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

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. 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 closely related species, and how many previously identified CPuORFs encode regulatory peptides is unclear. By using ESUCA, 1,517 (1,425 novel and 92 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 the 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.

Introduction

The human genome contains many regions encoding potential functional small peptides outside of the well-annotated protein-coding regions [1]. 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) [2-5]. During the translation of these regulatory uORFs, nascent peptides interact inside the ribosomal exit tunnel to cause ribosome stalling [6]. 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 [7]. 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 [8-13]. 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 [10] and ESUCA (an advanced version of BAIUCAS) [13] to genomes of *Arabidopsis*, rice, tomato, poplar, and grape.

ESUCA has many unique functions [13], 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 [14]. 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 [13]. By using these functions, ESUCA can efficiently identify CPuORFs likely to encode functional small peptides. In fact, our recent study demonstrated that poplar CPuORFs encoding regulatory peptides were efficiently identified by selecting ones conserved across diverse eudicots using ESUCA [13].

To date, only a few studies on genome-wide identification of animal CPuORFs have reported. In these previous studies, uORF sequences were compared between a limited number of closely related species, such as human and mouse or several species in dipteran, leading to identification of 204 and 198 CPuORFs in human and mouse, respectively [15], and 44 CPuORFs in fruit fly [16]. Additionally, the relationships between taxonomic ranges of CPuORF conservation and the likelihood of having a regulatory function have not been studied in animals.

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

Results

Genome-wide search for animal CPuORFs using ESUCA

Prior to ESUCA application (Fig. 1a and 1b), we counted the number of protein-coding genes for four species, i.e., fruit fly, zebrafish, chiken, and human. As shown in Supplementary Table S1, 13,938, 25,206, 14,697, and 19,956 genes were extracted for fruit fly, zebrafish, chiken, and human, respectively. After step 1 of ESUCA, we calculated the numbers of uORFs and protein-coding genes with any uORF for each species. As shown in Supplementary Table S1, 17,035 (7,066), 39,616 (14,453), 8,929 (3,535), and 44,085 (12,321) uORFs (genes) were extracted from fruit fly, zebrafish, chicken, and human genomes, respectively. In this analysis, when multiple uORFs from a gene shared the same stop or start codon, they were counted as one. Potential candidate CPuORFs were narrowed down by selection at step 2 of ESUCA in a step-by-step manner, as shown in Supplementary Table S1. The numbers of BLAST hits (expressed sequence tag [EST], transcriptome shotgun assembly [TSA], assembled EST/TSA, and RefSeq RNA sequences) extracted at step 3.2 are also shown in Supplementary Table S1. After the final step of ESUCA, 49, 192, 261, and 1,495 candidate CPuORFs were extracted from fruit fly, zebrafish, chiken, and human, respectively. We conducted manual validation for the extracted candidate CPuORFs as described in our previous study [13]. We selected CPuORFs conserved in at least two orders other than the order to which the original species belongs; subsequently, we classified these selected CPuORFs on the basis of animal taxonomic categories (Fig. 2) (see the Methods for details). In total, 1,517 animal CPuORFs (37 for fruit fly, 156 for zebrafish, 230 for chicken, and 1,094 for human) were identified (Fig. 3). Of these, 1,425 CPuORFs were newly identified in the current study. All alignments and detailed information on the identified CPuORFs are shown in Supplementary Figure S1 and Table S2, respectively. The identified CPuORF-containing genes were classified into 1,363 ortholog groups on the basis of similarities of mORF-encoded amino acid sequences, using OrthoFinder [17]. CPuORFs with similar amino acid sequences from the same ortholog groups were categorized as the same CPuORF families (homology groups [HGs]; see the Methods for details). The identified 1,517 CPuORFs were classified into 1,430 HGs. We assigned HG numbers to 1,430 HGs in an order based on numbers of orders in which any CPuORF belonging to each HG was extracted, the taxonomic range of the sequence conservation of each HG, and gene ID numbers. When multiple CPuORF families were identified in the same ortholog groups, the same HG number with a different subnumber was assigned to each of the families (e.g., HG0004.1 and HG0004.2; Supplementary Table S2).

Sequence-dependent effects of CPuORFs on mORF translation

To address the relationship between taxonomic ranges of CPuORF conservation and likelihood of having regulatory function, we selected 17 human CPuORFs conserved in various taxonomic ranges, including a previously identified sequence-dependent regulatory CPuORF, the *PTP4A1* CPuORF [18], as a positive control, and examined their sequence-dependent effects on the expression of the downstream reporter gene using transient expression assays (Fig. 4). Other uORFs overlapping any of the selected CPuORFs were eliminated by introducing mutations that changed the ATG codons of the overlapping uORFs to other codons but did not alter the amino acid sequences of the CPuORFs (Supplementary Figure S2). The resulting modified CPuORFs were used as CPuORFs bearing the wild-type amino acid sequences (WT-aa CPuORFs) (Fig. 4b). To assess the importance of amino acid sequences for the effects of these CPuORFs on mORF translation, frameshift mutations were introduced into the WT-aa CPuORFs such that the amino acid sequences of their conserved regions could be altered (see Methods and Supplementary Figure S2 for details). In eight of the 17 CPuORFs, the introduced frameshift mutations significantly upregulated the expression of the reporter gene, indicating that these CPuORFs repressed mORF translation in a sequence-dependent manner (Fig. 4c). One of the eight sequence-dependent regulatory CPuORFs, the *TMEM184C* CPuORF, is conserved only among Eutheria (Fig. 4a). This result suggests that CPuORFs conserved only among Eutheria can have seqence-dependent regulatory effects.

Discussion

In the current study, by applying ESUCA to four animal genomes, we identified 1,517 CPuORFs belonging to 1,430 HGs. Taxonomic ranges of sequence conservation of these CPuORFs largely vary, demonstrating that ESUCA can identify CPuORFs conserved in various taxonomic ranges (Supplementary Table S3). We examined the effects of 17 human CPuORFs conserved beyond Euarchontoglires on mORF translation, and identified seven novel sequence-dependent regulatory CPuORFs (in the *MKKS*, *SLC6A8*, *FAM13B*, *MIEF1*, *KAT6A*, *LRRC8B,* and *TMEM184C* genes). Of these, the *TMEM184* CPuORF is one of those conserved in the narrowest taxonomic range among the tested CPuORFs. This suggests that human CPuORFs conserved beyond Euarchontoglires are likely to be conserved because of functional constraints of their encoded peptides. Of the 1,094 CPuORFs extracted from the human genome, 1,082 are conserved beyond Euarchontoglires (Fig. 3 and Supplementary Table S3). Therefore, many CPuORFs encoding regulatory peptides are expected to be found in the human CPuORFs identified in this study.

Of the sequence-dependent regulatory CPuORFs identified here, the *MKKS* CPuORF has been previously reported to be a translational regulator that represses the production of a protein involved in McKusick-Kaufman syndrome [19]; however, the amino acid sequence dependence of the CPuORF function was not reported. Interestingly, the *MIEF1* CPuORF-encoded peptide is a functional peptide localized in the mitochondria [20]. Thus, the *MIEF1* CPuORF may have dual functions.

As shown in Fig. 1a and the Methods, we constructed a transcript sequence dataset with reduced redundancy, according to our previous study[13]. Numbers of bases and sequences of EST/TSA and RefSeq and their assembling results are shown in Supplementary Table S4. Although numbers of sequences were not reduced, the numbers of bases were reduced to approximately half. The calculation time of BLAST was proportional to the database size. Most of the calculation time for ESUCA was because of BLAST. Therefore, the calculation time for ESUCA could be reduced by using assembled EST/TSA+RefSeq datasets (transcript sequence datasets with reduced redundancy) instead of intact EST/TSA/RefSeq datasets. Although we could narrow down the assembled EST/TSA+RefSeq dataset by using an EST clustering method, such as CD-HIT[21], we did not conduct such a reduction, because there was a risk of selecting a sequence without a 5′-UTR as a representative sequence from a mixed cluster of one with the 5′-UTR and one without. Therefore, the assembled EST/TSA+RefSeq database was used at step 3.1 of ESUCA.

Supplementary Table S1 shows that the numbers of uORFs and genes with uORFs were greatly reduced at steps 1, 2, and 4.3 of ESUCA. Obviously, two steps, i.e., steps 1 and 4.3, were important because conservation of uORFs was estimated during these steps. Step 2 was newly implemented in ESUCA to distinguish between 'spurious' CPuORFs and 'true' CPuORFs[13]. In the case of CPuORF estimation without this step, we estimated the number of uORFs from which 'spurious' CPuORFs could be incorrectly identified as 'true' CPuORFs. As shown in Supplementary Table S5, approximately 20% of potential 'spurious' CPuORFs were found among uORFs that overlapped with mORFs of other splice variants according to the genomic information of the original species. Such 'spurious' uORFs were likely to remain in the final result as 'true' CPuORFs. Although 35 candiate CPuORFs were extracted by BAIUCAS in our previous study[10], of these 35, 12 uORFs were judged as 'spurious' CPuORFs by our manual validation. These results suggested that CPuORF determination based on sequence conservation of uORFs and mORFs, without filtering uORFs using the uORF-mORF fusion ratio, yielded approximately 30% 'spurious' CPuORFs. Therefore, step 2 of ESUCA is an important function for identification of CPuORFs. That is, ESUCA is superior to other conventional methods because it can exclude 'spurious' CPuORFs.

Chemical screening recently identified a compound that causes nascent peptide-mediated ribosome stalling in the mORF of the human *PCSK9* gene, resulting in specific translational inhibition of *PCSK9* and a reduction in total plasma cholesterol levels [22]. Nascent peptide-mediated ribosome stalling in some of the previously identified regulatory CPuORFs is promoted by metabolites, such as polyamine, arginine, and sucrose [4,23]. 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.

Methods

All procedures and protocols were approved by the Institutional Safety Committee for Recombinant DNA Experiments at Chiba University. All methods were carried out in accordance with approved guidelines.

Extraction of CPuORFs using ESUCA

ESUCA was developed as an advanced version of BAIUCAS [10] in our previous study [13]. ESUCA consists of six steps, and some of these steps are divided into substeps, as shown in Fig. 1a and 1b. To identify animal CPuORFs using ESUCA, the following eight-step procedures were conducted, including the six ESUCA steps: 0) data preparation for ESUCA, 1) uORF extraction from the 5′-UTR (Fig. 5), 2) calculation of uORF-mORF fusion ratios (Fig. 6), 3) uORF-tBLASTn against transcript sequence databases (Fig. 7a), 4) mORF-tBLASTn against downstream sequence datasets for each uORF (Fig. 7b and 7c), 5) calculation of K_a/K_s ratios (Fig. 8), 6) determination of the taxonomic range of uORF sequence conservation, and 7) manual validation after ESUCA. See the Materials and Methods in our previos study [13] for details.

Transcript dataset construction based on genome information (step 0.1)

To identify plant CPuORFs, data preparation for ESUCA (step 0.1) was conducted as described in our previous study [13]. We conducted data preparation for ESUCA to identify animal CPuORFs as follows. We used a genome sequence file in FASTA format and a genomic coordinate file in GFF3 format obtained from Ensemble Metazoa Release 33 (https://metazoa.ensembl.org/index.html)[24] to extract fruit fly (*Drosophila melanogaster*) uORF sequences. We used genome sequence files in FASTA format and genomic coordinate files in GFF3 format obtained from Ensemble Release 86 (https://metazoa.ensembl.org/index.html) [24] for zebrafish (*Danio rerio*), chicken (*Gallus gallus*), and human (*Homo sapiens*). We extracted exon sequences from genome sequences on the basis of genomic coordinate information and constructed transcript sequence datasets by combining exon sequences. On the basis of the transcription start site and the translation initiation codon of each transcript in the genomic coordinate files, we extracted 5′-UTR and mORF RNA sequences from the transcript sequence datasets, as shown in Fig. 1a (step 0.1). The 5′-UTR sequences were used at step 1 of ESUCA. The mORF RNA sequences were translated into amino acid sequences (mORF proteins) and used at step 4.1 of ESUCA.

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

To identify plant CPuORFs, data preparation for ESUCA (step 0.2) was conducted as described in our previous study [13]. We conducted data preparation for ESUCA to identify animal CPuORFs. As shown in Fig. 1b, Metazoa RefSeq RNA sequences were used at steps 2 and 3.1 of ESUCA. Assembled EST/TSA sequences generaged by using velvet [25] and Bowtie2 [26], were used at step 3.1 of ESUCA. Intact and merged EST/TSA/RefSeq sequences were used at step 4.2 of ESUCA. Taxomomy datasets derived from EST/TSA/RefSeq databases were used at steps 4.3 and 6 of ESUCA. See the Materials and Methods in our previos study [13] for details.

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

To automatically determine the taxonomic range of the sequence conservation of each CPuORF, we first defined 20 animal taxonomic categories (Fig. 2). The 20 taxonomic defined categories were Euarchontoglires, Eutheria other than Euarchontoglires, Mammalia other than Eutheria, Aves, Sauropsida other than Aves, Amphibia (Tetrapoda other than Sauropsida and Mammalia), Sarcopterygii other than Tetrapoda, Ostarioclupeomorpha, Actinopterygii other than Ostarioclupeomorpha, Vertebrata other than Euteleostomi (Actinopterygii and Sarcopterygii), Chordata other than Vertebrata, Deuterostomia other than Chordata, Insecta, Arthropoda other than Insecta, Ecdysozoa other than Arthropoda, Lophotrochozoa (Protostomia other than Ecdysozoa), Bilateria other than Protostomia and Deuterostomia, Cnidaria, Ctenophora (Eumetazoa other than Cnidaria and Bilateria), and Metazoa other than Eumetazoa. Based on taxonomic lineage information of EST, TSA, and RefSeq RNA sequences, which were provided by NCBI Taxonomy, the uORF-tBLASTn and mORF-tBLASTn hit sequences selected for K_a/K_s analysis were classified into the 19 taxonomic categories (Supplementary Table S3). The category ‘Ctenophora’ was omitted from animal taxonomic categories because no sequences were classified to this category. For each CPuORF, the numbers of transcript sequences classified into each category were counted and are shown in Supplementary Table S3. These numbers represent the number of orders in which the amino acid sequence of each CPuORF is conserved.

Classification of animal CPuORFs into HGs

Systemtic numbering of animal CPuORF families (HGs) has not been reported to date. Here, we defined systematic HG numbers for the identified 1,517 animal CPuORFs. Among these identified CPuORFs, those with both similar uORF and mORF amino acid sequences were classified into the same HGs. We first determined ortholog groups of CPuORF-containing genes, referred to as mORF clusters, based on similalities of mORF-encoded amino acid sequences, using OrthoFinder [17]. The identified CPuORF-containing genes were classified into 1,194 mORF clusters. CPuORFs contained in each ortholog group (mORF-cluster) were further classified into uORF clusters, as follows. We conducted a pairwise comparison of uORF peptide similary using BLASTp with *E*-values less than 2000 in each mORF cluster. Binarized distance matrixes consisting of 0 (hit) or 1 (no-hit) were generated by this comparison. Hierarchical clustering with single linkage with the cutoff parameter ($h = 0.5$) was applied to these matrixes for construction of uORF clusters. In total, 1,336 uORF-mORF clusters were generated automatically. We determined 1,430 clusters by manually checking alignments of uORFs and mORFs. We assigned HG numbers to the 1,430 clusters in an order based on the number of orders in which any CPuORF belonging to each HG was extracted, the taxonomic range of the sequence conservation of each HG, and gene ID numbers. The same HG number with a different sub-number was assigned to CPuORFs in genes of the same ortholog group with dissimilar uORF sequences (e.g., HG0004.1 and HG0004.2; Supplementary Table S2).

Plasmid construction and transient reporter assays

pSV40:Fluc was generated by inserting the SV40 promoter (BglII/HindIII fragment) from pRL-SV40 (Promega, Madison, WI, USA) into the KpnI site of pGL4.10[luc2] (Promega, Madison, WI, USA) by blunt-end cloning. The 5′-UTR sequences containing the selected CPuORFs (SacI/XhoI fragment) were fused to the Fluc coding sequence by subcloning the CPuORFs into the SacI/XhoI site of pSV40:luc2 to generate the WT-aa reporter construct (pSV40:UTR(WT-aa):Fluc, Fig. 4b, Supplementary Figure S2). To assess the importance of the amino acid sequences with regard to the effects of these CPuORFs on mORF translation, frameshift mutations were introduced into the CPuORFs so that the amino acid sequences of their conserved regions could be altered. A + 1 or − 1 frameshift was introduced upstream or within the conserved region of each CPuORF, and another frameshift was introduced before the stop codon to shift the reading frame back to the original frame (pSV40:UTR(fs):Fluc, Fig. 4b, Supplementary Figure S2). DNA fragments containing the CPuORFs of either WT-aa or fs mutants from the *PTP4A1*, *MKKS*, *SLC6A8*, *FAM13B*, *MIEF1*, *EIF5*, *MAPK6*, *MEIS2*, *KAT6A*, *SLC35A4*, *LRRC8B*, *CDH11*, *PNRC2*, *BACH2*, *FGF9*, *PNISR*, and *TMEM184C* genes were synthesized (GenScript , NJ, USA) and subcloned into the pSV40:Fluc, as shown in Fig. 4b and Supplementary Table S6. These reporter constructs were each transfected into human HEK293T cells. HEK293T cells (16,000/well) were cotransfected with 80 ng/well of a pSV40:UTR:Fluc reporter plasmid and 1.6 ng/well pGL4.74[hRluc/TK] plasmid (Promega, Madison, WI, USA). After 24 h, Firefly luciferase and Renilla luciferase activities were measured according to the Dual-Luciferase Reporter Assay protocol (Promega, Madison, WI, USA) using GloMaxR-Multi Detection System(Promega, Madison, WI, USA).

Statistical and informatic analyses

All programs, except for existing stand-alone programs, such as NCBI-BLAST+ ver. 2.6.0 [27], Clustal Omega (ClustalO) ver. 1.2.2[28], OrthoFinder ver. 1.1.4[17], velvet ver. 1.2.10[25], Bowtie2 ver. 2.2.9[26], and Jalview ver. 2.10.2 [29], were written in R (www.r-project.org). We also used R libraries, GenomicRanges ver. 1.32.7 [30], exactRankTests ver. 0.8.30, Biostrings ver. 2.48.0, and seqinr ver. 3.4.5 [31]. Statistical differences between the control (WT-aa) and fs constructs were determined by Student's *t*-tests in transient assays.

References

1 Ingolia, N. T. *et al.* Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep.* **8**, 1365-1379, doi:10.1016/j.celrep.2014.07.045 (2014).

2 Morris, D. R. & Geballe, A. P. Upstream open reading frames as regulators of mRNA translation. *Molecular and cellular biology* **20**, 8635-8642 (2000).

3 Cruz-Vera, L. R., Sachs, M. S., Squires, C. L. & Yanofsky, C. Nascent polypeptide sequences that influence ribosome function. *Current opinion in microbiology* **14**, 160-166, doi:10.1016/j.mib.2011.01.011 (2011).

4 Ito, K. & Chiba, S. Arrest peptides: cis-acting modulators of translation. *Annual review of biochemistry* **82**, 171-202, doi:10.1146/annurev-biochem-080211-105026 (2013).

5 Somers, J., Poyry, T. & Willis, A. E. A perspective on mammalian upstream open reading frame function. *The international journal of biochemistry & cell biology* **45**, 1690-1700, doi:10.1016/j.biocel.2013.04.020 (2013).

6 Bhushan, S. *et al.* Structural basis for translational stalling by human cytomegalovirus and fungal arginine attenuator peptide. *Molecular cell* **40**, 138-146, doi:10.1016/j.molcel.2010.09.009 (2010).

7 Wang, Z. & Sachs, M. S. Ribosome stalling is responsible for arginine-specific translational attenuation in *Neurospora crassa*. *Molecular and cellular biology* **17**, 4904-4913 (1997).

8 Hayden, C. A. & Jorgensen, R. A. Identification of novel conserved peptide uORF homology groups in *Arabidopsis* and rice reveals ancient eukaryotic origin of select groups and preferential association with transcription factor-encoding genes. *BMC biology* **5**, 32, doi:10.1186/1741-7007-5-32 (2007).

9 Tran, M. K., Schultz, C. J. & Baumann, U. Conserved upstream open reading frames in higher plants. *BMC genomics* **9**, 361, doi:10.1186/1471-2164-9-361 (2008).

10 Takahashi, H., Takahashi, A., Naito, S. & Onouchi, H. BAIUCAS: a novel BLAST-based algorithm for the identification of upstream open reading frames with conserved amino acid sequences and its

application to the *Arabidopsis thaliana* genome. *Bioinformatics* **28**, 2231-2241, doi:10.1093/bioinformatics/bts303 (2012).

11 Vaughn, J. N., Ellingson, S. R., Mignone, F. & Arnim, A. Known and novel post-transcriptional regulatory sequences are conserved across plant families. *Rna* **18**, 368-384, doi:10.1261/rna.031179.111 (2012).

12 van der Horst, S., Snel, B., Hanson, J. & Smeekens, S. Novel pipeline identifies new upstream ORFs and non-AUG initiating main ORFs with conserved amino acid sequences in the 5' leader of mRNAs in *Arabidopsis thaliana*. *Rna* **25**, 292-304, doi:10.1261/rna.067983.118 (2018).

13 Takahashi, H. *et al.* Comprehensive genome-wide identification of angiosperm upstream ORFs with peptide sequences conserved in various taxonomic ranges using a novel pipeline, ESUCA. *BMC genomics* **21**, 260, doi:10.1186/s12864-020-6662-5 (2020).

14 Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic acids research* **46**, D754-D761, doi:10.1093/nar/gkx1098 (2018).

15 Crowe, M. L., Wang, X. Q. & Rothnagel, J. A. Evidence for conservation and selection of upstream open reading frames suggests probable encoding of bioactive peptides. *BMC genomics* **7**, 16, doi:10.1186/1471-2164-7-16 (2006).

16 Hayden, C. A. & Bosco, G. Comparative genomic analysis of novel conserved peptide upstream open reading frames in *Drosophila melanogaster* and other dipteran species. *BMC genomics* **9**, 61, doi:10.1186/1471-2164-9-61 (2008).

17 Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157, doi:10.1186/s13059-015-0721-2 (2015).

18 Hardy, S. *et al.* Magnesium-sensitive upstream ORF controls PRL phosphatase expression to mediate energy metabolism. *Proceedings of the National Academy of Sciences of the United States of America* **116**, 2925-2934, doi:10.1073/pnas.1815361116 (2019).

19 Akimoto, C. *et al.* Translational repression of the McKusick-Kaufman syndrome transcript by unique upstream open reading frames encoding mitochondrial proteins with alternative polyadenylation sites. *Biochimica et biophysica acta* **1830**, 2728-2738 (2013).

20 Samandi, S. *et al.* Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins. *eLife* **6**, doi:10.7554/eLife.27860 (2017).

21 Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659, doi:10.1093/bioinformatics/btl158 (2006).

22 Lintner, N. G. *et al.* Selective stalling of human translation through small-molecule engagement of the ribosome nascent chain. *PLoS biology* **15**, e2001882, doi:10.1371/journal.pbio.2001882 (2017).

23 Yamashita, Y. *et al.* Sucrose sensing through nascent peptide-meditated ribosome stalling at the stop codon of Arabidopsis *bZIP11* uORF2. *FEBS letters* **591**, 1266-1277, doi:10.1002/1873-3468.12634 (2017).

24 Cunningham, F. *et al.* Ensembl 2019. *Nucleic Acids Res* **47**, D745-D751, doi:10.1093/nar/gky1113 (2019).

25 Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* **18**, 821-829, doi:10.1101/gr.074492.107 (2008).

26 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).

27 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **25**, 3389-3402 (1997).

28 Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology* **7**, 539, doi:10.1038/msb.2011.75 (2011).

29 Clamp, M., Cuff, J., Searle, S. M. & Barton, G. J. The Jalview Java alignment editor. *Bioinformatics* **20**, 426-427, doi:10.1093/bioinformatics/btg430 (2004).

30 Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS computational*

biology **9**, e1003118, doi:10.1371/journal.pcbi.1003118 (2013).

31 Charif, D. & Lobry, J. R. in *Structural Approaches to Sequence Evolution: Molecules, Networks, 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 to H.O., JP16K07387 to H.O., JP19K22892 to H.T., JP18H03330 to H.T., M.I, 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); the Naito Foundation (to H.O.) ; and the Research Foundation for the Electrotechnology of Chubu (to H.T.).

Author contributions

H.T., H.O., and M.I. designed the study. H.T. and S.M., performed experiments and analyzed the data supervised by 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 contribution of all coauthors.

Additional information

Supplementary information accompanies this paper.

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

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.

Figure 2. Defined animal taxonomic categories.

Figure 3. Numbers of CPuORFs extracted by ESUCA in each taxonomic ranges.

Figure 4. 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 at least three biological replicates are shown. $*p < 0.05$.

Figure 5. Extraction of the largest uORF sequences from the 5′-UTR. After data preparation for ESUCA (Fig. 1b), we conducted the extraction of uORF sequences by searching the 5′-UTR sequences for an ATG codon and its nearest downstream in-frame stop codon at step 1 of ESUCA (Fig. 1b). Sequences starting with an ATG codon and ending with the nearest in-frame stop codon were extracted as uORF sequences. When multiple uORFs shared the same stop codon in a transcript, only the longest uORF sequence was used for further analyses.

Figure 6. Outline for uORF-mORF fusion ratio calculations. For each original uORF-containing transcript sequence, RefSeq RNAs containing both sequences similar to the uORF and the mORF of each uORF-containing transcript were selected using uORF-tBLASTx and mORF-tBLASTx from the RefSeq RNA database (database (2) in Fig.1a). For example, the selected RNA sequences are RNA1, 2, 3...10, as illustrated.

Based on whether the uORF-tBLASTx-hit region was included in the largest RefSeq RNA ORF, the selected RefSeq RNAs were classified into two types, namely fusion (*X*) (RNA1 and 2) and separate types (*Y*) (RNA3-10). For each original uORF-containing transcript, the uORF-mORF fusion ratio was calculated as *X*/(*X* + *Y*).

Figure 7. Outline of homology searches for uORFs with amino acid sequences conserved between homologous genes. (**a**) For each original uORF-containing transcript, sequences containing both similar regions to the uORF and the mORF of uORF-containing transcripts were selected using uORF-tBLASTn (step 3.1 of ESUCA) and mORF-tBLASTn (step 4.1 of ESUCA). A transcript sequence database consisting of RefSeq RNAs (database (2) in Fig.1a) served as data source, while an assembled EST/TSA (database (3) in Fig.1a) was generated at step 0.2 of data preparation for ESUCA. Asterisks represent stop codons. At step 3.2 of ESUCA, the largest tBLASTn-hit region-overlapping uORF was extracted. (**b**) Detailed illustration of step 4.0 of ESUCA. Putative uORF extraction and downstream sequence dataset construction were conducted systematically for each uORF-tBLASTn hit sequence. (**c**) Detailed illustration of step 4.1 of ESUCA. After mORF-tBLASTn, the 5'-most in-frame ATG codon located downstream of the selected stop codon was identified as the initiation codon of the putative partial or intact mORF. uORF-mORF overlaps were discarded as fusion types, according to the positional relationship between them, when found in the hit-assembled EST/TSA+RefSeq sequences.

Figure 8. K_a/K_s simulation. (a) Putative uORF sequences in the selected transcripts were used for the generation of uORF amino acid sequence alignments and for K_a/K_s analysis. (b) ClustalO was used to generate multiple alignments. (c) For each candidate CPuORF, the median K_a/K_s ratios for all pairwise combinations of the original uORF and homologous putative uORFs were calculated using the LWL85 algorithm in the seqinR package. (d) For the K_a/K_s ratio statistical tests, we calculated mutation rate distributions between the original uORF and homologous putative uORFs; subsequently, we artificially generated mutants using the observed mutation rate distribution. Observed empirical K_a/K_s ratio distributions were then compared with null distributions (negative controls) using the Mann-Whitney *U* test to validate the statistical significance. The one-sided *U* test was used to investigate whether the observed distributions were significantly lower than the null distributions.

Figure 1

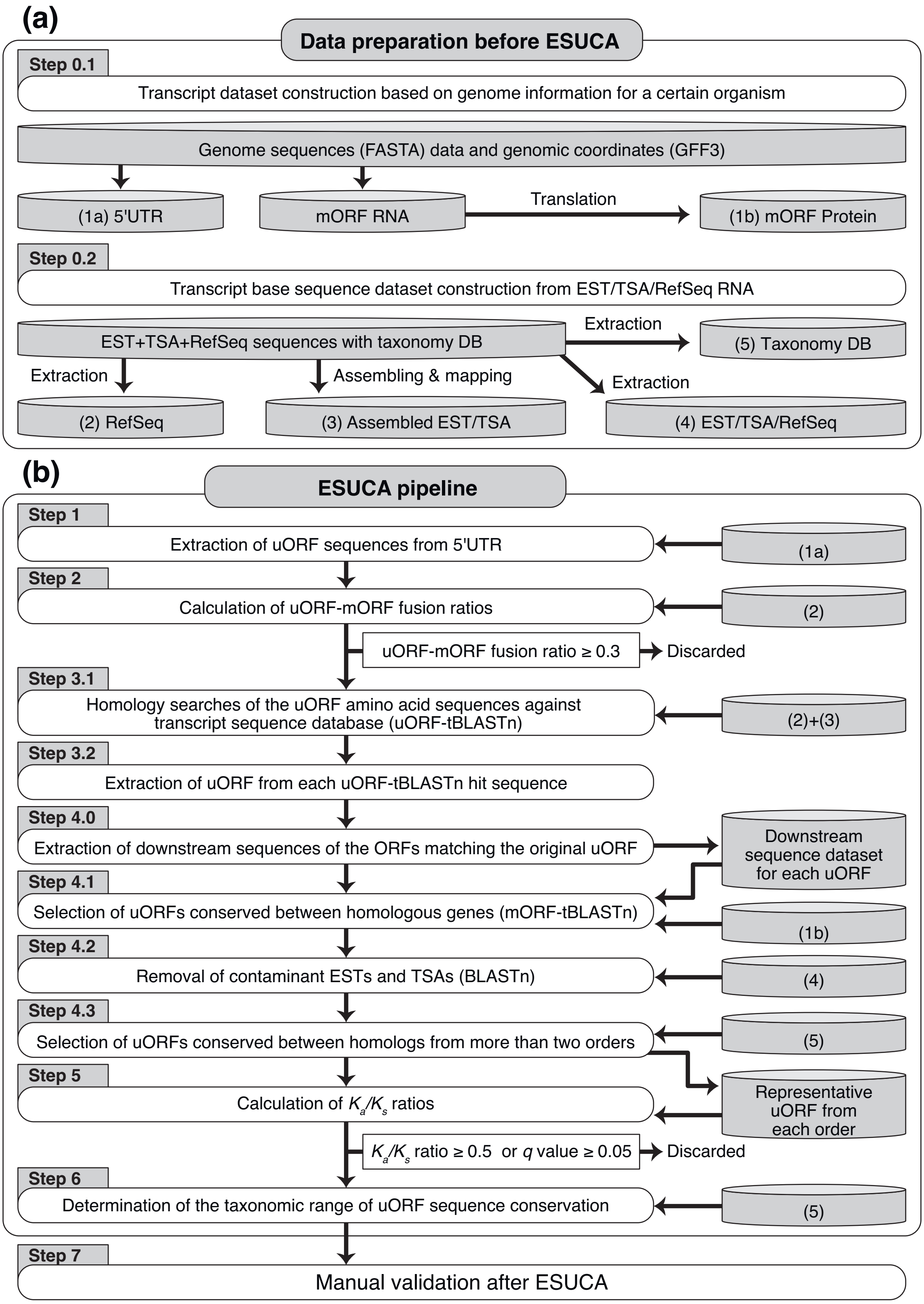

(a)
Data preparation before ESUCA
Step 0.1
Transcript dataset construction based on genome information for a certain organism
Genome sequences (FASTA) data and genomic coordinates (GFF3)
(1a) 5'UTR
mORF RNA
Translation
(1b) mORF Protein
Step 0.2
Transcript base sequence dataset construction from EST/TSA/RefSeq RNA
EST+TSA+RefSeq sequences with taxonomy DB
Extraction
(5) Taxonomy DB
Extraction
Assembling & mapping
Extraction
(2) RefSeq
(3) Assembled EST/TSA
(4) EST/TSA/RefSeq
(b)
ESUCA pipeline
Step 1
Extraction of uORF sequences from 5'UTR
(1a)
Step 2
Calculation of uORF-mORF fusion ratios
(2)
uORF-mORF fusion ratio ≥ 0.3
Discarded
Step 3.1
Homology searches of the uORF amino acid sequences against transcript sequence database (uORF-tBLASTn)
(2)+(3)
Step 3.2
Extraction of uORF from each uORF-tBLASTn hit sequence
Step 4.0
Extraction of downstream sequences of the ORFs matching the original uORF
Downstream sequence dataset for each uORF
Step 4.1
Selection of uORFs conserved between homologous genes (mORF-tBLASTn)
(1b)
Step 4.2
Removal of contaminant ESTs and TSAs (BLASTn)
(4)
Step 4.3
Selection of uORFs conserved between homologs from more than two orders
(5)
Representative uORF from each order
Step 5
Calculation of K_a/K_s ratios
K_a/K_s ratio ≥ 0.5 or q value ≥ 0.05
Discarded
Step 6
Determination of the taxonomic range of uORF sequence conservation
(5)
Step 7
Manual validation after ESUCA

Figure 2

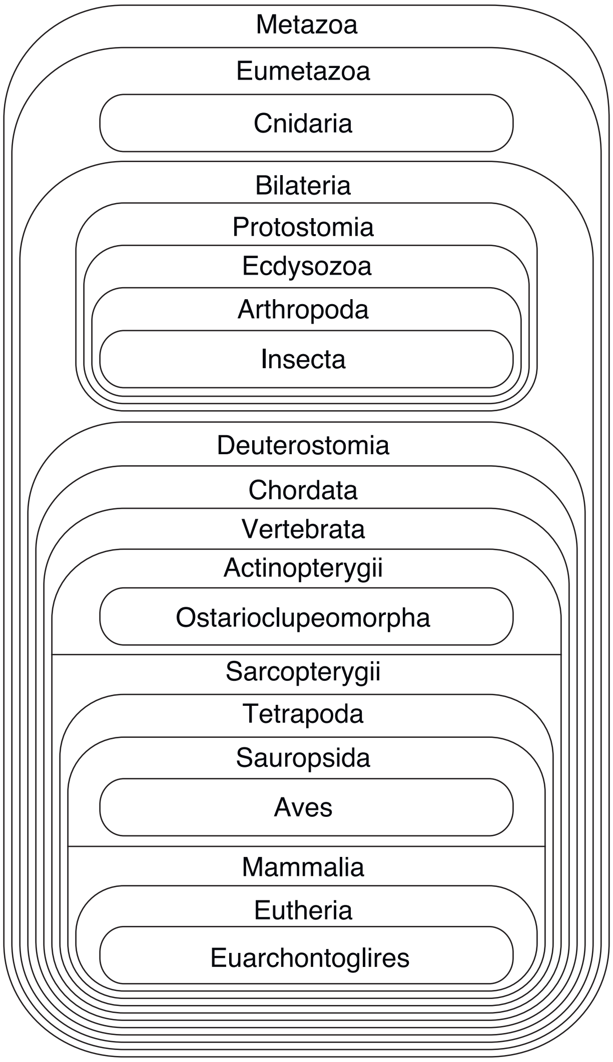

Metazoa
Eumetazoa
Cnidaria
Bilateria
Protostomia
Ecdysozoa
Arthropoda
Insecta
Deuterostomia
Chordata
Vertebrata
Actinopterygii
Ostarioclupeomorpha
Sarcopterygii
Tetrapoda
Sauropsida
Aves
Mammalia
Eutheria
Euarchontoglires

Figure 3

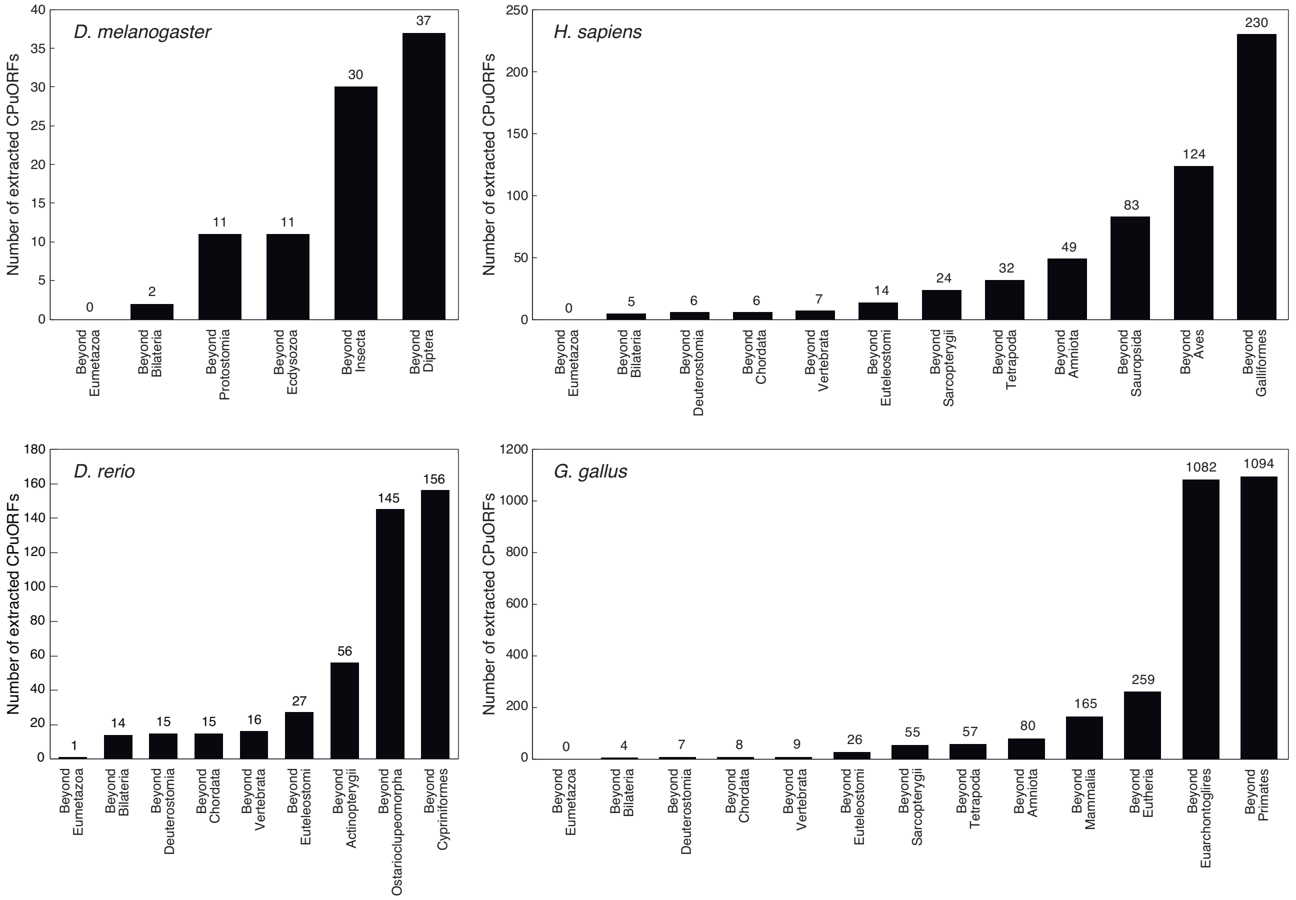
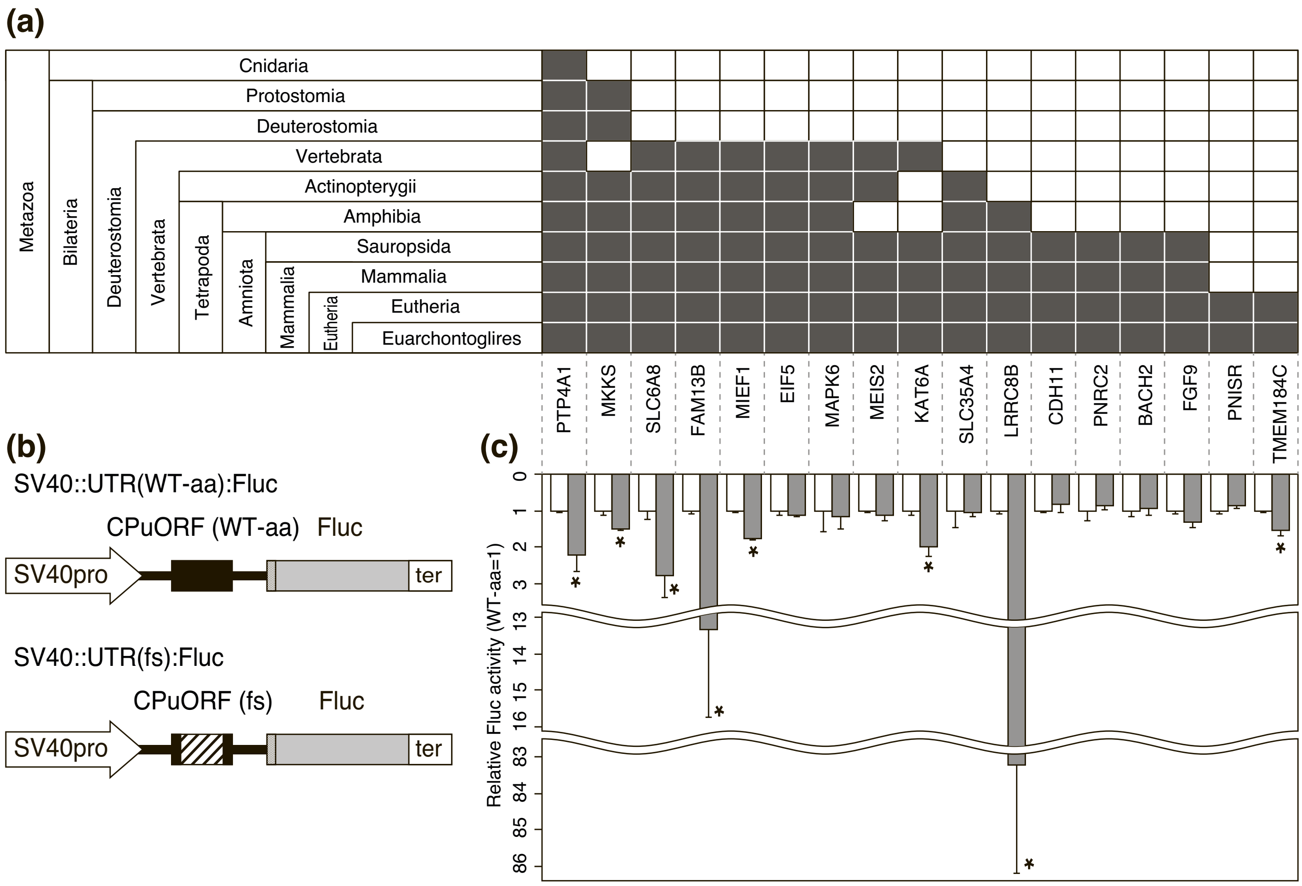

D. melanogaster
Number of extracted CPuORFs
0
5
10
15
20
25
30
35
40
0
2
11
11
30
37
Beyond Eumetazoa
Beyond Bilateria
Beyond Protostomia
Beyond Ecdysozoa
Beyond Insecta
Beyond Diptera
H. sapiens
Number of extracted CPuORFs
0
50
100
150
200
250
0
5
6
6
7
14
24
32
49
83
124
230
Beyond Eumetazoa
Beyond Bilateria
Beyond Deuterostomia
Beyond Chordata
Beyond Vertebrata
Beyond Euteleostomi
Beyond Sarcopterygii
Beyond Tetrapoda
Beyond Amniota
Beyond Sauropsida
Beyond Aves
Beyond Galliformes
D. rerio
Number of extracted CPuORFs
0
20
40
60
80
100
120
140
160
180
1
14
15
15
16
27
56
145
156
Beyond Eumetazoa
Beyond Bilateria
Beyond Deuterostomia
Beyond Chordata
Beyond Vertebrata
Beyond Euteleostomi
Beyond Actinopterygii
Beyond Ostarioclupeomorpha
Beyond Cypriniformes
G. gallus
Number of extracted CPuORFs
0
200
400
600
800
1000
1200
0
4
7
8
9
26
55
57
80
165
259
1082
1094
Beyond Eumetazoa
Beyond Bilateria
Beyond Deuterostomia
Beyond Chordata
Beyond Vertebrata
Beyond Euteleostomi
Beyond Sarcopterygii
Beyond Tetrapoda
Beyond Amniota
Beyond Mammalia
Beyond Eutheria
Beyond Euarchontoglires
Beyond Primates

Figure 4



	PTP4A1	MKKS	SLC6A8	FAM13B	MIEF1	EIF5	MAPK6	MEIS2	KAT6A	SLC35A4	LRRC8B	CDH11	PNRC2	BACH2	FGF9	PNISR	TMEM184C
Cnidaria	■	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
Protostomia	■	■	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
Deuterostomia	■	■	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
Vertebrata	■	□	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□
Actinopterygii	■	■	■	■	■	■	■	■	□	■	□	□	□	□	□	□	□
Amphibia	■	■	■	■	■	■	■	□	□	■	■	□	□	□	□	□	□
Sauropsida	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□
Mammalia	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□
Eutheria	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Euarchontoglires	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■

Figure 5

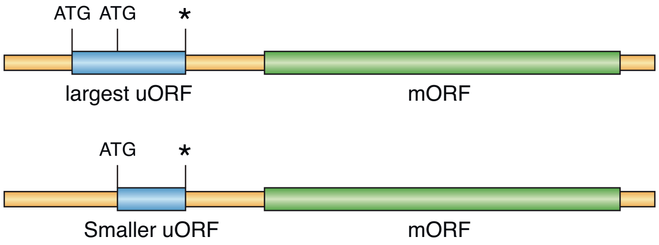


Figure 6

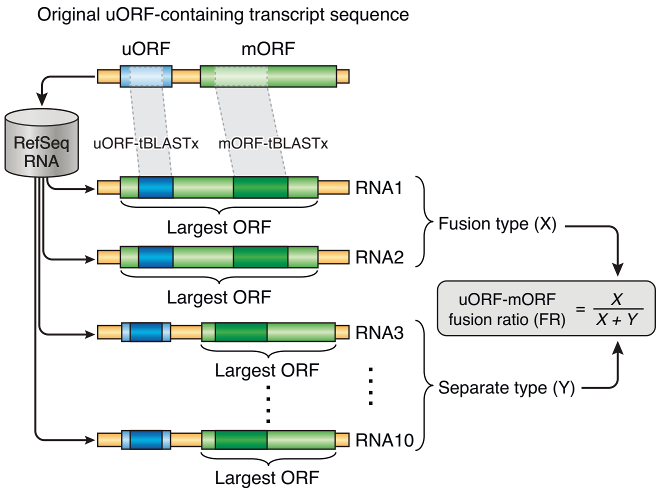

Original uORF-containing transcript sequence
uORF
mORF
RefSeq RNA
uORF-tBLASTx
mORF-tBLASTx
RNA1
Largest ORF
RNA2
Largest ORF
RNA3
Largest ORF
RNA10
Largest ORF
Fusion type (X)
Separate type (Y)
uORF-mORF fusion ratio (FR) = X / X + Y

Figure 7

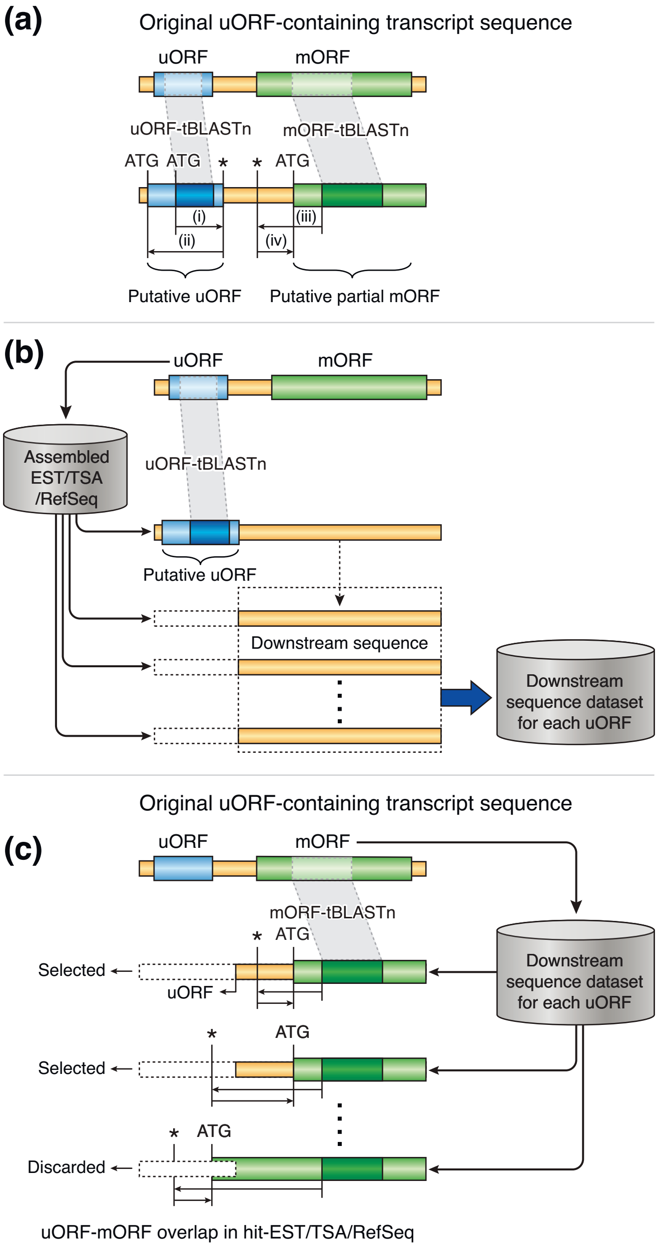
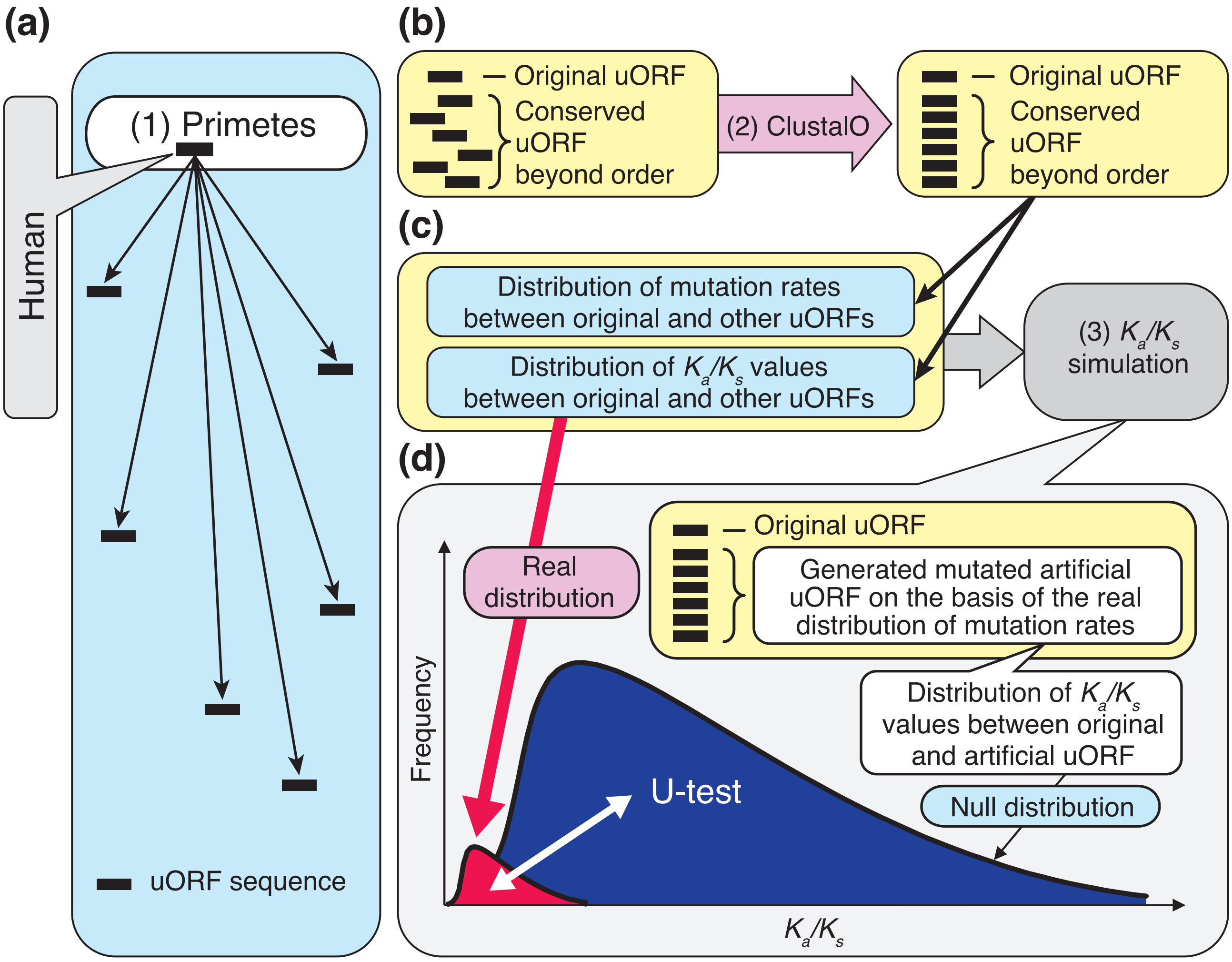

(a)
Original uORF-containing transcript sequence
uORF
mORF
uORF-tBLASTn
mORF-tBLASTn
ATG
ATG
*
*
ATG
(i)
(ii)
(iii)
(iv)
Putative uORF
Putative partial mORF
(b)
uORF
mORF
Assembled EST/TSA /RefSeq
uORF-tBLASTn
Putative uORF
Downstream sequence
Downstream sequence dataset for each uORF
Original uORF-containing transcript sequence
(c)
uORF
mORF
mORF-tBLASTn
*
ATG
Downstream sequence dataset for each uORF
Selected
uORF
*
ATG
Selected
*
ATG
Discarded
uORF-mORF overlap in hit-EST/TSA/RefSeq

Figure 8


(a)
Human
(1) Primetes
uORF sequence
(b)
Original uORF
Conserved uORF beyond order
(2) ClustalO
Original uORF
Conserved uORF beyond order
(c)
Distribution of mutation rates between original and other uORFs
Distribution of K_a/K_s values between original and other uORFs
(3) K_a/K_s simulation
(d)
Real distribution
Original uORF
Generated mutated artificial uORF on the basis of the real distribution of mutation rates
Distribution of K_a/K_s values between original and artificial uORF
Null distribution
U-test
Frequency
K_a/K_s