE* genes is quite variable (Jackson et al. 2010). For sources of expression data see respective sections of the Discussion.

Table 1A: Sequences marking the 3' and 5' ends of *COE-A* exons and the number of transcriptome fragments confirming each.

Exon-exon junction	3' of upstream exon	5' of downstream exon	count
Exon 1 - exon 2	GCCAGCACCGCGGCACAGAG	TGGCATTGCGCTAGCGCGGG	29
Exon 2 - exon 3	ACTTCGTGGAGAAGGATCGG	GAACCCAACAATGAAAAGAC	31
Exon 3 - exon 4	ACAGCTGTTGTACAGTAATG	GTGTGCGGACGGAGCAAGAT	29
Exon 4 - exon 5	TCGACTCCATGAACAAACAG	GCCATTATCTATGAAGGGCA	21
Exon 5 - exon 6	ACGCACGAGATCATGTGCAG	CCGCTGCTGCGACAAGAAGA	28
Exon 6 - exon 7	GATCCGGTGATAATAGACAG	ATTCTTTCTGAAGTTCTTTC	29
Exon 7 - exon 8	GGGACATGCGCCGATTTCAG	GTGGTTGTCTCGACGACAGT	38
Exon 8 short - exon 9 canonical	CAGAATCGACCCCTCCGAAG	CAGCCACACCGTGCATCAAA	8
Exon 8 short - exon 9 short	CAGAATCGACCCCTCCGAAG	CCACACCGTGCATCAAAGCC	2
Exon 8 canonical - exon 9 canonical	ACCGTCTTATCTGGACAATG	CAGCCACACCGTGCATCAAA	19
Exon 8 canonical - exon 9 short	ACCGTCTTATCTGGACAATG	CCACACCGTGCATCAAAGCC	0
Exon 9 - exon 10	CCATGCTGGTGTGGAGCGAG	CTCATCACGCCCCATGCCAT	33
Exon 10 - exon 11	AGGGCGATTCGTGTACACTG	CTCTGAATGAGCCAACAATT	23
Exon 11 - exon 13	ACCCAGAGAGGCTTCCCAAG	GAAATTATCCTGAAGCGAGC	31
Exon 13 - exon 14	CTCCCAGGCTGCTGACCAGG	GGTACACGCGCAACAGCAGC	27
Exon 14 short - exon 15	CGGCGGCTCTCCCTACGGCA	TGAAACAGAAGAGTGCATTC	1
Exon 14 canonical - exon 15	TGCGCCCCTCTTTTCATCGG	TGAAACAGAAGAGTGCATTC	26
Exon 15 - exon 17	CAACGCCAACGGTCTGCAAG	TCATGTCTGGACTGGTGGTC	17

Table 1B: Table 1A: Sequences marking the 3' and 5' ends of *COE-B* exons and the number of transcriptome fragments confirming each.

Exon-exon junction	3' of upstream exon	5' of downstream exon	count
Exon 1 - exon 2	GCCAACACGGCCGCCCAGAG	CGGAGTCGCTCTGGCGAGGG	12
Exon 2 - exon 3	ACTTTGTCGAGAAGGACAGA	GAACCAAACAGTGAAAAAAC	8
Exon 3 - exon 4	ACAGTTACTCTACAGCAATG	GCATCCGCACGGAGCAAGAC	15
Exon 4 - exon 5	TCGACTCCATGACTAAGCAG	GCGATCATTTACGATGGGCA	14
Exon 5 - exon 6 long	ACGCACGAGATCATGTGCAG	TGTGCAACTCACAGTCGCTG	0
Exon 5 - exon 6 canonical	ACGCACGAGATCATGTGCAG	TCGCTGCTGCGACAAGAAGA	18
Exon 6 - exon 7	GACCCGGTCATCATCGACAG	GTTTTTTCTCAAGTTCTTCC	8
Exon 6 - exon 9 short	GACCCGGTCATCATCGACAG	CTACGCCGTGCATCAAAGCA	1
Exon 7 - exon 8	GAGACATGAGGCGGTTCCAG	GTCGTCATCTCCACGACCGT	14
Exon 8 - exon 9 canonical	GAGACTCGACCCTTCGGAAG	CAGCTACGCCGTGCATCAAA	6
Exon 8 - exon 9 short	GAGACTCGACCCTTCGGAAG	CTACGCCGTGCATCAAAGCA	19
Exon 9 - exon 10 canonical	CCATGCTCGTGTGGAGTGAG	CTGATCACACCCCATGCCAT	24
Exon 9 - exon 10 short	GGGACGCTTCGTCTACACCG	GGTGCAGACACCTCCCCGCC	0
Exon 10 - exon 11	GGGACGCTTCGTCTACACCG	CTCTGAACGAGCCCACCATT	18
Exon 11 - exon 12	ACCCCGAGAGGCTACCCAAG	GAGGTGATCTTGAAGCGCGC	5
Exon 11 - exon 13	ACCCCGAGAGGCTACCCAAG	GAAATCATTCTGAAGAGAGC	8
Exon 12 - exon 13	GCCTGCCTCATTACAACCAG	GAAATCATTCTGAAGAGAGC	4
Exon 13 short - exon 14	CATGCTGAGCCACGCCCCCG	GTTACAGCCGCAATACGAGC	12
Exon 13 canonical - exon 14	CGCCCCCGGTCGGTACCCCT	GTTACAGCCGCAATACGAGC	8
Exon 14 - exon 15 canonical	CGCCAGCTCGCCCTATGCCA	TCATGCCGTCAAGCCCCCCC	6
Exon 14 - exon 15 short	CGCCAGCTCGCCCTATGCCA	AAACAGAAGAGCGCCTTCGC	3
Exon 14 - exon 16	CGCCAGCTCGCCCTATGCCA	GGAATTTTGGACATGGTCTT	2
Exon 14 - exon 17	CGCCAGCTCGCCCTATGCCA	CCATGTCCGGCTTGGTCGTC	6
Exon 15 - exon 16	TACCGGCAATGCCCTCCAAG	GGAATTTTGGACATGGTCTT	8
Exon 15 - exon 17	TACCGGCAATGCCCTCCAAG	CCATGTCCGGCTTGGTCGTC	4
Exon 16 - exon 17	GTTTAAAGATTCGGTTTTAG	CCATGTCCGGCTTGGTCGTC	5

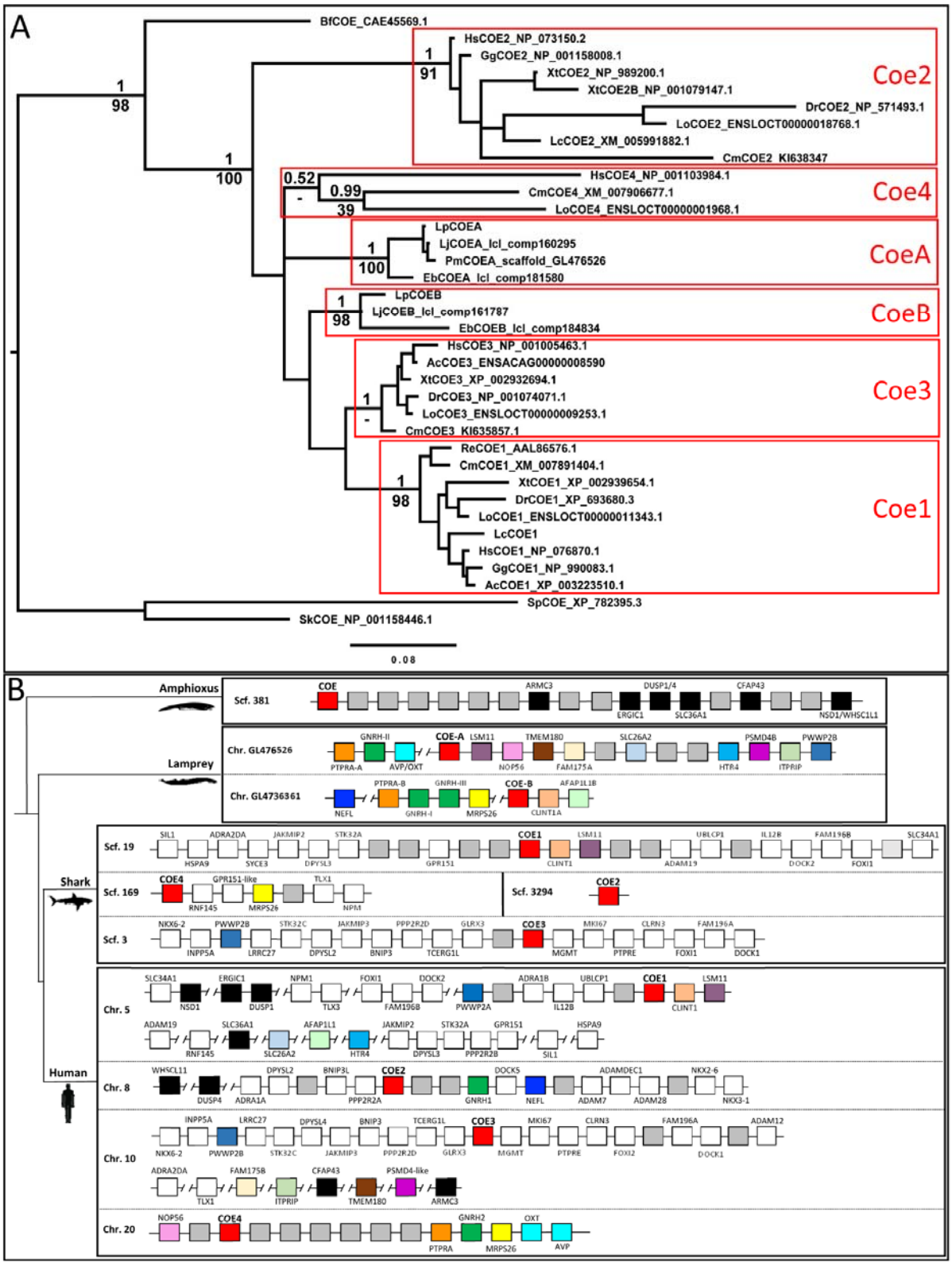
A
BfCOE_CAE45569.1
HsCOE2_NP_073150.2
GgCOE2_NP_001158008.1
XtCOE2_NP_989200.1
XtCOE2B_NP_001079147.1
DrCOE2_NP_571493.1
LoCOE2_ENSLOCT00000018768.1
LcCOE2_XM_005991882.1
CmCOE2_KI638347
Coe2
HsCOE4_NP_001103984.1
CmCOE4_XM_007906677.1
LoCOE4_ENSLOCT00000001968.1
Coe4
LpCOEA
LjCOEA_lcl_comp160295
PmCOEA_scaffold_GL476526
EbCOEA_lcl_comp181580
CoeA
LpCOEB
LjCOEB_lcl_comp161787
EbCOEB_lcl_comp184834
CoeB
HsCOE3_NP_001005463.1
AcCOE3_ENSACAG00000008590
XtCOE3_XP_002932694.1
DrCOE3_NP_001074071.1
LoCOE3_ENSLOCT00000009253.1
CmCOE3_KI635857.1
Coe3
ReCOE1_AAL86576.1
CmCOE1_XM_007891404.1
XtCOE1_XP_002939654.1
DrCOE1_XP_693680.3
LoCOE1_ENSLOCT00000011343.1
LcCOE1
HsCOE1_NP_076870.1
GgCOE1_NP_990083.1
AcCOE1_XP_003223510.1
Coe1
SpCOE_XP_782395.3
SkCOE_NP_001158446.1
0.08
B
Amphioxus
Lamprey
Shark
Human
Scf. 381
Chr. GL476526
Chr. GL4736361
Scf. 19
Scf. 169
Scf. 3294
Scf. 3
Chr. 5
Chr. 8
Chr. 10
Chr. 20

Figure 1

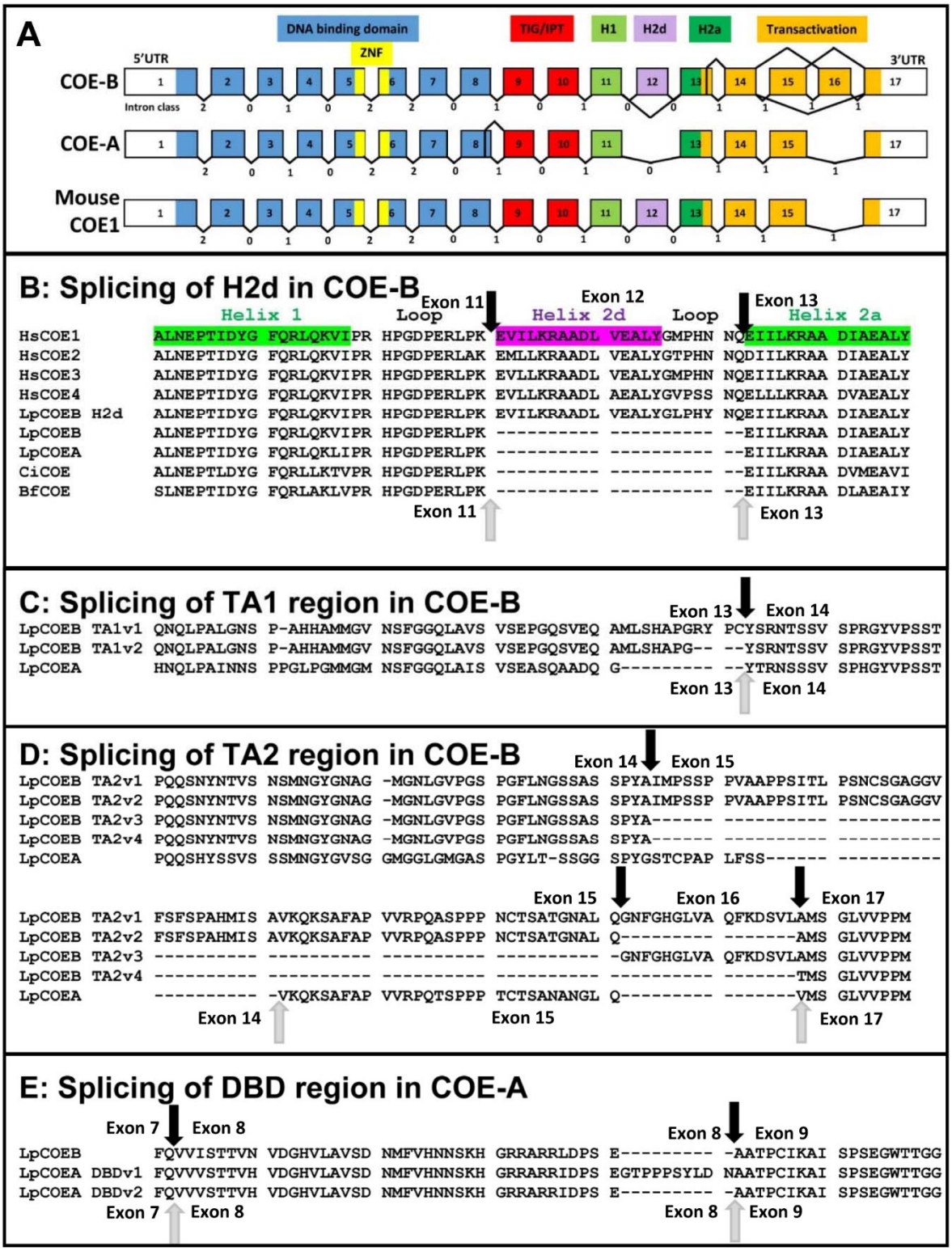
A
DNA binding domain
ZNF
TIG/IPT
H1
H2d
H2a
Transactivation
5'UTR
3'UTR
COE-B
COE-A
Mouse COE1
Intron class
B: Splicing of H2d in COE-B
Helix 1
Loop
Helix 2d
Loop
Helix 2a
Exon 11
Exon 12
Exon 13
HsCOE1 ALNEPTIDYG FQRLQKVIPR HPGDPERLPK EVILKRAADL VEALYGMPHN NQEIILKRAA DIAEALY
HsCOE2 ALNEPTIDYG FQRLQKVIPR HPGDPERLAK EMLLKRAADL VEALYGTPHN NQDIILKRAA DIAEALY
HsCOE3 ALNEPTIDYG FQRLQKVIPR HPGDPERLPK EVLLKRAADL VEALYGMPHN NQEIILKRAA DIAEALY
HsCOE4 ALNEPTIDYG FQRLQKVIPR HPGDPERLPK EVLLKRAADL AEALYGVPSS NQELLLKRAA DVAEALY
LpCOEB H2d ALNEPTIDYG FQRLQKVIPR HPGDPERLPK EVILKRAADL VEALYGLPHY NQEIILKRAA DIAEALY
LpCOEB ALNEPTIDYG FQRLQKVIPR HPGDPERLPK ---------- ---------- --EIILKRAA DIAEALY
LpCOEA ALNEPTIDYG FQRLQKLIPR HPGDPERLPK ---------- ---------- --EIILKRAA DIAEALY
CiCOE ALNEPTLDYG FQRLLKTVPR HPGDPERLPK ---------- ---------- --EIILKRAA DVMEAVI
BfCOE SLNEPTIDYG FQRLAKLVPR HPGDPERLPK ---------- ---------- --EIILKRAA DLAEAIY
Exon 11
Exon 13
C: Splicing of TA1 region in COE-B
Exon 13
Exon 14
LpCOEB TA1v1 QNQLPALGNS P-AHHAMMGV NSFGGQLAVS VSEPGQSVEQ AMLSHAPGRY PCYSRNTSSV SPRGYVPSST
LpCOEB TA1v2 QNQLPALGNS P-AHHAMMGV NSFGGQLAVS VSEPGQSVEQ AMLSHAPG-- --YSRNTSSV SPRGYVPSST
LpCOEA HNQLPAINNS PPGLPGMMGM NSFGGQLAIS VSEASQAADQ G--------- --YTRNSSSV SPHGYVPSST
Exon 13
Exon 14
D: Splicing of TA2 region in COE-B
Exon 14
Exon 15
LpCOEB TA2v1 PQQSNYNTVS NSMNGYGNAG -MGNLGVPGS PGFLNGSSAS SPYAIMPSSP PVAAPPSITL PSNCSGAGGV
LpCOEB TA2v2 PQQSNYNTVS NSMNGYGNAG -MGNLGVPGS PGFLNGSSAS SPYAIMPSSP PVAAPPSITL PSNCSGAGGV
LpCOEB TA2v3 PQQSNYNTVS NSMNGYGNAG -MGNLGVPGS PGFLNGSSAS SPYA------ ---------- ----------
LpCOEB TA2v4 PQQSNYNTVS NSMNGYGNAG -MGNLGVPGS PGFLNGSSAS SPYA------ ---------- ----------
LpCOEA PQQSHYSSVS SSMNGYGVSG GMGGLGMGAS PGYLT-SSGG SPYGSTCPAP LFSS------ ----------
Exon 15
Exon 16
Exon 17
LpCOEB TA2v1 FSFSPAHMIS AVKQKSAFAP VVRPQASPPP NCTSATGNAL QGNFGHGLVA QFKDSVLAMS GLVVPPM
LpCOEB TA2v2 FSFSPAHMIS AVKQKSAFAP VVRPQASPPP NCTSATGNAL Q--------- -------AMS GLVVPPM
LpCOEB TA2v3 ---------- ---------- ---------- ---------- -GNFGHGLVA QFKDSVLAMS GLVVPPM
LpCOEB TA2v4 ---------- ---------- ---------- ---------- ---------- -------TMS GLVVPPM
LpCOEA ---------- -VKQKSAFAP VVRPQTSPPP TCTSANANGL Q--------- -------VMS GLVVPPM
Exon 14
Exon 15
Exon 17
E: Splicing of DBD region in COE-A
Exon 7
Exon 8
Exon 8
Exon 9
LpCOEB FQVVISTTVN VDGHVLAVSD NMFVHNNSKH GRRARRLDPS E--------- -AATPCIKAI SPSEGWTTGG
LpCOEA DBDv1 FQVVVSTTVH VDGHVLAVSD NMFVHNNSKH GRRARRIDPS EGTPPPSYLD NAATPCIKAI SPSEGWTTGG
LpCOEA DBDv2 FQVVVSTTVH VDGHVLAVSD NMFVHNNSKH GRRARRIDPS E--------- -AATPCIKAI SPSEGWTTGG
Exon 7
Exon 8
Exon 8
Exon 9

Figure 2

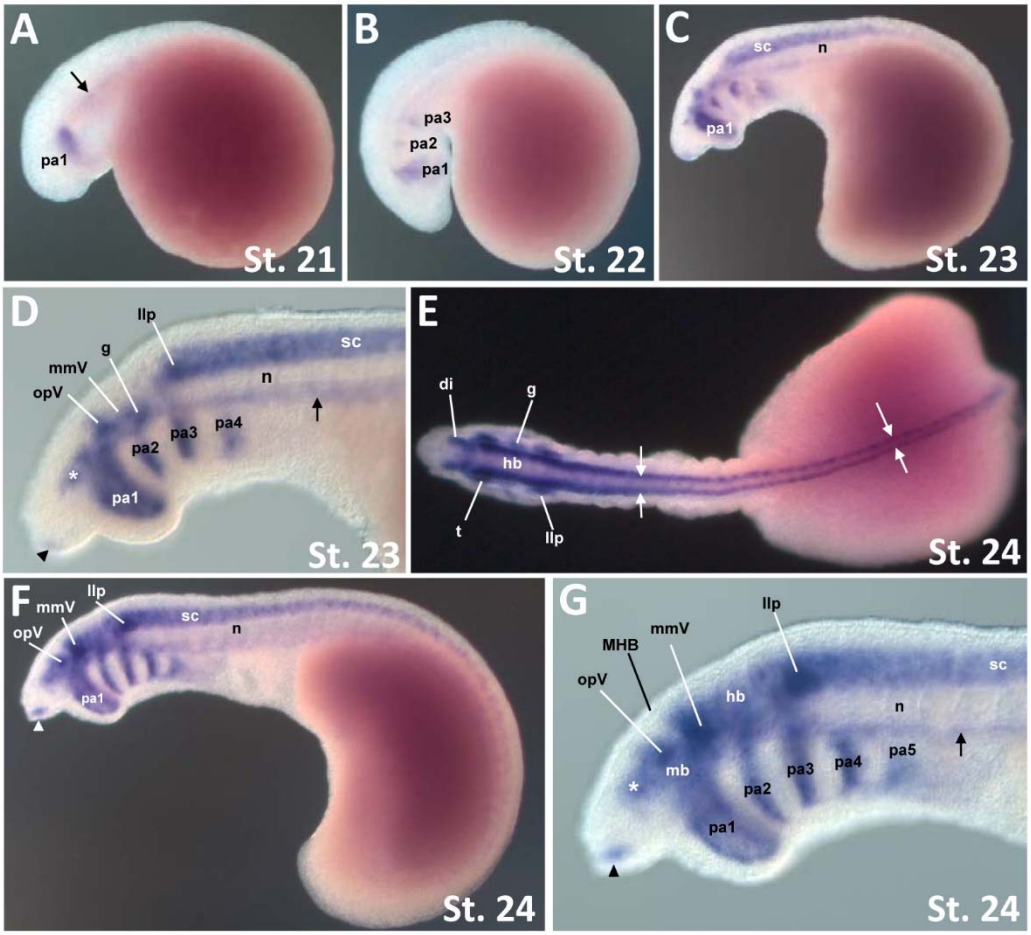
A
pa1
St. 21
B
pa3
pa2
pa1
St. 22
C
sc
n
pa1
St. 23
D
llp
g
sc
mmV
n
opV
pa3
pa4
pa2
*
pa1
St. 23
E
di
g
hb
t
llp
St. 24
F
mmV
llp
opV
sc
n
pa1
St. 24
G
llp
mmV
MHB
sc
opV
hb
n
pa5
mb
pa3
pa4
*
pa2
pa1
St. 24

Figure 3

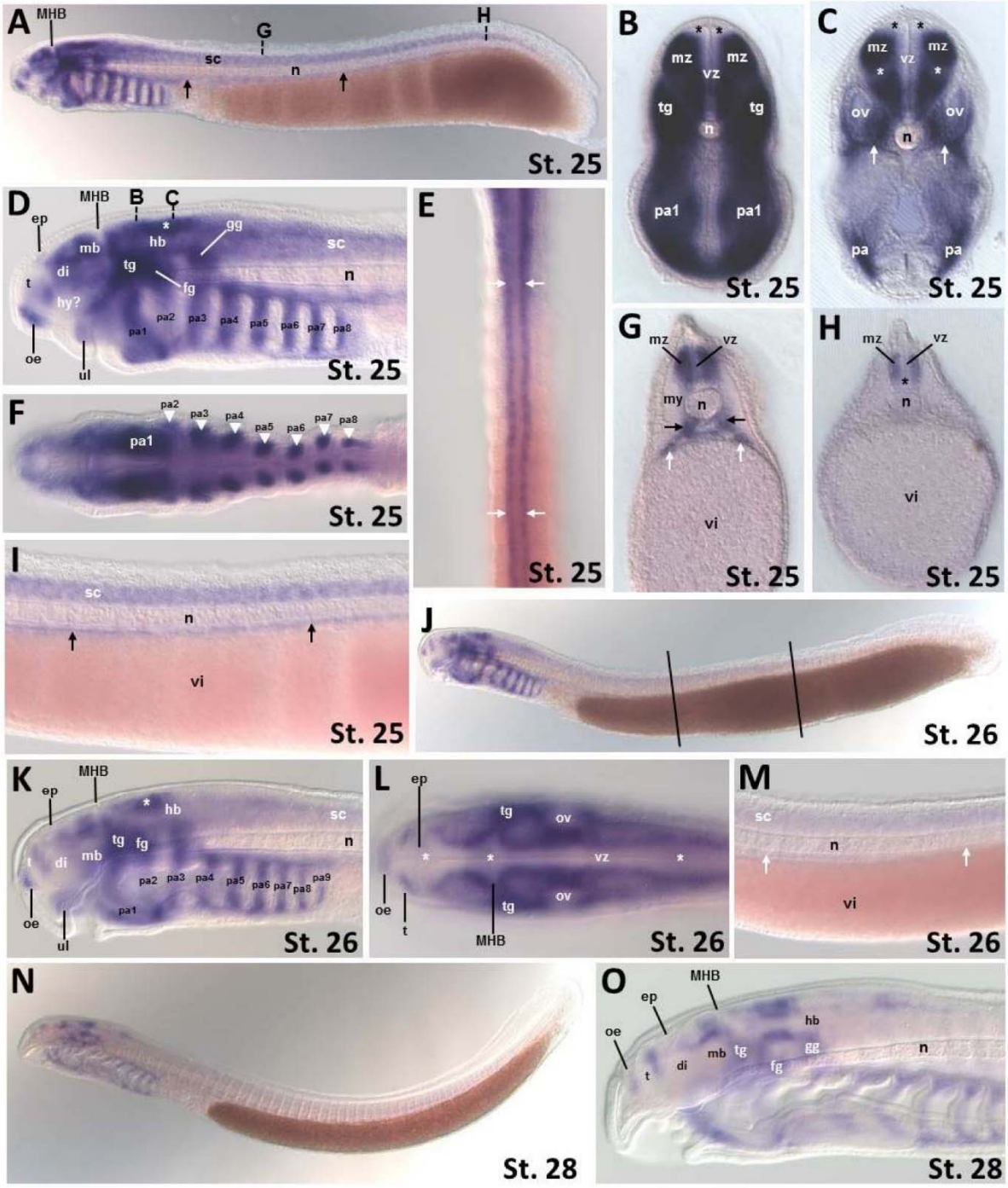
A
MHB
G
H
sc
n
St. 25
B
mz
vz
tg
n
pa1
St. 25
C
mz
vz
ov
n
pa
St. 25
D
MHB
B
C
ep
hb
gg
sc
mb
tg
di
t
fg
n
hy?
pa1
pa2
pa3
pa4
pa5
pa6
pa7
pa8
oe
ul
St. 25
E
St. 25
G
mz
vz
my
n
vi
St. 25
H
mz
vz
n
vi
St. 25
F
pa2
pa3
pa4
pa5
pa6
pa7
pa8
pa1
St. 25
I
sc
n
vi
St. 25
J
St. 26
K
MHB
ep
hb
sc
n
tg
fg
t
di
mb
pa2
pa3
pa4
pa5
pa6
pa7
pa8
pa9
pa1
oe
ul
St. 26
L
ep
tg
ov
vz
oe
t
MHB
St. 26
M
sc
n
vi
St. 26
N
St. 28
O
MHB
ep
oe
hb
gg
n
mb
tg
fg
di
t
St. 28

Figure 4

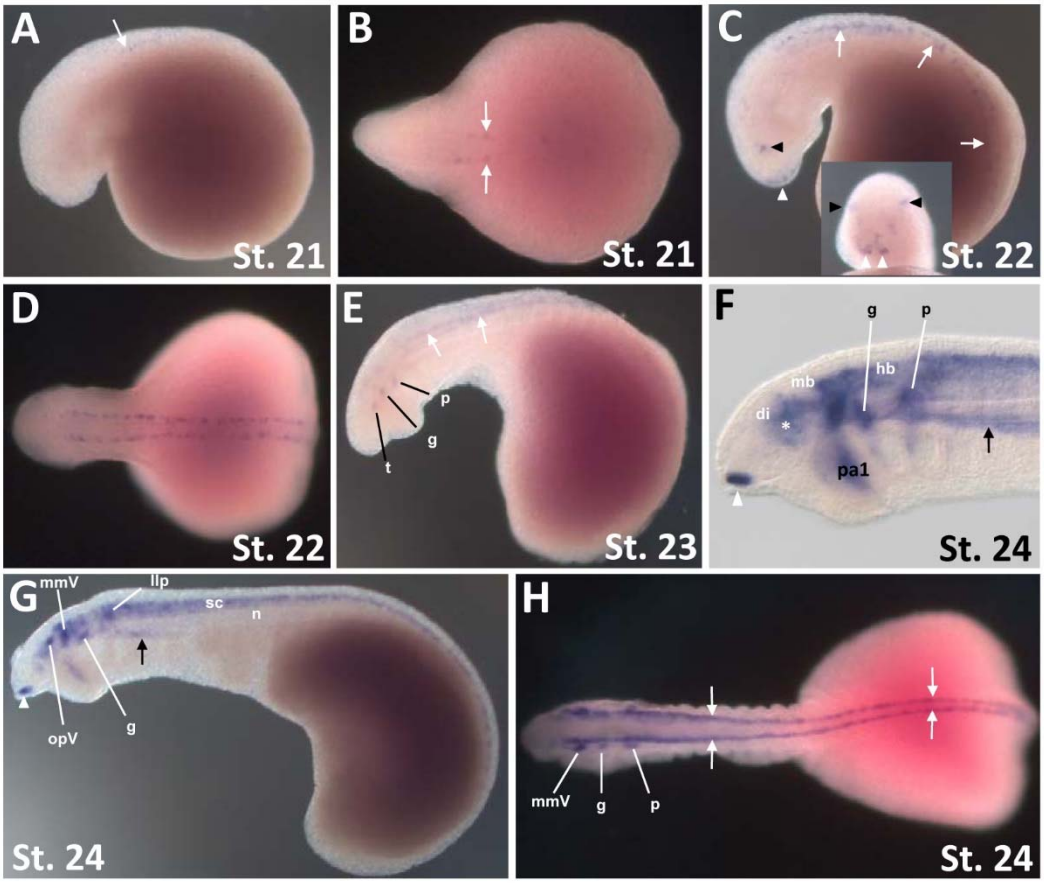
A
St. 21
B
St. 21
C
St. 22
D
St. 22
E
p
g
t
St. 23
F
g
p
mb
hb
di
*
pa1
St. 24
G
mmV
llp
sc
n
opV
g
St. 24
H
mmV
g
p
St. 24

Figure 5

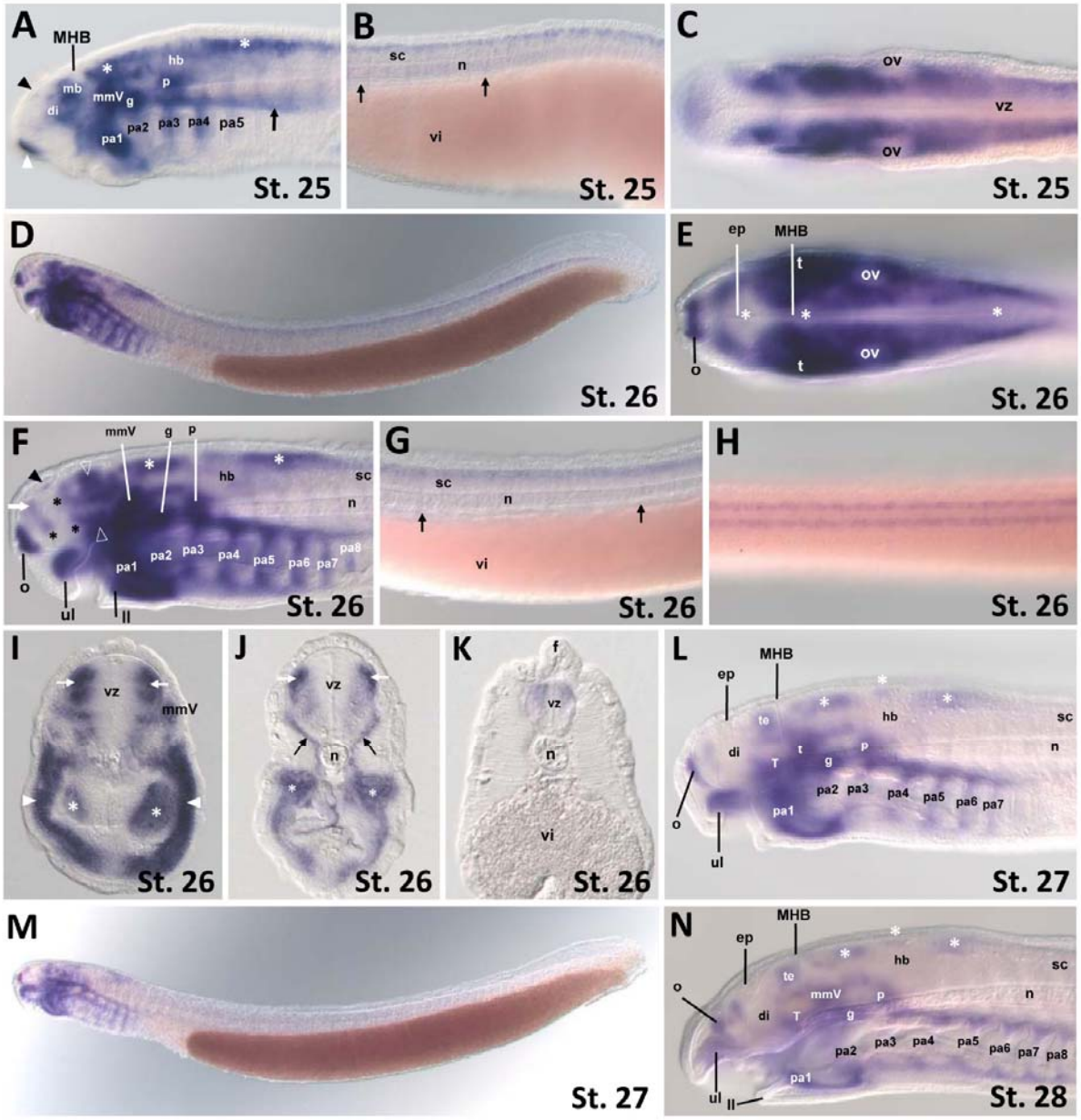
A
MHB
hb
mb
p
mmV
g
di
pa1
pa2
pa3
pa4
pa5
St. 25
B
sc
n
vi
St. 25
C
ov
vz
ov
St. 25
D
St. 26
E
ep
MHB
t
ov
t
ov
o
St. 26
F
mmV
g
p
hb
sc
n
pa1
pa2
pa3
pa4
pa5
pa6
pa7
pa8
o
ul
II
St. 26
G
sc
n
vi
St. 26
H
St. 26
I
vz
mmV
St. 26
J
vz
n
St. 26
K
f
vz
n
vi
St. 26
L
MHB
ep
te
hb
sc
di
t
p
n
T
g
pa2
pa3
pa4
pa5
pa6
pa7
pa1
o
ul
St. 27
M
St. 27
N
MHB
ep
hb
sc
te
o
mmV
p
n
di
T
g
pa2
pa3
pa4
pa5
pa6
pa7
pa8
pa1
ul
II
St. 28

Figure 6

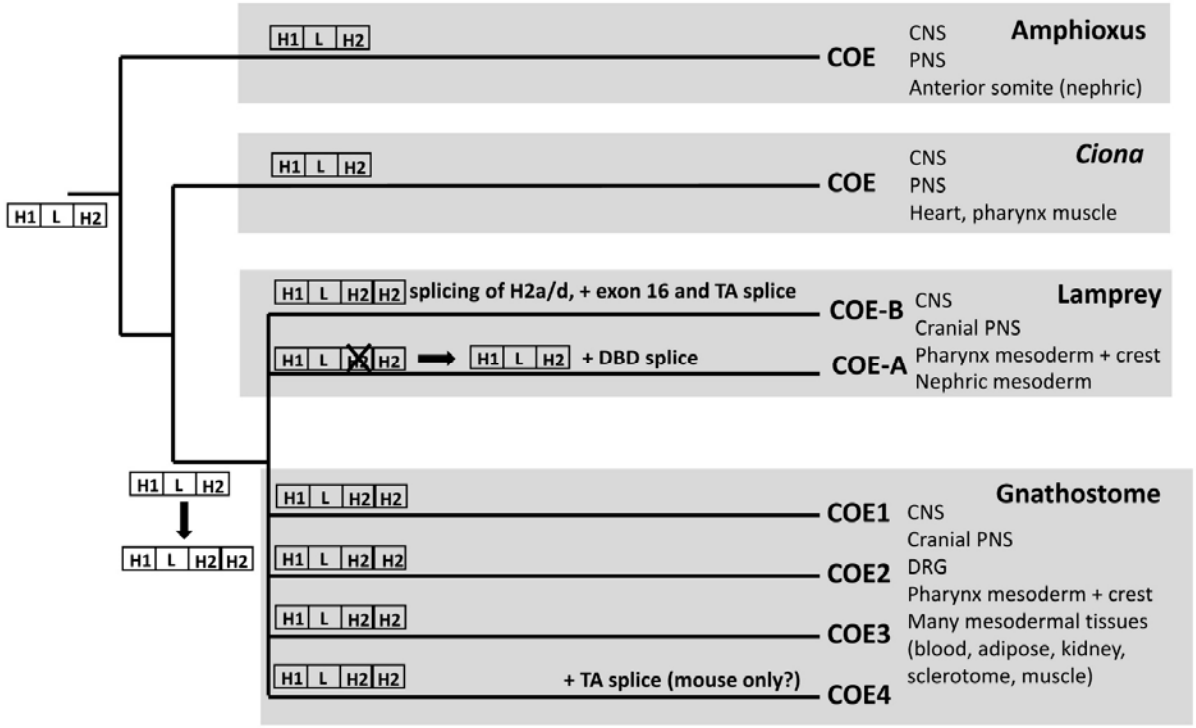
H1 L H2
COE
CNS
PNS
Anterior somite (nephric)
Amphioxus
H1 L H2
COE
CNS
PNS
Heart, pharynx muscle
Ciona
H1 L H2 H2
splicing of H2a/d, + exon 16 and TA splice
COE-B
CNS
Cranial PNS
Pharynx mesoderm + crest
H1 L H2
H1 L H2
+ DBD splice
COE-A
Nephric mesoderm
Lamprey
H1 L H2
H1 L H2 H2
COE1
CNS
Cranial PNS
DRG
Pharynx mesoderm + crest
Many mesodermal tissues
(blood, adipose, kidney,
sclerotome, muscle)
H1 L H2 H2
COE2
H1 L H2 H2
COE3
H1 L H2 H2
+ TA splice (mouse only?)
COE4
Gnathostome

Figure 7