Exhaustive identification of conserved upstream open reading frames with potential translational regulatory functions from animal genomes

Hiro Takahashi[1,2#]*, Shido Miyaki[2#], Hitoshi Onouchi[3#], Taichiro Motomura[1], Nobuo Idesako[2], Anna Takahashi[4], Masataka Murase[1], Shuichi Fukuyoshi[5], Toshinori Endo[6], Kenji Satou[7] , Satoshi Naito[3,8], and Motoyuki Itoh[9]*

[1]Graduate School of Medical Sciences, Kanazawa University, Kanazawa 920-1192, Japan

[2]Graduate School of Horticulture, Chiba University, Matsudo 271-8510, Japan

[3]Graduate School of Agriculture, Hokkaido University, Sapporo 060-8589, Japan

[4]Faculty of Information Technologies and Control, Belarusian State University of Informatics and Radio Electronics, Minsk 220013, Belarus

[5]Institute of Medical, Pharmaceutical and Health Sciences, Kanazawa University, Kakuma-machi, Kanazawa, Ishikawa 920-1192, Japan

[6]Graduate School of Information Science and Technology, Hokkaido University, Sapporo 060-0814, Japan

[7]Faculty of Biological Science and Technology, Institute of Science and Engineering, Kanazawa University, Kanazawa 920-1192, Japan

[8]Graduate School of Life Science, Hokkaido University, Sapporo 060-0810, Japan

[9]Graduate School of Pharmaceutical Science, Chiba University, Chuo-ku, Chiba 260-8675, Japan

*Correspondence. Tel: +81-76-234-4484; Fax: +81-76-234-4484; Email: takahasi@p.kanazawa-u.ac.jp

Correspondence may also be addressed to Motoyuki Itoh. Email: mito@chiba-u.jp

[#]Joint first authors.

Key words: upstream open reading frame; translational regulation; bioinformatics; nascent peptide

Abstract

Background: Upstream open reading frames (uORFs) are present in the 5′-untranslated regions of many eukaryotic mRNAs, and some peptides encoded in these regions play important regulatory roles in controlling main ORF (mORF) translation. To comprehensively identify uORFs encoding functional peptides, genome-wide searches for uORFs with conserved peptide sequences (CPuORFs) have been conducted in various organisms using comparative genomic approaches. However, in animals, CPuORFs have been identified only by comparing uORF sequences between a limited number of closely related species, and how many previously identified CPuORFs encode regulatory peptides is unclear.

Results: Here, we conducted exhaustive genome-wide searches for animal CPuORFs conserved in various taxonomic ranges, using the ESUCA pipeline, which we recently developed for efficient comprehensive identification of CPuORFs. ESUCA can efficiently compare uORF sequences between an unlimited number of species using BLAST and automatically determine the taxonomic ranges of uORF sequence conservation. By applying ESUCA to human, chicken, zebrafish, and fruit fly genomes, 1,517 (1,425 novel and 92 known) CPuORFs were identified. We examined the effects of 17 human CPuORFs on mORF translation using transient expression assays. Through these analyses, we identified seven novel regulatory CPuORFs that repressed mORF translation in a sequence-dependent manner, including the one conserved only among Eutheria.

Conclusions: We discovered a much higher number of animal CPuORFs than previously identified. Since most human CPuORFs identified in this study are conserved across a wide range of Eutheria or a wider taxonomic range, many CPuORFs encoding regulatory peptides are expected to be found in the identified CPuORFs.

Background

The human genome contains many regions encoding potential functional small peptides outside of the well-annotated protein-coding regions (Ingolia, et al., 2014). Some upstream open reading frames (uORFs), which are located in the 5′-untranslated regions (5′-UTRs) of mRNAs, have been shown to encode such functional small peptides. Most uORF-encoded peptides play regulatory roles in controlling the translation of protein-coding main ORFs (mORFs) (Cruz-Vera, et al., 2011; Ito and Chiba, 2013; Morris and Geballe, 2000; Somers, et al., 2013). During the translation of these regulatory uORFs, nascent peptides interact inside the ribosomal exit tunnel to cause ribosome stalling (Bhushan, et al., 2010). Ribosome stalling on a uORF results in translational repression of the downstream mORF because stalled ribosomes block scanning of subsequent pre-initiation complexes and prevent them from reaching the start codon of the mORF (Wang and Sachs, 1997). In some genes, uORF peptides are involved in translational regulation in response to metabolites (Ito and Chiba, 2013).

To comprehensively identify uORFs encoding functional peptides, genome-wide searches for uORFs with conserved peptide sequences (CPuORFs) have been conducted using comparative genomic approaches in plants (Hayden and Jorgensen, 2007; Takahashi, et al., 2019; Takahashi, et al., 2012; Tran, et al., 2008; van der Horst, et al., 2018; Vaughn, et al., 2012). To date, 157 CPuORF families have been identified by comparing 5′-UTR sequences between plant species. Of these, 101 families were identified in our previous studies by applying our original methods, BAIUCAS (Takahashi, et al., 2012) and ESUCA (an advanced version of BAIUCAS) (Takahashi, et al., 2019) to genomes of *Arabidopsis*, rice, tomato, poplar, and grape.

ESUCA has many unique functions (Takahashi, et al., 2019), such as efficient comparison of uORF sequences between an unlimited number of species using BLAST, automatic determination of taxonomic ranges of CPuORF sequence conservation, systematic calculation of K_a/K_s ratios of CPuORF sequences, and wide compatibility with any eukaryotic genome whose sequence database is registered in ENSEMBL (Zerbino, et al., 2018). More importantly, to distinguish between ‘spurious’ CPuORFs conserved because they encode parts of mORF-encoded proteins and ‘true’ CPuORFs conserved because of the functions of their encoded small peptides, ESUCA assesses whether a transcript containing a fusion of a uORF and an mORF is a major or minor form among homologous transcripts (Takahashi, et al., 2019). By using these functions, ESUCA can efficiently

73 identify CPuORFs likely to encode functional small peptides. In fact, our recent study demonstrated that poplar
74 CPuORFs encoding regulatory peptides were efficiently identified by selecting ones conserved across diverse
75 eudicots using ESUCA (Takahashi, et al., 2019).

76 To date, only a few studies on genome-wide identification of animal CPuORFs have reported. In these
77 previous studies, uORF sequences were compared between a limited number of closely related species, such as
78 human and mouse or several species in dipteran, leading to identification of 204 and 198 CPuORFs in human
79 and mouse, respectively (Crowe, et al., 2006), and 44 CPuORFs in fruit fly (Hayden and Bosco, 2008).
80 Additionally, the relationships between taxonomic ranges of CPuORF conservation and the likelihood of having
81 a regulatory function have not been studied in animals.

82 Accordingly, in this study, we applied ESUCA to genomes of fruit fly, zebrafish, chiken, and human to
83 exhaustively identify animal CPuORFs and to determine the taxonomic range of their sequence conservation.
84 Using ESUCA, we identified 1,517 animal (1,425 novel and 92 known) CPuORFs belonging to 1,430 CPuORF
85 families. We examined the effects of 17 CPuORFs conserved in various taxonomic ranges on mORF translation,
86 using transient expression assays. Through this analysis, we identified seven novel regulatory CPuORFs that
87 repress mORF translation in a sequence-dependent manner.

88

89

90 # Materials and methods

91 **Extraction of CPuORFs using ESUCA**

92 ESUCA was developed as an advanced version of BAIUCAS (Takahashi et al., 2012) in our previous study
93 (Takahashi et al., 2019). ESUCA consists of six steps, and some of these steps are divided into substeps, as
94 shown in Fig. 1A and 1B. To identify animal CPuORFs using ESUCA, the following eight-step procedures were
95 conducted, including the six ESUCA steps: 0) data preparation for ESUCA, 1) uORF extraction from the 5′-UTR,
96 2) calculation of uORF-mORF fusion ratios, 3) uORF-tBLASTn against transcript sequence databases, 4)
97 mORF-tBLASTn against downstream sequence datasets for each uORF, 5) calculation of K_a/K_s ratios, 6)

98 determination of the taxonomic range of uORF sequence conservation, and 7) manual validation after ESUCA.
99 See the Supplementary Materials and Methods for details.
100
101 **Determination of the taxonomic range of uORF sequence conservation for animal CPuORFs**
102 To apply ESUCA to animal genomes, we defined 20 animal taxonomic categories, as shown in Fig. 1C. See the
103 Supplementary Materials and Methods for details.
104
105 **Plasmids and reporter assays**
106 DNA fragments containing a CPuORF with the wild-type amino acid sequence or the frameshift mutant version
107 of the 17 selected genes were subcloned into pSV40:UTR:Fluc. Reporter assays were conducted using a
108 Dual-Luciferase Reporter Assay system (Promega, Madison, WI, USA). See the Supplementary Materials and
109 Methods for details.
110

111 Results

112 **Genome-wide search for animal CPuORFs using ESUCA**

113 Prior to ESUCA application (Fig. 1A and 1B), we counted the number of protein-coding genes for four species,
114 i.e., fruit fly, zebrafish, chiken, and human. As shown in Supplementary Table S1, 13,938, 25,206, 14,697, and
115 19,956 genes were extracted for fruit fly, zebrafish, chiken, and human, respectively. After step 1 of ESUCA, we
116 calculated the numbers of uORFs and protein-coding genes with any uORF for each species. As shown in
117 Supplementary Table S1, 17,035 (7,066), 39,616 (14,453), 8,929 (3,535), and 44,085 (12,321) uORFs (genes)
118 were extracted from fruit fly, zebrafish, chicken, and human genomes, respectively. In this analysis, when
119 multiple uORFs from a gene shared the same stop or start codon, they were counted as one. Potential candidate
120 CPuORFs were narrowed down by selection at step 2 of ESUCA in a step-by-step manner, as shown in
121 Supplementary Table S1. The numbers of BLAST hits (expressed sequence tag [EST], transcriptome shotgun
122 assembly [TSA], assembled EST/TSA, and RefSeq RNA sequences) extracted at step 3.2 are also shown in
123 Supplementary Table S1. After the final step of ESUCA, 49, 192, 261, and 1,495 candidate CPuORFs were

124 extracted from fruit fly, zebrafish, chiken, and human, respectively. We conducted manual validation for the
125 extracted candidate CPuORFs as described in our previous study (Takahashi, et al., 2019) and in the
126 Supplementary Materials and Methods. We selected CPuORFs conserved in at least two orders other than the
127 order to which the original species belongs. In total, 1,517 animal CPuORFs (37 for fruit fly, 156 for zebrafish,
128 230 for chicken, and 1,094 for human) were identified (Fig. 1D). Of these, 1,425 CPuORFs were newly
129 identified in the current study. Detailed information on the identified CPuORFs is shown in Supplementary Table
130 S2. The identified CPuORF-containing genes were classified into 1,363 ortholog groups on the basis of
131 similarities of mORF-encoded amino acid sequences, using OrthoFinder ver. 1.1.4 (Emms and Kelly, 2015).
132 CPuORFs with similar amino acid sequences from the same ortholog groups were categorized as the same
133 CPuORF families (homology groups [HGs]; Supplementary Materials and Methods). The identified 1,517
134 CPuORFs were classified into 1,430 HGs (Fig. 1D). We assigned HG numbers to 1,430 HGs in an order based
135 on numbers of orders in which any CPuORF belonging to each HG was extracted, the taxonomic range of the
136 sequence conservation of each HG, and gene ID numbers. When multiple CPuORF families were identified in
137 the same ortholog groups, the same HG number with a different subnumber was assigned to each of the families
138 (e.g., HG0004.1 and HG0004.2; Supplementary Table S2).
139
140 **Sequence-dependent effects of CPuORFs on mORF translation**
141 To address the relationship between taxonomic ranges of CPuORF conservation and likelihood of having
142 regulatory function, we selected 17 human CPuORFs conserved in various taxonomic ranges, including a
143 previously identified sequence-dependent regulatory CPuORF, the *PTP4A1* CPuORF (Hardy, et al., 2019), as a
144 positive control, and examined their sequence-dependent effects on the expression of the downstream reporter
145 gene using transient expression assays (Fig. 2). Other uORFs overlapping any of the selected CPuORFs were
146 eliminated by introducing mutations that changed the ATG codons of the overlapping uORFs to other codons but
147 did not alter the amino acid sequences of the CPuORFs (Supplementary Fig. S1). The resulting modified
148 CPuORFs were used as CPuORFs bearing the wild-type amino acid sequences (WT-aa CPuORFs) (Fig. 2B). To
149 assess the importance of amino acid sequences for the effects of these CPuORFs on mORF translation, frameshift

150 mutations were introduced into the WT-aa CPuORFs such that the amino acid sequences of their conserved
151 regions could be altered (see Supplementary Materials and Methods and Supplementary Fig. S1 for details). In
152 eight of the 17 CPuORFs, the introduced frameshift mutations significantly upregulated the expression of the
153 reporter gene, indicating that these CPuORFs repressed mORF translation in a sequence-dependent manner (Fig.
154 2C). One of the eight sequence-dependent regulatory CPuORFs, the *TMEM184C* CPuORF, is conserved only
155 among Eutheria (Fig. 2A). This result suggests that CPuORFs conserved only among Eutheria can have
156 seqence-dependent regulatory effects.
157

158 **Discussion**

159 In the current study, by applying ESUCA to four animal genomes, we identified 1,517 CPuORFs belonging to
160 1,430 HGs. Taxonomic ranges of sequence conservation of these CPuORFs largely vary, demonstrating that
161 ESUCA can identify CPuORFs conserved in various taxonomic ranges (Supplementary Table S3). We examined
162 the effects of 17 human CPuORFs conserved beyond Euarchontoglires on mORF translation, and identified
163 seven novel sequence-dependent regulatory CPuORFs (in the *MKKS*, *SLC6A8*, *FAM13B*, *MIEF1*, *KAT6A*,
164 *LRRC8B,* and *TMEM184C* genes). Of these, the *TMEM184* CPuORF is one of those conserved in the narrowest
165 taxonomic range among the tested CPuORFs. This suggests that human CPuORFs conserved beyond
166 Euarchontoglires are likely to be conserved because of functional constraints of their encoded peptides. Of the
167 1,094 CPuORFs extracted from the human genome, 1,082 are conserved beyond Euarchontoglires
168 (Supplementary Table S3). Therefore, many CPuORFs encoding regulatory peptides are expected to be found in
169 the human CPuORFs identified in this study.

170 Of the sequence-dependent regulatory CPuORFs identified here, the *MKKS* CPuORF has been
171 previously reported to be a translational regulator that represses the production of a protein involved in
172 McKusick-Kaufman syndrome (Akimoto, et al., 2013); however, the amino acid sequence dependence of the
173 CPuORF function was not reported. Interestingly, the *MIEF1* CPuORF-encoded peptide is a functional peptide
174 localized in the mitochondria (Samandi, et al., 2017). Thus, the *MIEF1* CPuORF may have dual functions.

175 Chemical screening recently identified a compound that causes nascent peptide-mediated ribosome
176 stalling in the mORF of the human *PCSK9* gene, resulting in specific translational inhibition of *PCSK9* and a

reduction in total plasma cholesterol levels (Lintner, et al., 2017). Nascent peptide-mediated ribosome stalling in some of the previously identified regulatory CPuORFs is promoted by metabolites, such as polyamine, arginine, and sucrose (Ito and Chiba, 2013; Yamashita, et al., 2017). Therefore, compounds that promote nascent peptide-mediated ribosome stalling in CPuORFs could be identified by chemical screening through a method similar to that used for the screening of the stall-inducing compound for *PCSK9*. The data from the current study may be useful for selection of CPuORFs as potential targets for pharmaceutical drugs and for identification of regulatory CPuORFs.

Funding: This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI (grant nos. JP16K07387 to H.O., JP19K22892 to H.T., JP18H03330 to H.T. and H.O., JP18H02568 to M.I.); the Ministry of Education, Culture, Sports, Science and Technology (MEXT) KAKENHI (grant nos. JP17H05658 to S.N., JP26114703 to H.T, JP17H05659 to H.T); and the Naito Foundation (to H.O.).

Competing interests: none declared.

205 # References

206 Akimoto, C., *et al.* (2013) Translational repression of the McKusick-Kaufman syndrome transcript by unique
207 upstream open reading frames encoding mitochondrial proteins with alternative polyadenylation sites,
208 *Biochim Biophys Acta*, **1830**, 2728-2738.

209 Bhushan, S., *et al.* (2010) Structural basis for translational stalling by human cytomegalovirus and fungal arginine
210 attenuator peptide, *Mol. Cell*, **40**, 138-146.

211 Crowe, M.L., Wang, X.Q. and Rothnagel, J.A. (2006) Evidence for conservation and selection of upstream open
212 reading frames suggests probable encoding of bioactive peptides, *BMC Genomics*, **7**, 16.

213 Cruz-Vera, L.R., *et al.* (2011) Nascent polypeptide sequences that influence ribosome function, *Curr. Opin.*
214 *Microbiol.*, **14**, 160-166.

215 Emms, D.M. and Kelly, S. (2015) OrthoFinder: solving fundamental biases in whole genome comparisons
216 dramatically improves orthogroup inference accuracy, *Genome Biol.*, **16**, 157.

217 Hardy, S., *et al.* (2019) Magnesium-sensitive upstream ORF controls PRL phosphatase expression to mediate
218 energy metabolism, *Proc Natl Acad Sci U S A*, **116**, 2925-2934.

219 Hayden, C.A. and Bosco, G. (2008) Comparative genomic analysis of novel conserved peptide upstream open
220 reading frames in *Drosophila melanogaster* and other dipteran species, *BMC Genomics*, **9**, 61.

221 Hayden, C.A. and Jorgensen, R.A. (2007) Identification of novel conserved peptide uORF homology groups in
222 *Arabidopsis* and rice reveals ancient eukaryotic origin of select groups and preferential association with
223 transcription factor-encoding genes, *BMC Biol.*, **5**, 32.

224 Ingolia, N.T., *et al.* (2014) Ribosome profiling reveals pervasive translation outside of annotated protein-coding
225 genes, *Cell Rep.*, **8**, 1365-1379.

226 Ito, K. and Chiba, S. (2013) Arrest peptides: cis-acting modulators of translation, *Annu. Rev. Biochem.*, **82**,
227 171-202.

228 Lintner, N.G., *et al.* (2017) Selective stalling of human translation through small-molecule engagement of the

229 ribosome nascent chain, *PLoS Biol*, **15**, e2001882.

230 Morris, D.R. and Geballe, A.P. (2000) Upstream open reading frames as regulators of mRNA translation, *Mol.*
231 *Cell Biol.*, **20**, 8635-8642.

232 Samandi, S., *et al.* (2017) Deep transcriptome annotation enables the discovery and functional characterization of
233 cryptic small proteins, *Elife*, **6**.

234 Somers, J., Poyry, T. and Willis, A.E. (2013) A perspective on mammalian upstream open reading frame
235 function, *Int. J. Biochem. Cell Biol.*, **45**, 1690-1700.

236 Takahashi, H., *et al.* (2019) Comprehensive genome-wide identification of angiosperm upstream ORFs with
237 peptide sequences conserved in various taxonomic ranges using a novel pipeline, ESUCA, *bioRxiv*, 524090.

238 Takahashi, H., *et al.* (2012) BAIUCAS: a novel BLAST-based algorithm for the identification of upstream open
239 reading frames with conserved amino acid sequences and its application to the *Arabidopsis thaliana* genome,
240 *Bioinformatics*, **28**, 2231-2241.

241 Tran, M.K., Schultz, C.J. and Baumann, U. (2008) Conserved upstream open reading frames in higher plants,
242 *BMC Genomics*, **9**, 361.

243 van der Horst, S., *et al.* (2018) Novel pipeline identifies new upstream ORFs and non-AUG initiating main ORFs
244 with conserved amino acid sequences in the 5' leader of mRNAs in *Arabidopsis thaliana*, *RNA*, **25**,
245 292-304.

246 Vaughn, J.N., *et al.* (2012) Known and novel post-transcriptional regulatory sequences are conserved across plant
247 families, *RNA*, **18**, 368-384.

248 Wang, Z. and Sachs, M.S. (1997) Ribosome stalling is responsible for arginine-specific translational attenuation
249 in *Neurospora crassa*, *Mol. Cell Biol.*, **17**, 4904-4913.

250 Yamashita, Y., *et al.* (2017) Sucrose sensing through nascent peptide-meditated ribosome stalling at the stop
251 codon of Arabidopsis *bZIP11* uORF2, *FEBS Lett.*, **591**, 1266-1277.

252 Zerbino, D.R., *et al.* (2018) Ensembl 2018, *Nucleic Acids Res.*, **46**, D754-D761.

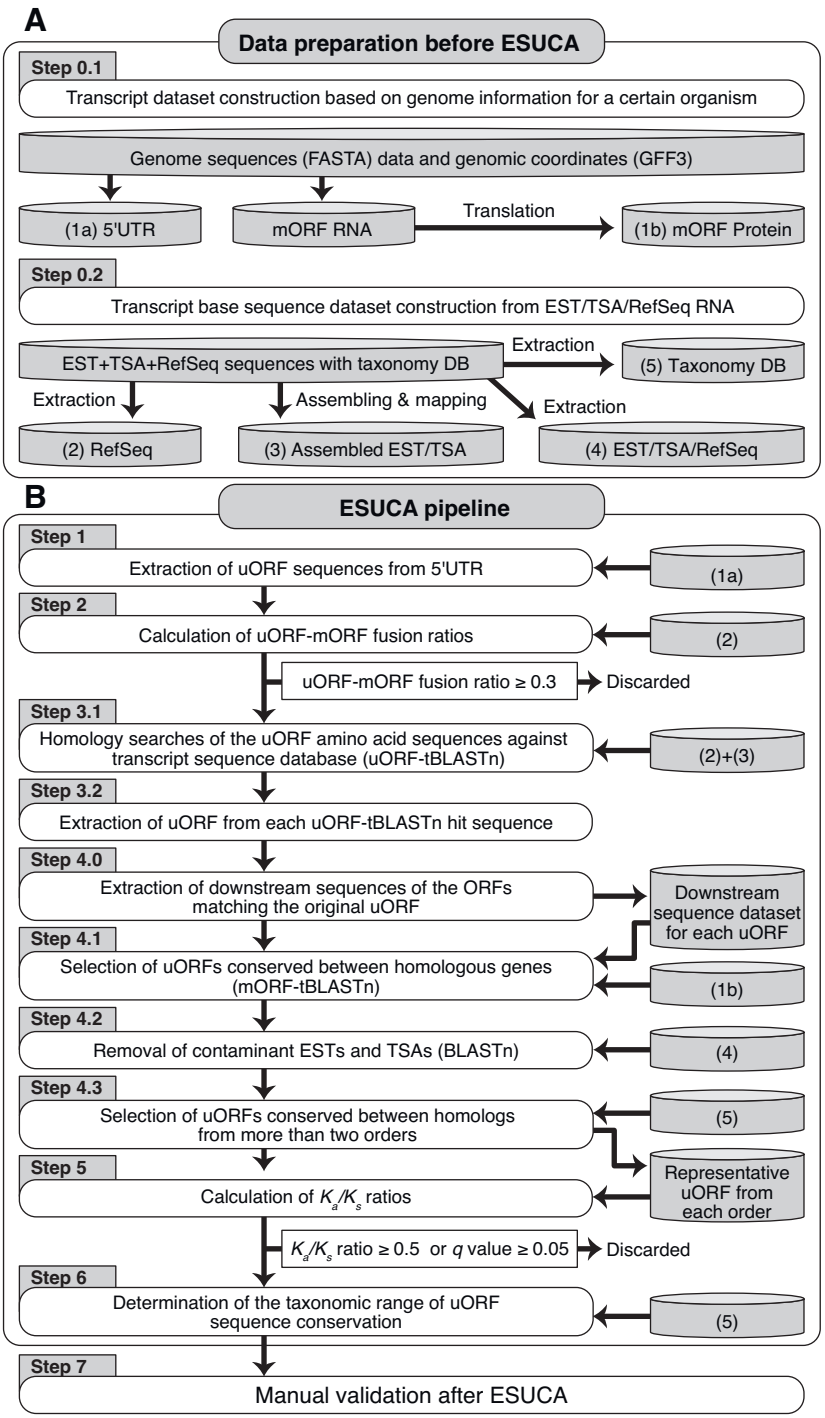
253

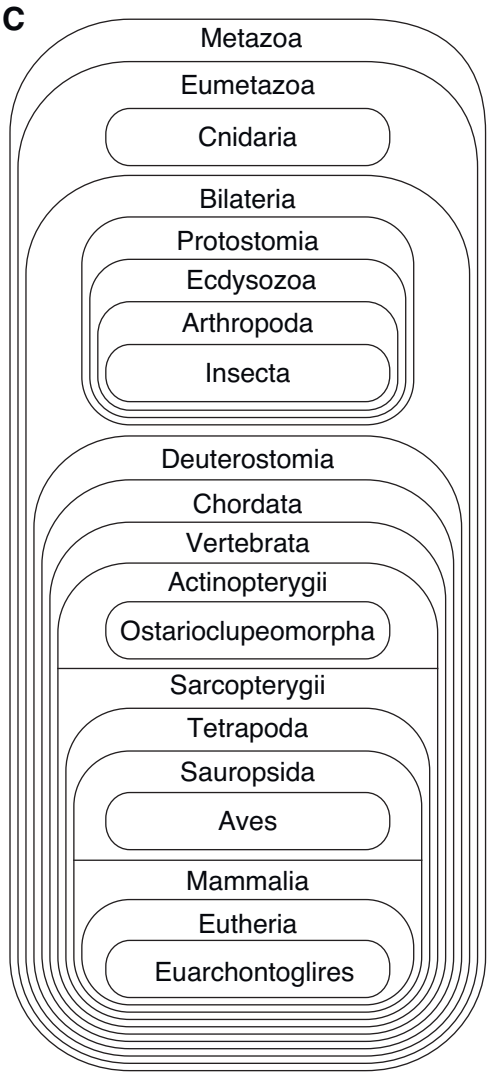
Figure Legends

Figure 1. Identification of animal CPuORFs using ESUCA. (A) Data preparation. (B) Outline of the ESUCA pipeline. Numbers with parenthesis indicate datasets labeled with the same numbers in A. (C) Defined animal taxonomic categories. (D) Numbers of identified CPuORFs.

Figure 2. Taxonomic conservation and experimental validation of 17 selected human CPuORFs. (A) Taxonomic ranges of conservation of CPuORFs examined in transient assays. Filled cells in each taxonomic category indicate the presence of uORF-tBLASTn and mORF-tBLASTn hits for CPuORFs of the indicated genes. (B) Reporter constructs used for transient assays. The hatched box in the frameshift (fs) mutant CPuORF indicates the frame-shifted region. Dotted boxes represent the first five nucleotides of the mORFs associated with the 17 human CPuORFs. (C) Relative luciferase activities of WT-aa (white) or frameshift (gray) CPuORF reporter plasmids. Means ± SDs of three biological replicates are shown. $*p < 0.05$.

Figure 1



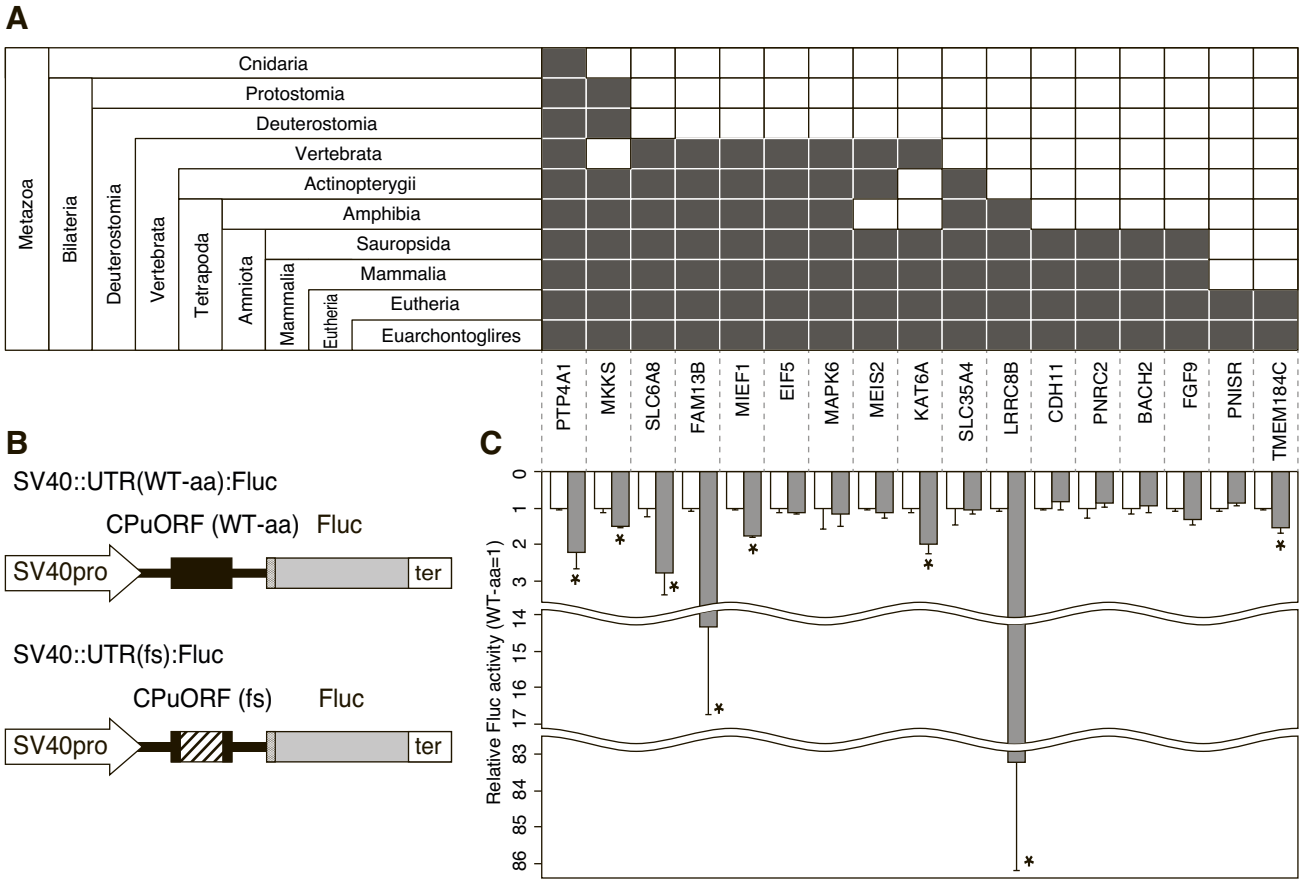


D

Species	CPuORF
D. melanogaster	37
D. rerio	156
G. gallus	230
H. sapiens	1094

Number of animal CPuORFs: 1517
Number of animal HGs: 1430

Figure 2


A
Metazoa
Cnidaria
Bilateria
Protostomia
Deuterostomia
Deuterostomia
Vertebrata
Vertebrata
Actinopterygii
Tetrapoda
Amphibia
Amniota
Sauropsida
Mammalia
Mammalia
Eutheria
Eutheria
Euarchontoglires
PTP4A1
MKKS
SLC6A8
FAM13B
MIEF1
EIF5
MAPK6
MEIS2
KAT6A
SLC35A4
LRRC8B
CDH11
PNRC2
BACH2
FGF9
PNISR
TMEM184C
B
SV40::UTR(WT-aa):Fluc
CPuORF (WT-aa)
Fluc
SV40pro
ter
SV40::UTR(fs):Fluc
CPuORF (fs)
Fluc
SV40pro
ter
C
Relative Fluc activity (WT-aa=1)
0
1
2
3
14
15
16
17
83
84
85
86