

uORFlight: a vehicle towards uORF-mediated translational regulation mechanisms in eukaryotes

Ruixia Niu^{1,4}, Yulu Zhou^{1,4}, Rui Mou¹, Zhijuan Tang¹, Zhao Wang¹, Guilong Zhou¹, Sibin Guo², Meng Yuan³ and Guoyong Xu^{1*}

Affiliations

¹State Key Laboratory of Hybrid Rice, Institute for Advanced Studies (IAS), Wuhan University, Wuhan, Hubei 430072, China

²Guangxi Key Laboratory of Rice Genetics and Breeding, Rice Research Institute, Guangxi Academy of Agricultural Science, Nanning, Guangxi 530007, China

³National Key Laboratory of Crop Genetic Improvement, National Centre of Plant Gene Research (Wuhan), Huazhong Agricultural University, Wuhan, Hubei 430070, China

⁴These authors contributed equally to this work

***Correspondence**

Dr. Guoyong Xu

Dangdai Building

Wuhan University

Wuhan, Hubei 430072

China

Phone: +8627-68758136

Fax: +8627-68758136

E-mail: guoyong.xu@whu.edu.cn

Abstract

Upstream open reading frames (uORFs) are prevalent in eukaryotic mRNAs. They act as a translational control element for precisely tuning the expression of the downstream major open reading frame (mORF) with essential cellular functionalities. uORF variation has been clearly associated with several human diseases. In contrast, natural uORF variants in plants have not ever been identified or linked with any phenotypic changes. The paucity of such evidence encouraged us to generate this database-uORFlight (<http://uorflight.whu.edu.cn>). It facilitates the exploration of uORF variation among different splicing models of Arabidopsis and rice genes. Most importantly, users can evaluate uORF frequency among different accessions at the population scale and find out the causal single nucleotide polymorphism (SNP) or insertion/deletion (INDEL) which can be associated with phenotypic variation through database mining or simple experiments. Such information will help to make hypotheses of uORF function in plant development or adaption to changing environments on the basis of the cognate mORF function. This database also curates plant uORF relevant literature into distinct groups. To be broadly interesting, our database expands uORF annotation into more species of fungi (*Botrytis cinerea*), plant (*Brassica napus*, *Glycine max*, *Gossypium raimondii*, *Medicago truncatula*, *Solanum lycopersicum*, *Solanum tuberosum*, *Triticum aestivum* and *Zea mays*), metazoan (*Caenorhabditis elegans* and *Drosophila melanogaster*) and vertebrate (*Homo sapiens*, *Mus musculus* and *Danio rerio*). Therefore, uORFlight will light up the runway toward how uORF genetic variation determines phenotypic diversity and advance our understanding of translational control mechanisms.

Introduction

Gene expression must be tightly regulated at transcription, translation and post-translation levels. The imperfect correlation between protein abundance and mRNA levels suggests translational efficiency regulated by translational control as one of the determinants of protein output from variable mRNA input. This layer of regulation is

mediated by the cooperative action between different mRNA elements and trans-acting factors (1). Upstream open reading frames (uORFs) are among the mRNA elements that can confer precise control of protein translation.

A uORF initiation codon resides upstream of the coherent mORF, and will be first encountered by 43S scanning ribosome (including 40S ribosomal subunit and eIF2 ternary complex). Sequentially, 60S subunit joins in and reconstitutes 80S ribosome for uORF translation elongation, after which the 40S and 60S are disjointed and 40S may remain associated with mRNA. Therefore, usually uORF translation is prioritized over mORF, leading to hindered translation of the mORF. Only in situations where the remaining 40S ribosomes regain fresh eIF2 ternary complex and other unknown reinitiation factors, or when uORF initiation codon is bypassed by the scanning ribosome, the downstream mORF has the chance to be translated (Figure 1a). The former situation is termed as reinitiation and the latter as leaky scanning, two mechanisms that have been accepted as an explanation of limited mORF expression under normal growth and developmental conditions (2-4). Most importantly, a uORF can confer selective mORF translation in response to a wide-range of cellular stimuli, such as metabolite and ion homeostasis, hormone changes, environmental signals, and immune induction (See uORF references on this website). This tight and temporal regulation pattern fine-tunes the translational efficiency of mORFs and thus guarantees appropriate protein quantity and quality for adaption to different physiological conditions. Because of those unique features, we have successfully utilized uORF-mediated translational control in engineering disease resistant plants without fitness costs by restricting toxic resistance protein translation under normal conditions but allowing transient induction under pathogen infection conditions (5).

However, once uORF-mediated precision control has been challenged by genetic variation or mis-regulation, it causes human diseases (6,7). By 2009, 509 human genes had been identified with polymorphic uORFs and some of them have been experimentally associated with different human diseases, including malignancies, metabolic or neurologic disorders, and inherited syndromes. This trend became more

1 striking recently with more genomic variation data released and analyzed (3,8). In
2 contrast, natural variation of plant uORFs has not yet been investigated, even though
3 there are abundant publicly accessible genetic and phenotypic variation data, especially
4 for *Arabidopsis* (*Arabidopsis thaliana*) and rice (*Oryza sativa* L.). Since the release of
5 the *Arabidopsis* reference genome of accession Col-0 in 2000, the rapid development
6 of sequencing technology has bolstered genome-wide association studies (GWAS) by
7 linkage disequilibrium of interesting phenotypic traits with the most probable genetic
8 variation, particularly using the genome sequences of 1135 accessions from the 1001
9 Genomes Project (9-15). Genetic variation of rice has long been used for molecular
10 breeding and recent re-sequencing of a large set of rice accessions, especially from the
11 3000 Rice Genomes Project, generated a wealth of genetic variation for the discovery
12 of useful alleles for agronomic trait improvement (16-24).

13 It is noteworthy that previous genotype-phenotype association studies mainly
14 focused on protein coding regions, while it is becoming more evident that the cis-
15 element variation, such as in the promoter regions, weighs a lot in determining
16 phenotypic variation (25). However, there is still less attention that has been paid on the
17 variation of mRNA regulatory elements, such as uORFs. In this study, we used public
18 resources to identify uORF variation for further experimental verification of phenotypic
19 diversity mediated by translational control.

20

21 **Methods**

22 A uORF is defined as the presence of an initiation codon in an annotated mRNA 5'
23 leader region and can be categorized into 'Types 1-3' based on the positions of uORF
24 stop codon, with Type1 in 5' leader, Type2 in mORF coding region and Type3 shared
25 with mORF (also known as an mORF N-extension). We term those mRNAs without a
26 uORF as 'Free'. It is obvious that reinitiation is impossible for translation of Type2
27 uORF-controlled mORFs (Figure 1b and c). Hereafter, ORF of both uORF and mORF
28 means that AUG is used as the initiation codon, unless specifically stated. We use the
29 term 5' leader sequence instead of 5' untranslated region (5' UTR), considering the

1 peptide-coding potential of uORFs (26,27). The sequence from -3 to +4 relative to AUG
2 initiation codon (A as +1) corresponding to Kozak consensus (A/GCCAUGG) position
3 is termed as initiation codon context (ICC) with ICCu and ICCm for uORF and mORF,
4 respectively (Figure 1d).

5 We chose the Arabidopsis Col-0 accession (Ensemble V39; TAIR11) and rice
6 Nipponbare cultivar (MSU V7) as reference genomes for dicot and monocot uORF
7 analysis, respectively. Arabidopsis representative gene models (27445 in total;
8 Supplementary Table 1 in **Download** menu) of the nuclear coding proteins contain
9 26713 genes from TAIR10 representative gene annotation file and 732 genes using their
10 ‘.1’ splicing models. Nipponbare representative gene models (38860 in total; Non-TE
11 Loci) are defined using their smallest numbered models (38618, 221, 16 and 5 genes
12 using the ‘.1’, ‘.2’, ‘.3’ and ‘.4’ splicing models, respectively; Supplementary Table 2
13 in **Download** menu). To calculate uORF variation, we downloaded VCF (variant call
14 format) files of Arabidopsis 1135 accessions from the 1001 Genomes Project and rice
15 3k varieties from the 3000 Rice Genomes Project (9,20,24,28). We filtered SNPs and
16 INDELs with low quality indicated in the VCF files and used the alleles with frequency
17 over 90% as suggested (28). We further removed genes with variants affecting the
18 annotated initiation codon of their mORF.

19 We replaced all the splicing models with filtered variants and searched uORFs of
20 all the accessions. We then transferred the relative positions of uORFs on the different
21 transcripts into absolute positions in the referenced genome and assigned continuous
22 index numbers starting with number 1.1 in the order of uORF ATG occurrence in the
23 genome. Those uORFs with the same genomic positions and sequences are assigned
24 the same index number to indicate no uORF changes. Otherwise, two leveled uORF
25 identifiers (Level1.Level2) were used to describe the variation. Level1 indicates major
26 differences that cause uORF creation, loss or length changes due to SNPs or INDELs.
27 Level2 indicates minor differences that lead to nucleotide and/or amino acid
28 substitution due to SNPs (Figure 1e). We grouped Arabidopsis accessions on the basis
29 of the latitude at which they were collected (15-degree interval; Supplementary Table 1

1 in **Download** menu). The frequency of an individual uORF in Arabidopsis was
 2 calculated based on its occurrence in the total population and in different latitude ranges.
 3 The frequency in rice was calculated as its occurrence in total population and nine
 4 subgroups (20). The associated SNP or INDEL identifiers are also recorded along with
 5 uORF variants.

6 MySQL database schema was used for uORF information storage and a user-
 7 friendly PHP web interface was designed to query and download. GO analysis was done
 8 using Omicshare online tools (<http://www.omicshare.com/>) with the default setting. The
 9 uORF annotation information of the other species can be found in the website.

10

11 **Results and Discussion**

12 With the recent recognition of the significance of uORFs within distinct
 13 physiological contexts, the following functionalities will help the community quickly
 14 overview the progress in this area and find out uORF variation to link with phenotypic
 15 diversity at the population level.

16 **uORF in the reference genomes**

17 uORFs are becoming increasingly attractive because of their capacity to fine-tune
 18 translational control and respond accurately to distinct extracellular and intracellular
 19 stimuli. However, the current understanding of uORFs is based on a small number of
 20 case studies. In an attempt to provide guidelines, we investigated natural patterns in
 21 uORF types, length distribution and ICC of Arabidopsis and rice representative gene
 22 models. uORFs are more prevalent in Arabidopsis (48.45%), and the lower frequency
 23 (20.65%) of uORFs in rice transcripts may arise from incomplete 5' leader annotation.
 24 Their prevalence is mostly due to overrepresentation of Type1 uORFs which account
 25 for 90.79% and 87.63%, in contrast to only 9.16% and 12.17% for Type2 uORFs of
 26 Arabidopsis and rice, respectively. The Type3 uORFs are the least common (19 uORFs
 27 in 17 Arabidopsis genes, and 74 uORFs in 41 rice genes), perhaps because they may
 28 give rise to N-extension and are likely to alter protein activities or molecular
 29 localizations as reported (29). Type2-containing genes tend to occur along with Type1

1 uORFs, and the significance of their co-existence needs further investigation (Figure
2 2a). GO analysis shows clearly different enriched terms among those uORF-Free,
3 Type1-only and Type2 genes in both Arabidopsis and rice, suggesting that uORF type
4 attributes could have an impact on different functional groups of genes (Figure 2b;
5 Supplementary Table1 and 2 in **Download** menu). This assumption appears to be
6 reasonable because both leaky scanning and reinitiation mechanisms could overcome
7 Type1 uORF inhibition, while only leaky scanning can overcome Type2 uORF
8 inhibition (Figure 1a and b), and leaky scanning seems to be of low efficiency in both
9 animals and plants (30,31). Importantly, uORF-containing genes are enriched in more
10 specific groups than uORF-Free genes. Moreover, even though rice has fewer enriched
11 GO-terms, most of them (89.47%) are shared with GO-terms enriched for Arabidopsis
12 uORF-containing genes. The 11 shared GO-terms for Type1 only and Type2 include
13 ‘signaling’ and ‘cellular response to stimulus’ (Figure 2c). All these results suggest that
14 uORFs are more frequently used by specific functional groups and these groups are
15 shared between Arabidopsis and rice. Next, we examined uORF length distribution and
16 found that Type2 uORFs are on average longer than Type1, and that rice uORFs are
17 generally longer than those of Arabidopsis (Figure 2d). A minimal Type1 uORF
18 consisting of just an AUG and a stop codon has been shown to be sufficient for
19 translational inhibition of three boron (B)-related genes and also sufficient for
20 translational responsiveness to low B stimulation (32). Such short uORFs (6.79%) are
21 more commonly found among Arabidopsis Type1 uORFs and must require other
22 synergistic cis-elements to allow translational responsiveness to specific stimuli.
23 Therefore, the attribute of uORF length appears not to be a definitive parameter for
24 prediction of uORF function. We then asked whether uORF surrounding sequences
25 display some informative patterns. We found that only 0.45% ICCu and 2.99% ICCm
26 in Arabidopsis and 1.93% ICCu and 11.88% ICCm in rice contain Kozak consensus
27 sequences, suggesting that more variable contexts are used to flexibly tune ORF
28 translation in nature. Enrichment analysis of ICCu and ICCm did not uncover any
29 obvious enriched sequence, suggesting the feasibility of engineering tailored protein

1 expression using ICC variants in plants.

2 From **uORF view** menu, users can browse and search uORF information for all
3 splicing models in the reference genomes, including eudicot Arabidopsis Col-0 and
4 monocot rice Nipponbare. The entirety of the uORF information is also saved in a table
5 format (Supplementary Table 1 and 2 in **Download** menu) which can be accessed in the
6 ‘Download’ Menu. Choosing the reference genome will pop out a search page, this
7 page requires users to input the splicing model identifier (ID; e.g. AT4G36990.1) and
8 the search results will return the summary information of the uORF, details of which
9 can be accessed by clicking the ID or downloaded in the top-right corner in the popped
10 out page. uORF annotation of other species can be obtained in a similar way under this
11 menu.

12 **uORF variation at the population level**

13 uORF variation has been studied at the genome-wide scale using human SNPs and
14 the results provide clear evidence of its association with genetic diseases. However,
15 these studies only assessed single SNP effects that result in uORF initiation codon
16 creation or stop codon loss (3,8). We analyzed uORF variation by considering the
17 integral effect of all homozygous SNPs and INDELs in each accession. Significantly,
18 we found that 54.17% of Arabidopsis uORF variants (64.52% in rice) of the
19 representative gene models are associated with at least two genetic changes. As
20 emphasized above, uORF creation, loss, length changes, and type switches are
21 considered as major variation and are defined as Level1 (Figure 1e). Level2 variation
22 is likely to affect uORF function, provided that its nucleotide sequence or encoded
23 peptide is able to cause ribosome stalling or mRNA decay (4,6,7). To simply show the
24 significance of uORF variation, we focused on three Level1 variants which alter uORF
25 type attributes between Type1-only and Type2, between uORF-Free and Type1-only,
26 between uORF-Free and Type2, which excludes Level1 changes (Figure 3a-c). We first
27 found that 31.03% and 20.40% of uORF-containing genes have altered uORF type
28 attributes among different splicing models in Col-0 and Nipponbare, respectively,
29 suggesting that uORF variation produced by alternative splicing may add translational

control regulation to specific processes (Figure 3c and d; Supplementary Table 3 and 4 in **Download** menu). We also identified the three Level1 variants of different accessions at the population level by firstly analyzing the representative gene models (Figure 3e; Supplementary Table 3 and 4 in **Download** menu). Alternatively, we identified the three Level1 variants of different accessions at the population level by considering all splicing models of a defined gene, and any splicing model of one accession different from the other accession are counted (Figure 3f; Supplementary Table 3 and 4 in **Download** menu). Significantly, these two different ways exemplify the importance of uORF variation on key genes of different cellular functions.

In **uORF variation** menu, users are asked to input the gene model ID (e.g. AT4G36990) to search uORF variation among different splicing models in the reference genomes. Users can also input the splicing model ID (e.g. AT4G36990.1) and choose the accessions to return uORF variation in the selected population. The upper summary table lists all uORFs and the lower table returns uORFs listed according to accession names. The detailed information of uORF variants contains its frequency and the causal SNP or INDEL number at the popped out page by clicking 'ALL' in the summary table. Downloading at the top right corner allows the retrieval of all uORF information, including associated SNP or INDEL identifiers (VariationID). This information can also be accessed by clicking the individual accession name. SNP VariationID can be directly used to search their associated phenotypic variation using online tools such as <https://aragwas.1001genomes.org> for Arabidopsis (15), and <http://snp-seek.irri.org/> and <http://ricevarmap.ncpgr.cn/> for rice (23,24,28). These databases use different VariationID which can be converted using ID converter under the Tool menu.

The current GWAS databases use SNPs as genetic variants to perform association studies. Methods using INDELs are still under active development (33). On the other hand, many variables, such as population size, may cause GWAS false positive or biased associations (34). Considering uORF variants are responsible for 30%-80% protein abundance changes in human, we recommend performing a simple Dual-Luciferase assay to compare the translation efficiency of uORF variants by

1 Agrobacterium-mediated transient expression in *Nicotiana benthamiana* (35,36). We
2 provide a Dual-Luciferase vector (pGX-5Dual) along with this database. Users can
3 easily clone 5' leaders containing different uORF alleles to the 5' of firefly luciferase
4 (FLUC) through ligation-independent cloning technology (37). Comparison of the ratio
5 of FLUC activity and mRNA level to the internal control Renilla luciferase (RLUC)
6 expressed from the same vector will indicate the effects of uORF variants on mRNA
7 stability and translation efficiency. The difference will provide support for the
8 association of uORF variation with its physiological roles deduced from its cognate
9 mORF function (Figure 3g).

10 **CPuORF information**

11 Analysis of evolutionary conservation suggests the existence of conserved peptide
12 uORFs (CPuORFs) (38-44) and 97 non-redundant AUG-initiated CPuORFs in
13 Arabidopsis have been confirmed here. It is suggested that most CPuORFs confer
14 peptide-sequence dependent regulation in a cis manner, as metabolite receptors or
15 sensors that function in the ribosome exit tunnel by stalling the ribosome and preventing
16 reinitiation (4). There are also exceptional reports on uORFs functioning in trans-
17 regulation (45,46). In 'uORF view' menu, we collected all Arabidopsis genes containing
18 CPuORF and also provided their rice homologues (Supplementary Table 5 in
19 **Download** menu). This information will advance our understanding of uORF sequence-
20 dependent regulation by facilitating the study of CPuORFs.

21 **Plant uORF reference curation**

22 Eukaryotic uORF-related literatures to year 2013 have been curated in uORFdb
23 through the Boolean search for key words in the NCBI PubMed database (47). To help
24 the community get the latest view of the progress in plant uORF research, we manually
25 curated all the relevant references on the basis of our knowledge and categorized the
26 references into Case study, Mechanism study, Practical study, Genome-wide study, and
27 Review. Users are invited to help us complete this section if missing or inappropriate
28 references are found. Clicking the reference link will direct the users to the associated
29 PubMed page or journal page.

Customized uORF analysis

In the current database, we only processed uORFs with AUG as the initiation codon. With the rapid development of ribosome footprinting, mounting evidence suggests the usage of Non-AUG initiation codons in translational control (48). Those Non-AUG-initiated ORF-encoded peptides are also detectable by mass spectrometry (26). However, Non-AUG-initiated uORFs may function in an opposite way to AUG uORFs in translational control (48). This nuance remains poorly understood and systematic variation analysis of Non-AUG-initiated uORFs will be included when more information is available in plants. Therefore, users who are interested in Non-AUG studies or uORF studies in species other than Arabidopsis and rice are encouraged to use our searching tool under the menu of Tools to obtain the basic uORF information by inputting different uORF initiation codons and input cDNA and CDS sequences separately.

Future Direction

uORFs are common in eukaryotes and information from more organisms will be useful additions to our database in the future. uORFs may encode functional peptides to act in either trans or cis manners, and this information will need to be evaluated by the combination of ribosome footprinting and mass-spectrometry data, which will be integrated as it becomes available. In addition, uORFs are RNA cis-elements which require trans-acting factors to regulate translation (36). Meanwhile, co-regulatory cis-elements, such as the R-motif identified in our previous study (36), may account for uORF regulation specificity and diversity. Information about regulatory trans-acting factors and co-acting cis-element variation will be integrated into the database progressively. Furthermore, a uORF calculator will be developed to predict the regulatory power of natural or synthetic uORFs for tailored protein expression after machine learning of large experimental data is achieved.

Supplementary Information (in Download menu)

- 1 Supplementary Table 1. uORF information in Arabidopsis
- 2 Supplementary Table 2. uORF information in rice
- 3 Supplementary Table 3. uORF variation in Arabidopsis
- 4 Supplementary Table 4. uORF variation in rice
- 5 Supplementary Table 5. CPuORFs in Arabidopsis and rice
- 6 Supplementary Method: Ligation-independent cloning of Dual-LUC vector

7

8 **Acknowledgements**

9 We thank Sophia Zebell and Paul J. Zwack at Duke University for comments.

10

11 **Funding**

12 This study was supported by grants from National Natural Science Foundation of China
13 to G. Xu.

14 Conflict of interest statement: None declared.

15

16 **References**

- 17 1. Merchante, C., Stepanova, A.N. and Alonso, J.M. (2017) Translation regulation in plants:
18 an interesting past, an exciting present and a promising future. *Plant J*, **90**, 628-653.
- 19 2. Yun, Y., Adesanya, T.M.A. and Mitra, R.D. (2012) A systematic study of gene expression
20 variation at single-nucleotide resolution reveals widespread regulatory roles for uAUGs.
21 *Genome research*, **22**, 1089-1097.
- 22 3. Calvo, S.E., Pagliarini, D.J. and Mootha, V.K. (2009) Upstream open reading frames cause
23 widespread reduction of protein expression and are polymorphic among humans. *Proc.*
24 *Natl. Acad. Sci. U.S.A.*, **106**, 7507-7512.
- 25 4. von Arnim, A.G., Jia, Q. and Vaughn, J.N. (2014) Regulation of plant translation by
26 upstream open reading frames. *Plant Sci.*, **214**, 1-12.
- 27 5. Xu, G., Yuan, M., Ai, C., Liu, L., Zhuang, E., Karapetyan, S., Wang, S. and Dong, X. (2017)
28 uORF-mediated translation allows engineered plant disease resistance without fitness
29 costs. *Nature*, **545**, 491-494.
- 30 6. Somers, J., Poyry, T. and Willis, A.E. (2013) A perspective on mammalian upstream open
31 reading frame function. *Int. J. Biochem. Cell Biol.*, **45**, 1690-1700.
- 32 7. Barbosa, C., Peixeiro, I. and Romao, L. (2013) Gene expression regulation by upstream
33 open reading frames and human disease. *Plos Genet.*, **9**, e1003529.
- 34 8. Whiffin, N., Karczewski, K.J., Zhang, X., Chothani, S., Smith, M.J., Evans, D.G., Roberts, A.M.,
35 Quaife, N.M., Schafer, S., Rackham, O. *et al.* (2019) Characterising the loss-of-function
36 impact of 5' untranslated region variants in whole genome sequence data from 15,708

- 1 individuals. *bioRxiv*, 543504.
- 2 9. (2016) 1,135 Genomes Reveal the Global Pattern of Polymorphism in *Arabidopsis thaliana*.
- 3 *Cell*, **166**, 481-491.
- 4 10. Ossowski, S., Schneeberger, K., Clark, R.M., Lanz, C., Warthmann, N. and Weigel, D. (2008)
- 5 Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome research*,
- 6 **18**, 2024-2033.
- 7 11. Schneeberger, K., Ossowski, S., Ott, F., Klein, J.D., Wang, X., Lanz, C., Smith, L.M., Cao, J.,
- 8 Fitz, J., Warthmann, N. *et al.* (2011) Reference-guided assembly of four diverse
- 9 *Arabidopsis thaliana* genomes. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 10249-10254.
- 10 12. Cao, J., Schneeberger, K., Ossowski, S., Gunther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C.,
- 11 Stegle, O., Lippert, C. *et al.* (2011) Whole-genome sequencing of multiple *Arabidopsis*
- 12 *thaliana* populations. *Nat. Genet.*, **43**, 956-963.
- 13 13. Long, Q., Rabanal, F.A., Meng, D., Huber, C.D., Farlow, A., Platzer, A., Zhang, Q.,
- 14 Vilhjalmsón, B.J., Korte, A., Nizhynska, V. *et al.* (2013) Massive genomic variation and
- 15 strong selection in *Arabidopsis thaliana* lines from Sweden. *Nat. Genet.*, **45**, 884-890.
- 16 14. Gan, X., Stegle, O., Behr, J., Steffen, J.G., Drewe, P., Hildebrand, K.L., Lyngsoe, R., Schultheiss,
- 17 S.J., Osborne, E.J., Sreedharan, V.T. *et al.* (2011) Multiple reference genomes and
- 18 transcriptomes for *Arabidopsis thaliana*. *Nature*, **477**, 419-423.
- 19 15. Togninalli, M., Seren, U., Meng, D., Fitz, J., Nordborg, M., Weigel, D., Borgwardt, K., Korte,
- 20 A. and Grimm, D.G. (2018) The AraGWAS Catalog: a curated and standardized *Arabidopsis*
- 21 *thaliana* GWAS catalog. *Nucleic Acids Res.*, **46**, D1150-D1156.
- 22 16. Yu, J., Hu, S., Wang, J., Wong, G.K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X. *et al.*
- 23 (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science*, **296**, 79-
- 24 92.
- 25 17. Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A.,
- 26 Oeller, P., Varma, H. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp.
- 27 *japonica*). *Science*, **296**, 92-100.
- 28 18. Project, T.I.R.G.S. (2005) The map-based sequence of the rice genome. *Nature*, **436**, 793-
- 29 800.
- 30 19. Huang, X., Zhao, Y., Wei, X., Li, C., Wang, A., Zhao, Q., Li, W., Guo, Y., Deng, L., Zhu, C. *et al.*
- 31 (2011) Genome-wide association study of flowering time and grain yield traits in a
- 32 worldwide collection of rice germplasm. *Nat. Genet.*, **44**, 32-39.
- 33 20. Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S., Wu, Z., Li, M., Zheng, T., Fuentes,
- 34 R.R., Zhang, F. *et al.* (2018) Genomic variation in 3,010 diverse accessions of Asian
- 35 cultivated rice. *Nature*, **557**, 43-49.
- 36 21. (2014) The 3,000 rice genomes project. *GigaScience*, **3**, 7.
- 37 22. Wing, R.A., Purugganan, M.D. and Zhang, Q. (2018) The rice genome revolution: from an
- 38 ancient grain to Green Super Rice. *Nat Rev Genet.*
- 39 23. Zhao, H., Yao, W., Ouyang, Y., Yang, W., Wang, G., Lian, X., Xing, Y., Chen, L. and Xie, W.
- 40 (2015) RiceVarMap: a comprehensive database of rice genomic variations. *Nucleic Acids*
- 41 *Res.*, **43**, D1018-1022.
- 42 24. Mansueto, L., Fuentes, R.R., Borja, F.N., Detras, J., Abriol-Santos, J.M., Chebotarov, D.,
- 43 Sanciango, M., Palis, K., Copetti, D., Poliakov, A. *et al.* (2017) Rice SNP-seek database
- 44 update: new SNPs, indels, and queries. *Nucleic Acids Res.*, **45**, D1075-D1081.

- 1 25. Rodríguez-Leal, D., Lemmon, Z.H., Man, J., Bartlett, M.E. and Lippman, Z.B. (2017)
2 Engineering quantitative trait variation for crop improvement by genome editing. *Cell*,
3 **171**, 470-480.
- 4 26. Slavoff, S.A., Mitchell, A.J., Schwaid, A.G., Cabili, M.N., Ma, J., Levin, J.Z., Karger, A.D., Budnik,
5 B.A., Rinn, J.L. and Saghatelian, A. (2012) Peptidomic discovery of short open reading
6 frame-encoded peptides in human cells. *Nat. Chem. Biol.*, **9**, 59.
- 7 27. Wu, H.-Y.L., Song, G., Walley, J.W. and Hsu, P.Y. (2019) Translational landscape in tomato
8 revealed by transcriptome assembly and ribosome profiling. *bioRxiv*, 534677.
- 9 28. Alexandrov, N., Tai, S., Wang, W., Mansueto, L., Palis, K., Fuentes, R.R., Ulat, V.J.,
10 Chebotarov, D., Zhang, G., Li, Z. *et al.* (2015) SNP-Seek database of SNPs derived from
11 3000 rice genomes. *Nucleic Acids Res.*, **43**, D1023-1027.
- 12 29. Xu, F., Huang, Y., Li, L., Gannon, P., Linster, E., Huber, M., Kapos, P., Bienvenut, W., Polevoda,
13 B., Meinel, T. *et al.* (2015) Two N-Terminal Acetyltransferases Antagonistically Regulate
14 the Stability of a Nod-Like Receptor in Arabidopsis. *The Plant cell*, **27**, 1547-1562.
- 15 30. Schleich, S., Strassburger, K., Janiesch, P.C., Koledachkina, T., Miller, K.K., Haneke, K., Cheng,
16 Y.S., Kuchler, K., Stoecklin, G., Duncan, K.E. *et al.* (2014) DENR-MCT-1 promotes translation
17 re-initiation downstream of uORFs to control tissue growth. *Nature*, **512**, 208-212.
- 18 31. Wang, L. and Wessler, S.R. (1998) Inefficient reinitiation is responsible for upstream open
19 reading frame-mediated translational repression of the maize R gene. *Plant Cell*, **10**,
20 1733-1745.
- 21 32. Tanaka, M., Sotta, N., Yamazumi, Y., Yamashita, Y., Miwa, K., Murota, K., Chiba, Y., Hirai,
22 M.Y., Akiyama, T., Onouchi, H. *et al.* (2016) The Minimum Open Reading Frame, AUG-
23 Stop, Induces Boron-Dependent Ribosome Stalling and mRNA Degradation. *The Plant*
24 *cell*, **28**, 2830-2849.
- 25 33. Song, B., Mott, R. and Gan, X. (2018) Recovery of novel association loci in *Arabidopsis*
26 *thaliana* and *Drosophila melanogaster* through leveraging INDELs association and
27 integrated burden test. *Plos Genet.*, **14**, e1007699.
- 28 34. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A. and Yang, J.
29 (2017) 10 years of GWAS discovery: biology, function, and translation. *Am J Hum. Genet.*,
30 **101**, 5-22.
- 31 35. Ferreira, J.P., Overton, K.W. and Wang, C.L. (2013) Tuning gene expression with synthetic
32 upstream open reading frames. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 11284-11289.
- 33 36. Xu, G., Greene, G.H., Yoo, H., Liu, L., Marques, J., Motley, J. and Dong, X. (2017) Global
34 translational reprogramming is a fundamental layer of immune regulation in plants.
35 *Nature*, **545**, 487-490.
- 36 37. Xu, G.Y., Sui, N., Tang, Y., Xie, K., Lai, Y.Z. and Liu, Y.L. (2010) One-step, zero-background
37 ligation-independent cloning intron-containing hairpin RNA constructs for RNAi in plants.
38 *New Phytol.*, **187**, 240-250.
- 39 38. Hayden, C.A. and Jorgensen, R.A. (2007) Identifican of novel conserved peptide uORF
40 homology groups in Arabidopsis and rice reveals ancient eukaryotic origin of select
41 groups and preferential associatiotion with transcription factor-encoding genes. *BMC*
42 *Biol.*, **5**, 32.
- 43 39. Tran, M.K., Schultz, C.J. and Baumann, U. (2008) Conserved upstream open reading frames
44 in higher plants. *BMC genomics*, **9**, 361.

- 1 40. Takahashi, H., Takahashi, A., Naito, S. and Onouchi, H. (2012) BAIUCAS: a novel BLAST-
2 based algorithm for the identification of upstream open reading frames with conserved
3 amino acid sequences and its application to the *Arabidopsis thaliana* genome.
4 *Bioinformatics*, **28**, 2231-2241.
- 5 41. Vaughn, J.N., Ellingson, S.R., Mignone, F. and Arnim, A. (2012) Known and novel post-
6 transcriptional regulatory sequences are conserved across plant families. *RNA*, **18**, 368-
7 384.
- 8 42. Ebina, I., Takemoto-Tsutsumi, M., Watanabe, S., Koyama, H., Endo, Y., Kimata, K., Igarashi,
9 T., Murakami, K., Kudo, R., Ohsumi, A. *et al.* (2015) Identification of novel *Arabidopsis*
10 *thaliana* upstream open reading frames that control expression of the main coding
11 sequences in a peptide sequence-dependent manner. *Nucleic Acids Res.*, **43**, 1562-1576.
- 12 43. Hsu, P.Y., Calviello, L., Wu, H.-Y.L., Li, F.-W., Rothfels, C.J., Ohler, U. and Benfey, P.N. (2016)
13 Super-resolution ribosome profiling reveals unannotated translation events in
14 *Arabidopsis*. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, E7126-7135.
- 15 44. van der Horst, S., Snel, B., Hanson, J. and Smekens, S. (2019) Novel pipeline identifies
16 new upstream ORFs and non-AUG initiating main ORFs with conserved amino acid
17 sequences in the 5' leader of mRNAs in *Arabidopsis thaliana*. *RNA*, **25**, 292-304.
- 18 45. Chang, K.S., Lee, S.H., Hwang, S.B. and Park, K.Y. (2000) Characterization and translational
19 regulation of the arginine decarboxylase gene in carnation (*Dianthus caryophyllus* L.).
20 *Plant J*, **24**, 45-56.
- 21 46. Combier, J.P., de Billy, F., Gamas, P., Niebel, A. and Rivas, S. (2008) Trans-regulation of the
22 expression of the transcription factor MtHAP2-1 by a uORF controls root nodule
23 development. *Genes Dev.*, **22**, 1549-1559.
- 24 47. Leutz, A., Barbosa-Silva, A., Andrade-Navarro, M.A. and Wethmar, K. (2013) uORFdb—a
25 comprehensive literature database on eukaryotic uORF biology. *Nucleic Acids Res.*, **42**,
26 D60-67.
- 27 48. Kearse, M.G. and Wilusz, J.E. (2017) Non-AUG translation: a new start for protein synthesis
28 in eukaryotes. *Genes Dev.*, **31**, 1717-1731.

29

Figure 1

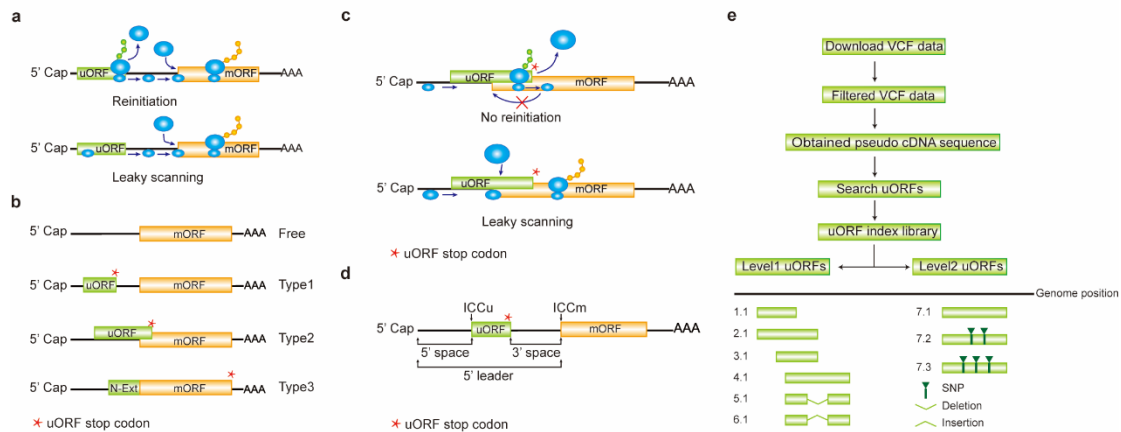


Figure 1. Computation of uORF variation. **a**, Reinitiation and leaky scanning models. In the reinitiation model, the 80S ribosomal subunit will separate after translating the uORF and the 40S subunit remains associated with mRNA, regaining fresh eIF2 ternary complex and other unknown reinitiation factors to translate the mORF. In the leaky scanning model, the uORF initiation codon is bypassed by the scanning complex, which will ignore the uORF and translate the mORF. **b**, uORF types. Genes without uORFs are grouped as ‘Free’. uORF are further divided into Types1-3 with respect to the position of uORF stop codon relative to the mORF. N-Ext, N extension. **c**, Type2 uORF-controlled mORF translation is only favored by leaky scanning. Overlap between the Type2 uORF and mORF makes reinitiation of the mORF after translation of the uORF impossible. **d**, uORF positional information on cDNA. The mORF is flanked by the 5’ leader and 3’ UTR (3’ untranslated region). 5’ and 3’ space are used to describe the distance from Cap to uORF AUG and from uORF stop codon to mORF AUG, respectively. The sequence from -3 to +4 relative to the AUG initiation codon (A as +1) corresponding to the Kozak consensus (A/GCCAUGG) position is termed as initiation codon context (ICC) with ICCu and ICCm for uORF and mORF, respectively. **e**, Workflow to identify uORF variants.

Figure 2

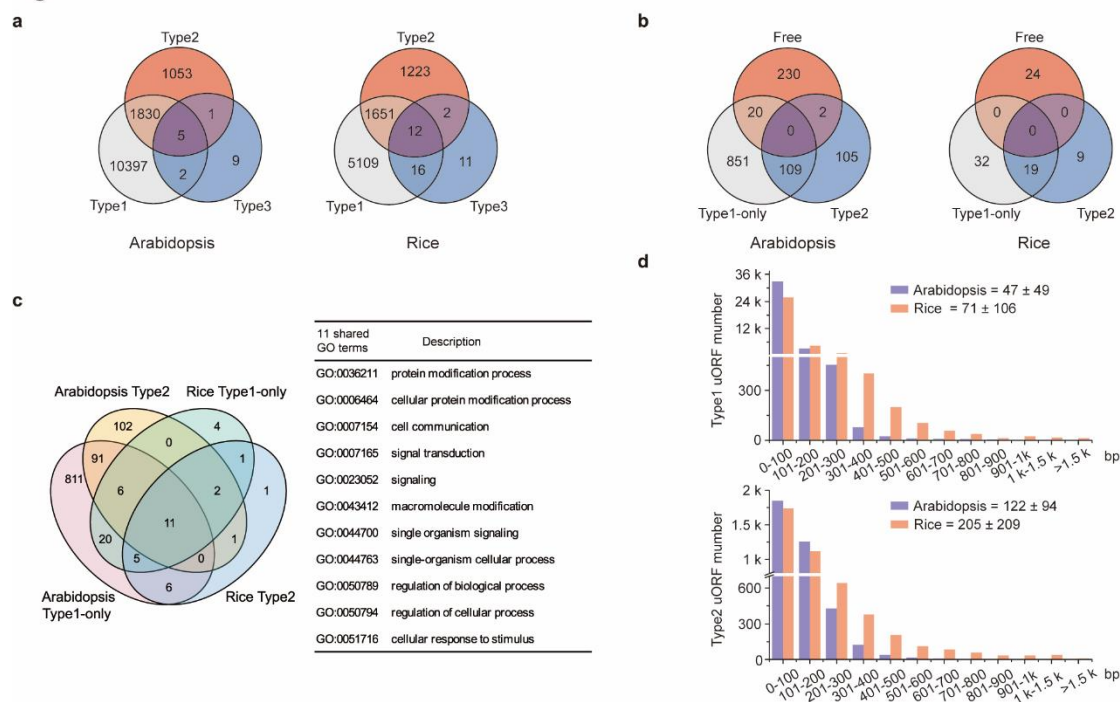


Figure 2. Characterization of uORFs in Arabidopsis and rice. **a**, Venn diagram of uORF type distribution among uORF-containing genes. **b**, Venn diagram of enriched Gene Ontology (GO) terms of uORF-Free-, Type1-only- and Type2- uORF-containing genes in Arabidopsis and rice. GO terms of FDR (false discovery rate) < 0.05 are used. **c**, Venn diagram to show GO-terms shared in Arabidopsis and rice uORF-containing genes. 51 out of 57 enriched GO-terms of rice uORF-containing genes are also found in those of Arabidopsis. 11 GO-terms enriched in both Type1-only and Type2 uORF-containing genes of Arabidopsis and rice are detailed in the table. **d**, Length distribution of Type1 and Type2 uORFs. Insert number shows Mean±SD of uORF length (bp). The representative gene models of Arabidopsis reference accession Col-0 and rice reference cultivar Nipponbare are used for analysis.

Figure 3

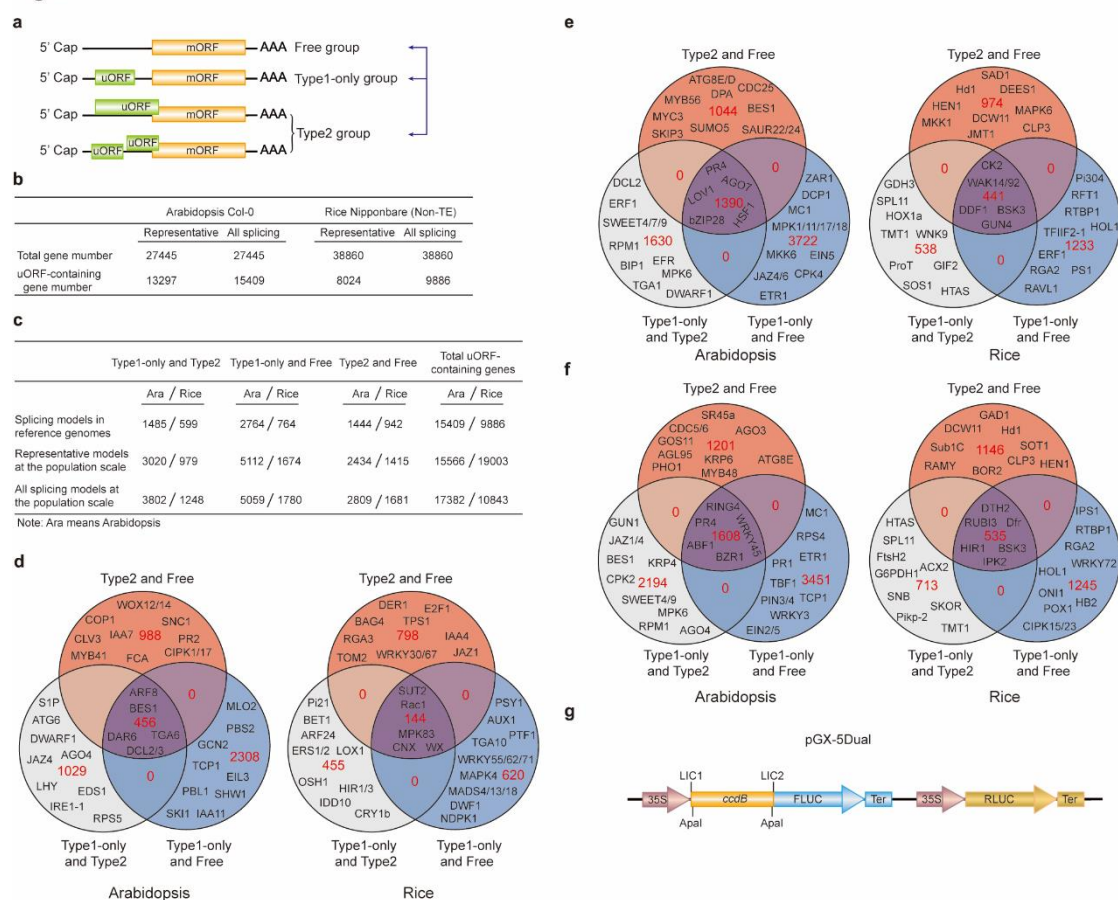


Figure 3. uORF variation. **a**, Schematic of altered uORF types analyzed in this study. Three Level1 uORF variants are focused by studying changes between Type1-only and Type2, between Type1-only and uORF-Free, and between Type2 and uORF-Free. Genes with both Type2 and Type1 on the same splicing model are grouped as Type2 in this analysis. **b**, Number of uORF-containing genes in Arabidopsis Col-0 and rice Nipponbare. The representative and all splicing gene models are calculated. Non-TE, none transposon element. **c**, Number of three Level1 uORF variants. Analysis is performed among different splicing models of reference genomes (Arabidopsis Col-0 and rice Nipponbare), and among accessions at the population scale using the representative and all splicing gene models. **d**, Venn diagram of three Level1 uORF variants within different splicing models of Arabidopsis Col-0 and rice Nipponbare. **e**, **f**, Venn diagram of three Level1 uORF variants of the representative gene models (e) and all splicing gene models at the population level (f). Example gene names are shown in different regions. **g**, Schematic of Dual-Luciferase vector pGX-5Dual. 5' leader

1 containing different uORF variants are cloned via LIC (ligation-independent cloning)
2 method to control the translation of firefly luciferase (FLUC). ApaI is used to linearize
3 the vector. Comparison of the ratio of FLUC activity and mRNA level to the internal
4 control Renilla luciferase (RLUC) expressed from the same vector will indicate the
5 effects of uORF variants on mRNA stability and translation efficiency.