Research and Reports in Biochemistry

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Comparative studies of vertebrate endothelin-converting enzyme-like 1 genes and proteins

This article was published in the following Dove Press journal:
Research and Reports in Biochemistry
4 January 2013
Number of times this article has been viewed

Roger S Holmes[1,2]
Laura A Cox[1]

[1]Department of Genetics and Southwest National Primate Research Center, Texas Biomedical Research Institute, San Antonio, TX, USA; [2]Eskitis Institute for Cell and Molecular Therapies and School of Biomolecular and Physical Sciences, Griffith University, Nathan, Queensland, Australia

Abstract: Endothelin-converting enzyme-like 1 (ECEL1) is a member of the M13 family of neutral endopeptidases which play an essential role in the neural regulation of vertebrate respiration. Genetic deficiency of this protein results in respiratory failure soon after birth. Comparative ECEL1 amino acid sequences and structures and *ECEL1* gene locations were examined using data from several vertebrate genome projects. Vertebrate ECEL1 sequences shared 66%–99% identity as compared with 30%–63% sequence identities with other M13-like family members, ECE1, ECE2, and NEP (neprilysin or MME). Three N-glycosylation sites were conserved among most vertebrate ECEL1 proteins examined. Sequence alignments, conserved key amino acid residues, and predicted secondary and tertiary structures were also studied, including cytoplasmic, transmembrane, and luminal sequences and active site residues. Vertebrate *ECEL1* genes usually contained 18 exons and 17 coding exons on the negative strand. Exons 1 and 2 of the human *ECEL1* gene contained 5′-untranslated (5′-UTR) regions, a large CpG island (CpG256), and several transcription factor binding sites which may contribute to the high levels of gene expression previously reported in neural tissues. Phylogenetic analyses examined the relationships and potential evolutionary origins of the vertebrate *ECEL1* gene with six other vertebrate neutral endopeptidase M13 family genes. These suggested that *ECEL1* originated in an ancestral vertebrate genome from a duplication event in an ancestral neutral endopeptidase *M13*-like gene.

Keywords: vertebrates, amino acid sequence, ECEL1, ECE1, ECE2, KELL, NEP, NEPL1, PHEX, evolution

Introduction

Endothelin-converting enzyme L1 (ECEL1, also called Xce protein; EC 3.4.24.-) is one of at least seven members of the M13 family of neutral endopeptidases which are zinc-containing type II transmembrane enzymes.[1–3] Neprilysin (also called atriopeptidase, common acute lymphocyte antigen, and enkephalinase; EC 3.4.24.11) is the best characterized M13 family member, and participates in the inactivation of signaling peptides involved in regulating blood pressure, neuronal activity, and the immune system.[4–6] Endothelin-converting enzyme 1 (ECE1; EC 3.4.24.71) and endothelin-converting enzyme 2 (ECE2; EC 3.4.24.71) are related M13 family members,[7–10] which play a role in the processing of regulatory peptides in the body, although *ECE2* also encodes a methyltransferase domain.[11]

The gene encoding ECEL1 (*ECEL1* in humans; *Ecel1* in mice) is expressed at very high levels in the central nervous system, sympathetic ganglia, uterine subepithelial cells, and in the primordia of the stratium, hypothalamus, and cranial motor nuclei.[12,13]

Correspondence: Roger S Holmes
Eskitis Institute for Cell and Molecular Therapies, School of Biomolecular and Physical Sciences,
Griffith University, Nathan,
QLD 4111, Australia
Tel +617 3735 6008
Email r.holmes@griffith.edu.au

ECEL1 gene expression is also induced in response to various forms of brain injury, although this induction is restricted to neurons.[13,14] This neuronal damage induction feature for mammalian *ECEL1* has led to another name for this gene, ie, *DINE* or damage-induced neuronal endopeptidase,[15] which is regulated by leukemia inhibitory factor and other transcription factors.[14,16] The upregulation of *ECEL1* following neural damage has provided an insight into the mechanism of protection against neuronal death by the activation of antioxidant enzymes, such as Cu/Zn superoxide dismutase, Mn superoxide dismutase, and glutathione peroxidase through the proteolytic activity of ECEL1.[13,16] Moreover, high expression of *ECEL1* is associated with a favorable prognosis in human neuroblastomas.[17]

Studies of *Ecel1*$^{-}$/*Ecel1*$^{-}$ knockout mice have shown that ECEL1 deficiency causes neonatal lethality as a result of respiratory failure soon after birth, even though there were no abnormalities in organs and tissues or pulmonary surfactant proteins.[18–20] Consequently, these studies have suggested that ECEL1 performs an essential role in the nervous control of respiration, although the natural substrate for ECEL1 has not been described. Structures of three mammalian *ECEL1* genes have been examined, including in man, rat, and mouse, and shown to contain 18 exons of DNA encoding ECEL1 sequences, which undergo exon shuffling, generating isoproteins in each case.[15,21,22]

This paper reports the predicted gene structures and amino acid sequences for several vertebrate *ECEL1* genes and proteins, the predicted structures for vertebrate ECEL1 proteins, a number of potential sites for regulating human *ECEL1* gene expression, and the structural, phylogenetic, and evolutionary relationships for these genes and enzymes with those for six other vertebrate M13 neutral endopeptidase gene families.

Materials and methods

Vertebrate *ECEL1* gene and protein identification

BLAST (Basic Local Alignment Search Tool) studies were undertaken using web tools from the National Center for Biotechnology Information (NCBI).[23,24] Protein BLAST analyses used vertebrate ECEL1, ECE1, and ECE2 amino acid sequences described previously (Table 1).[12,21,25] Nonredundant protein sequence databases for several mammalian genomes were examined using the BLASTP algorithm, including human (*Homo sapiens*), gorilla (*Gorilla gorilla*), orang-utan (*Pongo abelii*), marmoset (*Callithrix jacchus*), cow (*Bos taurus*), mouse (*Mus musculus*), rat (*Rattus norvegicus*), guinea pig (*Cavia porcellus*), pig (*Sus scrofa*), dog (*Canis familiaris*), opossum (*Monodelphis domestica*), chicken (*Gallus gallus*), lizard (*Anolis carolinensis*), frog (*Xenopus tropicalis* and *Xenopus laevis*), puffer fish (*Tetraodon nigroviridis*, *Fugu rubripes*), and zebrafish (*Danio rerio*).[26] This procedure produced multiple BLAST "hits" for each of the protein databases which were individually examined and retained in FASTA format, and a record kept of the sequences for predicted mRNAs and encoded ECE-like proteins. These records were derived from annotated genomic sequences using the gene prediction method, ie, GNOMON and predicted sequences with high similarity scores for human ECEL1, ECE1, and ECE2. Predicted ECE-like protein sequences were obtained in each case and subjected to analyses of predicted protein and gene structures.

BLAT analyses were undertaken subsequently for each of the predicted ECEL1, ECE1, ECE2, KELL, NEP, NEPL1, and PHEX amino acid sequences using the UC Santa Cruz Genome Browser with the default settings to obtain the predicted locations for each of the vertebrate *ECEL1* genes, including predicted exon boundary locations and gene sizes.[27] BLAT analyses were similarly undertaken for other vertebrate ECE-like genes using previously reported sequences in each case (see Table 1 and Supplementary Table 1). Structures for human and mouse isoforms (splicing variants) were obtained using the AceView website to examine predicted gene and protein structures.[22]

Predicted structures and properties of vertebrate ECEL1

Predicted secondary and tertiary structures for human and other vertebrate ECEL1, ECE1, and ECE2 proteins were obtained using the Swiss-Model web server with the reported tertiary structure for human ECE1 complexed with phosphoramidon (PDB:3dwbA) serving as the template in each case.[28,30] The following modeled residue ranges were observed: 103–775 for human ECEL1; 96–763 for chicken ECEL1; 101–770 for human ECE1; and 162–883 for human ECE2 (which excludes the methyltransferase region). Predicted secondary structures for human and mouse N-terminal regions of ECEL1 (residues 2–60), ECE1 (residues 2–68), and ECE2 (residues 2–178, including the methyltransferase domain) excluded from the Swiss-Model analyses, were obtained using PSIPRED prediction web tools (http://bioinf.cs.ucl.ac.uk/psipred/).[30] Molecular weights, N-glycosylation sites, and predicted transmembrane, cytosolic, and luminal sequences for vertebrate ECEL1, ECE1,

and ECE2 proteins were obtained using Expasy web tools (http://au.expasy.org/tools/pi_tool.html).[31] The identification of conserved domains for ECEL1 was conducted using NCBI web tools (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi).[32]

Comparative human (*ECEL1*) and mouse (*Ecel1*) gene expression

The genome browser (http://genome.ucsc.edu) was used to examine GNF Expression Atlas 2 data using various expression chips for human and mouse *ECEL1* genes (http://biogps.gnf.org).[33] Gene array expression "heat maps" were examined for comparative gene expression levels among human and mouse tissues showing high (red), intermediate (black), and low (green) expression levels.

Phylogeny studies and sequence divergence

Phylogenetic analyses were undertaken using the http://phylogeny.fr platform.[34] Alignments of vertebrate ECEL1 sequences, six other vertebrate M13 neutral endopeptidase sequences, and two nematode (*Caenorhabditis elegans*) neprilysin-like (*NEP*) sequences were assembled using PMUSCLE (Table 1).[35] Ambiguous alignment regions, including the methyltransferase region of ECE2, were excluded prior to phylogenetic analysis, yielding alignments for comparisons of vertebrate ECEL1 sequences with vertebrate neutral endopeptidase and nematode (*C. elegans*) M13-like sequences. The phylogenetic tree was constructed using the maximum likelihood tree estimation program, PHYML.[36,37]

Results and Discussion

Alignments of vertebrate ECEL1 amino acid sequences

The deduced amino acid sequences for pig (*S. scrofa*), chicken (*G. gallus*), and frog (*X. tropicalis*) ECEL1 are shown in Figure 1, together with previously reported sequences for man and mouse ECEL1 (Table 1).[12,21] Alignments of human with other vertebrate ECEL1 sequences examined were 66%–99% identical, suggesting that these are products of the same family of genes, whereas comparisons of sequence identities for vertebrate ECEL1 proteins with human ECE1 and ECE2 proteins exhibited lower levels of sequence identities (34%–37% and 30%–37% respectively), indicating that these are members of distinct *ECE*-like gene families (Table 1).

The amino acid sequences for vertebrate ECEL1 proteins contained between 763 (for chicken ECEL1) and 785 (for puffer fish [*F. rubripes*] ECEL1) residues whereas most mammalian ECEL1 sequences contained 775 amino acids (Figure 1 and Table 1). Previous studies have reported several key regions and residues for human and mouse ECEL1 proteins (human ECEL1 amino acid residues were identified in each case). These included an N-terminus cytoplasmic tail (58 residues excluding the N-terminus methionine) followed by a hydrophobic transmembrane 23-residue segment which may anchor the enzyme to the plasma membrane and the endoplasmic reticulum.[12,25] A comparison of 16 vertebrate ECEL1 sequences for these N-terminal regions revealed a high degree of conservation, especially for residues 5–23, which were substantially invariant among all vertebrate sequences examined (see Figures 1 and 2), yielding the following conserved sequence: Tyr-Ser-Leu/Met-Thr-Ala-His-Tyr-Asp-Glu-Phe-Gln-Glu-Val-Lys-Tyr-Val-Ser-Arg/Lys-Cys/Tyr. The biochemical role for this conserved N-terminal cytoplasmic tail sequence remains to be determined. A variable sequence region containing multiple glycine residues (residues 24–51 for human ECEL1) was followed by a predicted transmembrane spanning segment (human ECEL1 residues 60–82), which was predominantly invariant among the 16 vertebrate ECEL1 sequences examined. This transmembrane region was further characterized by conserved "book-end" sequences as Arg60-Arg61-Glu62 at the N-terminal end and Lys86-Tyr78-Leu79 at the C-terminal end of the membrane anchoring segment, which may contribute to the membrane spanning properties reported.[12,25] Residues 83–775 of the human ECEL1 sequence were identified using bioinformatics as a large peptidase family M13 neprilysin-like domain which is involved in proteolysis of neuroactive peptides in the body.[32] This C-terminal region is predicted to be localized in the luminal zone of the endoplasmic reticulum and to contain an active M13-like peptidase capable of metabolizing physiologically active peptides. Three N-glycosylation sites were consistently found for these vertebrate ECEL1 peptidase sequences, namely Asn255-Ser256-Ser257, Asn322-Ile323-Thr324n, and Asn656-Phe657-Thr658 (Figure 1). In addition, 11 conserved active site residues were observed for these vertebrate ECEL1 sequences, deduced from the M13 peptidase domain active site residues reported from NCBI domain studies.[32] These included an active site catalytic residue (Glu613); three residues involved in binding the active site, zinc (His612, His616, and Glu672), deduced from three-dimensional studies of the related enzymes, neprilysin and ECE1,[6,29] and seven other active site residues, also deduced from the neprilysin and ECE1 tertiary structures, namely

Table 1 Vertebrate *ECEL1*, *ECE1*, and *ECE2* genes and proteins

ECEL1 Gene	Species	RefSeq ID Ensembl/NCBI[a]	GenBank ID	UNIPROT ID	Amino acids	Chromosome location
Human	*Homo sapiens*	NM_004826	BC050453	O95672	775	2:233,344,866-233,351,363
Gorilla	*Gorilla gorilla*	ENSGGOT00000008202[a]	na	G3QYP2	775	2B:121,699,588-121,706,085
Orang-utan	*Pongo abelii*	XP_002813036[a]	na	H2P8X6	775	2B:124,886,954-124,893,595
Marmoset	*Callithrix jacchus*	XP_002749942[a]	na	na	773	6:148,896,501-148,903,400
Mouse	*Mus musculus*	NM_0213306	BC057569	Q9JMI0	775	1:89,044,565-89,051,564
Rat	*Rattus norvegicus*	NM_021776	AB026293	Q9JHL3	775	9:85,939,303-85,945,981
Cow	*Bos taurus*	XP_003585830[a]	na	E1BJE2	775	2:125,622,584-125,629,649
Pig	*Sus scrofa*	XP_003133775[a]	na	F1SMT6	775	15:125,282,534-125,289,804
Dog	*Canis familiaris*	XP_543287[a]	na	E2QY56	780	25:44,120,538-44,127,947
Rabbit	*Oryctolagus cuniculus*	XP_002721503[a]	na	G1SXE9	775	Un0038:101,642-107,854*
Guinea pig	*Cavia porcellus*	XP_003474642[a]	na	H0V3J9	765	13:29,247,384-29,253,642
Opossum	*Monodelphis domestica*	ENSMODT00000003255[a]	na	F7FUF8	776	2:537,252,473-537,269,989
Chicken	*Gallus gallus*	XP_422744[a]	na	F1NKL6	763	9:15,047,449-15,053,456
Lizard	*Anolis carolensis*	XP_003225531[a]	na	na	767	Un_GL3343304-586, 905-674,203*
Frog	*Xenopus tropicalis*	XP_002937386[a]	na	na	764	GL172921:408,042-446,553*
Puffer fish	*Tetraodon nigroviridis*	ENSTNIT00000017527[a]	na	H3C5L5	776	16:6,933,748-6,943,633
Puffer fish	*Fugu rubripes*	ENSTRUT00000032528[a]	na	H2U638	785	11:10,041,781-10,053,371
ECE1 gene						
Human	*Homo sapiens*	NM_001113347	BC117256	P42892	770	1:21,546,451-21,616,907
Mouse	*Mus musculus*	NM_199307	AB060648	Q4PZA2	769	4:137,469,641-137,518,818
Pig	*Sus scrofa*	XP_003356179[a]	na	F1SU04	771	6:54,780,088-54,836,122
Chicken	*Gallus gallus*	NM_204717	AF98287	Q9DGN6	752	21:6,600,258-6,607,612
Frog	*Xenopus laevis*	NM_001086909	BC108485	Q32NT6	766	GL174075:62,421-76,377^
Zebrafish	*Danio rerio*	NP_001071260	BC125952	F1RAS8	752	11:28,797,446-28,829,975
ECE2 gene						
Human	*Homo sapiens*	NM_014693	BC005835	O60344	883	3:183,967,483-184,010,023
Mouse	*Mus musculus*	NM_177940	BC115541	B2RQR8	881	16:20,611,698-20,645,306
Pig	*Sus scrofa*	XP_003358763[a]	na	I3LTQ3	883	6:54,780,088-54,805,725[#]
Chicken	*Gallus gallus*	XP_003641814[a]	na	F1NW61	768	9:15,356,256-15,361,432[#]
Frog	*Xenopus tropicalis*	XP_002935189[a]	na	F7CSC4	766	GL172768:567,488-594,018^
Zebrafish	*Danio rerio*	XP_00133328[a]	na	na	759	15:4,034,811-4,106,595
Worm NEP-like genes						
NEP2	*Caenorhabditis elegans*	NM_064089	CAA93782	Q18673	754	II:11,654,392-11,657,083
NEP22	*Caenorhabditis elegans*	NM_077127	na	Q22763	798	X:9,034,601-9,038,311

Notes: RefSeq: the reference amino acid sequence; [a]predicted Ensembl amino acid sequence; na, not available; GenBank IDs are derived from NCBI http://www.ncbi.nlm.nih.gov/genbank/; Ensembl ID was derived from Ensembl genome database http://www.ensembl.org; UNIPROT refers to UniprotKB/Swiss-Prot IDs for individual ECE-like proteins (see http://kr.expasy.org); Un-refers to unknown chromosome; GL refers to an unknown scaffold; bps refers to base pairs of nucleotide sequences; ^refers to incomplete gene sequence; pI refers to theoretical isoelectric points; the number of coding exons are listed; high % identities are shown in bold. High gene expression levels are in bold.

Asn571, Ala572, Ile609, Phe715, Ala716, His737, and Arg743 (Figures 1 and 2). Most of the other residues in the active site and C-terminal ECEL1 regions were also predominantly conserved among all of the vertebrate ECEL1 enzymes examined, reflecting strong functional roles for these sequences.

Predicted secondary and tertiary structures for vertebrate ECEL1

Predicted secondary structures for mammalian, chicken, and frog ECEL1 sequences were examined, particularly for the luminal sequences (Figure 1), using the known structure reported for an M13 family peptidase, human ECE1.[29] The α-helix and β-sheet structures were identical in each case, with 31 α-helices and 12 β-sheet structures being observed. Of particular interest were α-helices 24 and 27, which contained the predicted active site residues for human ECEL1, and the β1-sheet in the conserved cytoplasmic N-terminus sequence, the only secondary structure predicted for this region.

Predicted tertiary structures for human and chicken ECEL1 are shown in Figure 3, together with human ECE1 and the predicted structure for the M13 peptidase domain

Coding exons (strand)	Gene size bps	Subunit MW	Gene expression	pI	% identity with human ECE1	% identity with human ECE2	% identity with human ECEL1
17 (–ve)	6,498	87,791	0.4	6.6	36	32	**100**
17 (–ve)	6,498	87,858	na	7.1	36	32	**99**
17 (–ve)	6,642	87,758	na	6.7	36	32	**99**
17 (–ve)	6,900	87,624	na	6.6	36	32	**97**
17 (–ve)	7,000	87,993	0.7	7.9	36	32	**95**
17 (–ve)	6,679	87,970	0.1	7.6	36	32	**95**
17 (–ve)	7,066	87,782	na	7.3	36	32	**96**
17 (–ve)	7,271	87,995	na	7.6	36	32	**96**
17 (–ve)	7,410	88,073	na	7.3	36	32	**95**
17 (–ve)	6,213	87,886	na	7.3	36	32	**95**
17 (+ve)	6,259	86,968	na	7.9	36	32	**94**
17 (–ve)	17,517	86,954	na	7.1	34	30	**86**
17 (–ve)	6,008	87,744	na	6.5	37	36	**80**
17 (+ve)	87,299	88,872	na	6.6	36	31	**78**
17 (+ve)	38,512	88,637	na	6.4	36	37	**74**
18 (–ve)	9,886	89,111	na	6.9	34	30	**67**
17 (–ve)	11,591	91,637	na	6.9	34	30	**66**
19 (–ve)	70,457	87,163	**4.4**	5.6	**100**	54	36
18 (+ve)	49,178	87,085	**3.7**	5.6	**94**	54	37
19 (–ve)	56,035	87,360	na	5.6	**95**	55	36
17 (+ve)	7,355	84,986	na	5.1	**81**	55	37
(+ve)^	13,957^	87,211	na	5.6	**76**	53	37
18 (–ve)	32,530	85,206	na	5.8	**66**	51	35
19 (+ve)	42,451	99,773	**1.4**	5.0	54	**100**	32
19 (+ve)	33,609	99,480	**1.0**	5.0	54	**92**	32
7 (–ve)#	25,638#	99,824	na	5.0	54	**95**	32
19 (–ve)	5,177#	86,536	na	5.1	62	**68**	36
19 (–ve)	26,531	86,136	na	5.1	61	**64**	37
19 (–ve)	71,785	85,699	na	5.0	63	**63**	37
4 (–ve)	2,692	86,945	**1.9**	5.7	32	29	28
12 (+ve)	3,711	88,549	na	5.2	32	29	28

for human ECE2.[29] The tertiary structure of the extracellular domain (residues 90–770) for human ECE1 (with the metalloprotease inhibitor phosphoramidon) is also similar to that described for human neprilysin (NEP) as well as the predicted structures for human and chicken ECEL1.[6] Thirty-one α-helices and 12 β-sheet structures have been described for both M13-like metallopeptidases (ECE1 and NEP), which were also observed for the predicted structures of human and chicken ECEL1. In addition, two major domains for these enzymes were observed, that enclose a large cavity previously shown to contain the active site of the enzyme. The N-terminal of these two domains has been shown to have a fold similar to that of thermolysin and contains the active site residue, whereas the other domain may serve to control access of substrates to the active site.[29] Overall, the predicted human and chicken ECEL1 structures closely resemble that reported for human ECE1.

Alignments of human and mouse ECEL1, ECE1, and ECE2 amino acid sequences

The amino acid sequences for human and mouse ECEL1, ECE1, and ECE2 (see Table 1) are aligned in Figure 4.

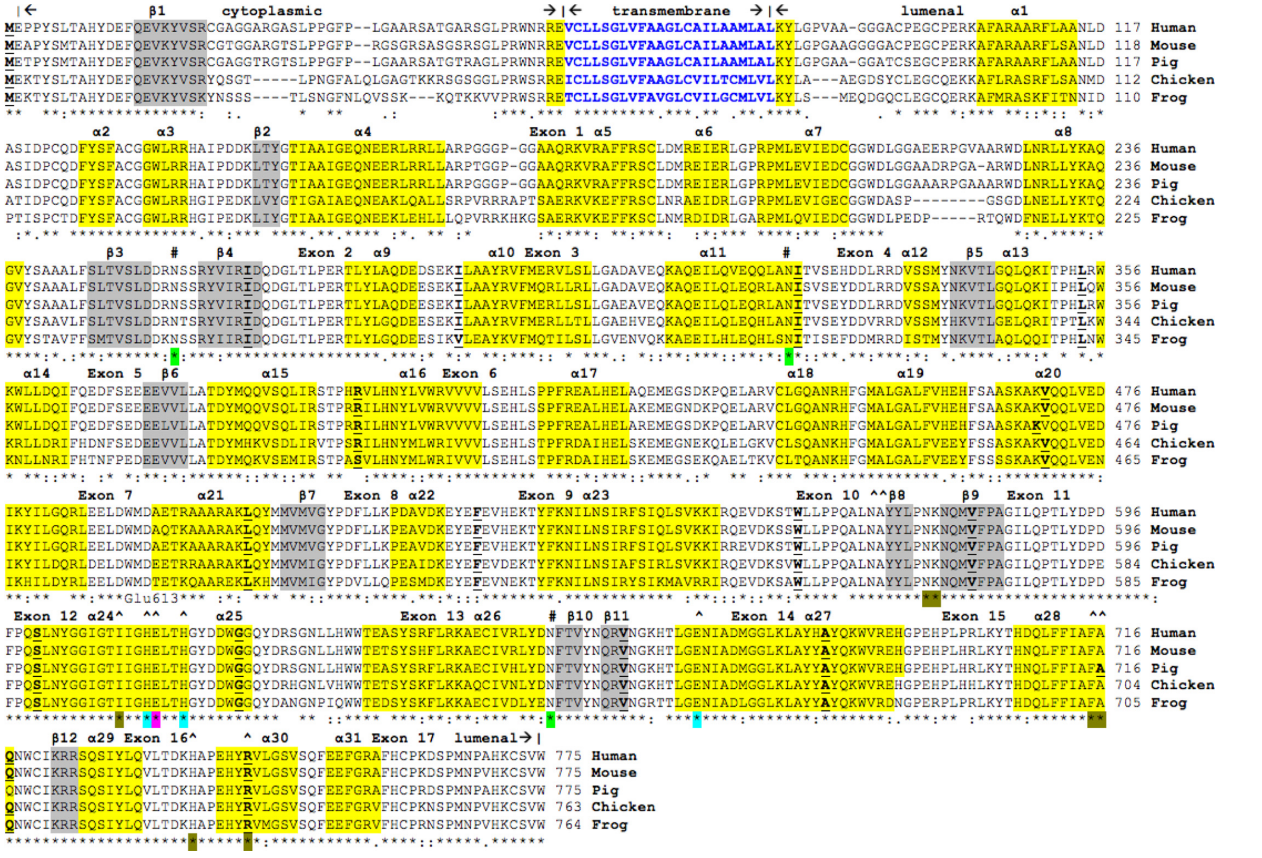

Figure 1 Amino acid sequence alignments for vertebrate ECEL1 sequences.

Notes: See Table 1 for sources of ECEL1 sequences; *shows identical residues for ECEL1 subunits; similar alternate residues; dissimilar alternate residues; predicted transmembrane residues are shown in blue; N-glycosylated and potential N-glycosylated Asn sites are in green and #; the endopeptidase active site Glu612 (human ECEL1) is shown in pink; other active site residues are shown in khaki and ^; three predicted zinc-binding residues are shown in blue; predicted α-helices for vertebrate ECEL1 are in shaded yellow and numbered in sequence from the start of the predicted transmembrane domain; predicted β-sheets are in shaded gray and also numbered in sequence from the N-terminus; bold underlined font shows residues corresponding to known or predicted exon start sites; exon numbers refer to human *ECEL1* gene exons; note the three major domains identified as cytoplasmic (N-terminal tail); transmembrane (for linking ECEL1 to the endoplasmic reticulum); and luminal regions (C-terminal tail localized in the lumen of the endoplasmic reticulum).

The sequences were 32%–54% identical and showed similarities in several key features and residues, including cytoplasmic N-terminal residues; transmembrane helical regions; predicted N-glycosylation sites for human and mouse ECEL1 (three sites), ECE1 (nine sites), and ECE2 (nine sites), of which one is shared between these sequences (human ECEL1 Asn322-Ile323-Thr324); and similar secondary structures previously identified for human ECE1 (Figures 1 and 3) observed also for ECEL1 and ECE2.[8] The methyltransferase domain previously reported for ECE2 was not observed for the other ECE-like sequences examined, ie, ECE1, ECEL1, and neprilysin.[11] In addition, the Cys428 residue previously identified as forming an interchain disulfide bond for human ECE1 was shared with human and mouse ECE2 sequences but not with ECEL1 sequences. Moreover, the residues previously identified for human ECE1, which are involved in signaling networks (phosphorylated 25Thr and 34Ser), were unique to ECE1 sequences.[38,39] Active site and zinc-binding residues previously identified for human ECE1[8] were identical in each case for the human and mouse ECEL1 and ECE2 sequences, with the exception of a conservative amino acid substitution observed for the human ECE1 active site residue (Val606→Ile) on the corresponding ECEL1 residue. These results suggest that human and mouse ECEL1, ECE1, and ECE2 enzymes share several important properties, features, and conserved residues, including being membrane-bound with cytoplasmic and transmembrane regions, and have similar secondary structures, but are sufficiently different to serve distinct functions.

Gene locations, exonic structures, and regulatory sequences for vertebrate *ECEL1* genes

Table 1 summarizes the predicted locations for vertebrate *ECEL1* genes based upon BLAT interrogations of several vertebrate genomes using the reported sequences for human

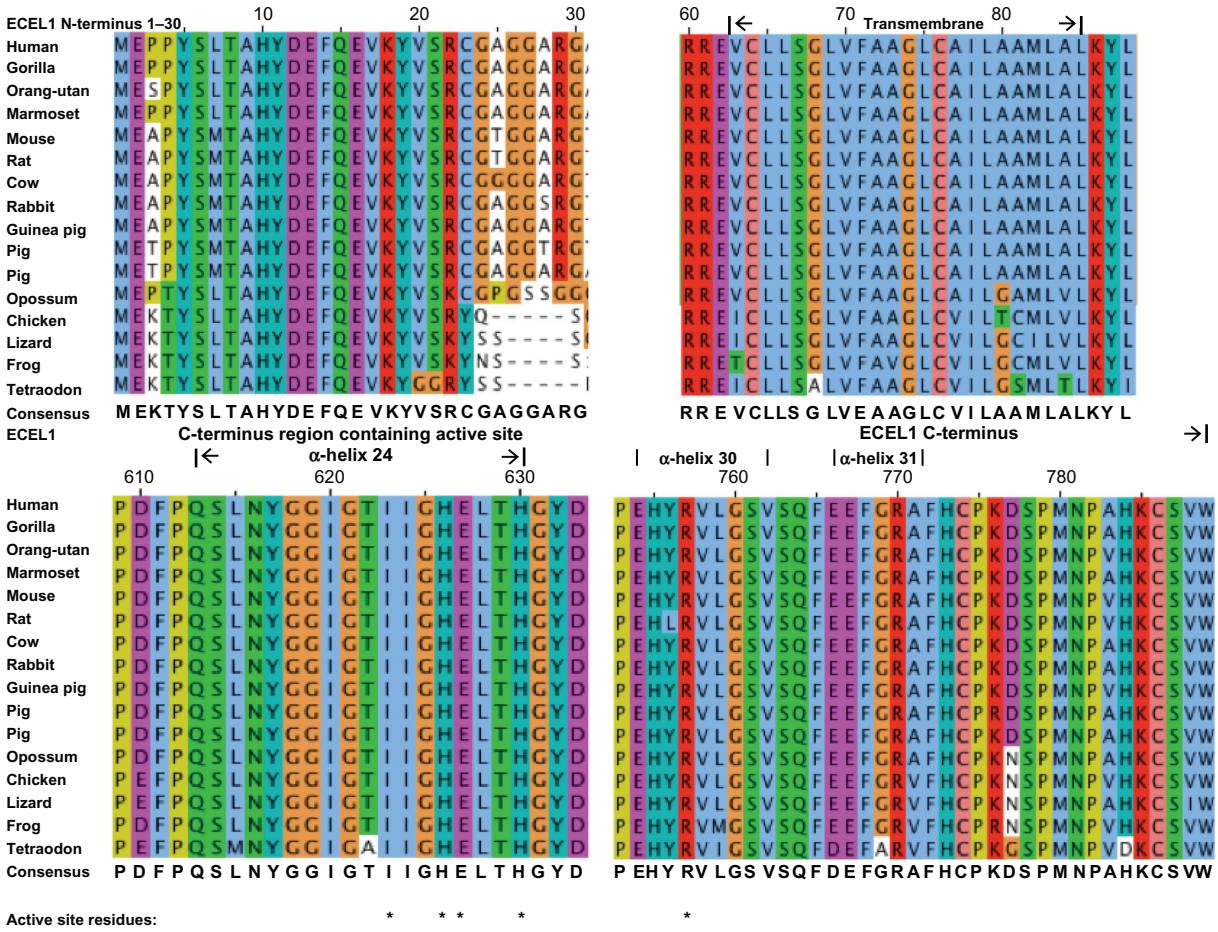

Figure 2 Amino acid sequence alignments for vertebrate ECEL1 domain sequences.

Notes: Amino acids are color coded: yellow for proline (P); S (serine); green for hydrophilic amino acids, S (serine), Q (glutamine), N (asparagine), and T (threonine); brown for glycine (G); light blue for hydrophobic amino acids, L (leucine), I (isoleucine), V (valine), M (methionine), W (tryptophan); dark blue for amino acids, T (tyrosine) and H (histidine); purple for acidic amino acids, E (glutamate), and D (aspartate); and red for basic amino acids, K (lysine) and R (arginine); four major domains were identified: N-terminus (cytoplasmic); transmembrane; C-terminus region containing the active site and alpha-helix 24; and ECEL1 C-terminus containing alpha-helices 30 and 31; active site residues are designated as *deduced consensus sequences are shown.

and mouse ECEL1 and the predicted sequences for other vertebrate ECEL1 enzymes and the UC Santa Cruz genome browser.[19,21,27] The predicted vertebrate *ECEL1* genes were predominantly transcribed on the negative strand, with the exception of guinea pig (*C. porcellus*), lizard (*A. carolensis*), and frog (*X. tropicalis*) *ECEL1* genes, which were transcribed on the positive strand. Figure 1 summarizes the predicted exonic start sites for human, mouse, pig, chicken, and frog *ECEL1* genes, with each having 17 coding exons in identical or similar positions to those predicted for the human *ECEL1* gene.

Figure 5 shows the predicted structure for the human *ECEL1* gene and exons with an extended CpG island (CpG256) and several transcription factor binding sites, located at the 5′ end of the gene, which is consistent with potential roles in regulating the transcription of this gene and forming part of the *ECEL1* gene promoter. The human *ECEL1* gene contained 18 exons, with exons 1 and 2 containing the 5′-untranslated (UTR) exon, and the 18th exon containing the translation stop site and the 3′-untranslated region (UTR). The human *ECEL1* genome sequence contained several predicted transcription factor binding sites (Supplementary Table 2) and a large CpG island (CpG256) located in the 5′-untranslated promoter region in exons 1 and 2 of human *ECEL1* on chromosome 2. CpG256 contained 2696 bps with a C plus G count of 1669 bps, a C or G content of 62%, and showed a ratio of observed to expected CpG of 0.92 (Supplementary Table 3). Comparative studies of CpG islands for other vertebrate *ECEL1* genes show that these are consistently found in the 5′-*ECEL1* promoters of varying CpG genome sizes, ranging from 306 in the lizard (*A. carolensis*) ECEL1 gene to 2999 bps in the dog (*C. familiaris*) *ECEL1* gene (Supplementary Table 3). Therefore, it is likely that the *ECEL1* CpG island plays a key role in regulating this

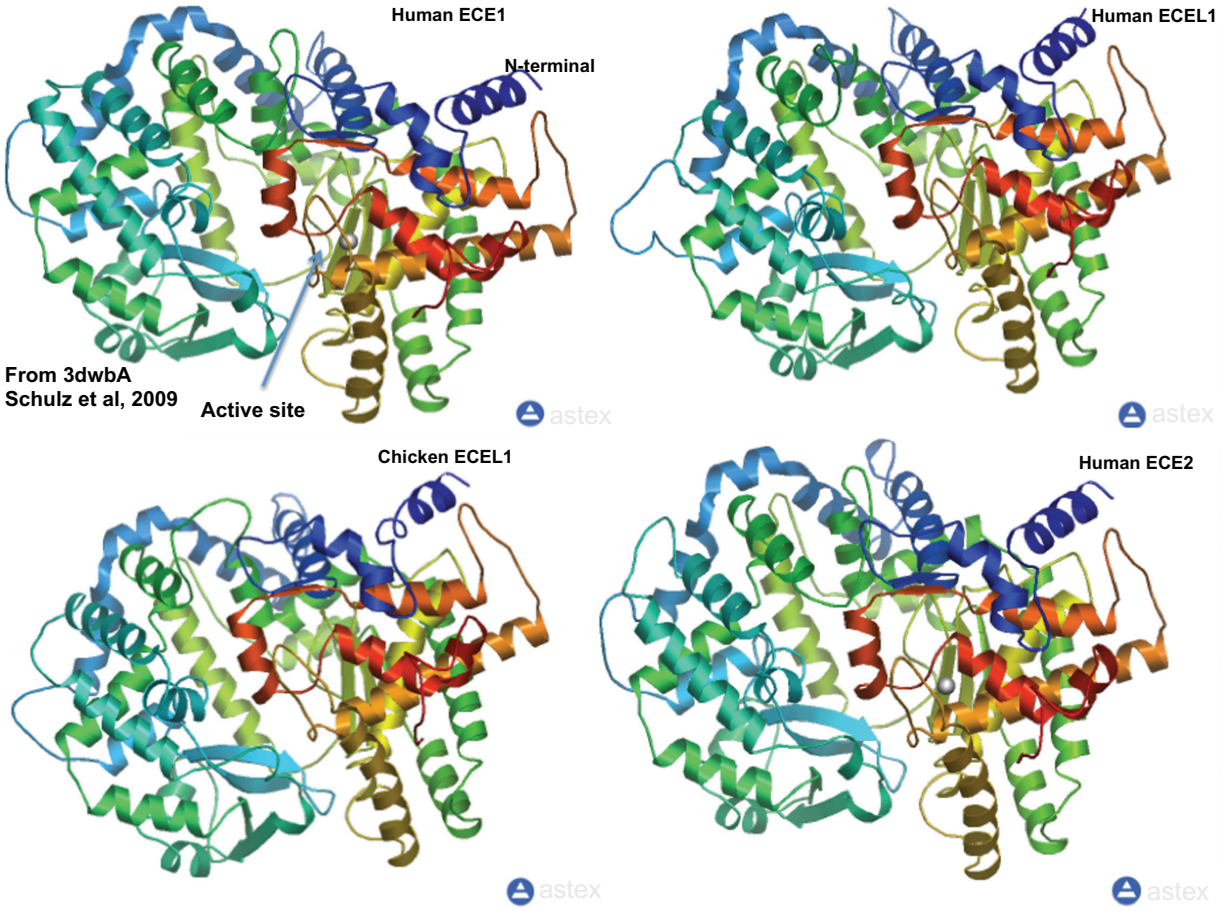

Figure 3 Comparisons of predicted tertiary structures for human and chicken ECEL1 and for human ECE2 with the known structure for human ECE1.
Notes: The predicted structures for the human and chicken ECEL1 subunits and the human ECE2 subunit are based on the reported structure for human ECE1 and obtained using the Swiss-Model web site based on PDB 3dwbA http://swissmodel.expasy.org/workspace/.[8] The rainbow color code describes the three-dimensional structures from the N-terminus (blue) to the C-terminus (red color); predicted α-helices, β-sheets, proposed active site cleft and the N-terminus and C-terminus are shown. The gray sphere for the human ECE1 and ECE2 structures represents the active site, zinc.

gene throughout vertebrate evolution, and may contribute to the very high level of gene expression observed in neural tissues (Supplementary Figure 1).[33,40] The *ECEL1* gene has also been identified as one of eight hypermethylation gene targets in a Chinese population.[41]

At least 10 transcription factor binding sites were colocated with CpG256 in the human *ECEL1* promoter region, which may also contribute to the neural tissue-specific expression (Supplementary Table 2). Of special interest among these identified *ECEL1* transcription factor binding sites are T-cell leukemia homeobox protein 2 which is required for the development of the enteric nervous system,[42] paired box protein Pax-2 which plays a critical role in the development of the central nervous system,[43] transcription factor E2alpha which is involved as an initiator of neuronal differentiation,[44] STAT5 which functions in signal transduction and activation of transcription,[45] and HMX1 which is a transcription factor involved in development of the hypothalamus.[46]

It is also relevant to report that the *ECEL1* gene is localized in a region of human chromosome 2 associated with QTL19 (COPD19_H) for chronic obstructive pulmonary disease and an OMIM disease phenotype associated with pulmonary disease, as shown in Figure 5 (http://genome.ucsc.edu).[47] In addition, Kiryu-Seo et al have reported that *ECEL1* (or *DINE*) is synergistically regulated by other transcriptional regulatory regions in the gene promoter of injured neurons,[16] including ATF3, c-Jun, and STAT3. It would appear that the *ECEL1* gene promoter is well endowed with gene regulatory sequences, including a large CpG island (CpG256) and several transcription factor binding sites which may contribute to the high levels of expression in mammalian neural tissues and to the response of neuronal cells following injury. In terms of genetic variability of *ECEL1* in human populations, recent studies have shown that this gene has 187 annotated single nucleotide polymorphisms, of which 41 are not synonymous, including four which occur in frequencies greater than 1%

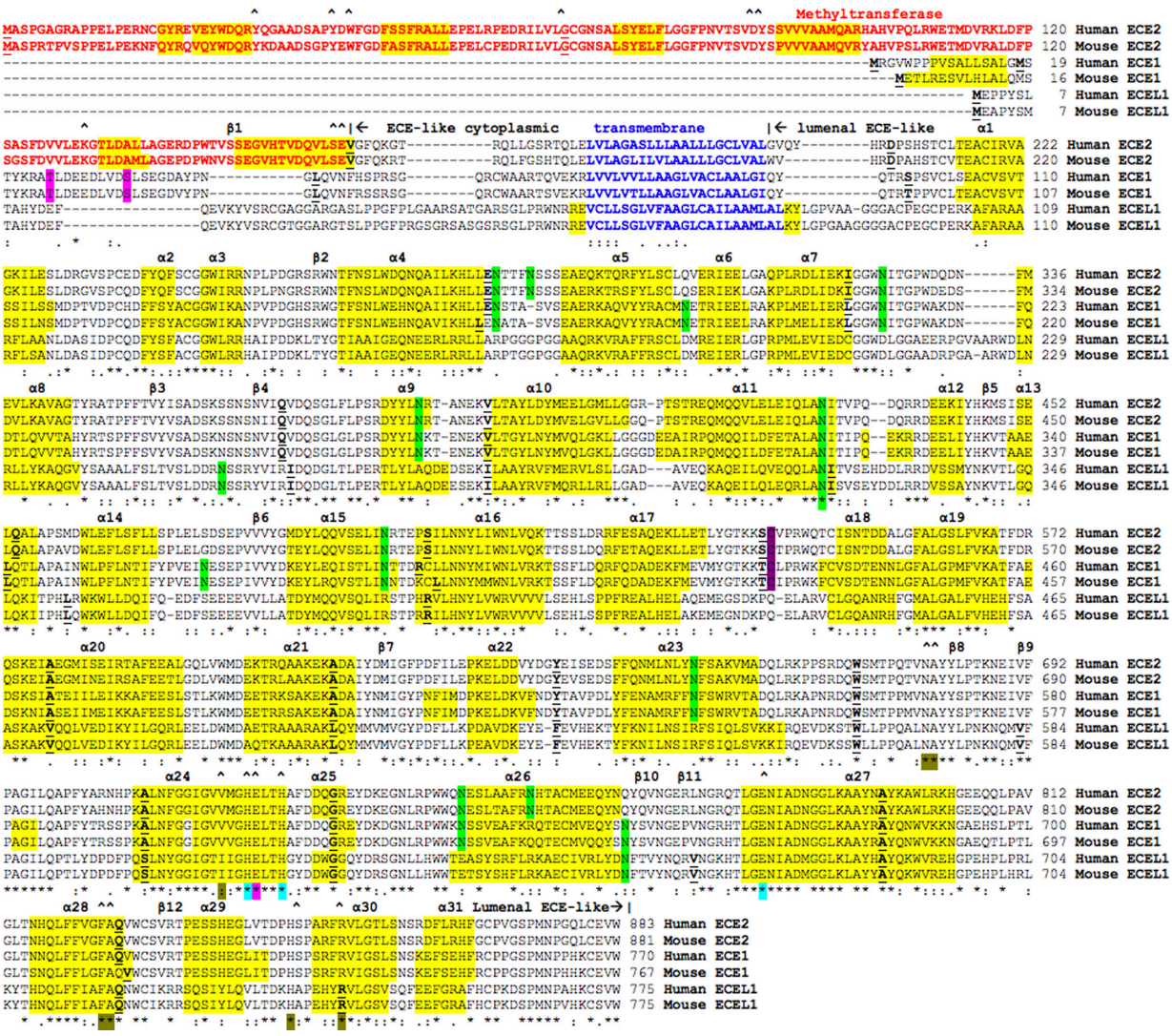

Figure 4 Amino acid sequence alignments for human and mouse ECEL1, ECE1, and ECE2 sequences.

Notes: See Table 1 for sources of human and mouse ECEL1, ECE1, and ECE2 sequences; *shows identical residues for ECE-like subunits; : shows similar alternate residues; . shows dissimilar alternate residues; the predicted methyltransferase domain for ECE2 is shown in red; predicted transmembrane residues are shown in blue; N-glycosylated and potential N-glycosylated Asn sites are in green; signal-phosphorylation sites for human and mouse ECE1 are shown in pink; for active site residues, the catalytic Glu is shown in *, zinc binding residues in * and other active site residues in * or :; active site residues for the methyltransferase domain for ECE2 are shown as ^; active site residues for the luminal ECE-like peptidase domain are shown as ^; predicted α-helices for vertebrate ECE-like sequences and for the ECE2 methyltransferase domain are in shaded yellow and numbered in sequence from the start of the predicted luminal ECE-like domain; predicted β-sheets are in shaded gray and also numbered in sequence for human ECEL1; ECE1 and ECE2 Cys residues involved in inter-subunit -S-S- shown as C; bold underlined font shows residues corresponding to known or predicted exon start sites.

(Figure 5). In particular, rs36038969 results in a Lys → Glu substitution within exon 14, a phylogenetically invariant region localized near the ECEL1 active site (Figure 1), which suggests functional variance for this enzyme within human populations. Even though there is a large number of single nucleotide polymorphisms in this gene, there are no single nucleotide polymorphisms in the 5′-UTR region or the conserved transcription factor binding sites, suggesting that expression of this gene is tightly regulated, which is consistent with evolutionary conservation of these regulatory regions.

Comparative human *ECEL1* and mouse *Ecel1* tissue expression

Supplementary Figure 1 contains "heat maps" showing comparative gene expression for various human and mouse tissues obtained from GNF Expression Atlas Data using GNF1H (human) and GNF1M (mouse) chips (http://genome.ucsc.edu; http://biogps.gnf.org).[33] These data support a high level of neural tissue expression for human *ECEL1* and mouse *Ecel1,* particularly for neural tissues, which is consistent with previous reports for these genes.[12,13] Overall, human *ECEL1* and mouse *Ecel1* tissue expression levels were 0.4–0.7 times

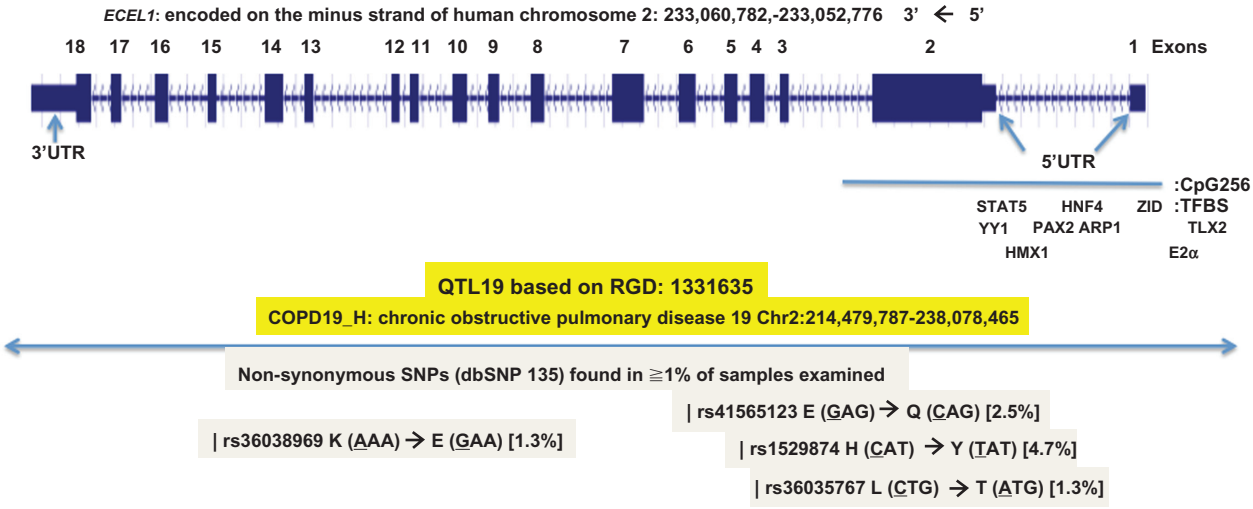

Figure 5 Gene structure and organization for the human *ECEL1* gene.

Notes: Derived from the AceView website http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/; shown with capped 5′-ends and 3′-ends for the predicted mRNA sequences;[22] NM refers to the NCBI reference sequence; exons are in pink; the direction for transcription is shown as 5′→3′; the solid blue line shows a large CpG256 island at or near the gene promoter; predicted transcription factor binding sites for human *ECEL1* are shown; see Supplementary Table 2 for details. The human genome browser (http://genome.ucsc.edu/) was used to identify a disease phenotype in the vicinity of the *ECEL1* region: QTL19 (based on RGD 1331635: COPD19_H; chronic obstructive pulmonary disease);[47] as well as single nucleotide polymorphisms for the human *ECEL1* gene. Individual nonsynonymous single nucleotide polymorphisms with variant allelic frequencies ≤ 1% in the population were identified (dbSNP 135). The variant nucleotide is underlined.

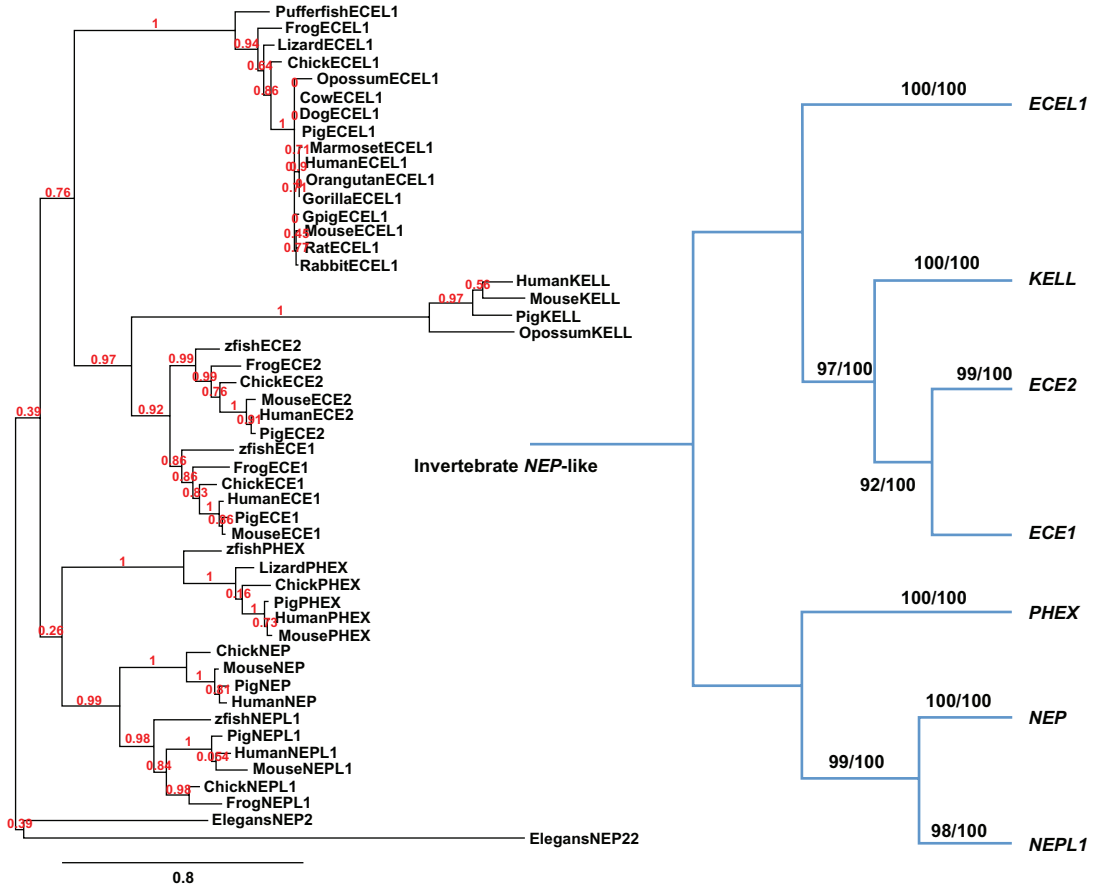

Figure 6 Phylogenetic tree of vertebrate ECEL1 amino acid sequences with other representative vertebrate M13 endopeptidase sequences.

Notes: The tree is labeled with the ECE-like name, and the name of the animal and is "rooted" with the worm (*Caenorhabditis elegans)* NEP2 and NEP22 sequences, which were used to "root" the tree. Note the seven major clusters corresponding to the *ECEL1*, *ECE1*, *ECE2*, *KELL*, *PHEX*, *NEP*, and *NEPL1* gene families. A genetic distance scale is shown. The number of times a clade (sequences common to a node or branch) occurred in the bootstrap replicates is shown. Replicate values of 0.9 (90/100) or more are highly significant with 100 bootstrap replicates performed in each case. A proposed sequence of gene duplication events is shown arising from an ancestral invertebrate *ECE*-like gene.

the average level of gene expression, supporting a key role for this enzyme as an endopeptidase, especially in nerve cells, the hypothalamus, sympathetic ganglion and pituitary gland, and in the neural regulation of respiration (http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/).[12,13,22]

Phylogeny and divergence of ECEL1 and other vertebrate ECE-like sequences

A phylogenetic tree (Figure 6) was calculated by the progressive alignment of 17 vertebrate ECEL1 amino acid sequences with several vertebrate ECE1 and ECE2 sequences, which was "rooted" with the worm (*C. elegans*) neprilysin (NEP2 and NEP22) sequences (see Table 1). The phylogram showed clustering of the ECEL1 sequences into groups, which was consistent with their evolutionary relatedness as well as groups for vertebrate ECE1, ECE2, KELL, PHEX, NEP, and NEPL1 sequences which were distinct from the worm NEP2 and NEP22 sequences. These groups were significantly different from each other (with bootstrap values $> 90/100$). It is apparent from this study of vertebrate *ECE*-like genes and proteins that this is an ancient protein for which a proposed common ancestor for the *ECEL1*, *ECE1*, *ECE2, KELL, PHEX, NEP*, and *NEPL1* genes may have predated the appearance of fish during vertebrate evolution. In addition, vertebrate *ECEL1* genes are seen to be more closely related to *ECE1*, *ECE2*, and *KELL* genes; whereas *PHEX*, *NEP*, and *NEPL1* genes form a distinct subgroup of vertebrate M13 endopeptidase genes.

Conclusion

The results of the present study indicate that vertebrate *ECEL1* genes and encoded proteins represent a distinct gene and protein family of *ECE*-like proteins which share key conserved sequences that have been reported for other ECE-like proteins (ECE1 and ECE2) studied previously.[7–10] ECEL1 has a unique property among these metalloendopeptidases in performing an essential role in the nervous control of respiration, although the natural substrate for ECEL1 has not been described. An association of the human and rat *ECEL1* gene region with pulmonary and neural diseases has been previously identified (http://genome.ucsc.edu).[47] *ECEL1* is encoded by a single gene among the vertebrate genomes studied, and is highly expressed in human and mouse neural tissues, particularly in the hypothalamus, pituitary, and sympathetic ganglion, and usually contains 17 coding exons on the negative strand. Human and other vertebrate *ECEL1* genes contain a large CpG island within the promoter region. In addition, several transcription factor binding sites are also located within the *ECEL1* gene promoter region, which may contribute to the high level of gene expression in neural tissues, such as the hypothalamus. Predicted secondary and tertiary structures for vertebrate ECEL1 proteins showed strong similarities with other ECE-like proteins, NEP, ECE1, and ECE2. Several major structural domains were apparent for vertebrate ECEL1, including the N-terminal cytoplasmic tail; the transmembrane domain which anchors the enzyme to the endoplasmic reticulum; and the luminal domain containing the active site (including a zinc binding site), which is responsible for endopeptidase activity; and three conserved N-glycosylation sites. Phylogenetic studies using 17 vertebrate ECEL1 sequences with several other vertebrate M13 endopeptidase sequences (ECE1, ECE2, KELL, NEP, NEPL1, and PHEX) indicate that the *ECEL1* gene appeared early in vertebrate evolution, prior to the appearance of bony fish, via a series of gene duplication events.

Acknowledgment

We acknowledge the expert assistance of Bharet Patel of Griffith University with the phylogeny studies.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Rawlings ND, Barrett AJ. Evolutionary families of peptidases. *Biochem J*. 1993;290:205–218.
2. Rawlings ND, O'Brien F, Barrett AJ. Merops: the protease database. *Nucleic Acids Res*. 2002;30:343–346.
3. Bland ND, Pinney JW, Thomas JE, Turner AJ, Isaac RE. Bioinformatic analysis of the neprilysin (M13) family of peptidases reveals complex evolutionary and functional relationships. *BMC Evol Biol*. 2008;8:16.
4. Letarte M, Vera S, Tran R, et al. Common acute lymphocytic anemia antigen is identical to neutral endopeptidase. *J Exp Med*. 1988;168:1247–1253.
5. Shipp MA, Richardson NE, Sayre PH, et al. Molecular cloning of the common acute lymphoblastic leukemia antigen (CALLA) identifies a type II integral membrane protein. *Proc Natl Acad Sci U S A*. 1988;85:4819–4823.
6. Oefner C, D'Arcy A, Hennig M, Winkler FK, Dale GE. Structure of human neutral endopeptidase (neprilysin) complexed with phosphoramidon. *J Mol Biol*. 2000;296:341–349.
7. Schmidt M, Kroger B, Jacob E, et al. Molecular characterization of human and bovine converting enzyme (ECE-1). *FEBS Lett*. 1994;356:238–243.
8. Schulz H, Dale GE, Karimi-Nejad Y, Oefner C. Structure of human endothelin-converting enzyme 1 complexed with phosphoramidon. *J Mol Biol*. 2009;385:178–187.
9. Lorenzo MN, Khan RY, Wang Y, et al. Human endothelin converting enzyme-2 (ECE2): characterization of mRNA species and chromosomal localization. *Biochim Biophys Acta*. 2001;1522:46–52.
10. Mzhavia N, Pan H, Che FY, et al. Characterization of endothelin-converting enzyme-2. Implication for a role in the nonclassical processing of regulatory peptides. *J Biol Chem*. 2003;278:14704–14711.

11. Tempel W, Wu H, Dombrovsky L, et al. An intact SAM-dependent methyltransferase fold is encoded by the human endothelin-converting enzyme-2 gene. *Proteins*. 2009;74:789–793.
12. Valdenaire O, Richards JG, Faull RL, Schweizer A. Xce, a new member of the endothelin-converting enzyme and neutral endopeptidase family, is preferentially expressed in the CNS. *Brain Res Mol Brain Res*. 1999; 74:211–221.
13. Nagata K, Kiryu-Seo S, Kiyama H. Localization and ontogeny of damage-induced neuronal endopeptidase mRNA-expressing neurons in the rat nervous system. *Neuroscience*. 2006;141:299–310.
14. Kato R, Kiryu-Seo S, Kiyama H. Damage-induced neuronal endopeptidase (DINE/ECEL) expression is regulated by leukemia inhibitory factor and deprivation of nerve growth factor in rat sensory ganglia after nerve injury. *J Neurosci*. 2002;22:9410–9418.
15. Kiryu-Seo S, Sasaki M, Yokohama H, et al. Damage-induced neuronal endopeptidase (DINE) is a unique metallopeptidase expressed in response to neuronal damage and activates superoxide scavengers. *Proc Natl Acad Sci U S A*. 2000;97:4345–4350.
16. Kiryu-Seo S. Identification and functional analysis of damage-induced neuronal endopeptidase (DINE), a nerve injury associated molecule. *Anat Sci Int*. 2006;81:1–6.
17. Kawamoto T, Ohira M, Hamano S, Hori T, Nakagawara A. High expression of the novel endothelin-converting enzyme genes, Nbla03145/ECEL1alpha and beta, is associated with favorable prognosis in human neuroblastomas. *Int J Oncol*. 2003;22:815–822.
18. Schweizer A, Valdenaire O, Koster A, et al. Neonatal lethality in mice deficient in XCE, a novel member of the enthelin-converting enzyme and neutral endopeptidase family. *J Biol Chem*. 1999;274: 20450–20456.
19. Valdenaire O, Rohrbacher E, Langeveld A, Schweizer A, Meijers C. Organization and chromosomal localization of the human ECEL1 (XCE) gene encoding a zinc metallopeptidase involved in the nervous control of respiration. *Biochem J*. 2000;343:611–616.
20. Valdenaire O, Schweizer A. Endothelin-converting enzyme-like 1 (ECEL1; 'XCE'): a putative metallopeptidase crucially involved in the nervous control of respiration. *Biochem Soc Trans*. 2000;28: 426–430.
21. Kiryu-Seo S, Kato R, Ogawa T, et al. Neuronal injury-inducible gene is synergistically regulated by ATF3, c-Jun and STAT3 through the interaction with Sp1 in damaged neurons. *J Biol Chem*. 2008;283: 6988–6996.
22. Thierry-Mieg D, Thierry-Mieg J. AceView: A comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol*. 2006;7: S12:1–14.
23. Altschul F, Vyas V, Cornfield A, et al. Basic local alignment search tool. *J Mol Biol*. 1990;215:403–410.
24. Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009;10:421.
25. Benoit A, Vergas MA, Desgroseillers L, Boileau. Endothelin-converting enzyme-like 1 (ECEL1) is present both in the plasma membrane and in the endoplasmic reticulum. *Biochem J*. 2004;380:881–888.
26. Holmes RS, Cox LA. Comparative studies of vertebrate scavenger receptor class B type 1: a high density lipoprotein binding protein. *Res Rep Biochem*. 2012;2:1–16.
27. Kent WJ, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. *Genome Res*. 2002;12:994–1006.
28. Schwede T, Kopp J, Guex N, Pietsch MC. Swiss-Model: an automated protein homology-modelling server. *Nucleic Acids Res*. 2003;31: 3381–3385.
29. Schulz H, Dale GE, Karimi-Nejad Y, Oefner C. Structure of human endothelin-converting enzyme 1 complexed with phosphoramidon. *J Mol Biol*. 2009;385:178–187.
30. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. *Bioinformatics*. 2000;16:404–405.
31. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol*. 2001;305:567–580.
32. Marchler-Bauer A, Lu S, Anderson JB, et al. CDD: a conserved domain database for the functional annotation of proteins. *Nucleic Acid Res*. 2011;39:D225–D229.
33. Su AI, Wiltshire T, Batalov S, et al. A gene atlas of the human and mouse protein encoding transcriptomes. *Proc Natl Acad Sci U S A*. 2004;101:6062–6067.
34. Dereeper A, Guignon V, Blanc G, et al. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res*. 2008;36: W465–W469.
35. Edgar RC. PMUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32:1792–1797.
36. Guidon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*. 2003;52: 696–704.
37. Anisimova M, Gascuel O. Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. *Syst Biol*. 2006;55: 539–552.
38. Olsen JV, Blagoev B, Gnad F, et al. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell*. 2006;127: 635–648.
39. Mayya V, Han DK. Phosphoproteomics by mass spectrometry: insights, implications, applications and limitations. *Expert Rev Proteomics*. 2009;6:605–618.
40. Saxonov S, Berg P, Brutlag DL. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci U S A*. 2006;103:1412–1417.
41. Zhao Y, Guo S, Sun J, et al. Methylcap-seq reveals novel DNA methylation markers for the diagnosis and recurrence prediction of bladder cancer in a Chinese population. *PloS One*. 2012;7:e35175.
42. Iitsuka Y, Shimizu H, Kang MM, et al. An enhancer element for expression of the Ncx (Enx, Hox11 L1) gene in neural crest-derived cells. *J Biol Chem*. 1999;274:24401–24407.
43. Eccles MR, Wallis LJ, Fidler AE, et al. Expression of the PAX2 gene in human fetal kidney and Wilms' tumor. *Cell Growth Differ*. 1992;3:279–289.
44. Ravanpay AC, Olsen JM. E protein dosage influences brain development more than family member identity. *J Neurosci Res*. 2008;86: 1472–1481.
45. Buitenhuis M, Coffer PJ, Koenderman L. Signal transducer and activator of transcription 5 (STAT5). *Int J Biochem*. 2004;36:2120–2124.
46. Bober E, Baum C, Braun T, Arnold H. A novel NK-related mouse homeobox gene: expression in central and peripheral nervous structures during embryonic development. *Dev Biol*. 1994;162:288–303.
47. Rapp JP. Genetic analysis of inherited hypertension in the rat. *Physiol Rev*. 2000;80:135–172.

Supplementary materials

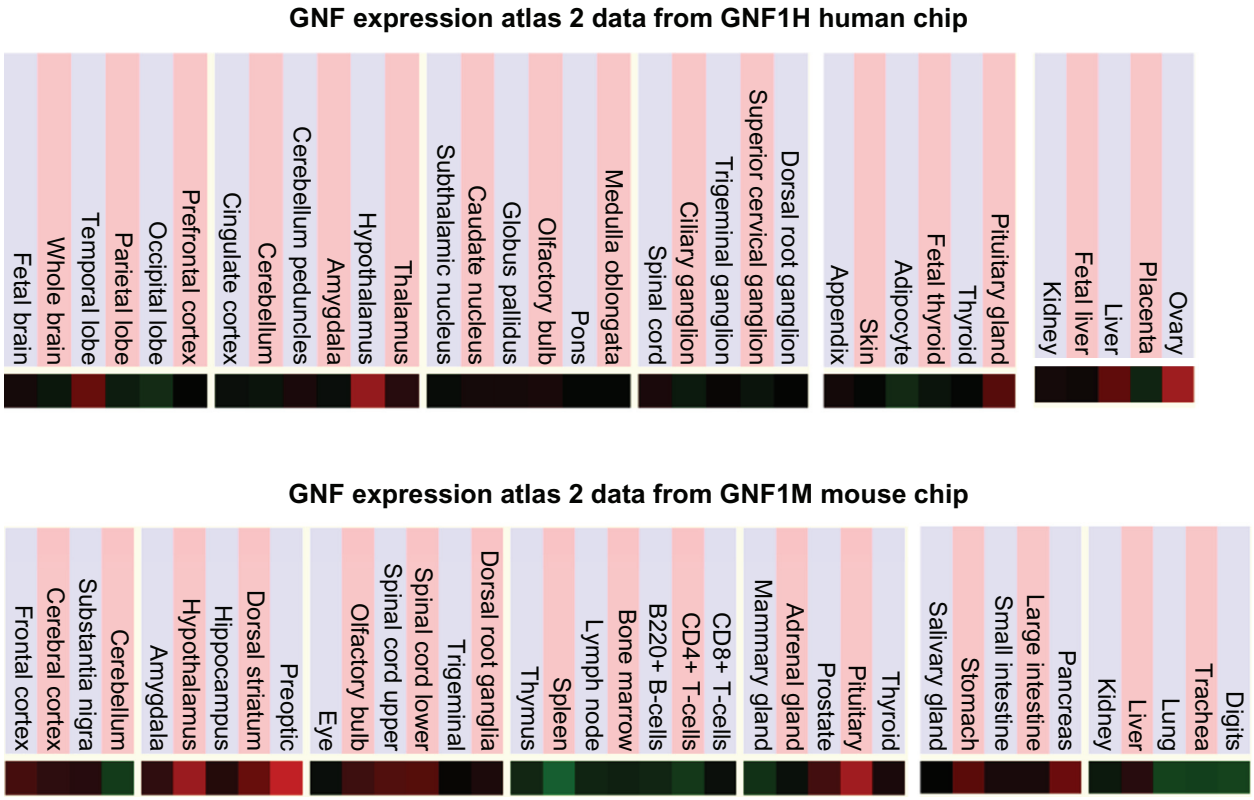

Figure S1 Comparative tissue expression for human and mouse *ECEL1* Gene Expression "heat maps" (GNF Expression Atlas 2 data) (http://biogps.gnf.org) were examined for comparative gene expression levels among human and mouse tissues for *ECEL1* genes showing high (red), intermediate (black), and low (green) expression levels.[33]
Note: Derived from human and mouse genome browsers (http://genome.ucsc.edu).

Table S1 Other vertebrate M13-like endopeptidase genes and proteins

ECE-like gene	Species	RefSeq ID Ensembl/NCBI[a]	GenBank ID	UNIPROT ID	Amino acids	Chromosome location	Coding exons (strand)	Gene size bps	Subunit MW	Gene expression	pI
KELL (KEL)											
Human	*Homo sapiens*	NM_000420	BC003135	P23276	732	7:142,638,342–142,658,959	18 (–ve)	20,618	82,823	0.6	8.1
Mouse	*Mus musculus*	NM_115929	BC099961	Q9EQF2	713	6:41,636,449–41,653,499	18 (–ve)	17,051	80,866	0.8	5.9
Pig	*Sus scrofa*	XP_003134648[a]	na	na	730	18:6,126,665–6,149,749	18 (–ve)	23,085	82,077	na	6.4
Opossum	*Monodelphis domestica*	XP_001364826	na	na	768	8:205,592,264–205,612,188	18 (–ve)	19,925	86,352	na	6.3
PHEX											
Human	*Homo sapiens*	NM_000444	BC105057	P78562	749	X:22,051,124–22,266,067	22 (+ve)	214,944	86,474	0.6	8.9
Mouse	*Mus musculus*	NM_011077	EF194891	A2ICR0	749	X:153,600,106–153,852,688	22 (–ve)	252,583	86,359	0.3	9.0
Pig	*Sus scrofa*	NM_001244594	na	na	749	X:16,927,787–17,142,215	22 (+ve)	214,429	86,307	na	9.0
Chicken	*Gallus gallus*	NM_001199277	na	E1BXR4	751	1:118,329,962–118,421,519	22 (–ve)	91,558	87,146	na	9.2
Lizard	*Anolis carolensis*	XP_003218886[a]	na	na	751	3:124,664,626–124,810,066	22 (+ve)	145,441	86,475	na	8.7
Zebrafish	*Danio rerio*	NM_001089349	BC139673	F1R6K1	745	24:26,230,575–26,253,987	22 (+ve)	23,413	85,271	na	7.1
NEPL1 (MMEL1)											
Human	*Homo sapiens*	NM_033467	BC032051	Q495T6	779	1:2,522,432–2,560,923	23 (–ve)	38,492	89,367		5.6
Mouse	*Mus musculus*	NM_013783	AF157105	Q9JLI3	765	4:154,245,762–154,269,331	23 (+ve)	23,570	88,700		6.1
Opossum	*Monodelphis domestica*	XP_001378153[a]	na	na	744	4:377,080,503–377,139,999	22 (–ve)	59,497	86,441	na	5.8
Chicken	*Gallus gallus*	XP_001233077[a]	na	na	745	21:1,388,408–1,409,376	22 (+ve)	20,969	85,614	na	5.6
Frog	*Xenopus tropicalis*	NM_001127095	na	F6UT14	745	GL173014:240,544–272,000^	22 (–ve)	31,347	86,073	na	5.5
Zebrafish	*Danio rerio*	XP_689191[a]	na	na	755	11:42,025,785–42,081,241	22 (+ve)	55,457	86,770	na	5.4

Notes: RefSeq: the reference amino acid sequence; [a]predicted Ensembl amino acid sequence; na, not available; GenBank IDs are derived NCBI http://www.ncbi.nlm.nih.gov/genbank/; Ensembl ID was derived from Ensembl genome database http://www.ensembl.org; UNIPROT refers to UniprotKB/Swiss-Prot IDs for individual endopeptidase M13-like proteins (see http://kr.expasy.org); *refers to incomplete gene sequence; GL refers to an unknown chromosome; bps refers to base pairs of nucleotide sequences; pI refers to theoretical isoelectric points; the number of coding exons are listed.

Dovepress

Table S2 Identification of transcription factor binding sites within the human *ECEL1* gene promoter

TFBS	Name	Strand	Chr 2 position	Function/role	Sequence	UNIPROT ID
ZID	Zinc finger and BTB domain-containing protein 6	(+ve)	233,352,605–617	Acts as a specific protein-protein interaction domain	CGGCTCCACCCTC	Q15916
TLX2	T-cell leukemia homeobox protein 2	(+ve)	233,352,429–443	Required for development of the enteric nervous system	CCGGCAGGTGCGCGC	O43763
E2alpha	Transcription factor E2-alpha	(+ve)	233,352,429–443	Transcriptional initiator of neuronal differentiation	CCGGCAGGTGCGCGC	P15923
HNF4A	Hepatocyte nuclear factor 4-alpha	(–ve)	233,351,932–946	Essential for liver, kidney and intestine development	GAGGACAAAGTTTGG	P41235
ARP1	COUP transcription factor 2	(–ve)	233,351,851–866	Regulation of apolipoprotein A-I gene transcription	CGCGCCCCTGACCGCA	P24468
PAX2	Paired box protein Pax-2	(+ve)	233,351,337–355	A critical role in the development of the nervous system	CGCCGTCAGCGAATACGGG	Q02962
STAT5	Signal transducer and activator of transcription 5A	(+ve)	233,351,313–327	Signal transduction and activation of transcription	GACCTCTTGGAACTC	P42229
HMX1	Homeobox protein Nkx-5.1	(–ve)	233,351,304–313	Transcription factor involved in hypothalamus development	CAAGTACGTG	P42581
PAX4	Paired box protein Pax-4	(+ve)	233,351,160–189	Key role in the development of pancreatic islet beta cells	GAACACCAGCCCCGACAGCAGGCACACCTC	O43316
YY1	Transcriptional repressor protein YY1	(–ve)	233,351,130–149	Transcription factor exhibiting both positive and negative control	TCTGCGCCATTCTGGCGGCT	P25490

Notes: Identification of *ECEL1* TFBS within the *ECEL1* promoter region was undertaken using the human genome browser (http://genome.ucsc.edu); UNIPROT refers to UniprotKB/Swiss-Prot IDs for individual transcription factor binding site sequences (see http://kr.expasy.org).

Table S3 Comparative vertebrate *ECEL1* CpG islands

Vertebrate	Species	CpG Island ID	Chromosomal position	Genome size	C count plus G count	Percentage C or G	Ratio of observed to expected CpG
Human	*Homo sapiens*	CpG 256	chr2:233,350,279–233,352,974	2,656	1,669	61.9	0.99
Mouse	*Mus musculus*	CpG 126	chr1:89,050,488–89,051,660	1,173	802	68.4	0.92
Rat	*Rattus norvegicus*	CpG 191	chr9:85,944,921–85,947,172	2,252	1,477	65.6	0.79
Cow	*Bos taurus*	CpG 236	chr2:125,628,627–125,631,112	2,486	1,576	63.4	0.95
Pig	*Sus scrofa*	CpG 194	chr15:125,288,762–125,291,009	2,247	1,380	61.4	0.92
Dog	*Canis familiaris*	CpG 300	chr25:44,126,533–44,129,531	2,999	1,787	59.6	1.13
Guinea pig	*Cavia porcellus*	CpG 240	Sc*13:29,245,735–29,248,257	2,523	1,622	64.3	0.92
Opossum	*Monodelphis domestica*	CpG 120	2:537,269,047–537,270,090	1,044	668	64	1.12
Chicken	*Gallus gallus*	CpG 40	9:15,054,664–15,055,043	380	237	62.4	1.11
Lizard	*Anolis carolensis*	CpG 25	chrUn^GL34304:691, 937–692,242	306	193	63.1	0.82
Puffer fish	*Tetraodon nigroviridis*	CpG 75	chr16:6,942,743–6,943,889	1,147	635	55.4	0.85

Notes: Identification of *ECEL1* CpG islands, sequences and properties was undertaken using various vertebrate genome browsers (http://genome.ucsc.edu).