ESUCA: a pipeline for genome-wide identification of upstream open reading frames with evolutionarily conserved sequences and determination of the taxonomic range of their conservation

Hiro Takahashi[1,2#]*, Noriya Hayashi[3#], Yui Yamashita[3], Satoshi Naito[3,4], Anna Takahashi[5], Kazuyuki Fuse[6], Toshinori Endo[7], Shoko Kojima[8], and Hitoshi Onouchi[3]*

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

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

[6]New Business Development Office, Churitsu Electric Corporation, Toyoake 470-1112, Japan

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

[8] Graduate School of Bioscience and Biotechnology, Chubu University, Kasugai 487-8501 Japan

*To whom correspondence should be addressed. Tel: +81-11-706-3887; Fax: +81-11-706-4932; Email: onouchi@abs.agr.hokudai.ac.jp

Correspondence may also be addressed to Hiro Takahashi. Email: takahasi@p.kanazawa-u.ac.jp

[#]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

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

Abstract

Background: Some upstream open reading frames (uORFs) in the 5′ leaders of eukaryotic mRNAs encode evolutionarily conserved functional peptides, such as cis-acting regulatory peptides that control main ORF (mORF) translation. To comprehensively identify uORFs with conserved peptide sequences (CPuORFs), we previously developed the BAIUCAS pipeline, in which uORF sequences are compared between a certain species and any others with available transcript sequence databases. However, further selection is needed to identify CPuORFs encoding functional peptides. The purpose of this study is to develop a novel pipeline to efficiently identify CPuORFs likely to encode functional peptides.

Results: Here, we present the ESUCA pipeline. In addition to the function of BAIUCAS, ESUCA has the following new functions: 1) to identify CPuORFs likely to be conserved due to the functions of their encoded small peptides, not due to encoding parts of the mORF-encoded proteins; 2) to systematically calculate K_a/K_s ratios to assess whether uORF sequences are conserved at the nucleotide or amino acid level; 3) to determine the taxonomic range of uORF sequence conservation. We applied ESUCA to five plant genomes and identified 88 novel CPuORF families. Using a transient expression assay, we examined the effects of eight CPuORFs conserved in various taxonomic ranges on mORF translation. Three of seven CPuORFs conserved across diverse eudicots showed sequence-dependent regulatory effects.

Conclusions: This study demonstrates that ESUCA can efficiently identify many CPuORFs conserved in various taxonomic ranges by applying it to many species and is highly useful to select CPuORFs likely to encode functional peptides.

Background

The 5′-untranslated regions (5′-UTRs) of many eukaryotic mRNAs contain upstream open reading frames (uORFs) [1-4]. Although most uORFs are not thought to encode functional proteins or peptides, some uORFs encode regulatory peptides that have roles in post-transcriptional regulation of gene expression [5-8]. During translation of some of these regulatory uORFs, nascent peptides act inside the ribosomal exit tunnel to cause ribosome stalling [9]. Ribosome stalling on a uORF results in translational repression of the downstream main ORF (mORF) because stalled ribosomes block subsequent scanning of 43S preinitiation complexes and prevent them from reaching the start codon of the mORF [10]. Additionally, if ribosome stalling occurs at the stop codon of a uORF, nonsense-mediated mRNA decay may be induced [11, 12]. In some genes, uORF-encoded nascent peptides cause ribosome stalling in response to metabolites to down-regulate mORF translation under specific cellular conditions [10, 12-17]. In contrast to the uORFs encoding cis-acting regulatory nascent peptides, a uORF in the *Medicago truncatula MtHAP2-1* gene encodes a trans-acting regulatory peptide, which binds to the 5′-UTR of *MtHAP2-1* mRNA. This binding causes degradation of *MtHAP2-1* mRNA in the infection zone of plant root nodules during the development of nodules formed by symbiosis of legume plants and nitrogen-fixing bacteria (rhizobia) [18].

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 various organisms [19-23]. In plants, approximately 40 CPuORF families have been identified by comparing the uORF-encoded amino acid sequences of orthologous genes in some of *Arabidopsis*, rice, cotton, orange, soybean, grape and tobacco or those of paralogous genes in *Arabidopsis* [20, 22, 23]. Recently, 29 additional CPuORF families, which include CPuORFs with non-canonical initiation codons, have been identified by comparing 5′-UTR sequences between *Arabidopsis* and 31 other plant species [24].

In conventional comparative genomic approaches, uORF sequences are compared between selected species. Therefore, homology detection depends on the selection of species for comparison. In the searches using this approach, if a uORF amino acid sequence is not conserved between the selected species, that uORF is not identified as a CPuORF, even if it is evolutionarily conserved between one of the selected species and other unselected species. To overcome this problem, we previously developed the BAIUCAS pipeline [25]. In

81 BAIUCAS, homology searches of uORF amino acid sequences are performed between a certain species and any
82 other species for which expressed sequence tag (EST) databases are available, and uORFs conserved beyond a
83 certain taxonomic range are selected. Using BAIUCAS, we searched for *Arabidopsis* CPuORFs conserved
84 beyond the order Brassicales, which *Arabidopsis* belongs to, and identified 13 novel CPuORF families [25]. We
85 examined the sequence-dependent effects of the CPuORFs identified by BAIUCAS on mORF translation using
86 a transient expression assay, and identified at least five regulatory uORFs that repress mORF translation in an
87 amino acid sequence-dependent manner [26, 27].

88 Here, to further efficiently identify CPuORFs likely to encode functional peptides, we developed a novel
89 pipeline and named it ESUCA (for evolutionary search for uORFs with conserved amino acid sequences) (Fig. 1).
90 In addition to the function of BAIUCAS, ESUCA has three new major functions. One newly added function is
91 the automatic determination of the taxonomic range within which each CPuORF is conserved. In our previous
92 study using BAIUCAS, CPuORFs conserved only among small taxonomic groups were extracted to
93 comprehensively identify CPuORFs if they were conserved beyond Brassicales [25]. However, CPuORFs
94 conserved among small taxonomic groups are thought to be less likely to encode functional peptides than are
95 those conserved across a wide taxonomic range. The automatic determination of the taxonomic range of
96 CPuORF conservation enables us to efficiently select CPuORFs likely to encode functional peptides. The second
97 new function is the automatic assessment of whether each transcript containing a uORF fused to the mORF is a
98 major or minor form among homologous transcripts containing sequences similar to that of the uORF. Among
99 previously identified CPuORF-containing genes, the presence of splice variants carrying a CPuORF-mORF
100 fusion has been reported in the At2g31280, At5g01710, and At5g03190 genes [20, 25, 28]. If uORF-mORF
101 fusion type transcripts related to a uORF-containing gene are found in only a few species and the uORF-mORF
102 fusion form is not representative of its orthologous transcripts, then the uORF-mORF fusion type transcripts are
103 considered minor forms. By contrast, if uORF-mORF fusion type transcripts related to a uORF-containing gene
104 are found in a majority of species with a transcript sequence similar to the uORF and its associated mORF, it is
105 likely that the amino acid sequence encoded within the uORF region is a part of the mORF-encoded protein.
106 Therefore, such a uORF should not be considered a CPuORF, even if its amino acid sequence is evolutionarily
107 conserved. In this study, we devised an algorithm to assess whether each transcript containing a uORF-mORF
108 fusion is a minor or major form. This algorithm uses the NCBI reference sequence (RefSeq) database, which

109 provides curated non-redundant transcript sequences [29], and calculates the ratio of RefSeq RNAs with a
110 uORF-mORF fusion to all RefSeq RNAs containing sequences similar to a uORF and its associated mORF (Fig.
111 2). The third new function is a systematic calculation of K_a/K_s ratios of CPuORF sequences. To confirm that each
112 CPuORF sequence is conserved at the amino acid level but not at the nucleotide level, ESUCA automatically
113 selects a uORF sequence from each order in which the CPuORF is conserved, and calculates the K_a/K_s ratio using
114 the selected sequences. We applied ESUCA to *Arabidopsis*, rice, tomato, poplar and grape and identified 88
115 novel CPuORF families. We examined the sequence-dependent effects of eight poplar CPuORFs on mORF
116 translation using transient expression assays, and identified three CPuORFs that repress mORF translation in a
117 sequence-dependent manner. These findings demonstrate that ESUCA is highly useful for comprehensive
118 identification of CPuORFs and for the selection of CPuORFs likely to encode functional peptides.
119

120 # Results

121 **ESUCA algorithm**

122 ESUCA consists of a six-step procedure (Fig. 1). The first step is extraction of uORF sequences from a transcript
123 sequence dataset. uORFs are extracted by searching the 5´-UTR sequence of each transcript for an ATG codon
124 and its nearest downstream in-frame stop codon. When there are splice variants of a gene, uORFs in all splice
125 variants are extracted. The second step assesses whether transcripts containing a uORF-mORF fusion are major
126 or minor forms among homologous transcripts containing sequences similar to each of the uORFs. In this step,
127 the ratio of RefSeq RNAs with a uORF-mORF fusion to all RefSeq RNAs containing both sequences similar to
128 the uORF and the mORF of each uORF-containing transcript is calculated (Fig. 2). We define this ratio as the
129 uORF-mORF fusion ratio. If the uORF-mORF fusion ratio of a uORF is equal to or greater than 0.3, then the
130 uORF is discarded. The third step is uORF amino acid sequence homology searches. In this step, tBLASTn
131 searches are performed against a transcript sequence database, using the amino acid sequences of the uORFs as
132 queries (uORF-tBLASTn analysis). uORFs are selected that have tBLASTn hits derived from other orders. The
133 fourth step is selection of uORFs conserved between homologous genes. To confirm whether each of the
134 uORF-tBLASTn hits is derived from a homolog of the gene containing the corresponding original uORF, the
135 downstream sequences of the ORFs matching the original uORF are subjected to another tBLASTn analysis

136 using the amino acid sequence of the mORF associated with the original uORF as a query (mORF-tBLASTn
137 analysis) (Fig. 3). If a uORF-tBLASTn hit sequence has a partial or intact ORF containing a sequence similar to
138 the mORF amino acid sequence downstream of the ORF matching the original uORF, it is considered to be
139 derived from a homolog of the original uORF-containing gene. If uORF-tBLASTn hits are found in homologous
140 genes from at least two orders other than that of the original uORF, then the uORF is selected for further analysis.
141 The fifth step is K_a/K_s analysis. In this step, a ratio of nonsynonymous to synonymous nucleotide
142 substitutions (K_a/K_s) for each of the selected uORFs is calculated to assess whether the selected uORF sequences
143 are conserved at the nucleotide or amino acid level. A K_a/K_s ratio close to 1 indicates neutral evolution,
144 whereas a K_a/K_s ratio close to 0 suggests that purifying selection acted on the amino acid sequences. If the K_a/K_s
145 ratio of a uORF is less than 0.5 and significantly different from that of the negative control with q less than 0.05,
146 then the uORF is selected as a CPuORF. The final step is to determine the taxonomic range of uORF sequence
147 conservation. In this step, the uORF-tBLASTn and mORF-tBLASTn hits extracted in the fourth step are
148 classified into taxonomic categories (Fig. 4). The taxonomic range of sequence conservation is determined for
149 each CPuORF based on the presence of the uORF-tBLASTn and mORF-tBLASTn hits in each taxonomic
150 category.
151
152 **Identification of angiosperm CPuORFs using ESUCA**
153 We applied ESUCA to five angiosperm species for which entire genomic DNA and transcript sequence datasets
154 were available: *Arabidopsis*, rice, tomato, poplar and grape. In the first step, we extracted uORF sequences from
155 the 5′-UTR sequence of each transcript, using the transcript sequence datasets described in the Materials and
156 Methods. In these datasets, different transcript IDs are assigned to each splice variant of the same gene. To extract
157 sequences of uORFs and their associated mORFs from all splice variants, we extracted uORF and mORF
158 sequences from each of the transcripts with different transcript IDs. In the second step, we calculated the
159 uORF-mORF fusion ratio of each uORF-containing transcript, using the extracted uORF and mORF sequences,
160 and removed uORFs with uORF-mORF fusion ratios equal to or greater than 0.3 (Supplementary Table S1). We
161 also discarded uORFs whose numbers of RefSeq RNAs containing both sequences similar to the uORF and its
162 associated mORF were less than 10, because appropriate evaluations of uORF-mORF fusion ratios were difficult
163 with a few related RefSeq RNAs and such uORFs are unlikely to be evolutionarily conserved. In the third step,

164 using the amino acid sequences of the remaining uORFs as queries, we performed uORF-tBLASTn searches
165 against the plant transcript sequence database containing assembled EST/transcriptome shotgun assembly (TSA)
166 contigs, singleton EST/TSA sequences and RefSeq RNAs. We selected uORF-tBLASTn hits derived from
167 different orders from that of each query uORF sequence. Plant EST and TSA databases can include contaminant
168 sequences from other organisms, such as parasites, plant-feeding insects and infectious microorganisms. We
169 checked the possibility that the selected uORF-tBLASTn hits included contaminant EST/TSA sequences, using
170 BLASTn searches. The BLASTn searches were performed using each uORF-tBLASTn hit EST/TSA sequence
171 as a query against EST/TSA and RefSeq RNA sequences from all organisms, with an *E*-value cutoff of 10^{-100} and
172 an identity threshold of 95%. Contaminant EST/TSA sequences were identified by this analysis, as described in
173 Materials and Methods, and removed from the uORF-tBLASTn hits. In the fourth step, the remaining
174 uORF-tBLASTn hit sequences were subjected to mORF-tBLASTn analysis, and transcript sequences
175 homologous to the corresponding original uORF-containing transcripts were extracted. We selected uORFs
176 whose tBLASTn hit sequences were found in homologs from at least two orders other than that of the original
177 uORF. Then, amino acid sequence alignments of the selected uORFs and their homologs were generated, using a
178 putative homologous uORF sequence from each order in which uORF-tBLASTn and mORF-tBLASTn hits
179 were found. When multiple original uORFs derived from splice variants of the same gene partially or completely
180 shared amino acid sequences, the one with the longest conserved region was manually selected based on the
181 uORF amino acid sequence alignments. In the fifth step, the remaining uORFs were subjected to K_a/K_s analysis.
182 The uORFs with a K_a/K_s ratio less than 0.5 showing a significant difference from that of the negative control ($q <$
183 0.05) were selected as candidate CPuORFs (Supplementary Table S1). In *Arabidopsis*, rice, tomato, poplar and
184 grape, 104, 59, 42, 148, and 79 candidate CPuORFs were extracted, respectively. Of these, 87 *Arabidopsis*, 53
185 rice, 30 tomato, 77 poplar, and 44 grape uORFs belong to the previously identified CPuORF families, homology
186 groups (HGs) 1 to 53 [20, 22-25] (Supplementary Table S1). The amino acid sequences of the remaining
187 candidate CPuORFs are not similar to those of the known CPuORFs. Therefore, 17, 6, 12, 71, and 35 novel
188 candidate CPuORFs were extracted from *Arabidopsis*, rice, tomato, poplar, and grape, respectively.
189
190 **Validation of the candidate CPuORFs**

191 If the amino acid sequence of a uORF is evolutionarily conserved because of the function of the uORF-encoded
192 peptide, it is expected that the amino acid sequence in the functionally important region of the peptide is
193 conserved among the uORF and its orthologous uORF sequences. Therefore, we manually checked whether the
194 amino acid sequences in the same region are conserved among uORF sequences in the alignment of each of the
195 novel candidate CPuORFs. We found that the alignments of 13 novel candidate CPuORFs contain sequences that
196 do not share the consensus amino acid sequence in the conserved region, and removed these sequences from the
197 alignments. We also removed sequences derived from genes not related to the corresponding original
198 uORF-containing gene from the alignments of five novel candidate CPuORFs. When these changes resulted in
199 the number of orders containing the uORF-tBLASTn and mORF-tBLASTn hits becoming less than two, the
200 candidate CPuORFs were discarded. Eight novel candidate CPuORFs were discarded for this reason.
201 Supplementary Figure S1 shows the uORF amino acid sequence alignments without the removed sequences.
202 K_a/K_s ratios were recalculated after the removal of the sequences (Supplementary Table S1), and six additional
203 novel candidate CPuORFs were discarded because their K_a/K_s ratios were greater than 0.5.

204 Using genomic position information from Ensembl Plants (http://plants.ensembl.org/index.html) [30] and
205 Phytozome v12.1 (https://phytozome.jgi.doe.gov/pz/portal.html) [31], we manually checked whether the
206 positions of the remaining novel candidate CPuORFs overlap with those of the mORFs of other genes or the
207 mORFs of splice variants of the same genes. We found that the genomic position of the candidate CPuORF of the
208 *Arabidopsis ROA1* (AT1G60200) gene overlaps with that of an intron in the mORF region of a splice variant.
209 Protein sequences with an N-terminal region similar to the amino acid sequence encoded by the 5′-extended region
210 of the mORF in this splice variant are found in most of orders from which the uORF-tBALASTn and
211 mORF-tBLASTn hits of this candidate CPuORF were extracted, suggesting that the splice variant carrying the
212 5′-extended mORF is not a minor transcript form among the homologs. Therefore, this candidate CPuORF was
213 discarded.

214 In the second ESUCA step, we excluded uORF sequences likely to encode parts of the mORF-encoded
215 proteins, by removing uORFs with high uORF-mORF fusion ratios. To confirm that the novel candidate
216 CPuORFs do not encode parts of the mORF-encoded proteins, each of the putative uORF sequences used for the
217 alignment and K_a/K_s analysis of each novel candidate CPuORF was queried against the UniProt protein database
218 (https://www.uniprot.org/), using BLASTx. When putative uORF sequences related to a candidate CPuORF

219 matched protein sequences with low *E*-values, we manually checked whether amino acid sequences similar to
220 those encoded by the putative uORFs were included within mORF-encoded protein sequences. In this analysis,
221 mORF-encoded proteins containing N-terminal sequences similar to the amino acid sequences encoded by the
222 candidate CPuORFs of the rice *OsUAM2* gene and its poplar ortholog, POPTR_0019s07850, were identified in
223 many orders. This suggests that the sequences encoded by these candidate CPuORFs are likely to function as
224 parts of the mORF-encoded proteins. Therefore, we excluded these candidate CPuORFs. For some other novel
225 candidate CPuORFs, mORF-encoded proteins containing sequences similar to those encoded by the candidate
226 CPuORF and/or its homologous putative uORFs were also found. However, we did not exclude these candidate
227 CPuORFs, because such uORF-mORF fusion type proteins were only found in a few species for each candidate.
228 After manual validation, 12, 4, 11, 70, and 34 uORFs were identified as novel CPuORFs in *Arabidopsis*, rice,
229 tomato, poplar and grape, respectively. Among these novel CPuORFs, those with both similar uORF amino acid
230 sequences and homologous mORF sequences were classified into the same HGs. Also, using OrthoFinder ver.
231 1.1.4 [32], an algorithm for ortholog group inference, we classified the genes containing novel CPuORFs and
232 those containing previously identified CPuORFs into orthologous groups. The same HG number with a different
233 sub-number was assigned to CPuORFs in genes of the same orthologous group with dissimilar uORF sequences
234 (e.g. HG56.1 and HG56.2). Of the newly identified CPuORF genes, six were classified into the same orthologous
235 groups as genes containing previously identified CPuORFs, but the amino acid sequences of these six CPuORFs
236 are dissimilar to those of the known CPuORFs. Including this type of CPuORF, we identified 131 novel
237 CPuORFs belonging to 88 novel HGs, HG2.2, HG9.2, HG16.2, HG43.2, HG50.2, HG52.2 and HG54-HG130
238 (Supplementary Table S1).
239
240 **Determination of the taxonomic range of CPuORF sequence conservation**
241 As the final ESUCA step, we determined the taxonomic range of the sequence conservation of each CPuORF
242 identified, including previously identified CPuORFs. For this purpose, the uORF-tBLASTn and
243 mORF-tBLASTn hit sequences selected for the K_a/K_s analysis and alignment generation and retained after
244 manual validation were classified into 13 plant taxonomic categories (See Materials and Methods for details.),
245 based on taxonomic lineage information of EST, TSA, and RefSeq RNA sequences. Supplementary Table S2
246 shows the taxonomic range of sequence conservation for each CPuORF. In this table, the presence of the

247 classified transcript sequences in each of the 13 categories is indicated by highlighting cells in red, and the
248 numbers of orders from which the classified transcript sequences are derived are indicated in cells of each
249 category. Figure 5 shows the taxonomic range of the sequence conservation of each HG. In this figure, the
250 conservation range of CPuORFs belonging to each HG is indicated as lines with colors showing species from
251 which the original uORF sequences are derived. In general, CPuORFs belonging to previously identified HGs
252 tend to be conserved in a wider range of taxonomic categories than those belonging to the newly identified HGs.
253 In 69 of the novel HGs, CPuORF sequences are conserved only among eudicots. In 12 of these, CPuORF
254 sequences are conserved in even narrower taxonomic ranges, only among rosids or asterids. In 44 of the 69
255 eudicot-conserved novel HGs, the CPuORF sequences are conserved both in rosids and asterids, indicating that
256 these CPuORFs are conserved across diverse eudicots. These results indicate that the taxonomic range of
257 CPuORF sequence conservation varies, and that ESUCA can identify CPuORFs conserved in a relatively narrow
258 taxonomic range.
259
260 **Sequence-dependent effects of CPuORFs on mORF translation**
261 To address the relationship between the taxonomic range of CPuORF sequence conservation and the
262 sequence-dependent effects of CPuORFs on mORF translation, we selected eight poplar CPuORFs and
263 examined their sequence-dependent effects on expression of the downstream reporter gene using a transient
264 expression assay. Of the selected CPuORFs, those belonging to HG46, HG55, HG57 and HG66 are conserved
265 across diverse angiosperms (Fig. 5, Supplementary Table S2). HG66 is even conserved in gymnosperm and
266 Polypodiopsida. The CPuORFs belonging to HG80, HG81 and HG87 are conserved in a wide range of eudicots,
267 whereas the CPuORF belonging to HG88 is conserved only among rosids (Fig. 5, Supplementary Table S2). The
268 5´-UTR sequences containing the selected CPuORFs were fused to the Fluc coding sequence and placed under
269 the control of the 35S promoter to generate the wild-type (WT) reporter constructs (Fig. 6a, Supplementary
270 Figure S2). To assess the importance of the amino acid sequences for the effects of these CPuORFs on mORF
271 translaion, frameshift mutations were introduced into the CPuORFs so that the amino acid sequences of their
272 conserved regions could be altered. A + 1 or − 1 frameshift was introduced upstream or within the conserved
273 region of each CPuORF, and another frameshift was introduced before the stop codon to shift the reading frame
274 back to the original frame (Supplementary Figure S2). These reporter constructs were each transfected into

275 protoplasts from *A. thaliana* MM2d suspension-cultured cells. After 24 h of incubation, cells were harvested and
276 disrupted to analyze luciferase activity. In three of the eight CPuORFs, the introduced frameshift mutations
277 significantly upregulated the expression of the reporter gene, indicating that these CPuORFs repress mORF
278 translation in a sequence-dependent manner (Fig. 6b). This result suggests that CPuORFs conserved only among
279 eudicots can have a sequence-dependent regulatory effect.
280

281 # Discussion

282 In this study, we developed ESUCA, a pipeline for genome-wide identification of CPuORFs and determination
283 of the taxonomic range of CPuORF sequence conservation. By applying ESUCA to five plant genomes, we
284 identified 131 CPuORFs belonging to 88 novel HGs. Furthermore, we showed that three of seven novel
285 CPuORFs conserved in a wide range of eudicots have a sequence-dependent regulatory effect on mORF
286 translation. Therefore, this study demonstrates that by applying ESUCA to many species it can be used to identify
287 many CPuORFs and efficiently select those likely to encode regulatory peptides.
288

289 **Identification of CPuORFs and determination of the taxonomic range of CPuORF sequence conservation**
290 **by ESUCA**

291 Of the 88 novel CPuORF HGs identified in this study, 70 were identified in ESUCA analysis of only one of
292 *Arabidopsis*, rice, tomato, poplar or grape (Fig. 5). This means that CPuORFs belonging to these HGs cannot be
293 identified by comparing uORF sequences of orthologous genes between two of these five species. This result
294 highlights that ESUCA can identify CPuORFs that are difficult to find by conventional comparative genomic
295 approaches in which uORF sequences are compared between selected species. In 41 of 70 HGs identified by
296 ESUCA analysis of only one of the five plant species, CPuORF sequences are conserved in both rosids and
297 asterids, two major groups of eudicots, indicating that these CPuORFs are conserved across diverse eudicots (Fig.
298 5). Nevertheless, CPuORFs belonging to 30 of these 41 HGs were extracted by ESUCA analysis of only one of
299 *Arabidopsis*, tomato or poplar, which belong to rosids or asterids (Fig. 5). In addition, 11 of 20 novel HGs
300 identified by only ESUCA analysis of grape, which belongs to eudicots other than rosids and asterids, are
301 conserved across diverse eudicots (Fig. 5). Also, among novel HGs identified by ESUCA analysis of only rice or

302 tomato, HG75 and HG77 are conserved in a wide range of eudicots (Fig. 5). One possible explanation of these
303 observations is that sequence conservation of uORFs is often lost in small taxonomic groups such as orders or
304 families. For example, while the CPuORF of the tomato LOC101264451 gene, which belongs to HG43.1, exerts
305 a sequence-dependent repressive effect on mORF translation, the CPuORF of its *Arabidopsis* ortholog,
306 *ANAC096*, lacks the C-terminal half of the amino acid sequence in the highly conserved region and does not have
307 a sequence-dependent regulatory effect, although it was extracted as a CPuORF due to its partially conserved
308 sequence [25-27]. In contrast to the *Arabidopsis ANAC096* CPuORF, the HG43.1 CPuORF of *Tarenaya*
309 *hassleriana*, which belongs to the same order as *Arabidopsis*, Brassicales, but a different family, has all the critical
310 amino acid residues in the highly conserved region. Another example is found in the CPuORFs belonging to
311 HG32. HG32 CPuORFs are widely conserved across Spermatophyta, including papaya (*Caryca papaya*) and *T.*
312 *hassleriana*, which belong to the Caricaceae and Cleomaceae families of the Brassicales order, respectively.
313 However, *Arabidopsis* orthologs of HG32 CPuORF-containing genes have no uORF with an amino acid
314 sequence similar to those of HG32 CPuORFs, although their 5′-UTR sequences contain nucleotide sequences
315 resembling parts of the papaya and *T. hassleriana* HG32 CPuORFs [25]. CPuORFs involved in preferable but
316 not essential post-transcriptional regulations may be lost in some taxonomic clades during evolution. Such
317 CPuORFs would be difficult to be identified by comparing uORF sequences between a few selected species if a
318 selected species is included in the clades where the CPuORF sequences were lost. Therefore, the ESUCA method
319 is advantageous for comprehensive identification of CPuORFs compared with conventional comparative
320 genomic approaches.

321 We examined effects of eight poplar CPuORFs on mORF translation using transient expression assays. Two
322 of four CPuORFs conserved beyond eudicots showed a sequence-dependent repressive effect, whereas one of
323 three CPuORFs conserved among diverse eudicots showed a sequence-dependent repressive effect (Fig. 6b).
324 Previously, we examined the effects of 16 CPuORFs, which belong to HG27-HG29, HG33-HG43.1, HG44, and
325 HG45, on mORF translation and identified six sequence-dependent regulatory uORFs. In this analysis, five of
326 eleven CPuORFs conserved beyond eudicots showed sequence-dependent repressive effects, whereas one of five
327 CPuORFs conserved among diverse eudicots showed a sequence-dependent repressive effect [26, 27]. These
328 results suggest that CPuORFs conserved among diverse eudicots are likely to encode functional peptides, and
329 that CPuORFs conserved beyond eudicots are even more likely to encode functional peptides. These studies

330 demonstrate that the automatic determination of CPuORF sequence conservation taxonomic range by ESUCA is
331 useful for the efficient selection of CPuORFs likely to encode functional peptides.
332 Of the CPuORFs analyzed for their sequence-dependent regulatory effects in the present study, the CPuORFs
333 belonging to HG55 and HG66 showed no significant sequence-dependent effect on mORF translation, despite
334 their widespread sequence conservation beyond eudicots (Figs. 5 and 6b). These CPuORFs might encode
335 peptides that have functions other than the control of mORF translation, or they might exert sequence-dependent
336 regulatory effects only under certain conditions. In fact, many known sequence-dependent regulatory uORFs
337 repress mORF translation in response to metabolites, such as polyamine, arginine, and sucrose (15-18,45).
338 Likewise, it is possible that the CPuORFs belonging to HG80 and HG81, which showed no significant
339 sequence-dependent regulatory effect despite their conservation across diverse eudicots (Figs. 5 and 6b), might
340 encode peptides that have other functions or might exert a sequence-dependent regulatory effect only under
341 specific conditions.
342
343 **Filtering using the uORF-mORF fusion ratio**
344 In many genes, there are splice variants in which the protein-coding mORF is split into two or more ORFs. The
345 ORF encoding the N-terminal region of the protein in each of such splice variants appears like a uORF and can
346 be extracted as a CPuORF if the amino acid sequence in the N-terminal region of the protein is evolutionarily
347 conserved. It is difficult to distinguish between this type of 'spurious' CPuORFs and 'true' CPuORFs because
348 even 'true' CPuORF-containing genes produce splice variants in which a CPuORF is fused to the mORF, as seen
349 in the At2g31280, At5g01710, and At5g03190 genes [20, 25, 28]. To automatically distinguish between
350 'spurious' and 'true' CPuORFs, we employed the criterion of the uORF-mORF fusion ratio in ESUCA and
351 discarded uORFs with uORF-mORF fusion ratios equal to or greater than 0.3. We checked how efficiently the
352 'spurious' CPuORFs were removed with this criterion, using a protein sequence database. Although
353 uORF-mORF fusion type protein sequences were found in OsUAM2 homologs from widespread angiosperm
354 species, no other candidate CPuORFs were extracted in which uORF-mORF fusion type protein sequences were
355 found in many of species with sequences related to the candidate CPuORFs. This indicates that the uORF-mORF
356 fusion ratio filtering worked efficiently to exclude 'spurious' CPuORFs that encode parts of the mORF-encoded
357 proteins.

358 Of the CPuORFs with previously demonstrated sequence-dependent regulatory effects, the HG3 CPuORFs,
359 associated with an mORF encoding an *S*-adenosylmethionine decarboxylase (AdoMetDC; EC 4.1.1.50) [20, 33],
360 tended to show relatively high uORF-mORF fusion ratios, although still below 0.3 (Supplementary Table S1). In
361 this study, eleven CPuORFs belonging to this HG were extracted using ESUCA and the uORF-mORF fusion
362 ratios of these CPuORFs were in the range of 0.17 to 0.27, with a median value of 0.26 (Supplementary Table
363 S1). Consistent with these relatively high uORF-mORF fusion ratios, uORF-mORF fusion type protein
364 sequences containing amino acid sequences resembling HG3 CPuORFs are found in widespread angiosperm
365 species. The uORF-mORF fusion ratios of the candidate CPuORFs of rice *OsUAM2* and its poplar ortholog are
366 0.28 (Supplementary Table S1). The rice *OsUAM2* gene encodes a protein similar to UDP-arabinopyranose
367 mutases (EC 5.4.99.30) [34]. Two protein isoforms, a uORF-mORF fusion type and one lacking the
368 uORF-encoded region, produced from splice variants of a *OsUAM2* orthologous gene are found in diverse
369 angiosperm species, but the functions of both of isoforms are not yet known. Considering the example of HG3
370 CPuORFs, we cannot rule out the possibility that a candidate CPuORF functions as a regulatory uORF even if
371 uORF-mORF fusion type protein sequences are found in widespread species. However, to avoid including
372 potential 'spurious' CPuORFs whose amino acid sequences are likely to be evolutionarily conserved due to their
373 function as N-terminal regions of the mORF-encoded proteins, we excluded the candidate CPuORFs of the rice
374 *OsUAM2* gene and its poplar ortholog. In conclusion, the criterion of the uORF-mORF fusion ratio used in this
375 study appears appropriate because all previously identified sequence-dependent regulatory CPuORFs were
376 extracted and most 'spurious' CPuORFs were removed.
377

378 ## Conclusion

379 The present study demonstrates that ESUCA enabled the identification of CPuORFs conserved in various
380 taxonomic ranges and allowed for the efficient selection of CPuORFs likely to function as regulatory uORFs.
381 This study also shows that CPuORF sequence conservation appears to be often lost during evolution, and that
382 ESUCA can identify CPuORFs conserved only among limited taxonomic groups. Therefore, it is expected that
383 additional novel angiosperm CPuORFs will be identified if ESUCA is applied to angiosperm species other than
384 those analyzed in this study. Moreover, applying ESUCA to other organisms will contribute to the

385 comprehensive identification of potential regulatory uORFs. Furthermore, in addition to CPuORFs, the
386 algorithms developed in this study can be applied to the identification of other sequences conserved in various
387 taxonomic ranges.
388

389 Materials and methods

390 **Extraction of uORF sequences**

391 We used genome sequence files in FASTA format and genomic coordinate files in GFF3 format obtained from
392 Ensemble Plants Release 33 (https://plants.ensembl.org/index.html) [30] to extract *Arabidopsis* (*Arabidopsis*
393 *thaliana*), tomato (*Solanum lycopersicum*), poplar (*Populus trichocarpa*), and grape (*Vitis vinifera*) uORF
394 sequences. We used a genome sequence file in FASTA format and a genomic coordinate files in GFF3 format
395 obtained from Phytozome v11 (https://phytozome.jgi.doe.gov/pz/portal.html) [31] for rice (*Oryza sativa*). We
396 extracted exon sequences from genome sequences, based on genomic coordinate information, and constructed
397 transcript sequence datasets by combining exon sequences. Based on the transcription start site and the translation
398 initiation codon of each transcript in the genomic coordinate files, we extracted 5′-UTR sequences from the
399 transcript sequence datasets. Then, we searched the 5′-UTR sequences for an ATG codon and its nearest
400 downstream in-frame stop codon. Sequences starting with an ATG codon and ending with the nearest in-frame
401 stop codon were extracted as uORF sequences. When multiple uORFs in a transcript shared the same stop codon,
402 only the longest uORF sequence was used for further analyses.
403

404 **Assembly of EST and TSA sequences**

405 EST, TSA, and RefSeq RNA sequence datasets were obtained from the International Nucleotide Sequence
406 Database Collaboration databases (NCBI release of 2016-12-03 and DDBJ release 106.0 for EST and TSA
407 sequences, NCBI release 79 for RefSeq RNA sequences). Based on taxonomic lineage information provided by
408 NCBI Taxonomy (https://www.ncbi.nlm.nih.gov/taxonomy), EST and TSA sequences derived from
409 Viridiplantae were extracted from the databases. EST and TSA sequences from the same species were assembled
410 using Velvet ver. 1.2.10 [35] with k-mer length of 99. The k-mer length was optimized using *A. thaliana* EST and
411 TSA sequences to minimize the total numbers of assembled contigs and unclustered singleton sequences.

412 Unclustered singleton EST/TSA sequences, derived from species for which RefSeq RNA sequences were
413 available, were mapped to the RefSeq RNA sequences using Bowtie2 ver. 2.2.9 [36] and the default parameters.
414 We discarded singleton EST/TSA sequences that matched any RefSeq RNA sequences from the same species.
415 We created a plant transcript sequence database for BLAST searches in our local computers by using
416 makeblastdb, a program contained in the NCBI-BLAST package, and the remaining singleton EST/TSA
417 sequences, assembled contigs, and Viridiplantae RefSeq RNA sequences.
418
419 **Calculation of the uORF-mORF fusion ratio**
420 The uORF-mORF fusion ratio of each of the extracted uORFs was assessed as follows. We performed tBLASTx
421 using each uORF sequence as a query against plant (Viridiplantae) RefSeq RNA sequences with an *E*-value
422 cutoff of 2000 (uORF-tBLASTx). We used standalone NCBI-BLAST + ver. 2.6.0 [37] for all BLAST analyses.
423 We next performed tBLASTx with an *E*-value threshold of 10^{-1} using the mORF sequence associated with each
424 uORF as a query against the uORF-tBLASTx hit sequences (mORF-tBLASTx). Using these two-step tBLASTx
425 searches, we selected RefSeq RNAs containing both sequences similar to the uORF and the mORF of each
426 uORF-containing transcript. Then, we examined whether the largest ORF of each of the selected RefSeq RNAs
427 included the region matching the original mORF in the same reading frame (Fig. 2). We also examined whether
428 the largest ORF included the region matching the original uORF, irrespective of the reading frame. The RefSeq
429 RNA was considered to have a uORF-mORF fusion if both regions matching the original uORF and mORF
430 were contained in the largest ORF (Fig. 2). The RefSeq RNA was considered to have a uORF separated from the
431 downstream mORF if the region matching the original mORF was contained in the largest ORF but the region
432 matching the original uORF was not (Fig. 2). RefSeq RNA numbers of the former and latter types were defined
433 as X and Y, respectively. We calculated a uORF-mORF fusion ratio as $X / (X + Y)$ for each of the original
434 uORF-containing transcript sequences.
435
436 **BLAST-based search for uORFs conserved between homologous genes**
437 To search for uORFs with amino acid sequences conserved between homologous genes, we first performed
438 tBLASTn searches against the assembled plant transcript sequence database, using the amino acid sequences of
439 the uORFs as queries. In these uORF-tBLASTn searches, we extracted transcript sequences matching a uORF

440 with an *E*-value less than 2000 and derived from orders other than that of the uORF. The downstream in-frame
441 stop codon closest to the 5′-end of the matching region of each uORF-tBLASTn hit sequence was selected (Fig.
442 3). Then, we looked for an in-frame ATG codon upstream of the selected stop codon, without any other in-frame
443 stop codon between them. uORF-tBLASTn hit sequences without such an ATG codon were discarded. If one or
444 more in-frame ATG codons were identified, the 5′-most ATG codon was selected. The ORF beginning with the
445 selected ATG codon and ending with the selected stop codon was extracted as a putative uORF (Fig. 3). The
446 downstream sequences of putative uORFs were subjected to another tBLASTn analysis to examine whether the
447 transcripts were derived from homologs of the gene containing the original uORF. In this analysis, the amino acid
448 sequence of the mORF associated with the original uORF was used as a query sequence, and transcript
449 sequences matching the mORF with an *E*-value less than 10^{-1} were extracted. For each of the uORF-tBLASTn
450 and mORF-tBLASTn hit sequences, the upstream in-frame stop codon closest to the 5′-end of the region
451 matching the original mORF was selected, and the 5′-most in-frame ATG codon located downstream of the
452 selected stop codon was identified as the putative mORF initiation codon (Fig. 3). uORF-tBLASTn and
453 mORF-tBLASTn hit sequences were discarded as uORF-mORF fusion type sequences if the putative mORF
454 overlapped with the putative uORF. The original uORF was selected as a candidate CPuORF if the remaining
455 uORF-tBLASTn and mORF-tBLASTn hits belonged to more than two orders, including that from which the
456 original uORF was derived.
457
458 **Identification of contaminant ESTs and TSAs**
459 In the uORF-tBLASTn searches described above, we excluded tBLASTn hit ESTs and TSAs derived from
460 contaminated organisms. To examine whether each uORF-tBLASTn hit sequence is derived from contaminated
461 organisms, we performed BLASTn searches against EST, TSA, and RefSeq RNA sequences from all organisms
462 except for those of metagenomes, using each uORF-tBLASTn hit EST/TSA sequence as a query. If a
463 uORF-tBLASTn hit EST/TSA sequence matched an EST, TSA, or RefSeq RNA sequence of a different order
464 from the species of the uORF-tBLASTn hit, with an *E*-value less than 10^{-100} and an identity equal to or greater
465 than 95%, it was considered a candidate contaminant sequence. In this case, either the uORF-tBLASTn hit or the
466 BLASTn hit may be a contaminant sequence. To distinguish these possibilities, we compared the ratio of the
467 BLASTn hit number to the total EST/TSA and RefSeq RNA sequence number between the species of each

468 uORF-tBLASTn hit and the species of its BLASTn hits. Appropriate comparisons are difficult unless species
469 used for this comparison have enough number of EST/TSA and RefSeq RNA sequences. Therefore, if the total
470 EST/TSA and RefSeq RNA sequence number of a species is less than 5000, BLASTn hits derived from the
471 species were not used for this analysis. If the ratio of the BLASTn hit number to the total EST/TSA and RefSeq
472 RNA sequence number of a uORF-tBLASTn hit species is less than that of any other BLASTn hit species, the
473 uORF-tBLAST hit sequence was identified as a contaminant sequence.
474
475 **K_a/K_s analysis**
476 For K_a/K_s analysis of each candidate CPuORF, a putative uORF sequence was selected from each order in which
477 uORF-tBLASTn and mORF-tBLASTn hits were found, using the following criteria. First, we selected transcript
478 sequences matching the mORF associated with the candidate CPuORF with an *E*-value less than 10^{-20} in
479 mORF-tBLASTn analysis. Of these sequences, we then selected the one with the smallest geometric means of
480 mORF-tBLASTn and uORF-tBLASTn *E*-values. When mORF-tBLASTn *E*-values of all uORF-tBLASTn and
481 mORF-tBLASTn hits in an order were equal to or greater than 10^{-20}, we selected the transcript sequence with the
482 smallest geometric means of mORF-tBLASTn and uORF-tBLASTn *E*-values in the order. Putative uORF
483 sequences in the selected transcript sequences were used to generate uORF amino acid sequence alignments as
484 well as for K_a/K_s analysis. ClustalW ver. 2.1 [38], on the DDBJ server, was used to generate multiple alignments
485 in our previous study [25]. Here, we used standalone Clustal Omega (ClustalO) ver. 1.2.2 [39] because it is
486 superior to ClustalW in terms of calculation speed and alignment accuracy [39]. For each candidate CPuORF, a
487 median K_a/K_s ratio for all pairwise combinations of the original uORF and its homologous putative uORFs was
488 calculated using the LWL85 algorithm [40] in the seqinR package [41].
489 For statistical tests of K_a/K_s ratios, we calculated distribution of mutation rates between the original uORF and
490 its homologous putative uORFs and those between the original uORF and its artificially generated mutants, using
491 the observed mutation rate distribution. Then, observed empirical K_a/K_s ratio distributions were compared with
492 null distributions (negative controls) using the Mann-Whitney *U* test to validate statistical significance. The
493 one-sided *U* test was used to investigate whether the observed distributions were significantly lower than the null
494 distributions. Adjustment for multiple comparisons was achieved by controlling the false discovery rate using the
495 Benjamini and Hochberg procedure [42].

496

497 **Determination of the taxonomic range of uORF sequence conservation**

498 To automatically determine the taxonomic range of the sequence conservation of each CPuORF, we first defined
499 13 plant taxonomic categories. The 13 taxonomic defined categories are lamiids, asterids other than lamiids,
500 mavids, fabids, eudicots other than rosids and asterids, commelinids, monocots other than commelinids,
501 Angiospermae other than eudicots and monocots, Gymnospermae, Polypodiopsida, Embryophyta other than
502 Euphyllophyta, Streptophyta other than Embryophyta, and Viridiplantae other than Streptophyta. Based on
503 taxonomic lineage information of EST, TSA, and RefSeq RNA sequences, which were provided by NCBI
504 Taxonomy, the uORF-tBLASTn and mORF-tBLASTn hit sequences selected for K_a/K_s analysis were classified
505 into the 13 taxonomic categories (Fig. 4). It should be noted that, in NCBI taxonomy, eudicots, Angiospermae,
506 and Ggymnospermae are referred to as eudicotyledons, Magnoliophyta, and Acrogymnospermae, respectively.
507 For each CPuORF, the numbers of transcript sequences classified into each category were counted and shown in
508 Supplementary Table S2. These numbers represent the numbers of orders in which the amino acid sequence of
509 each CPuORF is conserved.

510

511 **Statistical and informatic analyses**

512 All programs, except for existing stand-alone programs, such as BLAST [37], ClustalO [39] and Jalview [43],
513 were written in R (www.r-project.org). We also used R libraries, GenomicRanges [44], exactRankTests,
514 Biostrings and seqinr [41].

515

516 **Plasmid construction**

517 Plasmid pNH006 harbors the cauliflower mosaic virus 35S RNA (35S) promoter, the firefly luciferase (Fluc)
518 coding sequence, and the polyadenylation signal of the *A. thaliana HSP18.2* gene in pUC19. To construct this
519 plasmid, pMT61 [26] was digested with *Sal*I and *Sac*I, and the *Sal*I-*Sac*I fragment containing the Fluc coding
520 sequence was ligated into the *Sal*I and *Sac*I sites between the 35S promoter and the *HSP18.2* polyadenylation
521 signal of plasmid pKM56 [45]. To generate reporter plasmids (pNH92-pNH99) for transient expression assays,
522 the 5′-UTR sequences of eight poplar genes were amplified by PCR from poplar (*Populus nigra*) full-length
523 cDNA clones pds25559, pds10965, pds14390, pds13862, pds15817, pds28294, pds26157 and pds14623

524 (Supplementary Table S3), obtained from RIKEN [46]. Primer sets used are shown in Supplementary Tables S4
525 and S5. To construct pNH92, pNH94-pNH97 and pNH99, amplified fragments containing the 5′-UTR
526 sequences were digested with *Xba*I and *Sal*I and ligated between the *Xba*I and *Sal*I sites of pNH006. To create
527 pNH93 and pNH98, amplified fragments containing the 5′-UTR sequences were inserted between the *Xba*I and
528 *Sal*I sites of pNH006 using the SLiCE method [47]. Frameshift mutations were introduced into each CPuORF
529 using overlap extension PCR [48], with primers listed in Supplementary Tables S4 and S5, to yield
530 pNH100-pNH107. Sequence analysis confirmed the integrity of the PCR-amplified regions of all constructs.
531

532 **Transient expression assay**

533 Transient expression assays were performed as described in Hayashi *et al*. 2017 [45]. Protoplasts from
534 *Arabidopsis* MM2d suspension cells [49] were used, as were the reporter plasmids described above and the
535 pKM5 [45] internal control plasmid. pKM5 contains the 35S promoter, the *Renilla* luciferase (Rluc) coding
536 sequence, and the *NOS* polyadenylation signal in pUC19. For each experiment, 5 μg each of a reporter plasmid
537 and pKM5 were transfected into protoplasts.
538

539 Declarations

540 **Funding**

541 This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI [Grant Nos.
542 JP16H05063 to S.N., JP16K07387 to H.O., JP18H03330 to H.T and H.O., JP18K06297 to H.T]; the Ministry of
543 Education, Culture, Sports, Science and Technology (MEXT) KAKENHI [Grant Nos. JP26113519 to S.K,
544 JP16H01246 to S.K, JP17H05658 to S.N., JP26114703 to H.T, JP17H05659 to H.T]; the Research Foundation
545 for the Electrotechnology of Chubu to H.T; and the Naito Foundation to H.O.
546

547 **Authors' contributions**

548 HT, HO, SK, and SN conceived and designed the experiments. NH and YY performed the experiments. HT, HO,
549 TE, and AT analyzed the data. HT, NH, YY, AT, and KF contributed to reagents/materials/analysis tools. HO
550 and HT wrote the paper. All authors read and approved the final manuscript.

551 **Competing interests**

552 The authors declare that they have no competing interests.

553

554 **Ethics approval and consent to participate**

555 Not applicable.

556

557 **Consent for publication**

558 Not applicable.

559

560 **Availability of data and materials**

561 Not applicable.

562

563 **Acknowledgements**

564 We thank Ms. Kazuko Harada for general assistance. We thank, Rebecca Porter, PhD, from Edanz Group

565 (www.edanzediting.com/ac) for editing a draft of this manuscript.

566

567 **List of abbreviations**

568 35S: cauliflower mosaic virus 35S RNA; 5′-UTR: 5′-untranslated region; CPuORF: conserved peptide upstream

569 open reading frame; EST: expressed sequence tag; Fluc: firefly luciferase; HG: homology group; mORF: main

570 open reading frame; RefSeq: NCBI reference sequence; Rluc: Renilla luciferase; TSA: transcriptome shotgun

571 assembly; uORF: upstream open reading frame

572

573

574

575

576

577

578

579 # References

580 1. Churbanov A, Rogozin IB, Babenko VN, Ali H, Koonin EV. Evolutionary conservation suggests a
581 regulatory function of AUG triplets in 5'-UTRs of eukaryotic genes. Nucleic Acids Res. 2005;
582 33:5512-5520.

583 2. Galagan JE, Calvo SE, Cuomo C, Ma LJ, Wortman JR, Batzoglou S, et al. Sequencing of *Aspergillus*
584 *nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. Nature 2005; 438:1105-1115.

585 3. Kawaguchi R, Bailey-Serres J. mRNA sequence features that contribute to translational regulation in
586 *Arabidopsis*. Nucleic Acids Res. 2005; 33:955-965.

587 4. Rogozin IB, Kochetov AV, Kondrashov FA, Koonin EV, Milanesi L. Presence of ATG triplets in 5'
588 untranslated regions of eukaryotic cDNAs correlates with a 'weak' context of the start codon.
589 Bioinformatics 2001; 17:890-900.

590 5. Morris DR, Geballe AP. Upstream open reading frames as regulators of mRNA translation. Mol. Cell
591 Biol. 2000; 20:8635-8642.

592 6. Cruz-Vera LR, Sachs MS, Squires CL, Yanofsky C. Nascent polypeptide sequences that influence
593 ribosome function. Curr. Opin. Microbiol. 2011; 14:160-166.

594 7. Ito K, Chiba S. Arrest peptides: cis-acting modulators of translation. Annu. Rev. Biochem. 2013;
595 82:171-202.

596 8. Somers J, Poyry T, Willis AE. A perspective on mammalian upstream open reading frame function. Int.
597 J. Biochem. Cell Biol. 2013; 45:1690-1700.

598 9. Bhushan S, Meyer H, Starosta AL, Becker T, Mielke T, Berninghausen O, et al. Structural basis for
599 translational stalling by human cytomegalovirus and fungal arginine attenuator peptide. Mol. Cell 2010;
600 40:138-146.

601 10. Wang Z, Sachs MS. Ribosome stalling is responsible for arginine-specific translational attenuation in
602 *Neurospora crassa*. Mol. Cell Biol. 1997; 17:4904-4913.

603 11. Gaba A, Jacobson A, Sachs MS. Ribosome occupancy of the yeast *CPA1* upstream open reading frame
604 termination codon modulates nonsense-mediated mRNA decay. Mol. Cell 2005; 20:449-460.
605 12. Uchiyama-Kadokura N, Murakami K, Takemoto M, Koyanagi N, Murota K, Naito S, et al.
606 Polyamine-responsive ribosomal arrest at the stop codon of an upstream open reading frame of the
607 AdoMetDC1 gene triggers nonsense-mediated mRNA decay in *Arabidopsis thaliana*. Plant Cell
608 Physiol. 2014; 55:1556-1567.
609 13. Wang Z, Gaba A, Sachs MS. A highly conserved mechanism of regulated ribosome stalling mediated
610 by fungal arginine attenuator peptides that appears independent of the charging status of arginyl-tRNAs.
611 J. Biol. Chem. 1999; 274:37565-37574.
612 14. Law GL, Raney A, Heusner C, Morris DR. Polyamine regulation of ribosome pausing at the upstream
613 open reading frame of S-adenosylmethionine decarboxylase. J. Biol. Chem. 2001; 276:38036-38043.
614 15. Hanfrey C, Elliott KA, Franceschetti M, Mayer MJ, Illingworth C, Michael AJ. A dual upstream open
615 reading frame-based autoregulatory circuit controlling polyamine-responsive translation. J. Biol. Chem.
616 2005; 280:39229-39237.
617 16. Yamashita Y, Takamatsu S, Glasbrenner M, Becker T, Naito S, Beckmann R. Sucrose sensing through
618 nascent peptide-meditated ribosome stalling at the stop codon of Arabidopsis *bZIP11* uORF2. FEBS
619 Lett. 2017; 591:1266-1277.
620 17. Rahmani F, Hummel M, Schuurmans J, Wiese-Klinkenberg A, Smeekens S, Hanson J. Sucrose control
621 of translation mediated by an upstream open reading frame-encoded peptide. Plant Physiol. 2009;
622 150:1356-1367.
623 18. Combier JP, de Billy F, Gamas P, Niebel A, Rivas S. Trans-regulation of the expression of the
624 transcription factor MtHAP2-1 by a uORF controls root nodule development. Genes Dev. 2008;
625 22:1549-1559.
626 19. Crowe ML, Wang XQ, Rothnagel JA. Evidence for conservation and selection of upstream open
627 reading frames suggests probable encoding of bioactive peptides. BMC genomics 2006; 7:16.

628 20. Hayden CA, Jorgensen RA. Identification of novel conserved peptide uORF homology groups in
629 *Arabidopsis* and rice reveals ancient eukaryotic origin of select groups and preferential association with
630 transcription factor-encoding genes. BMC Biol. 2007; 5:32.
631 21. Hayden CA, Bosco G. Comparative genomic analysis of novel conserved peptide upstream open
632 reading frames in *Drosophila melanogaster* and other dipteran species. BMC genomics 2008; 9:61.
633 22. Tran MK, Schultz CJ, Baumann U. Conserved upstream open reading frames in higher plants. BMC
634 genomics 2008; 9:361.
635 23. Vaughn JN, Ellingson SR, Mignone F, Arnim A. Known and novel post-transcriptional regulatory
636 sequences are conserved across plant families. Rna 2012; 18:368-384.
637 24. van der Horst S, Snel B, Hanson J, Smeekens S. Novel pipeline identifies new upstream ORFs and
638 non-AUG initiating main ORFs with conserved amino acid sequences in the 5' leader of mRNAs in
639 *Arabidopsis thaliana*. RNA 2018:(in press).
640 25. Takahashi H, Takahashi A, Naito S, Onouchi H. BAIUCAS: a novel BLAST-based algorithm for the
641 identification of upstream open reading frames with conserved amino acid sequences and its application
642 to the *Arabidopsis thaliana* genome. Bioinformatics 2012; 28:2231-2241.
643 26. Ebina I, Takemoto-Tsutsumi M, Watanabe S, Koyama H, Endo Y, Kimata K, et al. Identification of
644 novel *Arabidopsis thaliana* upstream open reading frames that control expression of the main coding
645 sequences in a peptide sequence-dependent manner. Nucleic Acids Res. 2015; 43:1562-1576.
646 27. Noh AL, Watanabe S, Takahashi H, Naito S, Onouchi H. An upstream open reading frame represses
647 expression of a tomato homologue of Arabidopsis *ANAC096*, a NAC domain transcription factor gene,
648 in a peptide sequence-dependent manner. Plant Biotechnol. 2015; 32:157-163.
649 28. Jorgensen RA, Dorantes-Acosta AE. Conserved Peptide Upstream Open Reading Frames are
650 Associated with Regulatory Genes in Angiosperms. Front. Plant Sci. 2012; 3:191.
651 29. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant
652 sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2007; 35:D61-D65.

653 30. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, et al. Ensembl 2018. Nucleic
654 Acids Res. 2018; 46:D754-D761.
655 31. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, et al. Phytozome: a comparative
656 platform for green plant genomics. Nucleic Acids Res. 2012; 40:D1178-D1186.
657 32. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons
658 dramatically improves orthogroup inference accuracy. Genome Biol. 2015; 16:157.
659 33. Franceschetti M, Hanfrey C, Scaramagli S, Torrigiani P, Bagni N, Burtin D, et al. Characterization of
660 monocot and dicot plant S-adenosyl-l-methionine decarboxylase gene families including identification
661 in the mRNA of a highly conserved pair of upstream overlapping open reading frames. Biochem. J.
662 2001; 353:403-409.
663 34. Konishi T, Takeda T, Miyazaki Y, Ohnishi-Kameyama M, Hayashi T, O'Neill MA, et al. A plant
664 mutase that interconverts UDP-arabinofuranose and UDP-arabinopyranose. Glycobiology 2007;
665 17:345-354.
666 35. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs.
667 Genome Res. 2008; 18:821-829.
668 36. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat. Methods 2012; 9:357-359.
669 37. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and
670 PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;
671 25:3389-3402.
672 38. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive
673 multiple sequence alignment through sequence weighting, position-specific gap penalties and weight
674 matrix choice. Nucleic Acids Res. 1994; 22:4673-4680.
675 39. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of
676 high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 2011; 7:539.
677 40. Li WH. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol.

678 1993; 36:96-99.

679 41. Charif D, Lobry JR: SeqinR 1.0-2: a contributed package to the R project for statistical computing
680 devoted to biological sequences retrieval and analysis. In *Structural Approaches to Sequence Evolution:*
681 *Molecules, Networks, Populations.* Edited by Bastolla U, Porto M, Roman HE, Vendruscolo M. New
682 York: Springer Verlag; 2007: 207-232

683 42. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to
684 multiple testing. J. R. Statist. Soc. ser.B 1995; 57:298-300.

685 43. Clamp M, Cuff J, Searle SM, Barton GJ. The Jalview Java alignment editor. Bioinformatics 2004;
686 20:426-427.

687 44. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, et al. Software for computing
688 and annotating genomic ranges. PLoS Comput. Biol. 2013; 9:e1003118.

689 45. Hayashi N, Sasaki S, Takahashi H, Yamashita Y, Naito S, Onouchi H. Identification of *Arabidopsis*
690 *thaliana* upstream open reading frames encoding peptide sequences that cause ribosomal arrest. Nucleic
691 Acids Res. 2017; 45:8844-8858.

692 46. Nanjo T, Futamura N, Nishiguchi M, Igasaki T, Shinozaki K, Shinohara K. Characterization of
693 full-length enriched expressed sequence tags of stress-treated poplar leaves. Plant Cell Physiol. 2004;
694 45:1738-1748.

695 47. Motohashi K. Seamless Ligation Cloning Extract (SLiCE) Method Using Cell Lysates from Laboratory
696 Escherichia coli Strains and its Application to SLiP Site-Directed Mutagenesis. Methods Mol. Biol.
697 2017; 1498:349-357.

698 48. Ho SN, Hunt HD, Horton RM, Pullen JK, Pease LR. Site-directed mutagenesis by overlap extension
699 using the polymerase chain reaction. Gene 1989; 77:51-59.

700 49. Menges M, Murray JA. Synchronous *Arabidopsis* suspension cultures for analysis of cell-cycle gene
701 activity. Plant J. 2002; 30:203-212.

702

703 # Figure legends

704

705 **Figure 1.** Outline of the ESUCA pipeline.

706

707

708 **Figure 2.** Schematic representation of the algorithm to calculate uORF-mORF fusion ratios. For each original
709 uORF-containing transcript sequence, RefSeq RNAs are selected that match an original uORF sequence,
710 irrespective of the reading frame, and the associated original mORF sequence in the same reading frame as the
711 largest ORF of the RefSeq RNA, using tBLASTx. The shaded regions in the open boxes represent the matching
712 regions from the tBLASTx searches. If the original uORF-matching region is within the largest ORF, the RefSeq
713 RNA is considered a uORF-mORF fusion type, the number of which is defined as '*X*'. If the original
714 uORF-matching region is not within the largest ORF, the RefSeq RNA is considered a uORF-mORF separate
715 type, the number of which is defined as '*Y*'. For each of the original uORF-containing transcripts, the
716 uORF-mORF fusion ratio is calculated as $X / (X + Y)$.

717

718 **Figure 3.** Schematic representation of BLAST-based search for uORFs conserved between homologous genes.
719 In the third ESUCA step, tBLASTn searches are conducted against a transcript sequence database consisting of
720 assembled EST/TSA contigs, unclustered singleton EST/TSA sequences and RefSeq RNAs, using original
721 uORF sequences as queries (uORF-tBLASTn). The shaded regions in the open boxes show the matching regions
722 from the tBLASTn searches. Asterisks represent stop codons. (i) The downstream in-frame stop codon closest to
723 the 5′-end of the matching region of each uORF-tBLASTn hit sequence is selected. (ii) The 5′-most in-frame
724 ATG codon located upstream of the stop codon is selected. The ORF beginning with the selected ATG codon and
725 ending with the selected stop codon is extracted as a putative uORF. In the fourth ESUCA step, the downstream
726 sequences of putative uORFs in the transcript sequences are subjected to mORF-tBLASTn analysis. Transcript
727 sequences matching the mORF associated with the original uORF with an *E*-value less than 10^{-1} are extracted.
728 (iii) For each of the uORF-tBLASTn and mORF-tBLASTn hit sequences, the upstream in-frame stop codon
729 closest to the 5′-end of the matching region is selected. (iv) The 5′-most in-frame ATG codon located downstream

730 of the selected stop codon is identified as the initiation codon of the putative partial or intact mORF. If the putative
731 mORF overlaps with the putative uORF, the uORF-tBLASTn and mORF-tBLASTn hit sequence is discarded as
732 a uORF-mORF fusion type.
733
734 **Figure 4.** Schematic representation of the algorithm to determine the taxonomic range of uORF sequence
735 conservation. In the fifth ESUCA step, for K_a/K_s analysis of each of the remaining original uORF-containing
736 transcripts, a uORF-tBLASTn and mORF-tBLASTn hit sequence is selected from each order that has
737 uORF-tBLASTn and mORF-tBLASTn hits (See Materials and Methods for the criteria for the selection). In the
738 sixth ESUCA step, the selected transcript sequences are classified into the 13 plant taxonomic categories to
739 determine the taxonomic range of uORF sequence conservation, using taxonomic lineage information of EST,
740 TSA, and RefSeq RNA sequences.
741
742 **Figure 5.** Taxonomic range of the sequence conservation of the CPuORF families. The blue, yellow, red, green,
743 and purple lines show the conservation range of CPuORF HGs determined by applying ESUCA to *Arabidopsis*
744 (*Arabidopsis thaliana*), rice (*Oryza sativa*), tomato (*Solanum lycopersicum*), poplar (*Populus trichocarpa*), and
745 grape (*Vitis vinifera*), respectively. The presence of a line within a cell in each taxonomic category indicates the
746 presence of uORF-tBLASTn and mORF-tBLASTn hits for any of the CPuORFs belonging to each HG. In
747 taxonomic categories with a category name with an asterisk, uORF-tBLASTn and mORF-tBLASTn hits found
748 in lower taxonomic categories were excluded. In the case where no uORF-tBLASTn and mORF-tBLASTn hit is
749 found in the taxonomic category containing a species from which an original uORF sequence is derived, the line
750 showing the species was still drawn in the cell of the taxonomic category because the CPuORF sequence is
751 conserved in the species with the original uORF. *Arabidopsis*, rice, tomato, poplar and grape belong to malvids,
752 commelinids, lamiids, fabids and eudicots*, respectively. HG1-HG53 are previously identified HGs, except for
753 HG2.2, HG9.2, HG16.2, HG43.2, HG50.2 and HG52.2, whereas HG2.2, HG9.2, HG16.2, HG43.2, HG50.2,
754 HG52.2 and HG54-HG130 are newly identified HGs.
755
756 **Figure 6.** Sequence-dependent effects of novel CPuORFs on main ORF translation. (A) Schematic
757 representation of the WT (*35S::UTR(WT):Fluc*) and frameshift mutant (*35S::UTR(fs):Fluc*) reporter constructs.

758 The 5′-UTR sequence containing each of the eight poplar CPuORFs was inserted between the 35S promoter
759 (*35Spro*) and the *Fluc* coding sequence. The hatched box in the frameshift (fs) mutant CPuORF indicates the
760 frame-shifted region. The dotted boxes represent the first five nucleotides of the mORF associated with each of
761 the eight poplar CPuORFs. See Supplementary Figure S2 for the exact position and length of each CPuORF and
762 the exact frame-shifted region. The polyadenylation signal of the *A. thaliana HSP18.2* gene is designated as 'ter'.
763 (B) Transient expression assay. Each reporter plasmid containing a WT or fs mutant CPuORF, belonging to an
764 indicated HG, was co-transfected into MM2d protoplasts with the *35S::Rluc* internal control plasmid by PEG
765 treatment. After incubation for 24 h, cells were collected and disrupted for dual luciferase assay. Fluc activity was
766 normalized to Rluc activity, and the activity relative to that of the corresponding WT reporter construct was
767 calculated. Means ± SD of three biological replicates are shown. Single and double asterisks indicate significant
768 differences between the WT and fs constructs at $p < 0.05$ and $p < 0.01$, respectively, as determined by Student's
769 *t*-test. Representative results of two experiments using independently prepared protoplasts are shown.

Figure 1

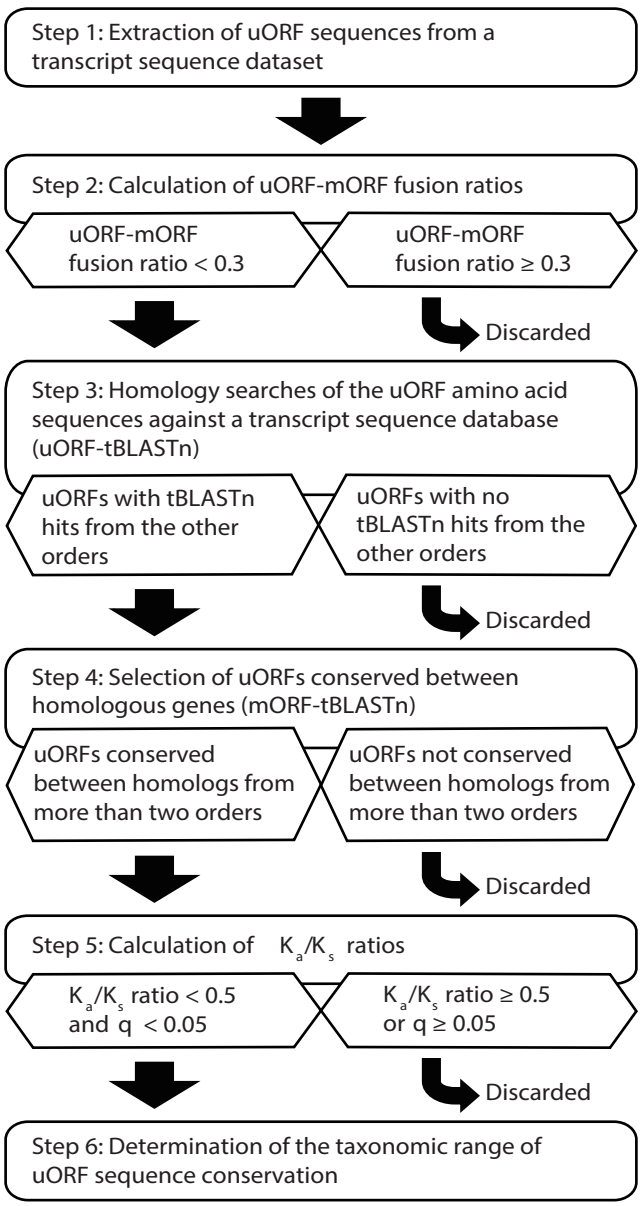

Step 1: Extraction of uORF sequences from a transcript sequence dataset
Step 2: Calculation of uORF-mORF fusion ratios
uORF-mORF fusion ratio < 0.3
uORF-mORF fusion ratio ≥ 0.3
Discarded
Step 3: Homology searches of the uORF amino acid sequences against a transcript sequence database (uORF-tBLASTn)
uORFs with tBLASTn hits from the other orders
uORFs with no tBLASTn hits from the other orders
Discarded
Step 4: Selection of uORFs conserved between homologous genes (mORF-tBLASTn)
uORFs conserved between homologs from more than two orders
uORFs not conserved between homologs from more than two orders
Discarded
Step 5: Calculation of K_a/K_s ratios
K_a/K_s ratio < 0.5 and q < 0.05
K_a/K_s ratio ≥ 0.5 or q ≥ 0.05
Discarded
Step 6: Determination of the taxonomic range of uORF sequence conservation

Figure 2

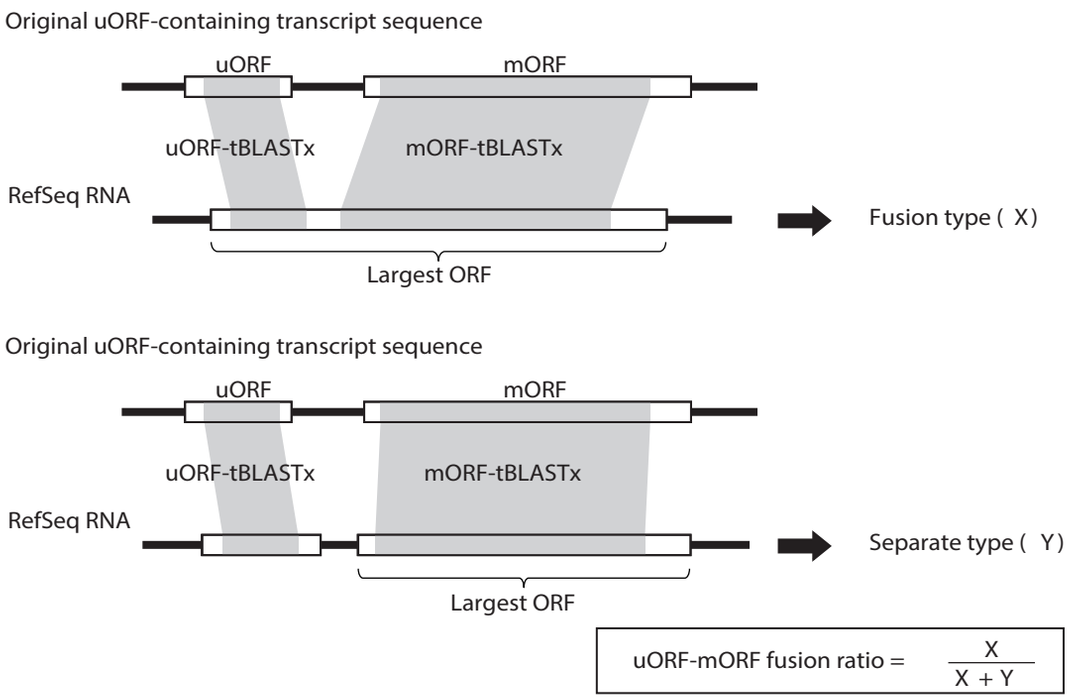

Original uORF-containing transcript sequence
uORF
mORF
uORF-tBLASTx
mORF-tBLASTx
RefSeq RNA
Fusion type (X)
Largest ORF
Original uORF-containing transcript sequence
uORF
mORF
uORF-tBLASTx
mORF-tBLASTx
RefSeq RNA
Separate type (Y)
Largest ORF
uORF-mORF fusion ratio = X / X + Y

Figure 3

Original uORF-containing transcript sequence

uORF-tBLASTn and mORF-tBLASTn hit sequence (Assembled EST/TSA and RefSeq RNA)

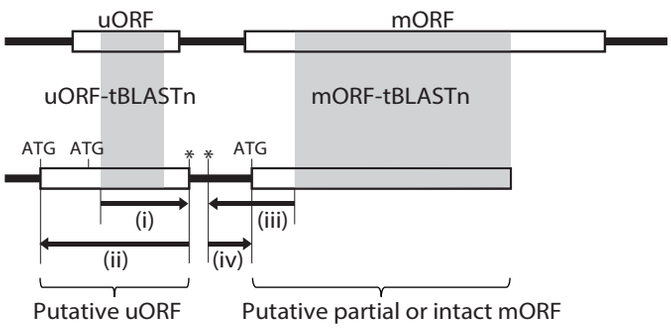


Figure 4

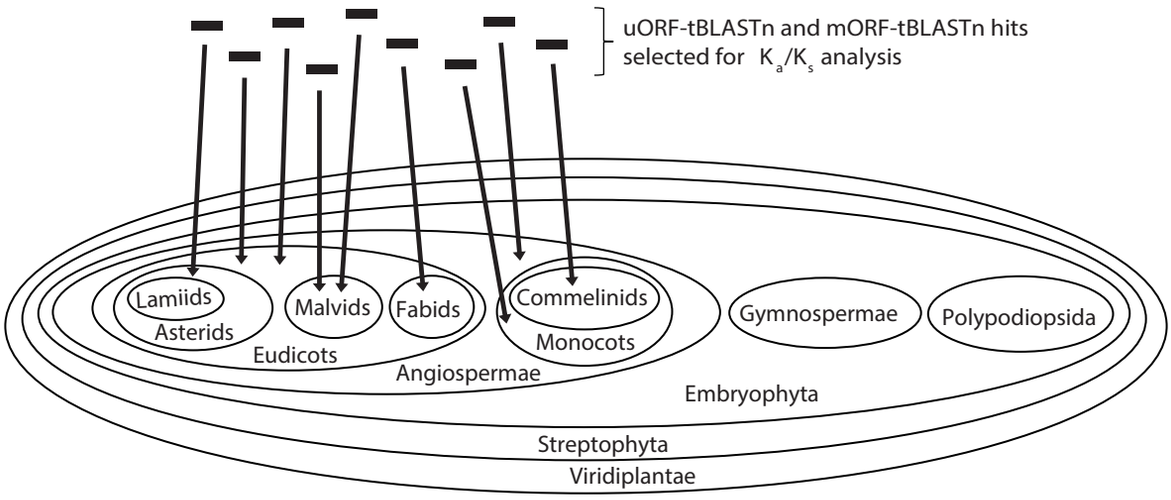

uORF-tBLASTn and mORF-tBLASTn hits
selected for K_a/K_s analysis
Lamiids
Asterids
Malvids
Fabids
Commelinids
Monocots
Gymnospermae
Polypodiopsida
Eudicots
Angiospermae
Embryophyta
Streptophyta
Viridiplantae

Figure 5

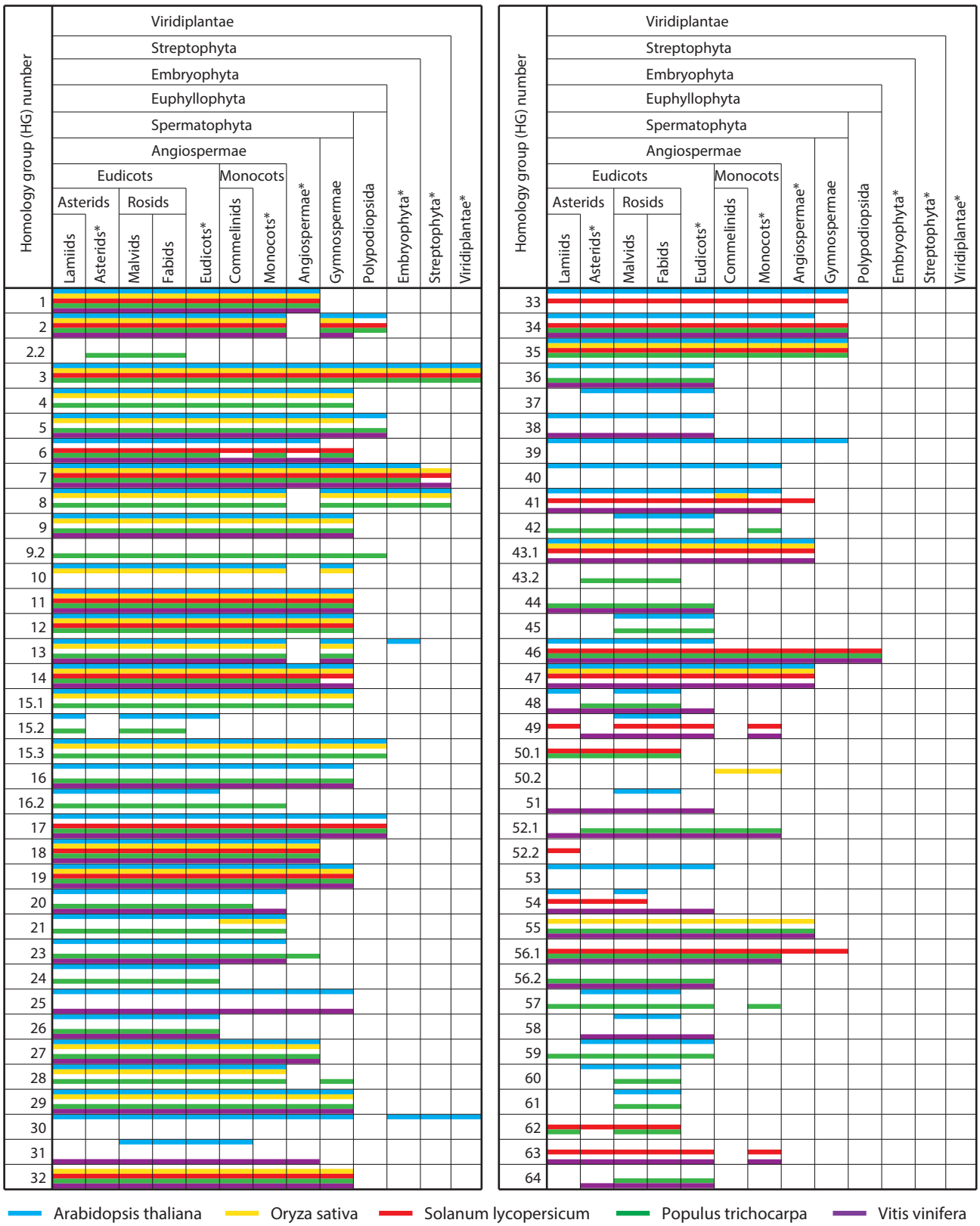

Homology group (HG) number
Viridiplantae
Streptophyta
Embryophyta
Euphyllophyta
Spermatophyta
Angiospermae
Eudicots
Monocots
Asterids
Rosids
Lamiids
Asterids*
Malvids
Fabids
Eudicots*
Commelinids
Monocots*
Angiospermae*
Gymnospermae
Polypodiopsida
Embryophyta*
Streptophyta*
Viridiplantae*
1
2
2.2
3
4
5
6
7
8
9
9.2
10
11
12
13
14
15.1
15.2
15.3
16
16.2
17
18
19
20
21
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43.1
43.2
44
45
46
47
48
49
50.1
50.2
51
52.1
52.2
53
54
55
56.1
56.2
57
58
59
60
61
62
63
64
Arabidopsis thaliana
Oryza sativa
Solanum lycopersicum
Populus trichocarpa
Vitis vinifera

Figure 5 (continued)

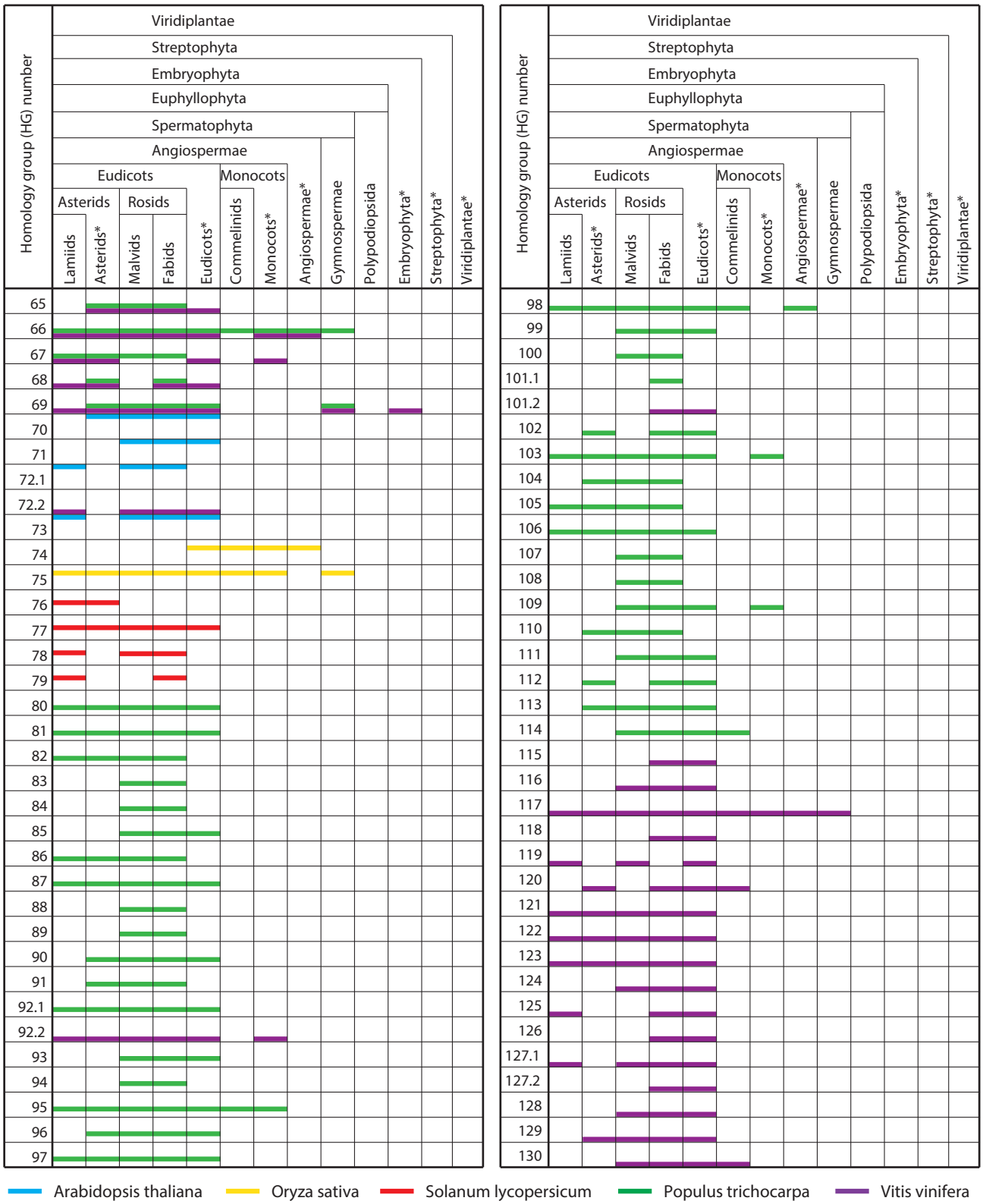
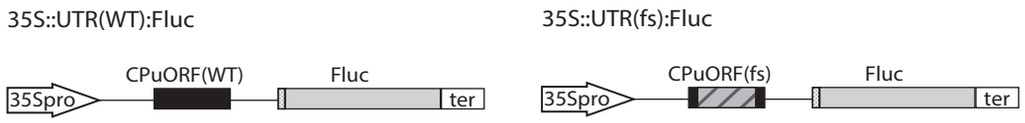
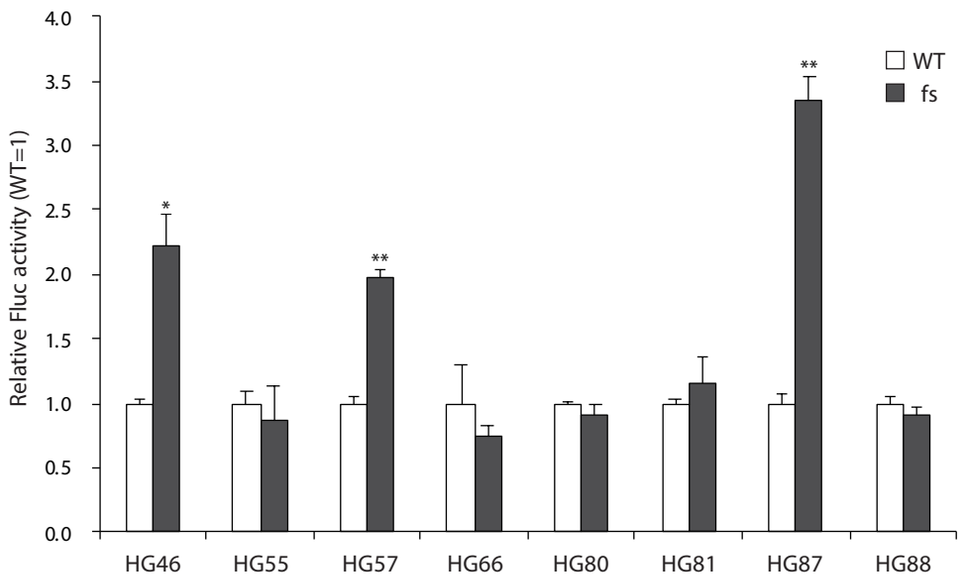


Figure 6

A


35S::UTR(WT):Fluc
35Spro
CPuORF(WT)
Fluc
ter
35S::UTR(fs):Fluc
35Spro
CPuORF(fs)
Fluc
ter

B


Relative Fluc activity (WT=1)
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
WT
fs
*
**
**
HG46
HG55
HG57
HG66
HG80
HG81
HG87
HG88