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

Hiro Takahashi[1,2,3#]*, Shido Miyaki[2#], Hitoshi Onouchi[4#], Taichiro Motomura[1], Nobuo Idesako[2], Anna Takahashi[5,6], Masataka Murase[1], Shuichi Fukuyoshi[7], Toshinori Endo[8], Kenji Satou[9] , Satoshi Naito[4,10], and Motoyuki Itoh[11]*

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

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

[3]Fundamental Innovative Oncology Core Center, National Cancer Center, Tokyo 104-0045, Japan

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

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

[6]College of Bioscience and Biotechnology, Chubu University, Kasugai 487-8501, Japan

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

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

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

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

[11]Graduate School of Pharmaceutical Science, Chiba University, 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

Upstream open reading frames (uORFs) are present in the 5′-untranslated regions of many eukaryotic mRNAs, and some peptides encoded by these regions play important regulatory roles in controlling main ORF (mORF) translation. We previously developed a novel pipeline, ESUCA, to comprehensively identify plant uORFs encoding functional peptides, based on genome-wide identification of uORFs with conserved peptide sequences (CPuORFs). Here, we applied ESUCA to diverse animal genomes, because animal CPuORFs have been identified only by comparing uORF sequences between a limited number of species, and how many previously identified CPuORFs encode regulatory peptides is unclear. By using ESUCA, 1,517 (1,373 novel and 144 known) CPuORFs were extracted from four evolutionarily divergent animal genomes. 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 one conserved only among Eutheria. 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.

42 # Introduction

43 The human genome contains many regions encoding potential functional small peptides outside the canonical
44 protein-coding regions [1]. Some upstream open reading frames (uORFs), which are located in the 5′-untranslated
45 regions (5′-UTRs) of mRNAs, have been shown to encode such functional small peptides [2-5]. uORFs are
46 cis-acting regulatory elements that control the translation of protein-coding main ORFs (mORFs) in various ways
47 [6,7]. In eukaryotes, 43S pre-initiation complexes (PICs) scan for a start codon along an mRNA from the 5′ end.
48 Therefore, PICs can recognize the start codon of a uORF and translate the uORF before reaching the downstream
49 mORF. In many cases, after translating a uORF, ribosomes dissociate from the mRNA or small ribosomal
50 subunits remain bound to the mRNA and resume scanning. When ribosomes dissociate from the mRNA after
51 uORF translation, ribosomes that have translated the uORF do not translate the downstream mORF. Therefore, if
52 the translation initiation efficiency of the uORF is high, the uORF exerts a substantial repressive effect on mORF
53 translation [6,7]. When a small ribosomal subunit resumes scanning after uORF translation, the ribosomes can
54 reinitiate translation at a downstream AUG codon. However, the reinitiation efficiency depends on the time
55 needed for the uORF translation and the distance between the uORF stop codon and the downstream start codon
56 [8-11]. The intercistronic distance required for efficient reinitiation depends on cellular availability of the ternary
57 complex that comprises eukaryotic initiation factor 2 (eIF2), GTP, and Met-tRNA$_i^{Met}$, and the level of the
58 available ternary complex is reduced under starvation or stress conditions [12]. These properties are utilized for the
59 translational regulation of yeast *GCN4* and mammalian *ATF4* and *ATF5* mRNAs [13-18]. In these mRNAs, there is
60 an inhibitory uORF downstream of the uORF that allows reinitiation. Under normal conditions, reinitiation
61 preferentially occurs at the start codon of the inhibitory uORF, and therefore, mORF translation is repressed. In
62 contrast, under starvation or stress conditions, reinitiation is delayed due to the reduced availability of the ternary
63 complex, and therefore, ribosomes more frequently bypass the inhibitory uORF and reinitiate translation at the
64 start codon of the mORF, resulting in enhanced mORF translation. Apart from mORF translation control, uORFs
65 can affect mRNA stability through the nonsense-mediated RNA decay pathway [19]. While the effects of most
66 uORFs on the expression of the mORF-encoded proteins are independent of the uORF-encoded sequences,
67 certain uORFs repress mORF translation in a peptide sequence-dependent manner. Most of these uORFs encode

68 regulatory peptides that cause ribosome stalling by interacting with components of the ribosomal exit tunnel
69 during uORF translation [4]. Ribosome stalling on a uORF results in translational repression of the downstream
70 mORF because the stalled ribosomes block the scanning of subsequently loaded PICs and prevent them from
71 reaching the start codon of the mORF [20]. In some genes, uORF-encoded peptides are involved in translational
72 regulation in response to metabolites or environmental stresses, whereby the uORF translation initiation
73 efficiency or the efficiency of ribosome stalling is regulated in a condition-dependent manner [4,21]. In the
74 sequence-dependent regulatory uORF of the mouse antizyme inhibitor (*AZIN1*) gene, which begins with a
75 non-canonical start codon [22], polyamine induces ribosome stalling, and the stalled ribosome causes ribosome
76 queuing by blocking the scanning of PICs [23]. This ribosome queuing promotes translation initiation at the
77 non-canonical start codon of the uORF by positioning PICs near the start codon, thereby enhancing the repressive
78 effect of the uORF on mORF translation. Apart from uORFs encoding regulatory peptides, some uORFs have
79 been reported to code for proteins with functions independent of the control of the downstream mORF [24-26].
80 To comprehensively identify uORFs encoding functional peptides or proteins, genome-wide searches
81 for uORFs with conserved peptide sequences (CPuORFs) have been conducted using comparative genomic
82 approaches in plants [27-32]. To date, 157 CPuORF families have been identified by comparing 5′-UTR sequences
83 among plant species. Of these, 101 families were identified in our previous studies by applying our original
84 methods, BAIUCAS [29] and ESUCA (an advanced version of BAIUCAS) [32] to the genomes of *Arabidopsis*, rice,
85 tomato, poplar, and grape.
86 ESUCA has many unique functions [32], such as efficient comparison of uORF sequences among an
87 unlimited number of species using BLAST, automatic determination of taxonomic ranges of CPuORF sequence
88 conservation, systematic calculation of $K_a/K_s$ ratios of CPuORF sequences, and wide compatibility with any
89 eukaryotic genome whose sequence database is registered in ENSEMBL [33]. By comparing uORF sequences
90 from certain species and those from many other species whose transcript sequence databases are available,
91 ESUCA enables more comprehensive identification of CPuORFs conserved in various taxonomic ranges than
92 conventional comparative genomic approaches, in which uORF sequences are compared among limited numbers
93 of selected species. In addition, to distinguish between "spurious" CPuORFs conserved because they encode

94 parts of mORF-encoded proteins and "true" CPuORFs conserved because of the functional constraints of their
95 encoded small peptides, ESUCA assesses whether a transcript containing a fusion of a uORF and an mORF is a
96 major or minor form among homologous transcripts [32]. By using these functions, ESUCA is able to efficiently
97 identify CPuORFs likely to encode functional small peptides. In fact, our recent study demonstrated that poplar
98 CPuORFs encoding regulatory peptides were efficiently identified using ESUCA by selecting ones conserved
99 across diverse eudicots [32].

100 Several studies on genome-wide identification of animal CPuORFs have been reported. By comparing
101 uORF sequences between human and mouse, 204 and 198 CPuORFs have been identified in human and mouse,
102 respectively [34]. In addition, by comparing uORF sequences among several species in dipteran, 44 CPuORFs have
103 been identified in fruit fly [35]. More recently, among translatable uORFs identified by ribosome profiling studies,
104 118, 80, 13, 50, and 37 CPuORFs in human, mouse, zebrafish, fruit fly, and nematode, respectively, have been
105 identified by Mackowiak *et al.* [36], and 97 CPuORFs in human have been identified by Samandi *et al.* [26]. In these
106 previous studies, uORF sequences were compared between a limited number of species. Therefore, further
107 comprehensive identification of animal CPuORFs was expected by applying the approach using ESUCA to
108 animal genomes. In addition, the relationships between the taxonomic ranges of CPuORF conservation and the
109 likelihood of having a regulatory function have not been studied in animals.

110 Accordingly, in this study, we applied ESUCA to the genomes of fruit fly, zebrafish, chicken, and
111 human to exhaustively identify animal CPuORFs and to determine the taxonomic range of their sequence
112 conservation. Using ESUCA, we identified 1,517 animal (1,373 novel and 144 known) CPuORFs belonging to
113 1,430 CPuORF families. Using transient expression assays, we examined the effects of 17 CPuORFs conserved
114 in various taxonomic ranges on mORF translation. Through this analysis, we identified seven novel regulatory
115 CPuORFs that repress mORF translation in a sequence-dependent manner.
116

117 ## Results

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

119 Prior to ESUCA application (Supplementary Fig. S1a and S1b), we counted the number of protein-coding genes
120 in four species, i.e., fruit fly, zebrafish, chicken, and human. The genes whose mORF-encoded amino acid
121 sequences were available in ENSEMBL (http://www.ensembl.org) were defined as protein-encoding genes in the
122 present study. As shown in Table 1, 13,938, 25,206, 14,697, and 19,956 genes were extracted for fruit fly,
123 zebrafish, chicken, and human, respectively. In step 1 of ESUCA, we extracted uORF sequences from the 5′-UTR
124 sequence of these genes, using the transcript sequence datasets described in the Methods section. In these datasets,
125 different transcript IDs are assigned to each splice variant from the same gene. To extract sequences of uORFs
126 and their downstream mORFs from all splice variants, we extracted uORF and mORF sequences from each of
127 the transcripts with different transcript IDs. The uORFs were extracted by searching the 5′-UTR sequence of each
128 transcript for an ATG codon and its nearest downstream in-frame stop codon (Supplementary Fig. S2a). As
129 shown in Table 1, 17,035, 39,616, 8,929, and 44,085 uORFs were extracted from 7,066, 14,453, 3,535, and
130 12,321 genes of fruit fly, zebrafish, chicken, and human genomes, respectively. In this analysis, when multiple
131 uORFs from splice variants of a gene shared the same stop or start codon, they were counted as one, but all
132 uORFs in splice variants were retained for further analyses. In step 2, we calculated the uORF–mORF fusion
133 ratio for each of the extracted uORFs. To assess whether transcripts bearing a uORF–mORF fusion are minor or
134 major forms among homologous transcripts, the ratio of the NCBI reference sequence (RefSeq) RNAs with a
135 uORF–mORF fusion to all RefSeq RNAs with both sequences similar to the uORF and its downstream mORF
136 was calculated as the uORF–mORF fusion ratio (Supplementary Fig. S2b). We discarded uORFs with
137 uORF–mORF fusion ratios equal to or greater than 0.3 (Supplementary Table S1). As shown in Table 1, the
138 numbers of uORFs were dramatically reduced after this step, suggesting that this step effectively excluded
139 "spurious" CPuORFs that were conserved because they encode parts of mORF-encoded proteins. In step 3.1, we
140 performed homology searches of the uORF amino acid sequences, using tBLASTn with an *E*-value cutoff of
141 2,000 (uORF-tBLASTn analysis). In this search, the amino acid sequence of each uORF was queried against an
142 animal transcript sequence database that contained contigs of assembled expressed sequence tags (ESTs) and
143 transcriptome shotgun assemblies (TSAs), singleton EST/TSA sequences, and RefSeq RNAs (See the Methods
144 for details). The uORFs with tBLASTn hits from other species were selected in this step. In step 3.2, an ORF
145 containing the amino acid sequence similar to the original uORF sequence was extracted from each of the
146 uORF-tBLASTn hits as putative uORFs (Supplementary Fig. S3a). In step 4.1, to confirm whether the

147 uORF-tBLASTn hits were derived from homologs of the original uORF-containing gene, the downstream
148 sequences of putative uORFs in the uORF-tBLASTn hits were subjected to another tBLASTn analysis
149 (Supplementary Fig. S3b and S3c). In this analysis, the amino acid sequence of the mORF in each original
150 uORF-containing transcript was used as a query, and the uORF-tBLASTn hits matching the mORF with an
151 *E*-value of less than $10^{-1}$ were extracted (mORF-tBLASTn analysis) (Supplementary Fig. S3c). If a
152 uORF-tBLASTn hit contained a partial or intact ORF sequence similar to the original mORF amino acid
153 sequence downstream of the putative uORF, it was considered to be derived from a homolog of the original
154 uORF-containing gene. In step 4.2, potential contaminant sequences derived from contaminating organisms, such
155 as parasites and infectious microorganisms, were excluded from the mORF-tBLASTn hits. In step 4.3, we
156 selected uORFs conserved in mORF homologs from at least two orders other than the order of the original
157 uORF; for example, in the search for human CPuORFs, we selected uORFs conserved in mORF homologs from at
158 least two orders other than Primates. When uORFs with identical sequences were extracted from different splice
159 variants of the same gene, we selected the one with the lowest median *E*-value in uORF-tBLASTn analysis. In
160 step 5, we calculated $K_a/K_s$ ratios of the uORFs to assess whether these uORF sequences were conserved at the
161 nucleotide or amino acid level. A $K_a/K_s$ ratio close to 1 indicates neutral evolution, whereas a $K_a/K_s$ ratio close to 0
162 suggests that purifying selection acted on the amino acid sequences [32]. For each of the uORFs extracted in step
163 4.3, one representative mORF-tBLASTn hit was selected from each order in which mORF-tBLASTn hits were
164 identified, and the putative uORFs in the selected mORF-tBLASTn hits and the original uORF sequence were
165 used for pairwise calculations of $K_a/K_s$ ratios. The uORFs with $K_a/K_s$ ratios less than 0.5 showing significant
166 differences from those of negative controls ($q < 0.05$) were selected as candidate CPuORFs (Supplementary
167 Table S1). In step 6, we determined the taxonomic range of sequence conservation of the candidate CPuORFs. In
168 this step, the representative mORF-tBLASTn hits selected in step 5 were classified into the 19 taxonomic
169 categories shown in Fig. 1a. On the basis of the presence of the mORF-tBLASTn hits in each taxonomic category,
170 the taxonomic range of sequence conservation was determined for each candidate CPuORF (Supplementary
171 Table S2). After the final step of ESUCA, 49, 192, 261, and 1,495 candidate CPuORFs were extracted from fruit
172 fly, zebrafish, chicken, and human, respectively (Table 1). To validate the sequence conservation of the candidate
173 CPuORFs, we generated multiple amino acid sequence alignment for each candidate CPuORF, using the original
174 uORF sequence and the representative putative uORF sequences used for calculating the $K_a/K_s$ ratio

175 (Supplementary Figs. S4 and S5). If the amino acid sequence of a uORF is evolutionarily conserved because of
176 functional constraints of the uORF-encoded peptide, the amino acid sequence in the functionally important region
177 of the peptide is expected to be conserved among the uORF and its orthologous uORFs. Therefore, we manually
178 checked whether the amino acid sequences in the same region are conserved among the uORF and putative
179 uORF sequences in the alignment of each candidate CPuORF. Then, we removed sequences that do not share the
180 consensus amino acid sequence in the conserved region. When this change resulted in the number of orders with
181 the mORF-tBLASTn hits becoming less than two, the candidate CPuORFs were discarded. In analyses of the
182 genomes of the four species, 66 candidate CPuORFs were discarded for this reason. In addition, when multiple
183 original uORFs in splice variants from the same gene partially shared amino acid sequences, the one with the
184 longest conserved region was manually selected on the basis of the uORF amino acid sequence alignments, and
185 the others were discarded. In the case where the length of the conserved regions is the same among splice variants,
186 the one with the lowest $K_a/K_s$ ratio was selected. When multiple original uORFs in splice variants from the same
187 gene overlapped each other in different reading frames, the one with the lowest $K_a/K_s$ ratio was selected, and the
188 others were discarded. After the manual validation, 1,517 animal CPuORFs (37 for fruit fly, 156 for zebrafish,
189 230 for chicken, and 1,094 for human) were identified as CPuORFs (Fig. 1). Of these, 1,373 CPuORFs have not
190 been reported and are therefore novel CPuORFs. The amino acid sequence alignments and detailed information
191 of the identified CPuORFs are shown in Supplementary Figs. S4, S5 and Table S1. It should be noted that the
192 sequence alignments shown in Supplementary Figs. S4 and S5 do not contain the putative uORF sequences
193 removed by manual validation. In the cases where one or more putative uORF sequences homologous to a
194 CPuORF were removed by the manual validation, we determined the taxonomic range of sequence conservation
195 of the CPuORF again after excluding the mORF-tBLASTn hits corresponding to the removed putative uORF
196 sequences. Figure 1b shows the numbers of the CPuORFs conserved in each of the 19 taxonomic categories. The
197 results shown in this figure indicate that CPuORFs conserved in various taxonomic ranges were identified by
198 ESUCA analyses of the fruit fly, zebrafish, chicken, and human genomes. The identified CPuORF-containing
199 genes were classified into 1,363 ortholog groups on the basis of similarities of mORF-encoded amino acid
200 sequences using OrthoFinder [37]. CPuORFs with similar amino acid sequences from the same ortholog groups
201 were categorized as the same CPuORF families (homology groups [HGs]; see the Methods for details). The
202 identified 1,517 CPuORFs were classified into 1,430 HGs. We assigned HG numbers to 1,430 HGs so that

203 CPuORF families conserved across wider taxonomic ranges and in higher numbers of orders could have smaller
204 HG numbers. When multiple CPuORF families were identified in the same ortholog groups, the same HG
205 number with a different subnumber was assigned to each of the families (e.g., HG0004.1 and HG0004.2;
206 Supplementary Table S1). It should be noted that the amino acid sequence alignments for each of the identified
207 CPuORFs were separately generated and are individually shown in Supplementary Figs. S4 and S5, even when
208 multiple homologous CPuORFs belonging to the same HG were identified.
209
210 **Sequence-dependent effects of CPuORFs on mORF translation**
211 To address the relationship between taxonomic ranges of CPuORF conservation and the likelihood of having a
212 regulatory function using transient expression assays, we selected 17 human CPuORFs conserved in various
213 taxonomic ranges, including a previously identified sequence-dependent regulatory CPuORF, the *PTP4A1*
214 CPuORF [38], as a positive control. Besides the *PTP4A1* CPuORF, we selected CPuORFs with well-conserved
215 C-terminal regions. This is because in many known sequence-dependent regulatory CPuORFs, the C-terminal
216 region of the CPuORF-encoded nascent peptide has been shown to be important to cause ribosome stalling [4]. Of
217 these 16 selected CPuORFs, those in the genes *eIF5*, *MKKS*, *MIEF1*, and *SLC35A4* have been reported to play
218 roles in regulating mORF translation [39-41]. However, the sequence dependence of their effects on mORF translation
219 have not been reported; therefore, we included these four CPuORFs. We examined the sequence-dependent effects
220 of these 17 CPuORFs on the expression of the downstream reporter gene using transient expression assays (Fig.
221 2). Other uORFs overlapping any of the selected CPuORFs were eliminated by introducing mutations that
222 changed the ATG codons of the overlapping uORFs to other codons but did not alter the amino acid sequences of
223 the CPuORFs (Supplementary Fig. S6). The resulting modified CPuORFs were used as CPuORFs bearing the
224 wild-type amino acid sequences (WT-aa CPuORFs) (Fig. 2b). To assess the importance of amino acid sequences
225 for the effects of these CPuORFs on mORF expression, frameshift mutations were introduced into the WT-aa
226 CPuORFs such that the amino acid sequences of their conserved regions could be altered (see the Methods and
227 Supplementary Figure S6 for details). In eight of the 17 CPuORFs including the *PTP4A1* CPuORF, the
228 introduced frameshift mutations significantly upregulated the expression of the reporter gene, indicating that these
229 CPuORFs repressed mORF expression in a sequence-dependent manner (Fig. 2c). One of the eight
230 sequence-dependent regulatory CPuORFs, the *TMEM184C* CPuORF, is conserved only among Eutheria (Fig.

231 2a). This result suggests that CPuORFs conserved only among Eutheria can have sequence-dependent regulatory
232 effects. This study identified five novel regulatory CPuORFs (in the genes *SLC6A8*, *FAM13B*, *KAT6A*, *LRRC8B*,
233 and *TMEM184C*). In addition, our results suggest that the repressive effects of the *MKKS* and *MIEF1* CPuORFs
234 on mORF expression at least partly depend on their encoding sequences.
235

236 **Discussion**

237 In the current study, we identified 1,517 (1,373 novel and 144 known) CPuORFs belonging to 1,430 HGs by
238 applying ESUCA to four animal genomes and selecting uORFs conserved across more than two orders. This
239 number of identified CPuORFs is much higher than those identified by previous studies on genome-wide
240 identification of CPuORFs [26,34-36]. Thus, our results demonstrate that the approach using ESUCA, in which uORF
241 sequences from certain species are compared with those from many other species, is highly effective in
242 comprehensively identifying CPuORFs conserved in various taxonomic ranges. Since 1,082 of the 1,094
243 CPuORFs identified from the human genome are conserved beyond Euarchontoglires (Supplementary Table S2),
244 and our transient expression assays suggested that CPuORFs conserved only among Eutheria can have
245 sequence-dependent regulatory functions, it is likely that many of the human CPuORFs identified in this study
246 are conserved because of functional constraints of their encoded peptides. It remains unknown why a much
247 higher number of CPuORFs was identified in the human genome by our search compared to the other three
248 species. One possible explanation is that in humans compared with the other three species, there are a higher
249 number of transcript isoforms with different 5′-UTR sequences, produced from the same genes via alternative
250 splicing or alternative transcription start site selection. The total numbers of such 5′-UTR variants of
251 protein-coding genes registered in the ENSEMBL transcript databases are 22,352, 32,974, 12,413, and 64,619 for
252 fruit fly, zebrafish, chicken, and human, respectively. The upregulation of mORF translation induced by the
253 removal of an inhibitory uORF via alternative splicing or alternative transcription start site selection has been
254 reported [42-44]. We speculate that the human genome might need many CPuORFs to differentially control the
255 translation efficiencies of the mORF among the 5′-UTR variants produced from the same gene.

256 The present study identified seven regulatory CPuORFs; however, the physiological roles of
257 translational regulation mediated by these CPuORFs remain to be elucidated. One of the known physiological

258 roles of CPuORFs is translational repression of the mORF in response to metabolites [45]. Alternatively, CPuORFs
259 are known to be involved in the promotion of mORF translation under specific conditions. In the latter case, a
260 CPuORF is efficiently translated by ribosomes under normal conditions, and therefore, mORF translation is
261 repressed, whereas ribosomes frequently bypass the CPuORF via leaky scanning under specific conditions such
262 as stress conditions, thereby promoting mORF translation [21]. The mORF of *SLC6A8*, one of the genes controlled
263 by the CPuORFs identified in this study, codes for a creatine transporter that transports creatine into cells [46].
264 Creatine is used as a readily available phosphate pool to regenerate ATP from ADP in cells. Therefore, the
265 CPuORF in *SLC6A8* could mediate feedback regulation of SLC6A8 expression in response to the cellular
266 creatine level to maintain creatine homeostasis. Alternatively, the CPuORF in *SLC6A8* could be involved in the
267 upregulation of SLC6A8 expression under ATP-deficient conditions, via the leaky scanning mechanism in which
268 the *SLC6A8* CPuORF is bypassed by ribosomes under these conditions. The *LRRC8B* mORF codes for a $Ca^{2+}$
269 leak channel localized in the endoplasmic reticulum and participates in intracellular $Ca^{2+}$ homeostasis [47].
270 Therefore, the CPuORF in *LRRC8B* could play a role in feedback regulation of LRRC8B expression in response
271 to the cytoplasmic $Ca^{2+}$ level to maintain $Ca^{2+}$ homeostasis. Interestingly, the *MKKS* and *MIEF1* CPuORFs have
272 been reported to code for mitochondrial proteins [24,26,39]. Therefore, these CPuORFs have been suggested to play two
273 roles, i.e., translational regulation and coding for mitochondrial proteins [26,39]. Our results additionally suggest that the
274 effects of the *MKKS* and *MIEF1* CPuORFs on mORF translation depend on their amino acid sequences, and
275 therefore, polypeptides encoded by these CPuORFs may have dual functions.

276 Nine of the 17 CPuORFs tested by transient expression assays exhibited no sequence-dependent effect
277 on mORF expression. One possible reason for this is that these CPuORFs might not have been translated in our
278 experimental conditions. The start codons of two of the nine CPuORFs are in a poor initiation context, with no
279 purine at −3 and no guanine at +4, where the A in ATG is +1. In contrast, the start codons of the remaining seven
280 CPuORFs are in an optimal or sub-optimal context, containing a purine at −3 and/or a guanine at +4
281 (Supplementary Fig. S6). If these CPuORFs were translated in our experimental conditions, other possible
282 explanations for the lack of sequence-dependent effects of these CPuORFs would be that these CPuORFs might
283 encode peptides with functions other than the control of mORF translation, or they might exert
284 sequence-dependent regulatory effects only under certain conditions.

285 Chemical screening recently identified a compound that causes nascent peptide-mediated ribosome
286 stalling in the mORF of the human *PCSK9* gene, resulting in specific translational inhibition of *PCSK9* and a
287 reduction in total plasma cholesterol levels [48]. Nascent peptide-mediated ribosome stalling in some of the
288 previously identified regulatory CPuORFs is promoted by metabolites, such as polyamine, arginine, and sucrose
289 [4,49]. Therefore, compounds that promote nascent peptide-mediated ribosome stalling in CPuORFs could be
290 identified by chemical screening through a method similar to that used for screening the stall-inducing compound
291 for *PCSK9*. The data from the current study may be useful for selecting CPuORFs as potential targets for
292 pharmaceuticals and for identifying regulatory CPuORFs.
293

294 # Methods

295 All procedures and protocols were approved by the Institutional Safety Committee for Recombinant DNA
296 Experiments at Chiba University. All methods were carried out in accordance with approved guidelines. All
297 human 5′-UTR sequences containing the CPuORFs used in the present study were artificially synthesized
298 (GenScript, Piscataway, NJ, USA) according to the RefSeq sequences (see "Plasmid construction and transient
299 reporter assays" for details).
300

301 **Extraction of CPuORFs using ESUCA**

302 ESUCA was developed as an advanced version of BAIUCAS [29] in our previous study [32]. ESUCA consists of six
303 steps, and some of these steps are divided into substeps, as shown in Supplementary Fig. S1a and S1b. To
304 identify animal CPuORFs using ESUCA, the following eight-step procedures were conducted, including the six
305 ESUCA steps: (0) data preparation for ESUCA, (1) uORF extraction from the 5′-UTR (Supplementary Fig. S2a),
306 (2) calculation of uORF–mORF fusion ratios (Supplementary Fig. S2b), (3) uORF-tBLASTn against transcript
307 sequence databases (Supplementary Fig. S3a), (4) mORF-tBLASTn against downstream sequence datasets for
308 each uORF (Supplementary Fig. S3b and S3c), (5) calculation of $K_a/K_s$ ratios, (6) determination of the taxonomic
309 range of uORF sequence conservation, and (7) manual validation after ESUCA. See the Materials and Methods
310 in our previous study [32] for details.
311

312 **Transcript dataset construction based on genome information (step 0.1)**

313 To identify animal CPuORFs, data preparation for ESUCA (step 0.1) was conducted as described in our previous
314 study [32]. We conducted data preparation for ESUCA to identify animal CPuORFs as follows. We used a genome
315 sequence file in FASTA format and a genomic coordinate file in GFF3 format obtained from Ensemble Metazoa
316 Release 33 (https://metazoa.ensembl.org/index.html) [50] to extract fruit fly (*Drosophila melanogaster*) uORF
317 sequences. We used genome sequence files in FASTA format and genomic coordinate files in GFF3 format
318 obtained from Ensemble Release 86 (https://metazoa.ensembl.org/index.html) [50] for zebrafish (*Danio rerio*),
319 chicken (*Gallus gallus*), and human (*Homo sapiens*). We extracted exon sequences from genome sequences on
320 the basis of genomic coordinate information and constructed transcript sequence datasets by combining exon
321 sequences. On the basis of the transcription start site and the translation initiation codon of each transcript in the
322 genomic coordinate files, we extracted 5′-UTR and mORF RNA sequences from the transcript sequence datasets,
323 as shown in Supplementary Fig. S1a (step 0.1). The 5′-UTR sequences were used at step 1 of ESUCA. The
324 mORF RNA sequences were translated into amino acid sequences (mORF proteins) and were used at step 4.1 of
325 ESUCA.

326

327 **Transcript base sequence dataset construction from EST/TSA/RefSeq RNA (step 0.2)**

328 To identify animal CPuORFs, data preparation for ESUCA (step 0.2) was conducted as described in our previous
329 study [32]. We conducted data preparation for ESUCA to identify animal CPuORFs. As shown in Supplementary
330 Fig. S1b, Metazoa RefSeq RNA sequences were used at steps 2 and 3.1 of ESUCA. Assembled EST/TSA
331 sequences generated by using Velvet [51] and Bowtie2 [52] were used at step 3.1 of ESUCA. Intact and merged
332 EST/TSA/RefSeq sequences were used at step 4.2 of ESUCA. Taxonomy datasets derived from
333 EST/TSA/RefSeq databases were used at steps 4.3 and 6 of ESUCA. See the Materials and Methods in our
334 previous study [32] for details.

335

336 **Determination of the taxonomic range of uORF sequence conservation for animal CPuORFs (step 6)**

337 To automatically determine the taxonomic range of the sequence conservation of each CPuORF, we first defined
338 20 animal taxonomic categories. The 20 defined taxonomic categories were Euarchontoglires, Eutheria other than
339 Euarchontoglires, Mammalia other than Eutheria, Aves, Sauropsida other than Aves, Amphibia (Tetrapoda other

340 than Sauropsida and Mammalia), Sarcopterygii other than Tetrapoda, Ostarioclupeomorpha, Actinopterygii other
341 than Ostarioclupeomorpha, Vertebrata other than Euteleostomi (Actinopterygii and Sarcopterygii), Chordata
342 other than Vertebrata, Deuterostomia other than Chordata, Insecta, Arthropoda other than Insecta, Ecdysozoa
343 other than Arthropoda, Lophotrochozoa (Protostomia other than Ecdysozoa), Bilateria other than Protostomia
344 and Deuterostomia, Cnidaria, Ctenophora (Eumetazoa other than Cnidaria and Bilateria), and Metazoa other than
345 Eumetazoa. Based on taxonomic lineage information of EST, TSA, and RefSeq RNA sequences, which were
346 provided by NCBI Taxonomy, the mORF-tBLASTn hit sequences selected for $K_a/K_s$ analysis were classified
347 into the 19 taxonomic categories (Fig. 1a and Supplementary Table S2). The category "Ctenophora" was omitted
348 from animal taxonomic categories because no sequences were classified to this category. For each CPuORF, the
349 numbers of transcript sequences classified into each category were counted and are shown in Supplementary
350 Table S2. These numbers represent the number of orders in which the amino acid sequence of each CPuORF is
351 conserved.
352
353 **Classification of animal CPuORFs into HGs**
354 Systematic numbering of animal CPuORF families (HGs) has not been reported to date. Here, we defined
355 systematic HG numbers for the identified 1,517 animal CPuORFs. Among these identified CPuORFs, those with
356 both similar uORF and mORF amino acid sequences were classified into the same HGs. We first determined
357 ortholog groups of CPuORF-containing genes, referred to as mORF clusters, based on similarities of
358 mORF-encoded amino acid sequences, using OrthoFinder [37]. The identified CPuORF-containing genes were
359 classified into 1,194 mORF clusters. CPuORFs contained in each ortholog group (mORF-cluster) were further
360 classified into uORF clusters as follows. We conducted a pairwise comparison of uORF peptide similarity using
361 BLASTp with *E*-values less than 2,000 in each mORF cluster. Binarized distance matrixes consisting of 0 (hit) or
362 1 (no-hit) were generated by this comparison. Hierarchical clustering with single linkage with the cutoff
363 parameter ($h$ = 0.5) was applied to these matrixes for the construction of uORF clusters. In total, 1,336
364 uORF–mORF clusters were generated automatically. We determined 1,430 clusters by manually checking the
365 alignments of uORFs and mORFs. We assigned HG numbers to the 1,430 clusters so that CPuORF families
366 conserved across wider taxonomic categories could have smaller HG numbers. When there were multiple
367 CPuORF families conserved in the same taxonomic categories, smaller HG numbers were assigned to CPuORF

368 families conserved in higher numbers of orders. The same HG number with a different sub-number was assigned
369 to CPuORFs in genes of the same ortholog group with dissimilar uORF sequences (e.g., HG0004.1 and
370 HG0004.2; Supplementary Table S1).
371
372 **Plasmid construction and transient reporter assays**
373 pSV40:Fluc was generated by inserting the SV40 promoter (*Bgl*II/*Hind*III fragment) from pRL-SV40 (Promega,
374 Madison, WI, USA) into the *Kpn*I site of pGL4.10[luc2] (Promega) by blunt-end cloning. The 5′-UTR sequences
375 containing the selected CPuORFs (*Sac*I/*Xho*I fragment) were fused to the Fluc-coding sequence by subcloning
376 the CPuORFs into the *Sac*I/*Xho*I site of pSV40:luc2 to generate the WT-aa reporter construct
377 (pSV40:UTR(WT-aa):Fluc; Fig. 2b, Supplementary Fig. S6). To assess the importance of the amino acid
378 sequences with regard to the effects of these CPuORFs on mORF translation, frameshift mutations were
379 introduced into the CPuORFs so that the amino acid sequences of their conserved regions could be altered. A +1
380 or −1 frameshift was introduced upstream or within the conserved region of each CPuORF, and another
381 frameshift was introduced before the stop codon to shift the reading frame back to the original frame
382 (pSV40:UTR(fs):Fluc; Fig. 2b, Supplementary Fig. S6). DNA fragments containing the CPuORFs of either
383 WT-aa or fs mutants from the *PTP4A1*, *MKKS*, *SLC6A8*, *FAM13B*, *MIEF1*, *EIF5*, *MAPK6*, *MEIS2*, *KAT6A*,
384 *SLC35A4*, *LRRC8B*, *CDH11*, *PNRC2*, *BACH2*, *FGF9*, *PNISR*, and *TMEM184C* genes were synthesized
385 (GenScript) and subcloned into pSV40:Fluc, as shown in Fig. 2b and Supplementary Table S5. These reporter
386 constructs were each transfected into human HEK293T cells. HEK293T cells (16,000/well) were co-transfected
387 with 80 ng/well of a pSV40:UTR:Fluc reporter plasmid and 1.6 ng/well pGL4.74[hRluc/TK] plasmid (Promega).
388 After 24 h, firefly luciferase (Fluc) and *Renilla* luciferase (Rluc) activities were measured according to the
389 Dual-Luciferase Reporter Assay protocol (Promega) using the GloMaxR-Multi Detection System (Promega).
390 Fluc activity was normalized by Rluc activity to correct for differences in cell viability and transfection efficiency.
391
392 **Statistical and informatics analyses**
393 All programs, except for existing stand-alone programs, such as NCBI-BLAST+ ver. 2.6.0 [53], Clustal Ω ver. 1.2.2
394 [54], OrthoFinder ver. 1.1.4 [37], Velvet ver. 1.2.10 [51], Bowtie2 ver. 2.2.9 [52], and Jalview ver. 2.10.2 [55], were written in
395 R (www.r-project.org). We also used R libraries, GenomicRanges ver. 1.32.7 [56], exactRankTests ver. 0.8.30,

Biostrings ver. 2.48.0, and seqinr ver. 3.4.5 [57]. Statistical differences between the control (WT-aa) and fs constructs were determined by Student's *t*-tests in transient assays.

**Data availability**

Individual FASTA-formatted files used in Supplementary Fig. S4 are available from http://www.p.kanazawa-u.ac.jp/~bukka/data/FASTA-Formatted.1517CPuORFs.zip or https://www.takahashi-lab.com/db/FASTA-Formatted.1517CPuORFs.zip. The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

424 # References


425 1 Ingolia, N. T. *et al.* Ribosome profiling reveals pervasive translation outside of annotated protein-coding
426 genes. *Cell Rep.* **8**, 1365-1379 (2014).
427 2 Morris, D. R. & Geballe, A. P. Upstream open reading frames as regulators of mRNA translation. *Mol.*
428 *Cell Biol.* **20**, 8635-8642 (2000).
429 3 Cruz-Vera, L. R., Sachs, M. S., Squires, C. L. & Yanofsky, C. Nascent polypeptide sequences that
430 influence ribosome function. *Curr. Opin. Microbiol.* **14**, 160-166 (2011).
431 4 Ito, K. & Chiba, S. Arrest peptides: cis-acting modulators of translation. *Annu. Rev. Biochem.* **82**,
432 171-202 (2013).
433 5 Somers, J., Poyry, T. & Willis, A. E. A perspective on mammalian upstream open reading frame
434 function. *Int. J. Biochem. Cell Biol.* **45**, 1690-1700 (2013).
435 6 Hood, H. M., Neafsey, D. E., Galagan, J. & Sachs, M. S. Evolutionary roles of upstream open reading
436 frames in mediating gene regulation in fungi. *Annu Rev Microbiol* **63**, 385-409 (2009).
437 7 Zhang, H., Wang, Y. & Lu, J. Function and Evolution of Upstream ORFs in Eukaryotes. *Trends*
438 *Biochem Sci* **44**, 782-794 (2019).
439 8 Child, S. J., Miller, M. K. & Geballe, A. P. Translational control by an upstream open reading frame in
440 the HER-2/neu transcript. *J Biol Chem* **274**, 24335-24341 (1999).
441 9 Kozak, M. Constraints on reinitiation of translation in mammals. *Nucleic Acids Res* **29**, 5226-5232
442 (2001).
443 10 Rajkowitsch, L., Vilela, C., Berthelot, K., Ramirez, C. V. & McCarthy, J. E. Reinitiation and recycling
444 are distinct processes occurring downstream of translation termination in yeast. *J Mol Biol* **335**, 71-85
445 (2004).
446 11 Vilela, C., Ramirez, C. V., Linz, B., Rodrigues-Pousada, C. & McCarthy, J. E. Post-termination
447 ribosome interactions with the 5′UTR modulate yeast mRNA stability. *EMBO J* **18**, 3139-3152 (1999).

448 12 Hinnebusch, A. G. Translational regulation of yeast GCN4. A window on factors that control
449 initiator-trna binding to the ribosome. *J Biol Chem* **272**, 21661-21664 (1997).
450 13 Hatano, M. *et al.* The 5'-untranslated region regulates ATF5 mRNA stability via nonsense-mediated
451 mRNA decay in response to environmental stress. *FEBS J* **280**, 4693-4707 (2013).
452 14 Hinnebusch, A. G. Translational regulation of GCN4 and the general amino acid control of yeast. *Annu*
453 *Rev Microbiol* **59**, 407-450 (2005).
454 15 Lu, P. D., Harding, H. P. & Ron, D. Translation reinitiation at alternative open reading frames regulates
455 gene expression in an integrated stress response. *J Cell Biol* **167**, 27-33 (2004).
456 16 Vattem, K. M. & Wek, R. C. Reinitiation involving upstream ORFs regulates ATF4 mRNA translation
457 in mammalian cells. *Proc Natl Acad Sci USA* **101**, 11269-11274 (2004).
458 17 Watatani, Y. *et al.* Stress-induced translation of ATF5 mRNA is regulated by the 5'-untranslated region.
459 *J Biol Chem* **283**, 2543-2553 (2008).
460 18 Zhou, D. *et al.* Phosphorylation of eIF2 directs ATF5 translational control in response to diverse stress
461 conditions. *J Biol Chem* **283**, 7064-7073 (2008).
462 19 Ramani, A. K. *et al.* High resolution transcriptome maps for wild-type and nonsense-mediated
463 decay-defective *Caenorhabditis elegans*. *Genome Biol* **10**, R101 (2009).
464 20 Wang, Z. & Sachs, M. S. Ribosome stalling is responsible for arginine-specific translational attenuation
465 in *Neurospora crassa*. *Mol. Cell Biol.* **17**, 4904-4913 (1997).
466 21 Young, S. K. & Wek, R. C. Upstream Open Reading Frames Differentially Regulate Gene-specific
467 Translation in the Integrated Stress Response. *J Biol Chem* **291**, 16927-16935 (2016).
468 22 Ivanov, I. P., Loughran, G. & Atkins, J. F. uORFs with unusual translational start codons autoregulate
469 expression of eukaryotic ornithine decarboxylase homologs. *Proc Natl Acad Sci USA* **105**, 10079-10084
470 (2008).
471 23 Ivanov, I. P. *et al.* Polyamine Control of Translation Elongation Regulates Start Site Selection on
472 Antizyme Inhibitor mRNA via Ribosome Queuing. *Mol Cell* **70**, 254-264 e256 (2018).

473 24 Brown, A. *et al.* Structures of the human mitochondrial ribosome in native states of assembly. *Nat Struct*
474 *Mol Biol* **24**, 866-869 (2017).
475 25 Lorenzo-Orts, L. *et al.* Concerted expression of a cell cycle regulator and a metabolic enzyme from a
476 bicistronic transcript in plants. *Nat Plants* **5**, 184-193 (2019).
477 26 Samandi, S. *et al.* Deep transcriptome annotation enables the discovery and functional characterization
478 of cryptic small proteins. *Elife* **6** (2017).
479 27 Hayden, C. A. & Jorgensen, R. A. Identification of novel conserved peptide uORF homology groups in
480 *Arabidopsis* and rice reveals ancient eukaryotic origin of select groups and preferential association with
481 transcription factor-encoding genes. *BMC Biol.* **5**, 32 (2007).
482 28 Tran, M. K., Schultz, C. J. & Baumann, U. Conserved upstream open reading frames in higher plants.
483 *BMC Genomics* **9**, 361 (2008).
484 29 Takahashi, H., Takahashi, A., Naito, S. & Onouchi, H. BAIUCAS: a novel BLAST-based algorithm for
485 the identification of upstream open reading frames with conserved amino acid sequences and its
486 application to the *Arabidopsis thaliana* genome. *Bioinformatics* **28**, 2231-2241 (2012).
487 30 Vaughn, J. N., Ellingson, S. R., Mignone, F. & Arnim, A. Known and novel post-transcriptional
488 regulatory sequences are conserved across plant families. *RNA* **18**, 368-384 (2012).
489 31 van der Horst, S., Snel, B., Hanson, J. & Smeekens, S. Novel pipeline identifies new upstream ORFs
490 and non-AUG initiating main ORFs with conserved amino acid sequences in the 5' leader of mRNAs in
491 *Arabidopsis thaliana*. *RNA* **25**, 292-304 (2018).
492 32 Takahashi, H. *et al.* Comprehensive genome-wide identification of angiosperm upstream ORFs with
493 peptide sequences conserved in various taxonomic ranges using a novel pipeline, ESUCA. *BMC*
494 *Genomics* **21**, 260 (2020).
495 33 Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46**, D754-D761 (2018).
496 34 Crowe, M. L., Wang, X. Q. & Rothnagel, J. A. Evidence for conservation and selection of upstream
497 open reading frames suggests probable encoding of bioactive peptides. *BMC Genomics* **7**, 16 (2006).

498 35 Hayden, C. A. & Bosco, G. Comparative genomic analysis of novel conserved peptide upstream open
499 reading frames in *Drosophila melanogaster* and other dipteran species. *BMC Genomics* **9**, 61 (2008).
500 36 Mackowiak, S. D. *et al.* Extensive identification and analysis of conserved small ORFs in animals.
501 *Genome Biol* **16**, 179 (2015).
502 37 Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons
503 dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).
504 38 Hardy, S. *et al.* Magnesium-sensitive upstream ORF controls PRL phosphatase expression to mediate
505 energy metabolism. *Proc Natl Acad Sci USA* **116**, 2925-2934 (2019).
506 39 Akimoto, C. *et al.* Translational repression of the McKusick-Kaufman syndrome transcript by unique
507 upstream open reading frames encoding mitochondrial proteins with alternative polyadenylation sites.
508 *Biochim Biophys Acta* **1830**, 2728-2738 (2013).
509 40 Andreev, D. E. *et al.* Translation of 5' leaders is pervasive in genes resistant to eIF2 repression. *Elife* **4**,
510 e03971 (2015).
511 41 Andreev, D. E. *et al.* Oxygen and glucose deprivation induces widespread alterations in mRNA
512 translation within 20 minutes. *Genome Biol* **16**, 90 (2015).
513 42 Hoang, H. D. *et al.* Induction of an Alternative mRNA 5' Leader Enhances Translation of the Ciliopathy
514 Gene Inpp5e and Resistance to Oncolytic Virus Infection. *Cell Rep* **29**, 4010-4023 e4015 (2019).
515 43 Kurihara, Y. *et al.* Transcripts from downstream alternative transcription start sites evade
516 uORF-mediated inhibition of gene expression in Arabidopsis. *Proc Natl Acad Sci USA* **115**, 7831-7836
517 (2018).
518 44 Ord, T., Ord, D., Koivomagi, M., Juhkam, K. & Ord, T. Human TRB3 is upregulated in stressed cells
519 by the induction of translationally efficient mRNA containing a truncated 5'-UTR. *Gene* **444**, 24-32
520 (2009).
521 45 van der Horst, S., Filipovska, T., Hanson, J. & Smeekens, S. Metabolite Control of Translation by
522 Conserved Peptide uORFs: The Ribosome as a Metabolite Multisensor. *Plant Physiol* **182**, 110-122

523 (2020).

524 46 Skelton, M. R. *et al.* Creatine transporter (CrT; Slc6a8) knockout mice as a model of human CrT
525 deficiency. *PLoS One* **6**, e16187 (2011).

526 47 Ghosh, A., Khandelwal, N., Kumar, A. & Bera, A. K. Leucine-rich repeat-containing 8B protein is
527 associated with the endoplasmic reticulum $Ca^{2+}$ leak in HEK293 cells. *J Cell Sci* **130**, 3818-3828
528 (2017).

529 48 Lintner, N. G. *et al.* Selective stalling of human translation through small-molecule engagement of the
530 ribosome nascent chain. *PLoS Biol* **15**, e2001882 (2017).

531 49 Yamashita, Y. *et al.* Sucrose sensing through nascent peptide-meditated ribosome stalling at the stop
532 codon of Arabidopsis *bZIP11* uORF2. *FEBS Lett.* **591**, 1266-1277 (2017).

533 50 Cunningham, F. *et al.* Ensembl 2019. *Nucleic Acids Res* **47**, D745-D751 (2019).

534 51 Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs.
535 *Genome Res.* **18**, 821-829 (2008).

536 52 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357-359
537 (2012).

538 53 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search
539 programs. *Nucleic Acids Res.* **25**, 3389-3402 (1997).

540 54 Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using
541 Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).

542 55 Clamp, M., Cuff, J., Searle, S. M. & Barton, G. J. The Jalview Java alignment editor. *Bioinformatics* **20**,
543 426-427 (2004).

544 56 Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9**,
545 e1003118 (2013).

546 57 Charif, D. & Lobry, J. R. in *Structural Approaches to Sequence Evolution: Molecules, Networks,*
547 *Populations* (eds U. Bastolla, M. Porto, H.E. Roman, & M. Vendruscolo) 207-232 (Springer

Verlag, 2007).

**Acknowledgement**

This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI (grant nos. JP19H02917 and JP16K07387 to H.O.; JP19K22892 to H.T.; JP18H03330 to H.T., M.I, and H.O.; and 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, and JP17H05659 to H.T); the Naito Foundation (to H.O.); and the Research Foundation for the Electrotechnology of Chubu (to H.T.). We would like to thank Editage (www.editage.com) for English language editing. This is a preprint of an article published in *Scientific Reports*. The final authenticated version is available online at: https://doi.org/10.1038/s41598-020-73307-6.

**Author contributions**

H.T., H.O., and M.I. designed the study. H.T. and S.M. performed experiments and analyzed the data under the supervision of S.F., T.E., K.S., S.N., and M.I. H.T., M.M., N.I., T.M., and A.T. contributed reagents/materials/analysis tools. H.T., H.O., M.I., and S.M. wrote the article with the contribution of all coauthors.

**Additional information**

Supplementary information accompanies this paper.

Competing financial interests: The authors declare no competing financial interests.

603 **Table 1**. Numbers of uORFs, protein-coding genes, and assembled EST/TSA and RefSeq sequences extracted at

604 each step of ESUCA.

| Step | *Drosophila melanogaster* | | | *Danio rerio* | | | *Gallus gallus* | | | *Homo sapiens* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | uORF[1] | Gene | EST/TSA +RefSeq | uORF[1] | Gene | EST/TSA +RefSeq | uORF[1] | Gene | EST/TSA +RefSeq | uORF[1] | Gene | EST/TSA +RefSeq |
| Before selection | - | 13,938 | - | - | 25,206 | - | - | 14,697 | - | - | 19,956 | - |
| Step 1 | 17,035 | 7,066 | - | 39,616 | 14,453 | - | 8,929 | 3,535 | - | 44,085 | 12,321 | - |
| Step 2 | 5,040 | 2,343 | - | 3,599 | 2,323 | - | 1,320 | 767 | - | 15,069 | 6,568 | - |
| Step 3.1 | 4,900 | 2,308 | 1,854,900 | 3,494 | 2,271 | 1,822,408 | 1,275 | 751 | 668,417 | 14,529 | 6,408 | 7,577,191 |
| Step 3.2 | 4,882 | 2,297 | 873,484 | 3,479 | 2,261 | 846,829 | 1,271 | 750 | 314,665 | 14,499 | 6,399 | 3,711,515 |
| Step 4.1 | 4,307 | 2,076 | 40,982 | 2,549 | 1,689 | 37,125 | 1,122 | 668 | 42,622 | 13,993 | 6,217 | 383,797 |
| Step 4.2 | 4,294 | 2,067 | 40,894 | 2,543 | 1,688 | 36,434 | 1,119 | 665 | 41,306 | 13,970 | 6,215 | 378,480 |
| Step 4.3 | 49 | 40 | 1,212 | 408 | 343 | 4,082 | 774 | 485 | 8,171 | 5,262 | 3,067 | 33,776 |
| Step 5 | 49 | 40 | 1,212 | 192 | 180 | 2,798 | 261 | 221 | 4,074 | 1,495 | 1,201 | 12,402 |
| Step 7 | 37 | 36 | 1,072 | 156 | 154 | 2,729 | 230 | 209 | 3,945 | 1,094 | 969 | 9,964 |

605 [1] When multiple uORFs in a transcript shared the same stop or start codon, they were counted as one.

606

607

608

609

610

611

612

613

614

615

616

617

618

619

620

621

622 # Figure Legends

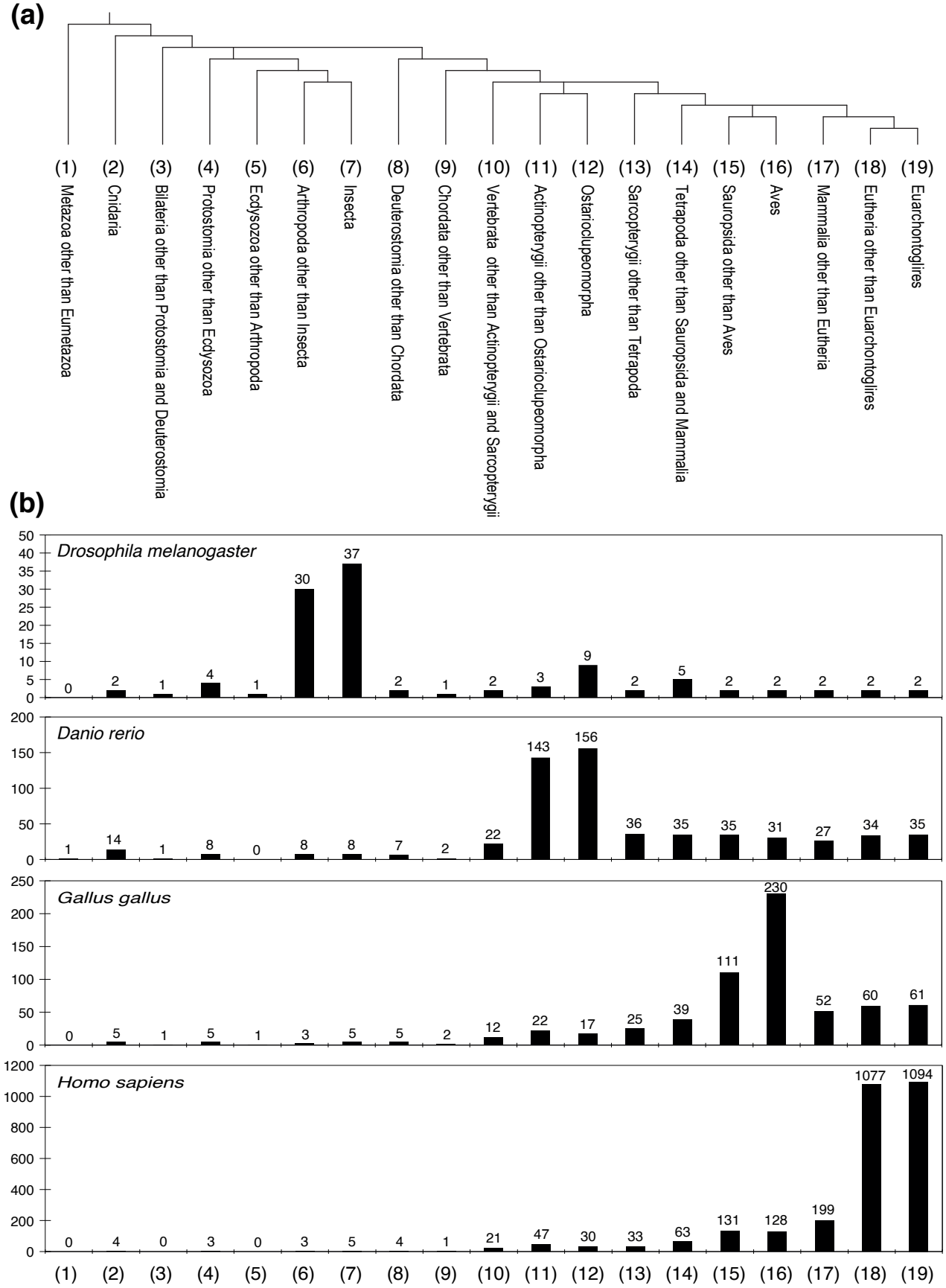
623 **Figure 1.** Numbers of CPuORFs extracted by ESUCA in each taxonomic category. (**a**) Cladogram showing the
624 relationship among the 19 taxonomic categories defined in this study. Fruit fly, zebrafish, chicken, and human
625 belong to Diptera, Cypriniformes, Galliformes, and Primates, respectively. Diptera, Cypriniformes, Galliformes,
626 and Primates belong to Insecta, Ostarioclupeomorpha, Aves, and Euarchontoglires, respectively. (**b**) Graphs
627 showing the numbers of CPuORFs extracted by ESUCA analyses of the genomes of the indicated species.

628

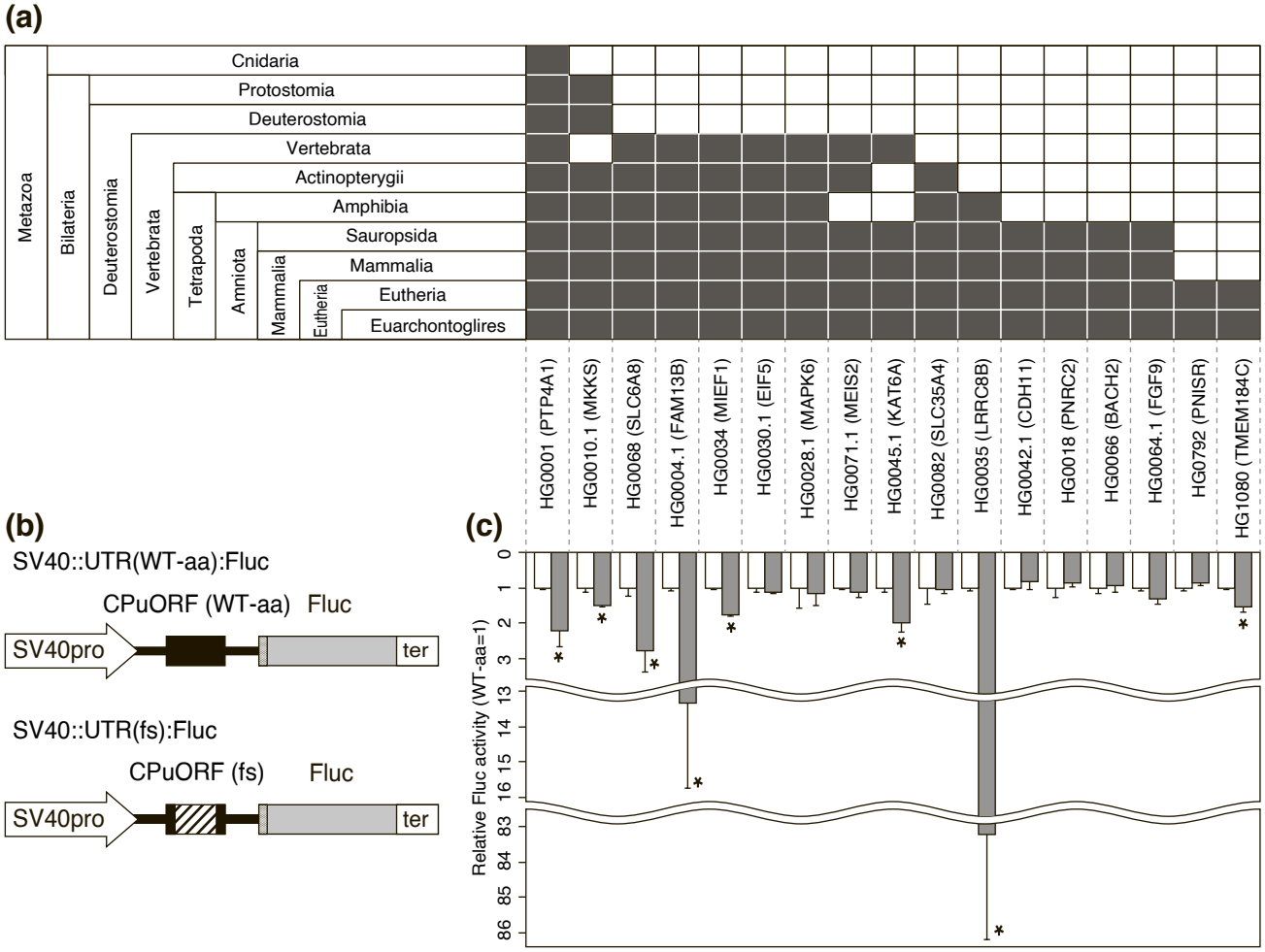
629 **Figure 2.** Taxonomic conservation and experimental validation of 17 selected human CPuORFs. (**a**) Taxonomic
630 ranges of conservation of CPuORFs examined in transient assays. Filled cells in each taxonomic category
631 indicate the presence of mORF-tBLASTn hits for CPuORFs of the indicated genes. (**b**) Reporter constructs used
632 for transient assays. The hatched box in the frameshift (fs) mutant CPuORF indicates the frame-shifted region.
633 Dotted boxes represent the first five nucleotides of the mORFs associated with the 17 human CPuORFs. See
634 Supplementary Fig. S6 for the exact position and the length of each CPuORF and the exact frame-shifted region.
635 (**c**) Relative luciferase (Fluc) activities of WT-aa (white) or fs (gray) CPuORF reporter plasmids. Means ± SDs of
636 at least three biological replicates are shown. *$p < 0.05$.

637

# Figure 1


(a)
(1) Metazoa other than Eumetazoa
(2) Cnidaria
(3) Bilateria other than Protostomia and Deuterostomia
(4) Protostomia other than Ecdysozoa
(5) Ecdysozoa other than Arthropoda
(6) Arthropoda other than Insecta
(7) Insecta
(8) Deuterostomia other than Chordata
(9) Chordata other than Vertebrata
(10) Vertebrata other than Actinopterygii and Sarcopterygii
(11) Actinopterygii other than Ostariocluepeomorpha
(12) Ostariocluepeomorpha
(13) Sarcopterygii other than Tetrapoda
(14) Tetrapoda other than Sauropsida and Mammalia
(15) Sauropsida other than Aves
(16) Aves
(17) Mammalia other than Eutheria
(18) Eutheria other than Euarchontoglires
(19) Euarchontoglires
(b)
Drosophila melanogaster
0 2 1 4 1 30 37 2 1 2 3 9 2 5 2 2 2 2 2
Danio rerio
1 14 1 8 0 8 8 7 2 22 143 156 36 35 35 31 27 34 35
Gallus gallus
0 5 1 5 1 3 5 5 2 12 22 17 25 39 111 230 52 60 61
Homo sapiens
0 4 0 3 0 3 5 4 1 21 47 30 33 63 131 128 199 1077 1094
(1) (2) (3) (4) (5) (6) (7) (8) (9) (10) (11) (12) (13) (14) (15) (16) (17) (18) (19)

# Figure 2


(a)
Metazoa
Cnidaria
Bilateria
Protostomia
Deuterostomia
Deuterostomia
Vertebrata
Vertebrata
Actinopterygii
Tetrapoda
Amphibia
Amniota
Sauropsida
Mammalia
Mammalia
Eutheria
Eutheria
Euarchontoglires
HG0001 (PTP4A1)
HG0010.1 (MKKS)
HG0068 (SLC6A8)
HG0004.1 (FAM13B)
HG0034 (MIEF1)
HG0030.1 (EIF5)
HG0028.1 (MAPK6)
HG0071.1 (MEIS2)
HG0045.1 (KAT6A)
HG0082 (SLC35A4)
HG0035 (LRRC8B)
HG0042.1 (CDH11)
HG0018 (PNRC2)
HG0066 (BACH2)
HG0064.1 (FGF9)
HG0792 (PNISR)
HG1080 (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
13
14
15
16
83
84
85
86