BMC Genomics

RESEARCH ARTICLE Open Access

# Comprehensive genome-wide identification of angiosperm upstream ORFs with peptide sequences conserved in various taxonomic ranges using a novel pipeline, ESUCA

Hiro Takahashi[1,2*†], Noriya Hayashi[3†], Yuta Hiragori[3], Shun Sasaki[3], Taichiro Motomura[1], Yui Yamashita[3], Satoshi Naito[3,4], Anna Takahashi[5], Kazuyuki Fuse[6], Kenji Satou[7], Toshinori Endo[8], Shoko Kojima[9] and Hitoshi Onouchi[3*]

## Abstract

**Background:** Upstream open reading frames (uORFs) in the 5′-untranslated regions (5′-UTRs) of certain eukaryotic mRNAs encode evolutionarily conserved functional peptides, such as cis-acting regulatory peptides that control translation of downstream main ORFs (mORFs). For genome-wide searches for uORFs with conserved peptide sequences (CPuORFs), comparative genomic studies have been conducted, in which uORF sequences were compared between selected species. To increase chances of identifying CPuORFs, we previously developed an approach in which uORF sequences were compared using BLAST between *Arabidopsis* and any other plant species with available transcript sequence databases. If this approach is applied to multiple plant species belonging to phylogenetically distant clades, it is expected to further comprehensively identify CPuORFs conserved in various plant lineages, including those conserved among relatively small taxonomic groups.

(Continued on next page)

* Correspondence: takahasi@p.kanazawa-u.ac.jp; onouchi@abs.agr.hokudai.ac.jp
†Hiro Takahashi and Noriya Hayashi contributed equally to this work.
[1]Graduate School of Medical Sciences, Kanazawa University, Kanazawa 920-1192, Japan
[3]Graduate School of Agriculture, Hokkaido University, Sapporo 060-8589, Japan
Full list of author information is available at the end of the article

(Continued from previous page)

**Results:** To efficiently compare uORF sequences among many species and efficiently identify CPuORFs conserved in various taxonomic lineages, we developed a novel pipeline, ESUCA. We applied ESUCA to the genomes of five angiosperm species, which belong to phylogenetically distant clades, and selected CPuORFs conserved among at least three different orders. Through these analyses, we identified 89 novel CPuORF families. As expected, ESUCA analysis of each of the five angiosperm genomes identified many CPuORFs that were not identified from ESUCA analyses of the other four species. However, unexpectedly, these CPuORFs include those conserved across wide taxonomic ranges, indicating that the approach used here is useful not only for comprehensive identification of narrowly conserved CPuORFs but also for that of widely conserved CPuORFs. Examination of the effects of 11 selected CPuORFs on mORF translation revealed that CPuORFs conserved only in relatively narrow taxonomic ranges can have sequence-dependent regulatory effects, suggesting that most of the identified CPuORFs are conserved because of functional constraints of their encoded peptides.

**Conclusions:** This study demonstrates that ESUCA is capable of efficiently identifying CPuORFs likely to be conserved because of the functional importance of their encoded peptides. Furthermore, our data show that the approach in which uORF sequences from multiple species are compared with those of many other species, using ESUCA, is highly effective in comprehensively identifying CPuORFs conserved in various taxonomic ranges.

**Keywords:** Upstream ORF, Translational regulation, Bioinformatics, Nascent peptide

## 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, certain uORFs encode regulatory peptides that have roles in post-transcriptional regulation of gene expression [5–9]. During translation of some of these regulatory uORFs, nascent peptides act inside the ribosomal exit tunnel to cause ribosome stalling [10]. Ribosome stalling on a uORF results in translational repression of the downstream main ORF (mORF) because stalled ribosomes block the access of subsequently loaded ribosomes to the mORF start codon [11]. Additionally, if ribosome stalling occurs at the stop codon of a uORF, nonsense-mediated mRNA decay (NMD) may be induced [12, 13]. In some genes, uORF-encoded nascent peptides cause ribosome stalling in response to metabolites to down-regulate mORF translation under specific cellular conditions [11, 13–18]. 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 and causes mRNA degradation [19].

To comprehensively identify uORFs that encode functional peptides, genome-wide searches for uORFs with conserved peptide sequences (CPuORFs) have been conducted using comparative genomic approaches in various organisms [20–24]. 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* [21, 23, 24]. 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 [25].

In conventional comparative genomic approaches, uORF sequences are compared among selected species. Therefore, homology detection depends on the selection of species for comparison. In searches using this approach, if a uORF amino acid sequence is not conserved among the selected species, this 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 (for BLAST-based algorithm for identification of uORFs with conserved amino acid sequences) pipeline [26]. In BAIUCAS, homology searches of uORF amino acid sequences are performed using BLAST between a certain species and any other species for which expressed sequence tag (EST) databases are available, and uORFs conserved beyond a certain taxonomic range are selected. Using BAIUCAS, we searched for *Arabidopsis* CPuORFs conserved beyond the order Brassicales, which *Arabidopsis* belongs to, and identified 13 novel CPuORF families [26]. We examined the sequence-dependent effects of the CPuORFs identified by BAIUCAS on mORF translation using a transient expression assay, and identified six regulatory CPuORFs that repress mORF translation in an amino acid sequence-dependent manner [27, 28]. These sequence-dependent regulatory CPuORFs include ones conserved only among relatively small taxonomic groups, such as a part of eudicots. Therefore, it is expected that sequence-dependent regulatory CPuORFs conserved in various plant lineages, including narrowly conserved ones, will be more comprehensively identified if BAIUCAS is applied to many plant species.

Before applying BAIUCAS to many species, improvement of BAIUCAS was desired to more efficiently identify CPuORFs that were conserved because of the functional importance of their encoded peptides. One major problem with identifying CPuORFs is that there are cases where a uORF found in the 5′-UTR of a transcript is fused to the mORF in an isoform of the transcript, and in some of these cases, such uORF sequences are conserved because they actually encode parts of mORF-encoded protein sequences. In other words, there are cases where the protein-coding mORF is split into multiple ORFs in a splice variant and the ORF coding for the N-terminal region of the protein appears like a uORF. Such an ORF can be extracted as a CPuORF if the amino acid sequence in the N-terminal region of the protein is evolutionarily conserved. It is difficult to distinguish between this type of 'spurious' CPuORFs and 'true' CPuORFs because even 'true' CPuORF-containing genes produce splice variants in which a CPuORF is fused to the mORF, as seen in the At2g31280, At5g01710, and At5g03190 genes [21, 26, 29]. Another major point to be improved is the method of calculating nonsynonymous to synonymous nucleotide substitution ($K_a/K_s$) ratios for CPuORF sequences. These $K_a/K_s$ ratios are used to evaluate whether uORF sequences are conserved at the amino acid level or at the nucleotide level [30]. However, $K_a/K_s$ ratios largely depend on the selection of uORF sequences used for their calculations. If uORF sequences used for the calculation of a $K_a/K_s$ ratio include many sequences from closely related species, the $K_a/K_s$ ratio tends to be high. For appropriate calculations of $K_a/K_s$ ratios, uORF sequences need to be selected using proper criteria.

Here, we present an improved BAIUCAS version ESUCA (for evolutionary search for uORFs with conserved amino acid sequences) and genome-wide identification of CPuORFs from five angiosperm genomes using ESUCA. To distinguish between 'spurious' CPuORFs conserved because they code for parts of mORF-encoded proteins and 'true' CPuORFs conserved because of functional constraints of their encoded small peptides, ESUCA includes an algorithm to assess whether, for each uORF, transcripts bearing a uORF-mORF fusion are minor or major forms among orthologous transcripts. Another new function of ESUCA is systematic calculations of $K_a/K_s$ ratios for CPuORF sequences. ESUCA includes an algorithm to select one uORF sequence from each order for calculation of the $K_a/K_s$ ratio of each CPuORF. Additionally, ESUCA is capable of determining the taxonomic range within which each CPuORF is conserved. Although ESUCA can identify CPuORFs conserved only among a small taxonomic group because ESUCA compares uORF sequences between a certain species and any other species with available transcript databases, CPuORFs conserved among a small taxonomic group may be less likely to encode functional peptides than those conserved across a wide taxonomic range. The automatic determination of the taxonomic range of CPuORF conservation provides useful information for the selection of CPuORFs likely to encode functional peptides. The current study demonstrates that ESUCA efficiently identifies CPuORFs likely to be conserved because of functional constraints of their encoded peptides. Furthermore, the data presented here show that the approach in which uORF sequences from multiple species are compared with those of many other species, using ESUCA, is highly effective in comprehensively identifying CPuORFs conserved in various taxonomic lineages.

## Results

### The ESUCA pipeline

In this study, to efficiently identify CPuORFs likely to be conserved because of functional importance of their encoded peptides, we developed a novel pipeline, ESUCA, which consists of a six-step procedure (Fig. 1). The first step is extraction of uORF sequences from a transcript sequence dataset. The uORFs are extracted by searching the 5′-UTR sequence of each transcript for an ATG codon and its nearest downstream in-frame stop codon. Although uORFs overlapping their downstream mORFs are also usually considered uORFs, we focus on the type of uORFs that has both the start and stop codons within the 5′-UTR to avoid including uORFs whose sequences are conserved because of functional constraints of mORF-encoded proteins. When there are splice variants of a gene, uORFs in all splice variants are extracted. The second step assesses whether, for each uORF, uORF-mORF fusion type transcripts are minor or major forms among orthologous transcripts. If transcripts with a uORF-mORF fusion are found as a major form in a majority of species with their orthologs, the uORF sequence is likely to code for a part of the mORF-encoded protein. Therefore, such a uORF should be discarded as a 'spurious' uORF. In contrast, if transcripts with a uORF-mORF fusion are found in only a small proportion of species with their orthologs, the uORF-mORF fusion type transcripts are considered minor form transcripts and therefore can be ignored. For this assessment, the NCBI reference sequence (RefSeq) database is used, which provides curated non-redundant transcript sequences [31]. For each uORF, the ratio of RefSeq RNAs with a uORF-mORF fusion to all RefSeq RNAs with both sequences similar to the uORF and its downstream mORF is calculated (Fig. 2). We define this ratio as the uORF-mORF fusion ratio. If the uORF-mORF fusion ratio of a uORF is equal to or greater than 0.3, then the uORF is discarded. The third step is uORF amino acid sequence homology searches. In this step, tBLASTn searches are performed against a transcript sequence database, using the amino acid sequences of the uORFs as queries (uORF-tBLASTn analysis). The uORFs with tBLASTn hits from other species are selected. The fourth step is selection of

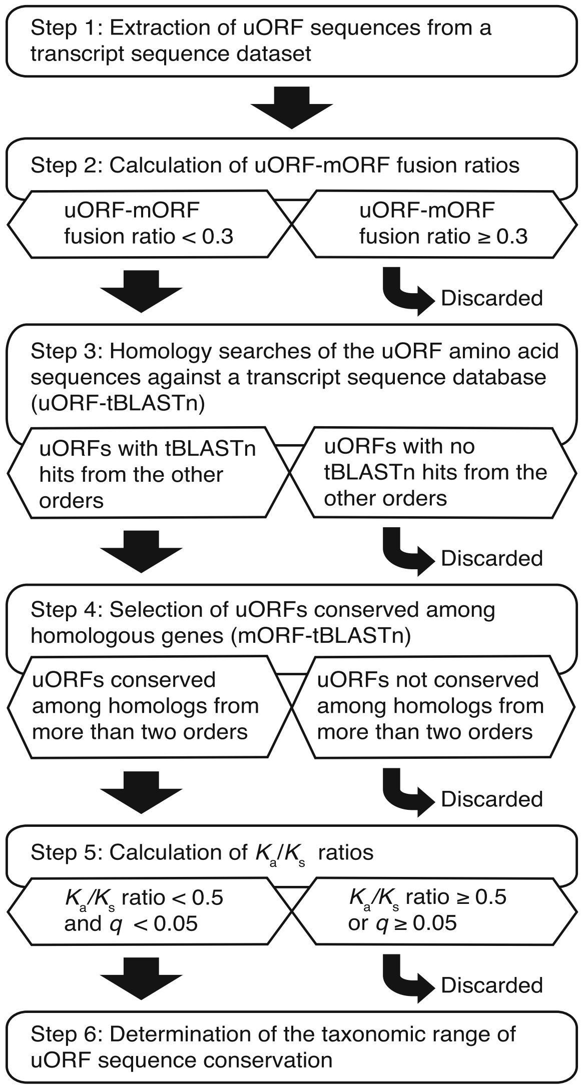
Step 1: Extraction of uORF sequences from a transcript sequence dataset
Step 2: Calculation of uORF-mORF fusion ratios
uORF-mORF fusion ratio < 0.3
uORF-mORF fusion ratio ≥ 0.3
Discarded
Step 3: Homology searches of the uORF amino acid sequences against a transcript sequence database (uORF-tBLASTn)
uORFs with tBLASTn hits from the other orders
uORFs with no tBLASTn hits from the other orders
Discarded
Step 4: Selection of uORFs conserved among homologous genes (mORF-tBLASTn)
uORFs conserved among homologs from more than two orders
uORFs not conserved among 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

**Fig. 1** Outline of the ESUCA pipeline

uORFs conserved among homologous genes. To confirm whether the uORF-tBLASTn hits are derived from homologs of the original uORF-containing gene, the downstream sequences of putative uORFs in the uORF-tBLASTn hits are subjected to another tBLASTn analysis, which uses the mORF amino acid sequence of the original uORF-containing transcript as a query (mORF-tBLASTn analysis) (Fig. 3). If a uORF-tBLASTn hit has a partial or intact ORF that contains a sequence similar to the mORF amino acid sequence downstream of the putative uORF, it is considered to be derived from a homolog of the original uORF-containing gene. If uORF-tBLASTn and mORF-tBLASTn hits are found in at least two orders other than that of the original uORF, then the uORF is selected as a candidate CPuORF. This is because at least three uORF sequences from different orders are necessary to confirm at the later manual validation step that the same region is conserved among homologous uORF sequences. The fifth step is $K_a/K_s$ analysis. In this step, $K_a/K_s$ ratios for the selected candidate CPuORFs is calculated to assess whether the candidate CPuORF sequences are conserved at the nucleotide or amino acid level. A $K_a/K_s$ ratio close to 1 indicates neutral evolution, whereas a $K_a/K_s$ ratio close to 0 suggests that purifying selection acted on the amino acid sequences. For each candidate CPuORF, a representative uORF-tBLASTn and mORF-tBLASTn hit is selected from each order, and the putative uORF sequences in the representative uORF-tBLASTn and mORF-tBLASTn hits are used for the calculation of the $K_a/K_s$ ratio (Fig. 4). If the $K_a/K_s$ ratio of a candidate CPuORF is less than 0.5 and significantly different from that of the negative control with $q$ less than 0.05, then the candidate CPuORF is selected for further analysis. The final step is to determine the taxonomic range of uORF sequence conservation. In this step, the representative uORF-tBLASTn and mORF-tBLASTn hits selected in the fifth step are classified into taxonomic categories (Fig. 4). On the basis of the presence of the uORF-tBLASTn and mORF-tBLASTn hits in each taxonomic category, the taxonomic range of sequence conservation is determined for each CPuORF.

## Identification of angiosperm CPuORFs using ESUCA

We applied ESUCA to five angiosperm species, *Arabidopsis*, rice, tomato, poplar and grape, which belong to phylogenetically distant clades of angiosperm, and for which entire genomic DNA and transcript sequence datasets were available. Rice is a monocot, whereas the others are eudicots. *Arabidopsis* and poplar belong to two different groups of rosids (marvids and fabids), whereas tomato belongs to asterids. Grape belongs to neither rosids nor asterids. In the first step of ESUCA, we extracted uORF sequences from the 5′-UTR sequence of each transcript of these species, using the transcript sequence datasets described in the Materials and Methods. In these datasets, different transcript IDs are assigned to each splice variant from the same gene. To extract sequences of uORFs and their downstream mORFs from all splice variants, we extracted uORF and mORF sequences from each of the transcripts with different transcript IDs. In the second step, we calculated the uORF-mORF fusion ratio of each uORF-containing transcript, using the extracted uORF and mORF sequences, and removed uORFs with uORF-mORF fusion ratios equal to or greater than 0.3 (Supplementary Table S1). We also discarded uORFs whose numbers of RefSeq RNAs containing both sequences similar to the uORF and its downstream mORF were less than 10. This was done because appropriate evaluations of uORF-mORF fusion ratios were difficult with a few related RefSeq RNAs and such uORFs are unlikely to be evolutionarily conserved. In the third step, using the amino acid sequences of the remaining uORFs as queries, we performed uORF-tBLASTn searches

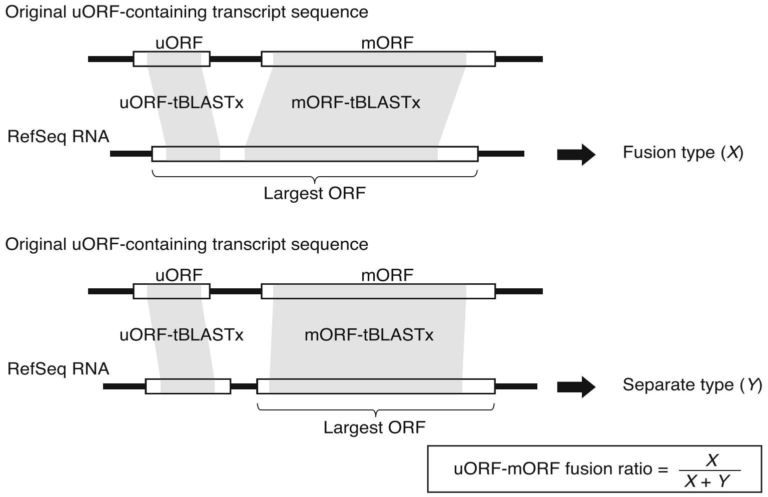

**Fig. 2** Schematic representation of the algorithm to calculate uORF-mORF fusion ratios. For each original uORF-containing transcript sequence, RefSeq RNAs are selected that match an original uORF sequence, irrespective of the reading frame, and the original mORF sequence in the same reading frame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

against a plant transcript sequence database that contained contigs of assembled EST and transcriptome shotgun assembly (TSA), singleton EST/TSA sequences, and RefSeq RNAs (See Materials and Methods for details). In the fourth step, the uORF-tBLASTn hits were subjected to mORF-tBLASTn analysis, and uORF-tBLASTn and mORF-tBLASTn hits were extracted. Plant EST and TSA databases can include contaminant sequences from other organisms, such as parasites, plant-feeding insects and infectious microorganisms. We checked the possibility that the extracted uORF-tBLASTn and mORF-tBLASTn hits included contaminant EST/TSA sequences, using BLASTn searches. The BLASTn searches were performed using each uORF-tBLASTn and mORF-tBLASTn hit EST/TSA sequence as a query against EST/TSA and RefSeq RNA sequences from all organisms, with an *E*-value cutoff of $10^{-100}$ and an identity threshold of 95%. Contaminant EST/TSA sequences were identified by this analysis, as described in Materials and Methods, and were removed from the uORF-tBLASTn and mORF-tBLASTn hits. We selected uORFs whose remaining uORF-tBLASTn and mORF-tBLASTn hits were found in homologs from at least two orders other than that of the original uORF. Thereafter, we generated multiple amino acid sequence

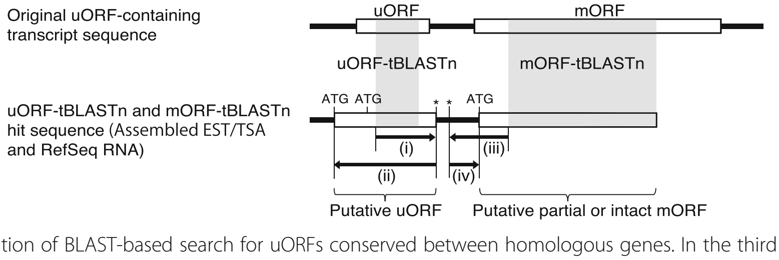
Original uORF-containing transcript sequence
uORF
mORF
uORF-tBLASTn
mORF-tBLASTn
uORF-tBLASTn and mORF-tBLASTn hit sequence (Assembled EST/TSA and RefSeq RNA)
ATG ATG
* *
ATG
(i)
(iii)
(ii)
(iv)
Putative uORF
Putative partial or intact mORF

**Fig. 3** Schematic representation of BLAST-based search for uORFs conserved between homologous genes. In the third step of ESUCA, tBLASTn searches are conducted against a transcript sequence database that consists of assembled EST/TSA contigs, unclustered singleton EST/TSA sequences and RefSeq RNAs, using original uORF sequences as queries (uORF-tBLASTn). The shaded regions in the open boxes show the tBLASTn-matching regions. Asterisks represent stop codons. (i) The downstream in-frame stop codon closest to the 5′-end of the matching region of each uORF-tBLASTn hit is selected. (ii) The 5′-most in-frame ATG codon located upstream of the stop codon is selected. The ORF beginning with the selected ATG codon and ending with the selected stop codon is extracted as a putative uORF. In the fourth step of ESUCA, the downstream sequences of putative uORFs in the transcript sequences are subjected to mORF-tBLASTn analysis. Transcript sequences matching the original mORF with an *E*-value less than $10^{-1}$ are extracted. (iii) For each of the uORF-tBLASTn and mORF-tBLASTn hits, the upstream in-frame stop codon closest to the 5′-end of the matching region is selected. (iv) The 5′-most in-frame ATG codon located downstream of the selected stop codon is identified as the initiation codon of the putative partial or intact mORF. If the putative mORF overlaps with the putative uORF, the uORF-tBLASTn and mORF-tBLASTn hit is discarded as a uORF-mORF fusion type

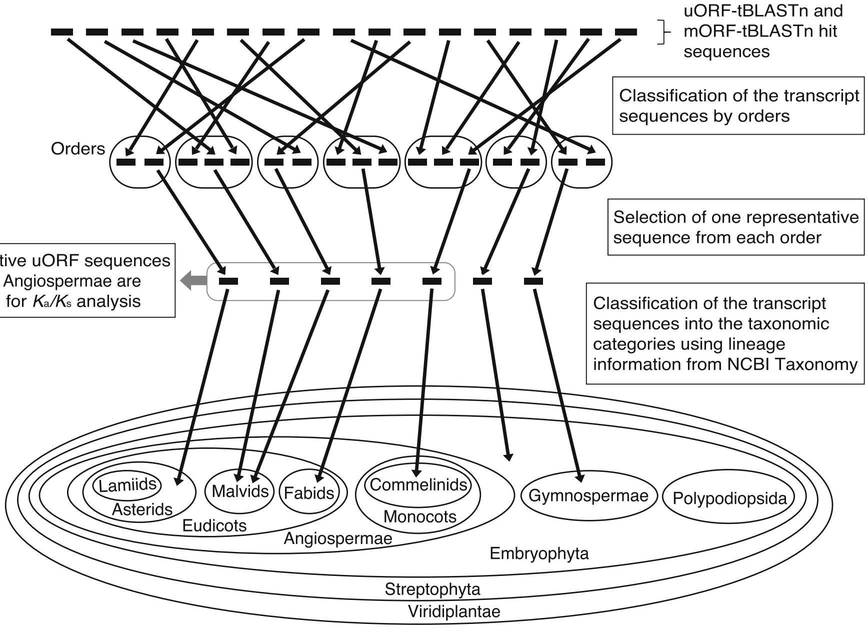
uORF-tBLASTn and mORF-tBLASTn hit sequences
Classification of the transcript sequences by orders
Orders
Selection of one representative sequence from each order
Putative uORF sequences from Angiospermae are used for $K_a/K_s$ analysis
Classification of the transcript sequences into the taxonomic categories using lineage information from NCBI Taxonomy
Lamiids
Asterids
Malvids
Fabids
Commelinids
Monocots
Eudicots
Angiospermae
Gymnospermae
Polypodiopsida
Embryophyta
Streptophyta
Viridiplantae

**Fig. 4** Schematic representation of the algorithms to select putative uORF sequences used for $K_a/K_s$ analysis and to determine the taxonomic range of uORF sequence conservation. Horizontal short black bars depict uORF-tBLASTn and mORF-tBLASTn hit sequences selected in the fourth step of ESUCA. In the fifth step, the uORF-tBLASTn and mORF-tBLASTn hit sequences are classified by orders, using taxonomic lineage information of EST, TSA, and RefSeq RNA sequences from NCBI Taxonomy, and one sequence is selected from each order (See Materials and Methods for the criteria for the selection). The putative uORF sequences in the selected transcript sequences are used for generating the multiple alignments of the uORF amino acid sequences. For $K_a/K_s$ analysis, only putative uORF sequences from orders belonging to Angiospermae are used. In the sixth step of ESUCA, the selected transcript sequences are classified into the 13 plant taxonomic categories to determine the taxonomic range of uORF sequence conservation, using taxonomic lineage information of EST, TSA, and RefSeq RNA sequences

alignments of each selected uORF and its homologs, using a putative homologous uORF sequence from each order in which uORF-tBLASTn and mORF-tBLASTn hits were found (Fig. 4). When multiple original uORFs derived from splice variants of the same gene partially or completely shared amino acid sequences, the one with the longest conserved region was manually selected on the basis of the uORF amino acid sequence alignments. In the fifth step, the remaining uORFs were subjected to $K_a/K_s$ analysis. The uORFs with $K_a/K_s$ ratios less than 0.5 showing significant differences from those of negative controls ($q <$ 0.05) were selected as candidate CPuORFs (Supplementary Table S1). Through ESUCA analyses of *Arabidopsis*, rice, tomato, poplar, and grape genomes, 105, 57, 42, 149, and 78 candidate CPuORFs were extracted, respectively. Of these, 87 *Arabidopsis*, 51 rice, 29 tomato, 76 poplar, and 43 grape uORFs belong to the previously identified CPuORF families, homology groups (HGs) 1 to 53 [21, 23–26] (Supplementary Table S1). The amino acid sequences of the remaining candidate CPuORFs are not similar to those of the known CPuORFs. Therefore, 18, 6, 13, 73, and 35 novel candidate CPuORFs were extracted from *Arabidopsis*, rice, tomato, poplar, and grape genomes, respectively.

## Validation of candidate CPuORFs

If the amino acid sequence of a uORF is evolutionarily conserved because of functional constraints of the uORF-encoded peptide, it is expected that the amino acid sequence in the functionally important region of the peptide is conserved among the uORF and its orthologous uORFs. Therefore, we manually checked whether the amino acid sequences in the same region are conserved among uORF sequences in the alignment of each novel candidate CPuORF. We found that the alignments of 17 novel candidate CPuORFs contain sequences that do not share the consensus amino acid sequence in the conserved region, and removed these sequences from the alignments. We also removed sequences derived from genes not related to the corresponding original uORF-containing gene from the alignments of five novel candidate CPuORFs. When these changes resulted in the number of orders with the uORF-tBLASTn and mORF-tBLASTn hits becoming less than two, the candidate CPuORFs were discarded. Ten novel candidate CPuORFs were discarded for this reason. Supplementary Figure S1 shows the uORF amino acid sequence alignments without the removed sequences. The $K_a/K_s$ ratios were recalculated after the manual removal of the sequences (Supplementary Table S1), and eight additional novel candidate CPuORFs were discarded because their $K_a/K_s$ ratios were greater than 0.5.

Using genomic position information from Ensembl Plants (http://plants.ensembl.org/index.html) [32] and Phytozome v12.1 (https://phytozome.jgi.doe.gov/pz/portal.html) [33], we manually checked whether the positions of the remaining novel candidate CPuORFs overlap with those of the mORFs of other genes or the mORFs of splice variants of the same genes. We found that the genomic position of the candidate CPuORF of the *Arabidopsis ROA1* (AT1G60200) gene overlaps with that of an intron in the mORF region of a splice variant. Protein sequences with an N-terminal region similar to the amino acid sequence encoded by the 5′-extended region of the mORF in this splice variant are found in most orders from which the uORF-tBLASTn and mORF-tBLASTn hits of this candidate CPuORF were extracted, suggesting that the splice variant with the 5′-extended mORF is not a minor form among orthologous transcripts. Therefore, this candidate CPuORF was discarded.

In the second step of ESUCA, we excluded uORF sequences likely to encode parts of the mORF-encoded proteins, by removing uORFs with high uORF-mORF fusion ratios. To confirm that the novel candidate CPuORFs do not code for parts of the mORF-encoded proteins, each of the putative uORF sequences used for the alignment and $K_a/K_s$ analysis was queried against the UniProt protein database (https://www.uniprot.org/), using BLASTx. When putative uORF sequences matched protein sequences with low *E*-values, we manually checked whether amino acid sequences similar to those encoded by the putative uORFs were contained within mORF-encoded protein sequences. In this analysis, mORF-encoded proteins with N-terminal sequences similar to the amino acid sequences encoded by the candidate CPuORFs of the rice *OsUAM2* gene and its poplar ortholog, POPTR_0019s07850, were identified in many orders. This suggests that the sequences encoded by these candidate CPuORFs are likely to function as parts of the mORF-encoded proteins. Therefore, we discarded these candidate CPuORFs. For some other novel candidate CPuORFs, mORF-encoded proteins with sequences similar to those encoded by the candidate CPuORF and/or its homologous putative uORFs were also found. However, we did not exclude these candidate CPuORFs, because such uORF-mORF fusion type proteins were found in only a few species for each candidate.

After manual validation, 13, 4, 11, 70, and 34 uORFs were identified as novel CPuORFs in *Arabidopsis*, rice, tomato, poplar and grape, respectively. Among these novel CPuORFs, those of orthologous genes with similar CPuORF amino acid sequences were classified into the same HGs. It should be noted that no apparent sequence similarity was found between the novel CPuORFs of non-orthologous genes. Also, using OrthoFinder ver. 1.1.4 [34], an algorithm for ortholog group inference, we classified the genes with novel CPuORFs and those with previously identified CPuORFs into ortholog groups. The same HG number with a different sub-number was assigned to CPuORFs of genes in the same ortholog group with dissimilar uORF sequences (e.g. HG56.1 and HG56.2). Of the newly identified CPuORF genes, six were classified into the same ortholog groups as previously identified CPuORF genes, but the amino acid sequences of these six CPuORFs are dissimilar to those of the known CPuORFs. Including this type of CPuORFs, we identified 132 novel CPuORFs that belong to 89 novel HGs (HG2.2, HG9.2, HG16.2, HG43.2, HG50.2, HG52.2, HG54-HG83, HG86-HG130 and HG149–151) (Supplementary Table S1).

### Determination of the taxonomic range of CPuORF sequence conservation

As the final step of ESUCA, we determined the taxonomic range of the sequence conservation of each CPuORF identified, including previously identified CPuORFs. For this purpose, the uORF-tBLASTn and mORF-tBLASTn hits selected for generating the multiple amino acid sequence alignments and retained after manual validation were classified into 13 plant taxonomic categories (See Materials and Methods for details.), on the basis of taxonomic lineage information of EST, TSA, and RefSeq RNA sequences (Fig. 4). Figure 5 and Supplementary Table S2 show the taxonomic range of sequence conservation for each HG and each CPuORF, respectively. In general, CPuORFs belonging to previously identified HGs tend to be conserved in a wider range of taxonomic categories than those belonging to the newly identified HGs. For 19 of the novel HGs, CPuORF sequences are conserved both in eudicots and monocots or in wider taxonomic ranges. In contrast, for 70 of the novel HGs, CPuORF sequences are conserved only among eudicots. For 12 of these, CPuORF sequences are conserved in narrower taxonomic ranges, only among rosids or asterids. These results indicate that the taxonomic range of CPuORF sequence conservation varies, and that ESUCA can identify CPuORFs conserved in a relatively narrow taxonomic range.

### Sequence-dependent effects of CPuORFs on mORF translation

To address the relationship between the taxonomic range of CPuORF sequence conservation and the sequence-dependent effects of CPuORFs on mORF translation, we selected 11 poplar CPuORFs and examined their sequence-dependent effects on expression of the downstream reporter gene using a transient expression assay. Of the selected CPuORFs, those belonging to HG46, HG55, HG57, HG66 and HG103 are conserved in diverse angiosperms or in wider taxonomic ranges

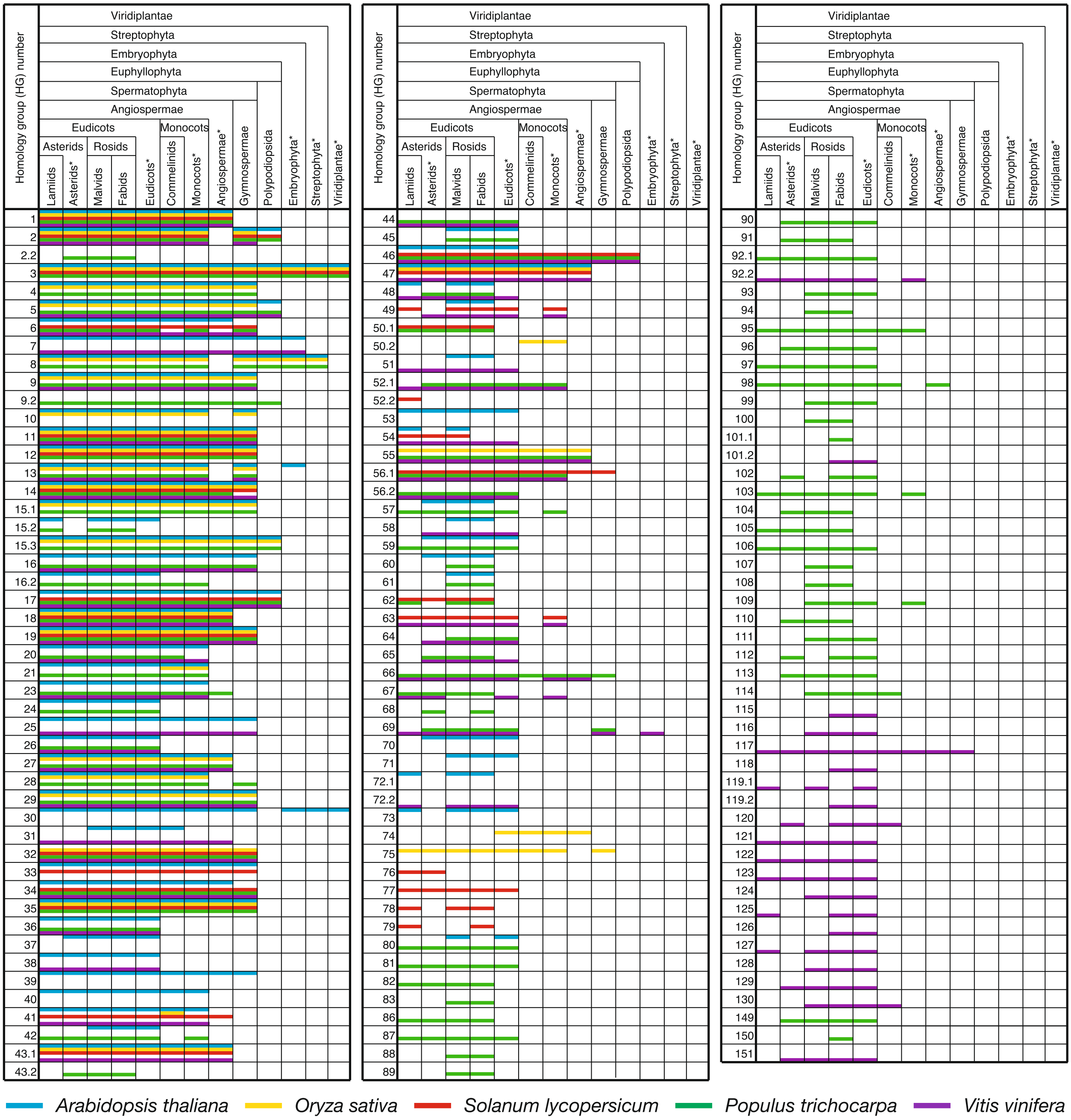

**Fig. 5** Taxonomic range of the sequence conservation of the CPuORF families. The blue, yellow, red, green, and purple lines show the conservation range of CPuORF HGs determined by applying ESUCA to *Arabidopsis* (*Arabidopsis thaliana*), rice (*Oryza sativa*), tomato (*Solanum lycopersicum*), poplar (*Populus trichocarpa*), and grape (*Vitis vinifera*) genomes, respectively. The presence of a line within a cell in each taxonomic category indicates the presence of uORF-tBLASTn and mORF-tBLASTn hits for any of the CPuORFs that belong to each HG. In taxonomic categories with a category name with an asterisk, uORF-tBLASTn and mORF-tBLASTn hits found in lower taxonomic categories were excluded. In the case where no uORF-tBLASTn and mORF-tBLASTn hit was found in the taxonomic category that contain a species from which the original uORF was derived, the line showing the species was still drawn in the cell of the taxonomic category because this category contained the species with the original uORF. *Arabidopsis*, rice, tomato, poplar and grape belong to malvids, commelinids, lamiids, fabids and eudicots*, respectively. HG1-HG53 are previously identified HGs, except for HG2.2, HG9.2, HG16.2, HG43.2, HG50.2 and HG52.2, whereas HG2.2, HG9.2, HG16.2, HG43.2, HG50.2, HG52.2, HG54-HG83, HG86-HG130 and HG149–151 are newly identified HGs

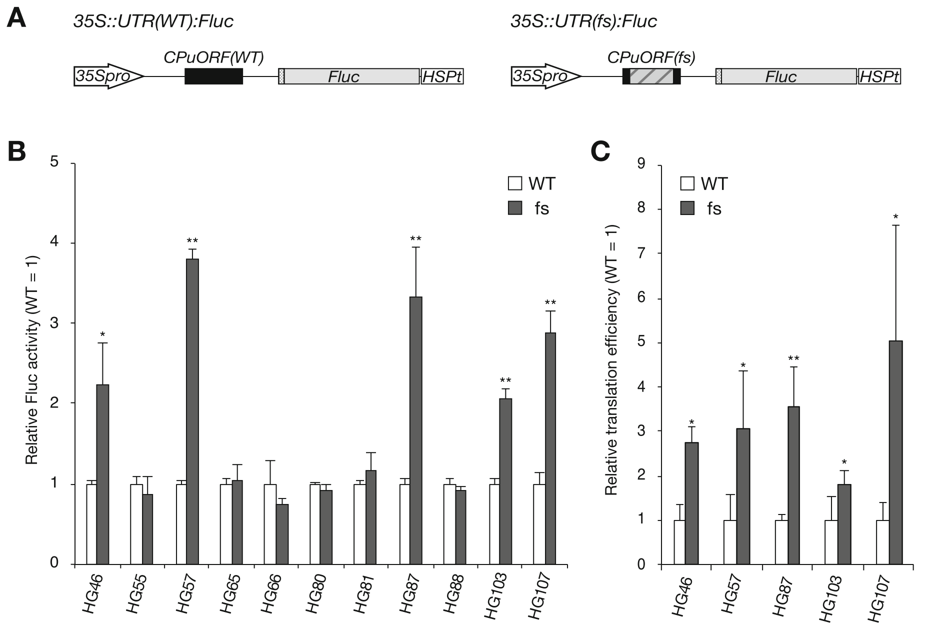

**Fig. 6** Sequence-dependent effects of novel CPuORFs on main ORF translation. **a** Schematic representation of the WT (*35S::UTR (WT):Fluc*) and frameshift (fs) mutant (*35S::UTR (fs):Fluc*) reporter constructs. The 5′-UTR containing each CPuORF tested was inserted between the 35S promoter (*35Spro*) and the *Fluc* coding sequence. The hatched box in the CPuORF (fs) indicates the frame-shifted region. The dotted boxes represent the first five nucleotides of the mORF. See Supplementary Figure S2 for the exact position and length of each CPuORF and the exact frame-shifted region. HSPt: the *AtHSP18.2* polyadenylation signal. **b** Transient expression assay for luciferase activity. Each reporter plasmid containing a WT or fs CPuORF, belonging to an indicated HG, was co-transfected into MM2d protoplasts with the *35S::Rluc* internal control plasmid by PEG treatment. After 24-h incubation, dual luciferase assays were performed. Fluc activity was normalized to Rluc activity, and the normalized activity relative to that of the corresponding WT reporter construct is shown. **c** Transient expression assay for translation efficiency. For the five CPuORFs that showed a significant sequence-dependent effect in (**b**), the WT and fs reporter plasmids were each co-transfected into MM2d protoplasts with the *35S::Rluc* plasmid again. After 24-h incubation, the Fluc and Rluc activities and the *Fluc* and *Rluc* mRNA levels were measured. Fluc activity was normalized to Rluc activity, and the *Fluc* mRNA level was normalized to the *Rluc* mRNA level. The normalized Fluc activity was divided by the normalized *Fluc* mRNA level to calculate the translation efficiency of *Fluc*. The *Fluc* translation efficiency relative to that of the corresponding WT reporter construct was calculated to determine the relative translation efficiency. In (**b**) and (**c**), means ± SD of at least three biological replicates are shown. Single and double asterisks indicate significant differences between the WT and fs constructs at $p < 0.05$ and $p < 0.01$, respectively, as determined by Student's *t*-test

(Fig. 5, Supplementary Table S2). The CPuORFs belonging to HG65, HG80, HG81 and HG87 are conserved in a wide range of eudicots, whereas the CPuORFs belonging to HG88 and HG107 are conserved only among rosids (Fig. 5, Supplementary Table S2). In the 5′-UTR of the poplar gene with the HG107 CPuORF, there is another uORF immediately upstream of the CPuORF (Supplementary Figure S2K). To focus on the sequence-dependent effect of the CPuORF on mORF translation, the upstream uORF was eliminated by mutating its initiation codon because the presence of the immediate upstream uORF may reduce the translation efficiency of the CPuORF and therefore potentially make the effect of the CPuORF ambiguous. The 5′-UTR sequences containing the selected CPuORFs were fused to the firefly luciferase (Fluc) coding sequence and were placed under the control of the 35S promoter to generate the wild-type (WT) reporter constructs (Fig. 6a, Supplementary Figure S2). To assess the importance of the amino acid sequences for 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 (Supplementary Figure S2). These reporter constructs were each transfected into protoplasts from *Arabidopsis thaliana* MM2d suspension-cultured cells. After 24 h of incubation, cells were harvested and disrupted to analyze luciferase activity. In five of the 11 CPuORFs, the introduced frameshift mutations significantly increased Fluc activity, indicating that these CPuORFs repress expression of the *Fluc* reporter gene expression in a sequence-dependent manner (Fig. 6b). As mentioned in the Background section, sequence-dependent regulatory uORFs may cause not only translational repression but also NMD, via ribosome stalling at the uORF stop codons [12, 13]. Therefore, the sequence-dependent repressive effects observed for the five CPuOFs could be due to NMD rather than translational repression. To confirm that the sequence-dependent effects of these CPuORFs on the repoerter gene expression were exerted at the translational level, we examined the mRNA level of the *Fluc* reporter gene by quantitative reverse transcription PCR and devided the Fluc activity by the *Fluc* mRNA level to calculate the relative translation efficiency of the *Fluc* reporter gene. As shown in Fig. 6c, for all the five CPuORFs tested, the frameshift mutations significantly enhanced the relative translation efficiency. This result suggests that these five CPuORFs cause translational repression of the mORFs in a sequence-dependent manner, although we cannot rule out the possibility that these CPuORF sequences induce NMD in addition to the translational repression. These five novel sequence-dependent regulatory CPuORFs include the HG107 CPuORF, which is one of the CPuORFs conserved only among rosids. Therefore, this result suggests that CPuORFs conserved only among rosids can have sequence-dependent regulatory effects.

## Discussion

### Comprehensive identification of angiosperm CPuORFs by ESUCA analyses of multiple species' genomes

In this study, we developed ESUCA, a pipeline for efficient genome-wide identification of CPuORFs. By applying ESUCA to five angiosperm genomes, we identified 132 CPuORFs that belong to 89 novel HGs. Of these HGs, 71 were identified through ESUCA analysis of only one of the *Arabidopsis*, rice, tomato, poplar or grape genomes (Fig. 5). This means that CPuORFs belonging to these HGs cannot be identified by comparing uORF sequences of orthologous genes between these five species. Therefore, this result demonstrates that the approach used in this study, in which uORF sequences from multiple species are compared with those of many other species, is highly effective in comprehensively identifying CPuORFs. We expected this approach to be particularly useful for comprehensive identification of CPuORFs conserved in relatively narrow taxonomic ranges. Therefore, we used the five angiosperm species belonging to relatively distant lineages in order to identify CPuORFs conserved only among the taxonomic categories to which each of the five species belong. However, unexpectedly, in 43 of 71 HGs identified through ESUCA analysis of only one of the five plant genomes, CPuORF sequences are conserved in both rosids and asterids, two major groups of eudicots, indicating that these CPuORFs are conserved across diverse eudicots (Fig. 5). One possible explanation of this observation is that sequence conservation of uORFs is often lost in small taxonomic groups, such as orders or families, during evolution. For example, while the CPuORF of the tomato LOC101264451 gene, which belongs to HG43.1, exerts a sequence-dependent repressive effect on mORF translation, the CPuORF of its *Arabidopsis* ortholog, *ANAC096*, lacks the C-terminal half of the amino acid sequence in the highly conserved region and does not have a sequence-dependent regulatory effect [26–28]. In contrast to the *Arabidopsis ANAC096* CPuORF, all the critical amino acid residues in the highly conserved region are retained in the HG43.1 CPuORF of *Tarenaya hassleriana*, which belongs to the same order as *Arabidopsis*, Brassicales, but a different family. CPuORFs involved in preferable but not essential post-transcriptional regulations may be lost in some taxonomic clades during evolution. Such CPuORFs would be difficult to be identified by comparing uORF sequences between a few selected species if a selected species is included in clades where the CPuORF sequences were lost. Therefore, the ESUCA method is advantageous for comprehensive identification of CPuORFs compared with conventional comparative genomic approaches. Our results reveal that the approach used here (i.e. ESUCA analyses of multiple species' genomes) is further advantageous and is highly useful not only for comprehensive identification of narrowly conserved CPuORFs but also for that of widely conserved CPuORFs.

The transient expression assays in the current study identified five novel poplar regulatory CPuORFs that exert sequence-dependent repressive effects on mORF translation. One of the identified regulatory CPuORFs is conserved only among rosids (i.e. only among fabids and malvids). This result suggests that at least, CPuORFs conserved only among fabids and malvids can have sequence-dependent regulatory effects, although we cannot rule out the possibility that CPuORFs conserved in narrower taxonomic ranges can have sequence-dependent regulatory effects. Of the 91 CPuORF HGs identified through ESUCA analysis of the poplar genome, HG101.1 and HG150 are conserved only among fabids, whereas the others are conserved in at least one other taxonomic category in addition to fabids (Fig. 5). Likewise, all the HGs identified through ESUCA analysis of the other four plant genomes are conserved in multiple taxonomic categories. Altogether, these results suggest that most the CPuORFs identified in this study, in which CPuORFs conserved in more than two orders were extracted, are likely to be conserved because of functional constraints of their encoded peptides. Of the 11 poplar CPuORFs analyzed by the transient expression assays, five are conserved beyond eudicots, and three of them exhibited sequence-dependent repressive effects (Figs. 5 and 6b). In addition, we previously examined the effects of 16 CPuORFs, which belong to HG27-HG29, HG33-HG43.1, HG44, and HG45, on mORF translation and identified six sequence-dependent regulatory CPuORFs. In this analysis, five of 11 CPuORFs conserved beyond eudicots showed sequence-dependent repressive effects, whereas one of five CPuORFs conserved only among eudicots showed a sequence-dependent repressive effect [27, 28].

These results may suggest that CPuORFs conserved beyond eudicots are more likely to encode functional peptides than CPuORFs conserved only among eudicots. ESUCA is capable of selecting CPuORFs conserved in certain taxonomic ranges on the basis of two criteria, the numbers of orders in which CPuORFs are conserved and/or taxonomic categories in which CPuORFs are conserved. Our results demonstrate that this function of ESUCA is highly useful for the efficient selection of CPuORFs likely to encode functional peptides.

Of the CPuORFs analyzed for their sequence-dependent regulatory effects in the present study, the CPuORFs belonging to HG55 and HG66 showed no significant sequence-dependent effect on mORF translation, despite their widespread sequence conservation beyond eudicots (Figs. 5 and 6b). These CPuORFs might encode peptides that have functions other than the control of mORF translation, or they might exert sequence-dependent regulatory effects only under certain conditions. In fact, many known sequence-dependent regulatory uORFs repress mORF translation in response to metabolites, such as polyamine, arginine, and sucrose [11, 15, 16, 18]. Likewise, the other CPuORFs that exhibited no significant sequence-dependent regulatory effect might encode peptides that have other functions or exert regulatory effects only under specific conditions.

### Filtering using the uORF-mORF fusion ratio

To distinguish between 'spurious' CPuORFs conserved because they code for parts of mORF-encoded proteins and 'true' CPuORFs conserved because of functional constraints of their encoded small peptides, we employed the criterion of the uORF-mORF fusion ratio and discarded uORFs with uORF-mORF fusion ratios equal to or greater than 0.3. We checked how effectively the 'spurious' CPuORFs were removed with this criterion, using a protein sequence database. Although uORF-mORF fusion type protein sequences were found in OsUAM2 homologs from widespread angiosperm species, no other candidate CPuORFs were extracted in which uORF-mORF fusion type protein sequences were found in many species that have sequences similar to the candidate CPuORFs. This indicates that the uORF-mORF fusion ratio filtering worked effectively to exclude 'spurious' CPuORFs that code for parts of the mORF-encoded proteins.

ESUCA extracted all known plant cis-acting regulatory CPuORFs that control mORF translation in a sequence-dependent manner (i.e. HG1, HG3, HG6, HG12, HG13, HG14, HG15.3, HG18, HG19, HG27, HG34, HG36, HG41, HG42, and HG43.1 CPuORFs [16, 18, 27, 28, 35–40]) (Supplementary Table S1). Of these, the HG3 CPuORFs, associated with an mORF coding for an *S*-adenosylmethionine decarboxylase (AdoMetDC; EC 4.1.1.50) [21, 41], showed relatively high uORF-mORF fusion ratios, although still below 0.3 (Supplementary Table S1). In this study, 11 CPuORFs belonging to this HG were extracted using ESUCA and the uORF-mORF fusion ratios of these CPuORFs were in the range of 0.17 to 0.27, with a median value of 0.26 (Supplementary Table S1). Consistent with these relatively high uORF-mORF fusion ratios, uORF-mORF fusion type protein sequences that contain amino acid sequences resembling HG3 CPuORFs are found in widespread angiosperm species. The uORF-mORF fusion ratios of the candidate CPuORFs of rice *OsUAM2* and its poplar ortholog were 0.28. The rice *OsUAM2* gene codes for a protein similar to UDP-arabinopyranose mutases (EC 5.4.99.30) [42]. Two protein isoforms, a uORF-mORF fusion type and one lacking the uORF-encoded region, produced from splice variants of an *OsUAM2* orthologous gene, are found in diverse angiosperm species. The functions of both of isoforms are not yet known. Considering the example of HG3 CPuORFs, we cannot rule out the possibility that a candidate CPuORF functions as a regulatory uORF even if uORF-mORF fusion type protein sequences are found in widespread species. However, to avoid including potential 'spurious' CPuORFs whose amino acid sequences are likely to be evolutionarily conserved because of their function as N-terminal regions of the mORF-encoded proteins, we excluded the candidate CPuORFs of the rice *OsUAM2* gene and its poplar ortholog. In conclusion, the criterion of the uORF-mORF fusion ratio used in this study appears appropriate because all known cis-acting sequence-dependent regulatory CPuORFs were extracted and most 'spurious' CPuORFs were removed.

## Conclusion

The present study demonstrates that the approach in which uORF sequences from multiple species are compared with those of many other species, using ESUCA, is highly effective in comprehensive identification of CPuORFs. Using this approach, we identified many novel angiosperm CPuORFs, which include CPuORFs conserved among limited clades and widely conserved CPuORFs. Our results also showed that ESUCA is capable of efficiently selecting CPuORFs likely to be conserved because of functional importance of their encoded peptides. The approach used here can be applied to any eukaryotic organism with available genome and transcript sequence databases and therefore is expected to contribute to the comprehensive identification of CPuORFs encoding functional peptides in various organisms. Furthermore, besides CPuORFs, the algorithms developed for ESUCA and the approach used here can be applied to the identification of other sequences conserved in various taxonomic ranges.

## Materials and methods

### Extraction of uORF sequences

We used genome sequence files in FASTA format and genomic coordinate files in GFF3 format obtained from Ensembl Plants Release 33 (https://plants.ensembl.org/index.html) [32] to extract *Arabidopsis* (*Arabidopsis thaliana*), tomato (*Solanum lycopersicum*), poplar (*Populus trichocarpa*), and grape (*Vitis vinifera*) uORF sequences. We used a genome sequence file in FASTA format and a genomic coordinate files in GFF3 format obtained from Phytozome v11 (https://phytozome.jgi.doe.gov/pz/portal.html) [33] for rice (*Oryza sativa*). 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 sequences from the transcript sequence datasets. Then, we searched the 5′-UTR sequences for an ATG codon and its nearest downstream in-frame stop codon. Sequences starting with an ATG codon and ending with the nearest in-frame stop codon were extracted as uORF sequences. When multiple uORFs from a gene shared the same stop codon, only the longest uORF sequence was used for further analyses.

### Assembly of EST and TSA sequences

EST, TSA, and RefSeq RNA sequence datasets were obtained from the International Nucleotide Sequence Database Collaboration databases (NCBI release of 2016-12-03 and DDBJ release 106.0 for EST and TSA sequences, NCBI release 79 for RefSeq RNA sequences). On the basis of taxonomic lineage information provided by NCBI Taxonomy (https://www.ncbi.nlm.nih.gov/taxonomy), EST and TSA sequences derived from Viridiplantae were extracted from the databases. EST and TSA sequences from the same species were assembled using Velvet ver. 1.2.10 [43] with k-mer length of 99. The k-mer length was optimized using *A. thaliana* EST and TSA sequences to minimize the total numbers of assembled contigs and unclustered singleton sequences. Unclustered singleton EST/TSA sequences, derived from species for which RefSeq RNA sequences were available, were mapped to the RefSeq RNA sequences using Bowtie2 ver. 2.2.9 [44] and the default parameters. We discarded singleton EST/TSA sequences that matched any RefSeq RNA sequences from the same species. We created a plant transcript sequence database for BLAST searches in our local computers by using the remaining singleton EST/TSA sequences, assembled contigs, Viridiplantae RefSeq RNA sequences, and makeblastdb, a program contained in the NCBI-BLAST package.

### Calculation of the uORF-mORF fusion ratio

The uORF-mORF fusion ratio for each of the extracted uORFs was assessed as follows. We performed tBLASTx using each uORF sequence as a query against plant (Viridiplantae) RefSeq RNA sequences with an *E*-value cutoff of 2000 (uORF-tBLASTx). We used standalone NCBI-BLAST+ ver. 2.6.0 [45] for all BLAST analyses. We next performed tBLASTx with an *E*-value threshold of $10^{-1}$ using the mORF sequence associated with each uORF as a query against the uORF-tBLASTx hit sequences (mORF-tBLASTx). Using these two-step tBLASTx searches, we selected RefSeq RNAs that contain both sequences similar to the original uORF and its downstream mORF. Then, we examined whether the largest ORF of each of the selected RefSeq RNAs included the region that matched the original mORF in the same reading frame (Fig. 2). We also examined whether the largest ORF included the region that matched the original uORF, irrespective of the reading frame. The RefSeq RNA was considered to have a uORF-mORF fusion if the largest ORF contained both regions that matched the original uORF and mORF (Fig. 2). The RefSeq RNA was considered to have a uORF separated from the downstream mORF if the largest ORF contained the region that matched the original mORF but not the region that matched the original uORF (Fig. 2). RefSeq RNA numbers of the former and latter types were defined as *X* and *Y*, respectively. We calculated a uORF-mORF fusion ratio as $X\ /\ (X+Y)$ for each of the original uORF-containing transcript sequences.

### BLAST-based search for uORFs conserved between homologous genes

To search for uORFs with amino acid sequences conserved between homologous genes, we first performed tBLASTn searches against the assembled plant transcript sequence database, using the amino acid sequences of the uORFs as queries. In these uORF-tBLASTn searches, we extracted transcript sequences that matched a uORF with an *E*-value less than 2000 and derived from species other than that of the original uORF. The downstream in-frame stop codon closest to the 5′-end of the matching region of each uORF-tBLASTn hit was selected (Fig. 3). Then, we looked for an in-frame ATG codon upstream of the selected stop codon, without any other in-frame stop codon between them. uORF-tBLASTn hits without such an ATG codon were discarded. If one or more in-frame ATG codons were identified, the 5′-most ATG codon was selected. The ORF beginning with the selected ATG codon and ending with the selected stop codon was extracted as a putative uORF (Fig. 3). The downstream sequences of putative uORFs were subjected to another tBLASTn analysis to examine whether the transcripts were derived from homologs of the original uORF-containing gene. In this analysis, the amino acid sequence of the mORF associated with the original uORF was used as a query sequence, and transcript sequences matching the mORF with an *E*-value less than $10^{-1}$ were extracted. For each of the uORF-tBLASTn and mORF-tBLASTn hits, the upstream in-frame stop codon closest to the 5′-end of the region matching the original mORF was selected, and the 5′-most in-frame ATG codon located downstream of the selected stop codon was identified as the putative mORF initiation codon (Fig. 3). uORF-tBLASTn and mORF-tBLASTn hits were discarded as uORF-mORF fusion type sequences if the putative mORF overlapped with the putative uORF. The original uORF was selected as a candidate CPuORF if the remaining uORF-tBLASTn and mORF-tBLASTn hits belonged to at least two orders other than that from which the original uORF was derived.

### Identification of contaminant ESTs and TSAs

In the uORF-tBLASTn and mORF-tBLASTn analyses described above, we excluded tBLASTn hit ESTs and TSAs derived from contaminated organisms. To examine whether each uORF-tBLASTn and mORF-tBLASTn hit sequence is derived from contaminated organisms, we performed BLASTn searches against EST, TSA, and RefSeq RNA sequences from all organisms except for those of metagenomes, using each uORF-tBLASTn and mORF-tBLASTn hit EST/TSA sequence as a query. If a uORF-tBLASTn hit EST/TSA sequence matched an EST, TSA, or RefSeq RNA sequence of a different order from the species of the uORF-tBLASTn and mORF-tBLASTn hit, with an *E*-value less than $10^{-100}$ and an identity equal to or greater than 95%, it was considered a candidate contaminant sequence. In this case, either the uORF-tBLASTn and mORF-tBLASTn hit or the BLASTn hit may be a contaminant sequence. To distinguish these possibilities, we compared the ratio of the BLASTn hit number to the total EST/TSA and RefSeq RNA sequence number between the species of each uORF-tBLASTn and mORF-tBLASTn hit and the species of its BLASTn hits. Appropriate comparisons are difficult unless species used for this comparison have enough number of EST/TSA and RefSeq RNA sequences. Therefore, if the total EST/TSA and RefSeq RNA sequence number of a species is less than 5000, BLASTn hits derived from the species were not used for this analysis. If the ratio of the BLASTn hit number to the total EST/TSA and RefSeq RNA sequence number of a uORF-tBLASTn and mORF-tBLASTn hit species is less than that of any other BLASTn hit species, the uORF-tBLASTn and mORF-tBLASTn hit sequence was identified as a contaminant sequence.

### $K_a/K_s$ analysis

For $K_a/K_s$ analysis of each candidate CPuORF, one putative uORF sequence was selected from each order in which uORF-tBLASTn and mORF-tBLASTn hits were found, using the following criteria. First, we selected transcript sequences that matched the mORF associated with the candidate CPuORF with an *E*-value less than $10^{-20}$ in mORF-tBLASTn analysis. Of these sequences, we then selected the one with the smallest geometric means of mORF-tBLASTn and uORF-tBLASTn *E*-values. When mORF-tBLASTn *E*-values of all uORF-tBLASTn and mORF-tBLASTn hits in an order were equal to or greater than $10^{-20}$, we selected the transcript sequence with the smallest geometric means of mORF-tBLASTn and uORF-tBLASTn *E*-values in the order. Putative uORF sequences in the selected transcript sequences were used for generating multiple uORF amino acid sequence alignments presented in Supplementary Figure S1 and Supplementary Table S3. Only putative uORF sequences selected from orders belonging to Angiospermae were used for $K_a/K_s$ analysis. Multiple alignments of the uORF amino acid sequences were generated by using standalone Clustal Omega (ClustalO) ver. 1.2.2 [46] with the default parameters. On the basis of the multiple uORF amino acid sequence alignments, codon-based multiple alignments (also referred to as codon-delimited multiple alignments) [47] of the uORF nucleotide sequences were generated (Supplementary Table S3). For each candidate CPuORF, a median $K_a/K_s$ ratio for all pairwise combinations of the original uORF and its homologous putative uORFs was calculated using the codon-based multiple alignment and the kaks function in the seqinR package (ver. 3.4.5) [48] with the parameter setting 'rmgap = FALSE'.

For statistical tests of $K_a/K_s$ ratios, we calculated the distribution of mutation rates between the original uORF and its homologous putative uORFs and those between the original uORF and its artificially generated mutants, using the observed mutation rate distribution. Then, observed empirical $K_a/K_s$ ratio distributions were compared with null distributions (negative controls) using the Mann-Whitney *U* test to validate statistical significance. The one-sided *U* test was used to investigate whether the observed distributions were significantly lower than the null distributions. Adjustment for multiple comparisons was achieved by controlling the false discovery rate using the Benjamini and Hochberg procedure [49].

### Determination of the taxonomic range of uORF sequence conservation

To automatically determine the taxonomic range of the sequence conservation of each CPuORF, we first defined 13 plant taxonomic categories. The 13 defined taxonomic categories are lamiids, asterids other than lamiids, mavids, fabids, eudicots other than rosids and asterids, commelinids, monocots other than commelinids, Angiospermae other than eudicots and monocots, Gymnospermae, Polypodiopsida, Embryophyta other than Euphyllophyta, Streptophyta other than Embryophyta, and Viridiplantae other than Streptophyta. On the basis of 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 generating the multiple uORF amino acid sequence alignments were classified into the 13 taxonomic categories (Fig. 4). It should be noted that, in NCBI Taxonomy, eudicots, Angiospermae, and Ggymnospermae are referred to as eudicotyledons, Magnoliophyta, and Acrogymnospermae, respectively. For each CPuORF, the numbers of transcript sequences classified into each category were counted and shown in Supplementary Table S2. These numbers represent the numbers of orders in which the amino acid sequence of each CPuORF is conserved.

### Statistical and informatic analyses

All programs, except for existing stand-alone programs, such as BLAST [45], ClustalO [46] and Jalview [50], were written in R (www.r-project.org). We also used R libraries, GenomicRanges ver. 1.32.7 [51], exactRankTests ver. 0.8.30, Biostrings ver. 2.48.0 and seqinr ver. 3.4.5 [48].

### Plasmid construction

Plasmid pNH006 harbors the cauliflower mosaic virus 35S RNA (35S) promoter, the Fluc coding sequence, and the polyadenylation signal of the *A. thaliana HSP18.2 (AtHSP 18.2)* gene in pUC19. To construct this plasmid, pMT61 [27] was digested with *Sal*I and *Sac*I, and the *Sal*I-*Sac*I fragment containing the Fluc coding sequence was ligated into the *Sal*I and *Sac*I sites between the 35S promoter and the *HSP18.2* polyadenylation signal of plasmid pKM56 [40]. To generate reporter plasmids pNH92-pNH101 (Supplementary Table S4) for transient expression assays, the 5′-UTR sequences of 10 poplar genes were amplified by PCR from poplar (*Populus nigra*) full-length cDNA clones pds25559, pds10965, pds14390, pds12940, pds13862, pds15817, pds28294, pds26157, pds14623 and pds23234 (Supplementary Table S5), obtained from RIKEN [52]. Primer sets used are shown in Supplementary Tables S4 and S6. To construct pNH92, pNH94-pNH98 and pNH100-pNH101, amplified fragments containing the 5′-UTR sequences were digested with *Xba*I and *Sal*I and ligated between the *Xba*I and *Sal*I sites of pNH006. To create pNH93 and pNH99, amplified fragments containing the 5′-UTR sequences were inserted between the *Xba*I and *Sal*I sites of pNH006 using the SLiCE method [53]. To make pHN102, the 5′-UTR sequence of the *Populus trichocarpa* POPTR_0013s08000 gene with a mutation at the initiation codon of the uORF located immediately upstream of the HG107 CPuORF was synthesized by Fasmac (Atsugi, Japan) on the basis of NCBI RefSeq accession no. XM_002319213.3. The synthesized 5′-UTR sequence was amplified by PCR using primers 35S_XbaI_SLiCE-F and FLUC_SalI_SLiCE-R (Supplementary Table S6) and were inserted between the *Xba*I and *Sal*I sites of pNH006 using the SLiCE method [53]. Frameshift mutations were introduced into each CPuORF using overlap extension PCR [54], with primers listed in Supplementary Tables S4 and S6, to yield pNH103-pNH112. Sequence analysis confirmed the integrity of the PCR-amplified regions of all constructs.

### Transient expression assay

Transient expression assays for measuring only luciferase activities were performed as described in Hayashi et al. 2017 [40]. Protoplasts from *A. thaliana* MM2d suspension cells [55] were used, as were the reporter plasmids described above and the pKM5 [40] internal control plasmid. pKM5 contains the 35S promoter, the *Renilla* luciferase (Rluc) coding sequence, and the *NOS* polyadenylation signal in pUC19. For each experiment, 5 μg each of a reporter plasmid and pKM5 were transfected into protoplasts.

Transient expression assays for measuring luciferase activities and the mRNA levels of the reporter genes were carried out with the following modifications. Reporter plasmid DNA (10 μg) and pKM5 DNA (10 μg) were mixed with $3.0 \times 10^5$ MM2d protoplasts in 100 μl of MaMg solution (5 mM morpholinoethanesulfonic acid, 15 mM $MgCl_2$, and 0.4 M mannitol, pH 5.8) and 120 μl of polyethylene glycol (PEG) solution (40% PEG4000, 0.1 mM $CaCl_2$ and 0.2 M mannitol). After 15-min incubation at room temperature, the mixture was diluted by adding 800 μl of wash buffer [0.4 M mannitol, 5 mM $CaCl_2$, and 0.5 M 2-(*N*-morpholino) ethanesulfonic acid, pH 5.8]. The protoplasts were centrifuged and resuspended in 400 μl of modified Linsmaier and Skoog medium [56] containing 0.4 M mannitol. The protoplasts were incubated for 24 h at 22 °C in the dark. For measurement of luciferase activities, 100 μl of cells were harvested and disrupted in 50 μl of extraction buffer [100 mM $(NaH_2/Na_2H)PO_4$ and 5 mM DTT, pH 7] by vortexing. A Dual-Luciferase Reporter Assay kit (Promega) was used to measure the Fluc and Rluc activities. For RNA analysis, 300 μl of cells were harvested and frozen in liquid nitrogen. Total RNAs were extracted using TRIzol (Thermo Fisher Scientific), following the manufacturer's protocol. To remove DNA, 500 ng of total RNAs were treated with 1 unit of RQ1 RNase-Free DNase (Promega) for 30 min at 37 °C. DNase was inactivated by adding 1 μl of RQ1 DNase stop buffer (Promega). The DNase-treated RNAs were subsequently reverse transcribed using SuperScript III Reverse Transcriptase (Thermo Fisher Scientific) and an oligo (dT) primer, according to the manufacturer's instructions, and the synthesized cDNAs were used as templates for quantitative real-time PCR. Quantitative real-time PCR was performed on a LightCycler 480 System II (Roche Applied Science) using a LightCycler 480 SYBR Green I Master kit (Roche Applied Science), following the manufacturer's instructions, except that the extension step of each PCR cycle was performed for 30 s. The *Fluc* and *Rluc* mRNA levels were measured using a primer set FLUCqPCRf (5′-GTCGATGTACACGTTCGTCA-3′) and FLUCqPCRr (5′-GACACCTTTAGGCAGACCA-3′) and a primer set RLUCqPCRf (5′-GGTGAAGTTCGTCGTCCA-3′) and RLUCqPCRr (5′-GGCACCTTCAACAATAGCA-3′), respectively.

## Supplementary information

**Supplementary information** accompanies this paper at https://doi.org/10.1186/s12864-020-6662-5.

**Additional file 1** : **Supplementary Table S1**. CPuORFs extracted from *A. thaliana*, *O. sativa*, *S. lycopersicum*, *P. trichocarpa* and *V. vinifera*.

**Additional file 2** : **Supplementary Table S2**. Taxonomic range of sequence conservation of the CPuORFs.

**Additional file 3** : **Supplementary Table S3**. Multiple CPuORF sequence alignments.

**Additional file 4** : **Supplementary Table S4**. Plasmids used in this study and primers used for plasmid construction.

**Additional file 5** : **Supplementary Table S5**. Resourse numbers of *Poplus nigra* full-length cDNA clones used for cloning of 5′-UTRs.

**Additional file 6** : **Supplementary Table S6**. Primers used in this study.

**Additional file 7** : **Supplementary Figure S1.** Alignments of the newly identified CPuORF sequences.

**Additional file 8** : **Supplementary Figure S2.** 5′-UTR nucleotide and deduced amino acid sequences of the poplar CPuORFs analyzed in the transient expression study.

### Abbreviations

35S: Cauliflower mosaic virus 35S RNA; 5′-UTR: 5′-untranslated region; CPuORF: Conserved peptide upstream open reading frame; EST: Expressed sequence tag; Fluc: Firefly luciferase; HG: Homology group; mORF: Main open reading frame; RefSeq: NCBI reference sequence; NMD: Nonsense-mediated mRNA decay; Rluc: Renilla luciferase; TSA: Transcriptome shotgun assembly; uORF: Upstream open reading frame

### Acknowledgements

We thank Ms. Kazuko Harada for general assistance.

### Authors' contributions

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

### Funding

This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI [Grant Nos. JP16H05063 to S.N., JP16K07387 to H.O., JP18H03330 to H. T and H.O., JP18K06297 to H. T, JP19H02917 to H.O., JP19K22892 to H.T., JP19K22299 to H.O.]; the Ministry of Education, Culture, Sports, Science and Technology (MEXT) KAKENHI [Grant Nos. JP26113519 to S. K, JP16H01246 to S. K, JP17H05658 to S.N., JP26114703 to H. T, JP17H05659 to H.T]; the Research Foundation for the Electrotechnology of Chubu to H.T; and the Naito Foundation to H.O.

### Availability of data and materials

The Ensembl and Phytozome transcript IDs of the transcript sequences on which the identified CPuORF sequences were based are shown in Supplementary Table S1. The NCBI GenBank accession numbers of RefSeqs, ESTs and TSAs on which the sequence used for generating the multiple uORF sequence alignments and calculating the $K_a/K_s$ ratios were based are shown in Supplementary Table S3.

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

### Author details

[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, 220013 Minsk, Belarus. [6]New Business Development Office, Churitsu Electric Corporation, Toyoake 470-1112, Japan. [7]Faculty of Biological Science and Technology, Institute of Science and Engineering, Kanazawa University, Kanazawa 920-1192, Japan. [8]Graduate School of Information Science and Technology, Hokkaido University, Sapporo 060-0814, Japan. [9]Graduate School of Bioscience and Biotechnology, Chubu University, Kasugai 487-8501, Japan.

Received: 23 July 2019 Accepted: 10 March 2020
Published online: 30 March 2020

### References

1. Churbanov A, Rogozin IB, Babenko VN, Ali H, Koonin EV. Evolutionary conservation suggests a regulatory function of AUG triplets in 5′-UTRs of eukaryotic genes. Nucleic Acids Res. 2005;33:5512–20.
2. Galagan JE, Calvo SE, Cuomo C, Ma LJ, Wortman JR, Batzoglou S, et al. Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. Nature. 2005;438:1105–15.
3. Kawaguchi R, Bailey-Serres J. mRNA sequence features that contribute to translational regulation in *Arabidopsis*. Nucleic Acids Res. 2005;33:955–65.
4. Rogozin IB, Kochetov AV, Kondrashov FA, Koonin EV, Milanesi L. Presence of ATG triplets in 5′ untranslated regions of eukaryotic cDNAs correlates with a 'weak' context of the start codon. Bioinformatics. 2001;17:890–900.
5. Morris DR, Geballe AP. Upstream open reading frames as regulators of mRNA translation. Mol Cell Biol. 2000;20:8635–42.
6. Cruz-Vera LR, Sachs MS, Squires CL, Yanofsky C. Nascent polypeptide sequences that influence ribosome function. Curr Opin Microbiol. 2011;14:160–6.
7. Ito K, Chiba S. Arrest peptides: cis-acting modulators of translation. Annu Rev Biochem. 2013;82:171–202.
8. Somers J, Poyry T, Willis AE. A perspective on mammalian upstream open reading frame function. Int J Biochem Cell Biol. 2013;45:1690–700.
9. van der Horst S, Filipovska T, Hanson J, Smeekens SCM. Metabolite control of translation by conserved peptide uORFs: the ribosome as a metabolite multi-sensor. Plant Physiol. 2020;182:110–22.
10. Bhushan S, Meyer H, Starosta AL, Becker T, Mielke T, Berninghausen O, et al. Structural basis for translational stalling by human cytomegalovirus and fungal arginine attenuator peptide. Mol Cell. 2010;40:138–46.
11. Wang Z, Sachs MS. Ribosome stalling is responsible for arginine-specific translational attenuation in *Neurospora crassa*. Mol Cell Biol. 1997;17:4904–13.
12. Gaba A, Jacobson A, Sachs MS. Ribosome occupancy of the yeast *CPA1* upstream open reading frame termination codon modulates nonsense-mediated mRNA decay. Mol Cell. 2005;20:449–60.
13. Uchiyama-Kadokura N, Murakami K, Takemoto M, Koyanagi N, Murota K, Naito S, et al. Polyamine-responsive ribosomal arrest at the stop codon of an upstream open reading frame of the AdoMetDC1 gene triggers nonsense-mediated mRNA decay in *Arabidopsis thaliana*. Plant Cell Physiol. 2014;55:1556–67.
14. Wang Z, Gaba A, Sachs MS. A highly conserved mechanism of regulated ribosome stalling mediated by fungal arginine attenuator peptides that appears independent of the charging status of arginyl-tRNAs. J Biol Chem. 1999;274:37565–74.
15. Law GL, Raney A, Heusner C, Morris DR. Polyamine regulation of ribosome pausing at the upstream open reading frame of S-adenosylmethionine decarboxylase. J Biol Chem. 2001;276:38036–43.
16. Hanfrey C, Elliott KA, Franceschetti M, Mayer MJ, Illingworth C, Michael AJ. A dual upstream open reading frame-based autoregulatory circuit controlling polyamine-responsive translation. J Biol Chem. 2005;280:39229–37.

17. Yamashita Y, Takamatsu S, Glasbrenner M, Becker T, Naito S, Beckmann R. Sucrose sensing through nascent peptide-meditated ribosome stalling at the stop codon of Arabidopsis *bZIP11* uORF2. FEBS Lett. 2017;591:1266–77.
18. Rahmani F, Hummel M, Schuurmans J, Wiese-Klinkenberg A, Smeekens S, Hanson J. Sucrose control of translation mediated by an upstream open reading frame-encoded peptide. Plant Physiol. 2009;150:1356–67.
19. Combier JP, de Billy F, Gamas P, Niebel A, Rivas S. Trans-regulation of the expression of the transcription factor MtHAP2-1 by a uORF controls root nodule development. Genes Dev. 2008;22:1549–59.
20. Crowe ML, Wang XQ, Rothnagel JA. Evidence for conservation and selection of upstream open reading frames suggests probable encoding of bioactive peptides. BMC Genomics. 2006;7:16.
21. Hayden CA, Jorgensen RA. 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 Biol. 2007;5:32.
22. Hayden CA, Bosco G. Comparative genomic analysis of novel conserved peptide upstream open reading frames in *Drosophila melanogaster* and other dipteran species. BMC Genomics. 2008;9:61.
23. Tran MK, Schultz CJ, Baumann U. Conserved upstream open reading frames in higher plants. BMC Genomics. 2008;9:361.
24. Vaughn JN, Ellingson SR, Mignone F, Arnim A. Known and novel post-transcriptional regulatory sequences are conserved across plant families. RNA. 2012;18:368–84.
25. 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. 2019;25:292–304.
26. 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. 2012;28:2231–41.
27. Ebina I, Takemoto-Tsutsumi M, Watanabe S, Koyama H, Endo Y, Kimata K, et al. Identification of novel *Arabidopsis thaliana* upstream open reading frames that control expression of the main coding sequences in a peptide sequence-dependent manner. Nucleic Acids Res. 2015;43:1562–76.
28. Noh AL, Watanabe S, Takahashi H, Naito S, Onouchi H. An upstream open reading frame represses expression of a tomato homologue of Arabidopsis *ANAC096*, a NAC domain transcription factor gene, in a peptide sequence-dependent manner. Plant Biotechnol. 2015;32:157–63.
29. Jorgensen RA, Dorantes-Acosta AE. Conserved peptide upstream open Reading frames are associated with regulatory genes in angiosperms. Front Plant Sci. 2012;3:191.
30. Li WH. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol. 1993;36:96–9.
31. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2007;35:D61–5.
32. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, et al. Ensembl 2018. Nucleic Acids Res. 2018;46:D754–61.
33. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, et al. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 2012;40:D1178–86.
34. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 2015;16:157.
35. Alatorre-Cobos F, Cruz-Ramirez A, Hayden CA, Perez-Torres CA, Chauvin AL, Ibarra-Laclette E, et al. Translational regulation of Arabidopsis XIPOTL1 is modulated by phosphocholine levels via the phylogenetically conserved upstream open reading frame 30. J Exp Bot. 2012;63:5203–21.
36. Guerrero-Gonzalez ML, Ortega-Amaro MA, Juarez-Montiel M, Jimenez-Bremont JF. Arabidopsis polyamine oxidase-2 uORF is required for downstream translational regulation. Plant Physiol Biochem. 2016;108:381–90.
37. Imai A, Hanzawa Y, Komura M, Yamamoto KT, Komeda Y, Takahashi T. The dwarf phenotype of the Arabidopsis acl5 mutant is suppressed by a mutation in an upstream ORF of a bHLH gene. Development. 2006;133:3575–85.
38. Ribone PA, Capella M, Arce AL, Chan RL. A uORF represses the transcription factor AtHB1 in aerial tissues to avoid a deleterious phenotype. Plant Physiol. 2017;175:1238–53.
39. Tabuchi T, Okada T, Azuma T, Nanmori T, Yasuda T. Posttranscriptional regulation by the upstream open reading frame of the phosphoethanolamine N-methyltransferase gene. Biosci Biotechnol Biochem. 2006;70:2330–4.
40. Hayashi N, Sasaki S, Takahashi H, Yamashita Y, Naito S, Onouchi H. Identification of *Arabidopsis thaliana* upstream open reading frames encoding peptide sequences that cause ribosomal arrest. Nucleic Acids Res. 2017;45:8844–58.
41. Franceschetti M, Hanfrey C, Scaramagli S, Torrigiani P, Bagni N, Burtin D, et al. Characterization of monocot and dicot plant S-adenosyl-l-methionine decarboxylase gene families including identification in the mRNA of a highly conserved pair of upstream overlapping open reading frames. Biochem J. 2001;353:403–9.
42. Konishi T, Takeda T, Miyazaki Y, Ohnishi-Kameyama M, Hayashi T, O'Neill MA, et al. A plant mutase that interconverts UDP-arabinofuranose and UDP-arabinopyranose. Glycobiology. 2007;17:345–54.
43. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18:821–9.
44. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. Nat Methods. 2012;9:357–9.
45. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–402.
46. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal omega. Mol Syst Biol. 2011;7:539.
47. Steinway SN, Dannenfelser R, Laucius CD, Hayes JE, Nayak S. JCoDA: a tool for detecting evolutionary selection. BMC Bioinformatics. 2010;11:284.
48. Charif D, Lobry JR. SeqinR 1.0–2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: Bastolla U, Porto M, Roman HE, Vendruscolo M, editors. Structural approaches to sequence evolution: molecules, networks, populations. New York: Springer Verlag; 2007. p. 207–32.
49. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B. 1995;57:298–300.
50. Clamp M, Cuff J, Searle SM, Barton GJ. The Jalview Java alignment editor. Bioinformatics. 2004;20:426–7.
51. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, et al. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013;9:e1003118.
52. Nanjo T, Futamura N, Nishiguchi M, Igasaki T, Shinozaki K, Shinohara K. Characterization of full-length enriched expressed sequence tags of stress-treated poplar leaves. Plant Cell Physiol. 2004;45:1738–48.
53. Motohashi K. Seamless ligation cloning extract (SLiCE) method using cell lysates from laboratory Escherichia coli strains and its application to SLiP site-directed mutagenesis. Methods Mol Biol. 2017;1498:349–57.
54. Ho SN, Hunt HD, Horton RM, Pullen JK, Pease LR. Site-directed mutagenesis by overlap extension using the polymerase chain reaction. Gene. 1989;77:51–9.
55. Menges M, Murray JA. Synchronous *Arabidopsis* suspension cultures for analysis of cell-cycle gene activity. Plant J. 2002;30:203–12.
56. Nagata T, Nemoto Y, Hasezawa S. Tobacco BY-2 cell line as the “HeLa” cell in the cell biology of higher plants. Int Rev Cytol. 1992;132:1–30.

## Publisher’s Note