Supplemental Material

Supplemental Methods

Supplemental Figures

Supplemental Figure S14. The distribution of 5′UTR lengths for genes with and without 5′UTR AUG and nonAUG TISs in tomato and *Arabidopsis*.

Supplemental Figure S15. The correlation of the feature values between two different strategies of random sampling.

Supplemental Tables

Supplemental Table S1. Supplementary Table S1. The mean F1 measure, Matthews correlation coefficient (MCC; indicated in column) and Area Under The Curve Receiver Operating Characteristics scores (AUROC, indicated in column) scores showing the performances of models predicting translation initiation sites (TISs) based on the indicated features.

Supplemental Table S2. List of primers used in this study.

Supplemental Methods

Generation of the true-positive (TP) and true-negative (TN) translation initiation sites (TISs)

The LTM treatment blocks the initial rounds of translating ribosomes, resulting in ribosomes stalling at initiation sites, and thus increase the sensitivity and precision of profiling TISs used *in vivo* (Schneider-Poetsch et al. 2010; Lee et al. 2012; Li and Liu 2020). The TP TISs were defined as TISs with significant translation initiation activities and were identified by analysis of the LTM- and CHX-treated ribosome profiling datasets using a TIS-finding pipeline as described previously (Machkovech et al. 2019). The LTM- and CHX-treated ribosome profiling datasets generated from *Arabidopsis* suspension cells and tomato leaves were retrieved from Gene Expression Omnibus database (GSE88790 and GSE143311) (Willems et al. 2017; Li and Liu 2020) and used to profile the translating ribosome positions on transcripts as performed previously (Li and Liu 2020). The TIS-finding pipeline (Machkovech et al. 2019) with default parameters was used to call the TIS peaks by employing the LTM and CHX datasets to identify the TISs used *in vivo*, referred as the TP TISs. Briefly, a zero-truncated negative binomial distribution (ZTNB) was performed to statistically model (1) the background distribution of LTM + CHX pooled counts in genomes (i.e., all non-zero positions with LTM + CHX pooled counts across all transcripts) to obtain a global threshold with a p-value < 0.05 and (2) the background distribution of LTM + CHX pooled counts and CHX counts in transcripts with more than 50 positions with non-zero counts to obtain the local p-values for each position on a transcript. The candidate start site examined was called a TIS peak based on the following criteria: (1) an LTM count and CHX count both >0; (2) an LTM + CHX pooled count > the global threshold; (3) a local p-value of the LTM + CHX pooled counts <0.01 and 1000-fold higher than the local p-value of the CHX counts at the same location or a local p-value of the LTM + CHX pooled counts less than 10^{-7}. The LTM signals in both *Arabidopsis* and tomato LTM-treated Ribo-seq datasets showed significantly enriched at annotated TISs (left panels in Supplemental Fig. S1A). Compared to tomato CHX-treated Ribo-seq datasets (right top panel in Supplemental Fig. S1A), the *Arabidopsis* one showed CHX signals in UTRs (right bottom panel in Supplemental Fig. S1A). Note that the TIS identification pipeline was applied in *Arabidopsis* and tomato, separately. The ML-derived TIS features from Arabidopsis and tomato are comparable and consistent across various samples and species (Supplemental Fig. S1 and Figs. 2-4). The ML models built from tomato can predict the TISs revealed based on the LTM-seq datasets from *Arabidopsis* (Fig. 1) and the CHX-seq datasets from *Arabidopsis*, rice and maize (Fig. 7). The consistency of the observations between species (Figs. 1-4; Supplemental Figs. S1,S3-S6) and the findings were validated via experimental approaches (Fig. 5) suggest that the quality of *Arabidopsis* CHX ribosome profiling did not significantly impact the findings in this study and likely affected the identification of few, if any, TISs.

To generate reliable TN TIS datasets, we focused on AUG and near-cognate codons and only searched for the TN sites in transcripts with identified TP TISs. The candidate codon sites were referred as to TN sites based on the following criteria: (1) LTM and CHX counts both = 0; (2) the site is located in the upstream region of the most downstream TP TIS on the same transcript as descripted previously (Reuter et al. 2016). Note that the TPs and TNs were identified from the same set of the genes. While the 5′ UTR annotation of tomato genes is poor, it may not be a critical factor affecting our findings in this study. In addition, the 5′UTR lengths for the genes with and without 5′UTR-AUG and 5′UTR-nonAUG TISs were similar between *Arabidopsis* and tomato (Supplemental Fig. S14).

The machine-learning workflow

First, we collected 2657 features, comprising 8 known (Kozak and previously reported nearby flanking sequences (Kozak 1984; Kozak 1989; Noderer et al. 2014; Reuter et al. 2016; de Arce et al. 2018; Li and Liu 2020)), 23 ORF (mononucleotide contents and secondary RNA structures upstream of or within ORFs and ORF sizes), and 2626 contextual features (nucleotide/amino acid frequency of *k*-mers in the 200-nt region centered on a TIS) for each TIS (right panel, Fig. 1A) and generated a balanced dataset with equal numbers of TPs and TNs by random sampling (Fig. 1D, step 1, see Methods). Second, we imputed and scaled the feature data to make them comparable. Third, we selected the final feature set for training (70%) and testing (30%) data based on the correlation between features and the significance of enrichment between TPs and TNs, as too many features can interfere with prediction performance (Bzdok et al. 2018) (Fig. 1D, step 2, see Methods). Fourth, we generated ML models using four algorithms and compared their performance to assess the predictive power of different feature categories (Fig. 1D, step 3, see Methods). To reveal the robustness of the ML prediction performance and the important features identified, we run this ML workflow with the 10 randomly balanced TP and TN datasets (Fig. 1D).

Feature collection

Feature collection was based on a previous report (Reuter et al. 2016) with slight modifications unless specified otherwise. The definition, generation and slight modification of known/open-reading frame (ORF)/contextual/TIS codon usage feature categories are described below.

-- Known features:

(1) Position-weight-matrix (PWM): multiple PWM-related features representing the relationship between the flanking sequence context and the TIS translational efficiency were determined as follows. First, the features "PWM_{TP}" and "$\text{PWM}_{\text{annotated}}$" were determined by a PWM matrix generated based on the flanking sequences (positions -15 to +10) of a given TP TIS group and of the annotated TISs with *in vivo* translation initiation activity. Second, the feature "Noderer translational efficiency", representing the relationship between the flanking sequences (position -6 to +5) and the TIS translational efficiency in mammalian cells, was derived from a previous report (Noderer et al. 2014). In brief, all possible flanking sequence (positions -6 to +5) contexts around the AUG translational start were identified as features.

(2) Kozak sequence context: The Kozak sequence context was discretized into strong (A or G at -3 and G at +4), intermediate (A or G at -3 and no G at +4), weak (no A and no G at -3 and G at +4), and no Kozak context. These categories were presented as the values 1 (no), 2 (weak), 3 (intermediate), and 4 (strong).

-- ORF features:

(1) ORF length: arbitrary start sites in the mRNA sequence, the lengths of the TP/TN TIS- and annotated TIS-initiated ORFs, and the A/T/C/G mononucleotide contents in their upstream regions were considered.

(2) Minimum free energy (MFE) of mRNA secondary structure: the MFEs in the 80-bp regions centered on TISs were calculated using the RNAfold program (Lorenz et al. 2011) in a sliding window with a 20-nt window size and a 10-nt step size. In addition, to summarize the magnitudes of the MFE difference, the 20-bp upstream and downstream regions flanking the TISs were also computed by normalizing the MEF values of the region at positions -20 to +0 to those of the regions at positions -10 to +10 and the MEF values of the region at positions +10 to +30 to those of the regions at positions +0 to +20.

-- Contextual features

We counted the frequency of all possible *k*-mers of length *k* = 1 (position-specific *k*-mers) and *k* = 3 (codon and respective amino acid *k*-mers) in a window from -99 to +99 around the start site. The *k*-mers were defined as all possible combinations of subsequence of length *k*, given one of the four nucleotides A, U, C, and G. The in-frame and out-of-frame *k*-mers as well as *k*-mers upstream and/or downstream of the start site were considered. In addition, we also considered the frequency of all possible amino acids with length =1 and 2 in the 99-nt regions downstream of the start site and within a TIS-initiated ORF. These generated 2626 contextual features in total.

-- TIS codon usage

The TIS codon usage of all 64 codons was determined as described previously (Zhang et al. 2017). Briefly, for a given codon, the proportion of the target codon sites among all the identified TP TISs were normalized to the proportion of the target codon sites among all codon sites found in the transcript regions of all annotated genes; then the corresponding $\log_2$ ratio was computed and referred to as the feature value of the TIS codon usage bias.

Feature selection and data imputation

The size of the TN/TP dataset was balanced via random sampling without the replacement to contain the same number of TN and TP sites by randomly under-sampling from the larger dataset. We also generated balanced TN datasets by randomly undersampling the TNs with replacement (i.e., the bootstrapping) and observed a strong correlation of feature value (Supplemental Figure S15), suggesting that these two underdamping approaches (i.e., with and without replacement) would not bias the findings. Thus for the following analyses, we employed the strategy of random sampling without the replacement.

We applied Wilcoxon-rank sum test and Bonferroni correction for all features to test for the statistical significance of differences between TP and TN sites. We then calculated Pearson correlation coefficients among the contextual and ORF features. The 50 most significant (with smallest adjusted *p*-value) and uncorrelated (r < 0.7) contextual features and the uncorrelated (r < 0.7; adjusted *p*-value < 0.01) ORF features were kept for the model training step. All features were normalized and scaled to ensure comparability. We repeated this selection 10 times to evaluate the model robustness for each TIS group.

Reference

Bzdok D, Altman N, Krzywinski M. 2018. Statistics versus machine learning. *Nat Methods* **15**: 233-234.

de Arce AJD, Noderer WL, Wang CL. 2018. Complete motif analysis of sequence requirements for translation initiation at non-AUG start codons. *Nucleic Acids Res* **46**: 985-994.

Kozak M. 1984. Point mutations close to the AUG initiator codon affect the efficiency of translation of rat preproinsulin in vivo. *Nature* **308**: 241-246.

Kozak M. 1989. Context Effects and Inefficient Initiation at Non-Aug Codons in Eukaryotic Cell-Free Translation Systems. *Mol Cell Biol* **9**: 5073-5080.

Lee S, Liu B, Lee S, Huang SX, Shen B, Qian SB. 2012. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proceedings of the National Academy of Sciences of the United States of America* **109**: E2424-2432.

Li YR, Liu MJ. 2020. Prevalence of alternative AUG and non-AUG translation initiators and their regulatory effects across plants. *Genome Res* **30**: 1418-1433.

Lorenz R, Bernhart SH, Honer Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA Package 2.0. *Algorithms Mol Biol* **6**: 26.

Machkovech HM, Bloom JD, Subramaniam AR. 2019. Comprehensive profiling of translation initiation in influenza virus infected cells. *Plos Pathog* **15**.

Noderer WL, Flockhart RJ, Bhaduri A, de Arce AJD, Zhang JJ, Khavari PA, Wang CL. 2014. Quantitative analysis of mammalian translation initiation sites by FACS-seq. *Molecular Systems Biology* **10**.

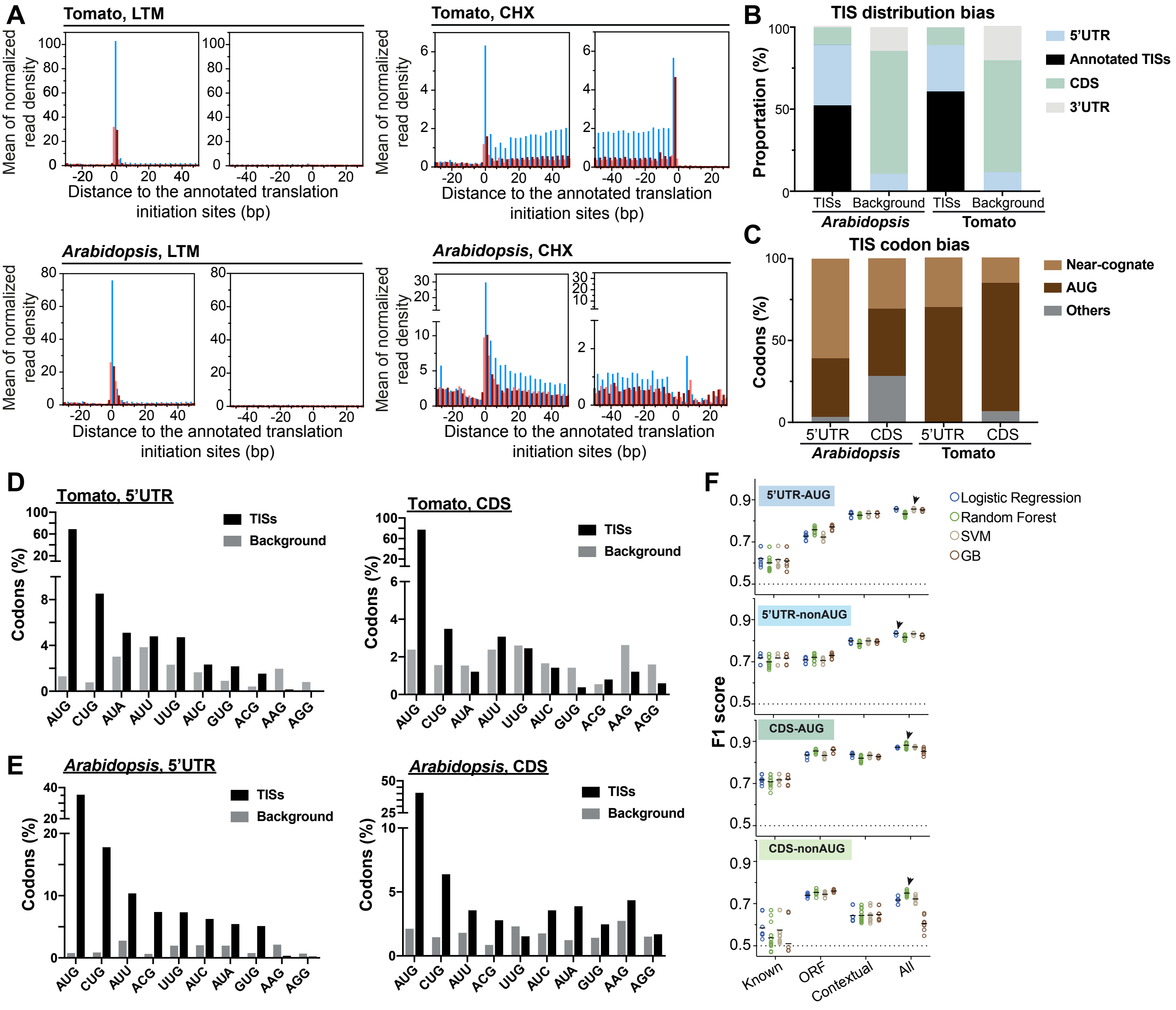
Reuter K, Biehl A, Koch L, Helms V. 2016. PreTIS: A Tool to Predict Non-canonical 5' UTR Translational Initiation Sites in Human and Mouse. *PLoS Comput Biol* **12**: e1005170.

Schneider-Poetsch T, Ju J, Eyler DE, Dang Y, Bhat S, Merrick WC, Green R, Shen B, Liu JO. 2010. Inhibition of eukaryotic translation elongation by cycloheximide and lactimidomycin. *Nat Chem Biol* **6**: 209-217.

Willems P, Ndah E, Jonckheere V, Stael S, Sticker A, Martens L, Van Breusegem F, Gevaert K, Van Damme P. 2017. N-terminal Proteomics Assisted Profiling of the Unexplored Translation Initiation Landscape in *Arabidopsis* thaliana. *Mol Cell Proteomics* **16**: 1064-1080.

Zhang S, Hu H, Jiang T, Zhang L, Zeng J. 2017. TITER: predicting translation initiation sites by deep learning. *Bioinformatics* **33**: i234-i242.

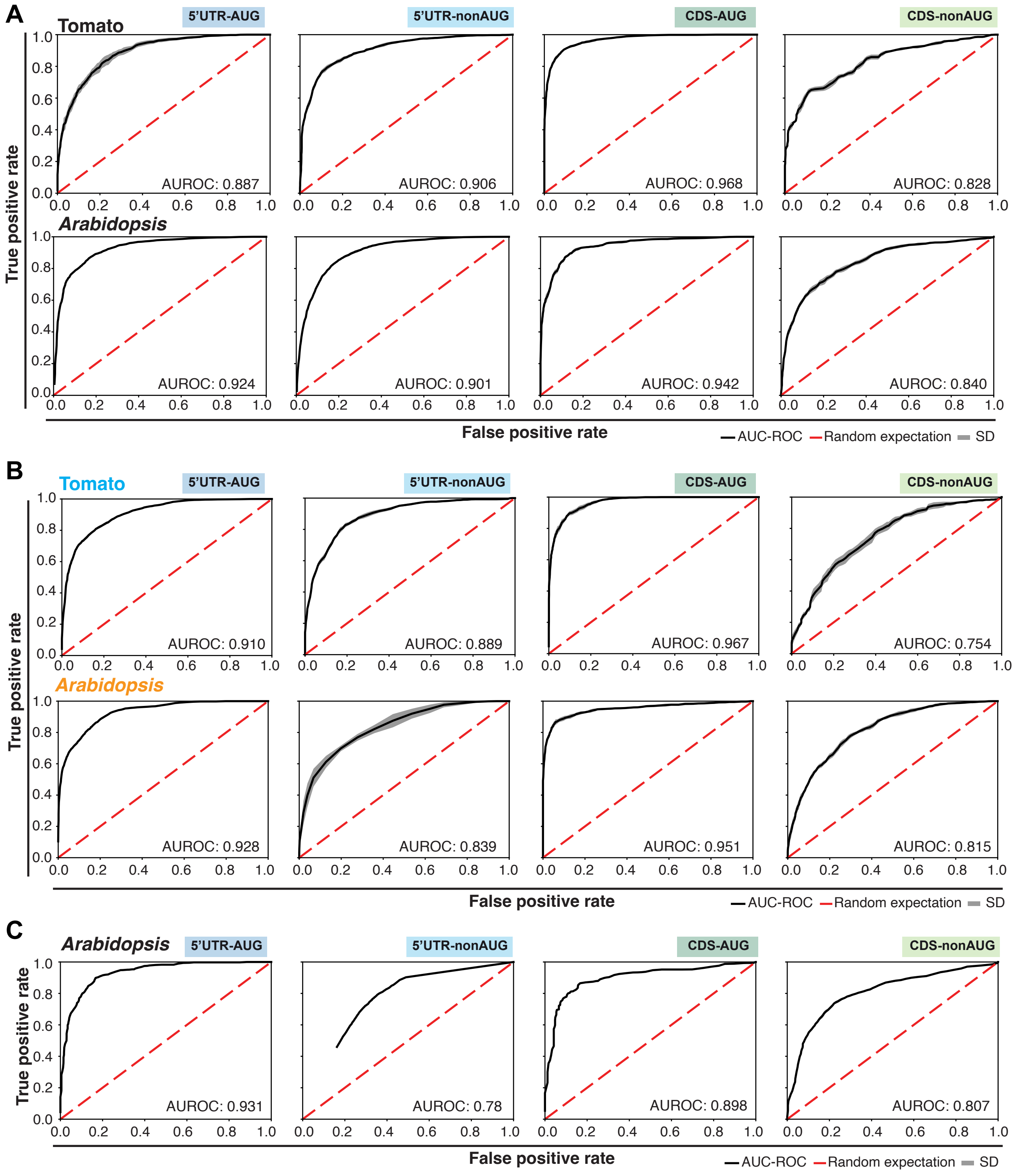
Supplemental_Figure_S1



Supplemental Figure S1. The characteristics of the identified translation initiation sites (TISs) in tomato and *Arabidopsis*.

(A) Metagene plots of mean normalized read densities in regions around translation initiation sites and translation stop sites of genes calculated for LTM-treated (left) and CHX-treated (right) samples generated from tomato and *Arabidopsis*. Normalized read density of a gene was calculated by normalizing the read count per base to the average read density for the entire CDS. Blue, green and orange bars indicate the reads whose assigned P-sites map to the codon positions 0, 1, and 2 (i.e. phases 1–3, respectively) relative to the annotated AUG codon. (B) The positional distributions of the identified TISs mapping to different genic regions of transcripts including the 5' UTRs, annotated AUG TIS sites, CDS and 3' UTRs. Background: the locations of all 64 codons in transcripts. (C) The codon compositions of the identified TISs located in the 5′ UTRs and CDS. Near-cognate: the codons that differ from AUG by one base. Others: codons other than AUG and near-cognate codons. (D) The codon compositions of the identified TISs containing AUG and near-cognate codons for the tomato TISs located in the 5′ UTRs (left) and CDS (right). Background: the codon compositions based on all transcript sequences. (E) As described in (C), but for the identified TISs in *Arabidopsis*. (F) The F1 scores showing the performances of four different ML algorithms based on different sets of features in predicting the four *Arabidopsis* TIS groups (i.e., the AUG and nonAUG TISs located in the 5′ UTRs and CDSs). Circle: the performance of a model on a randomly balanced TP and TN TIS dataset; black line: mean of the F1 scores for a given ML algorithm. Arrow: the best model (i.e., for the ML algorithm with highest mean performance, the model with highest F1 score); dashed line: the baseline performance expected by random guessing. The corresponding AUC-ROC curves and the AUROC and MCC scores were showed in Supplemental Fig. S2.

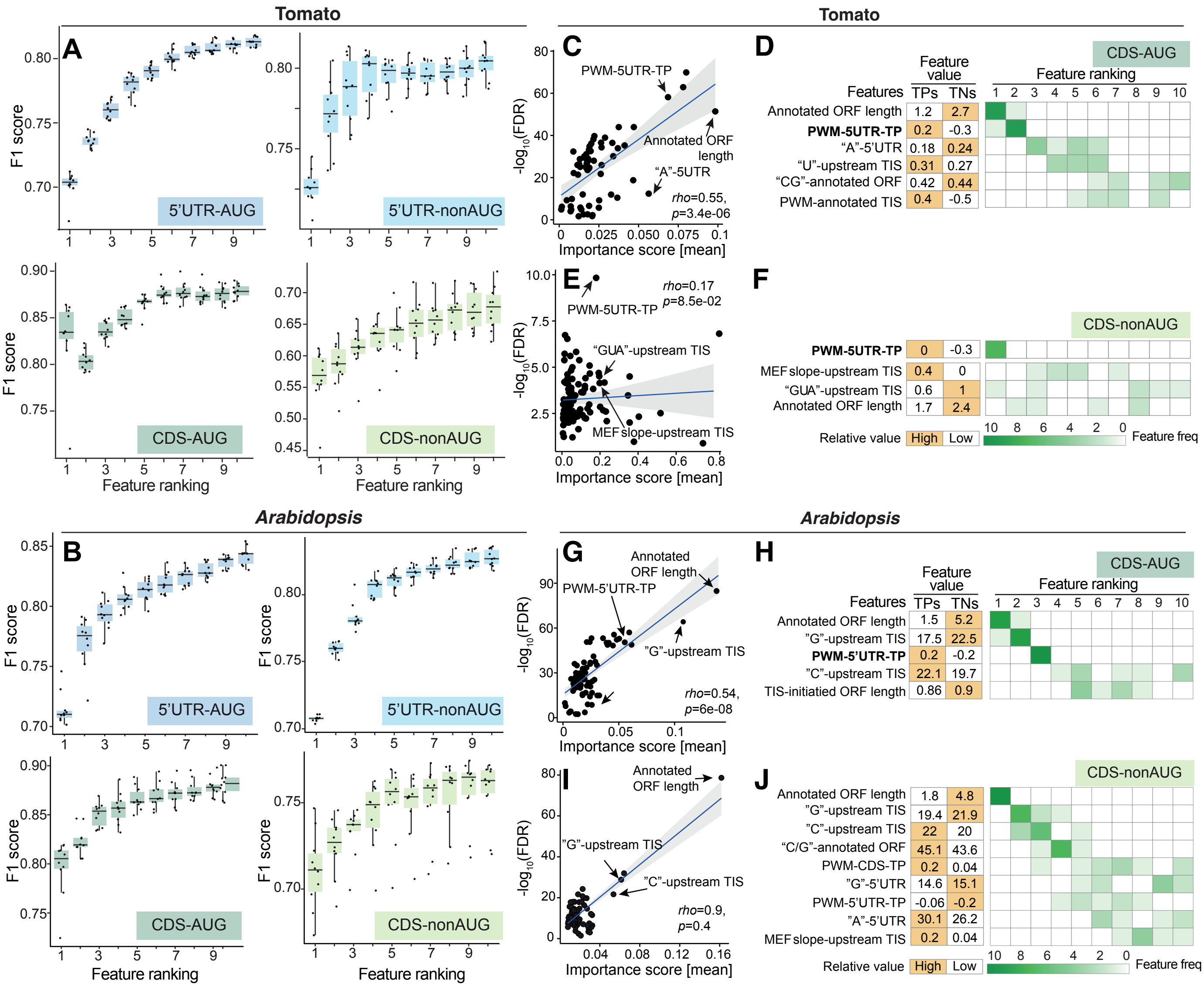
Supplemental_Figure_S2



Supplemental Figure S2. The AUC-ROC (Area Under The Curve Receiver Operating Characteristics) curves showing the performance of models in predicting four TIS groups.

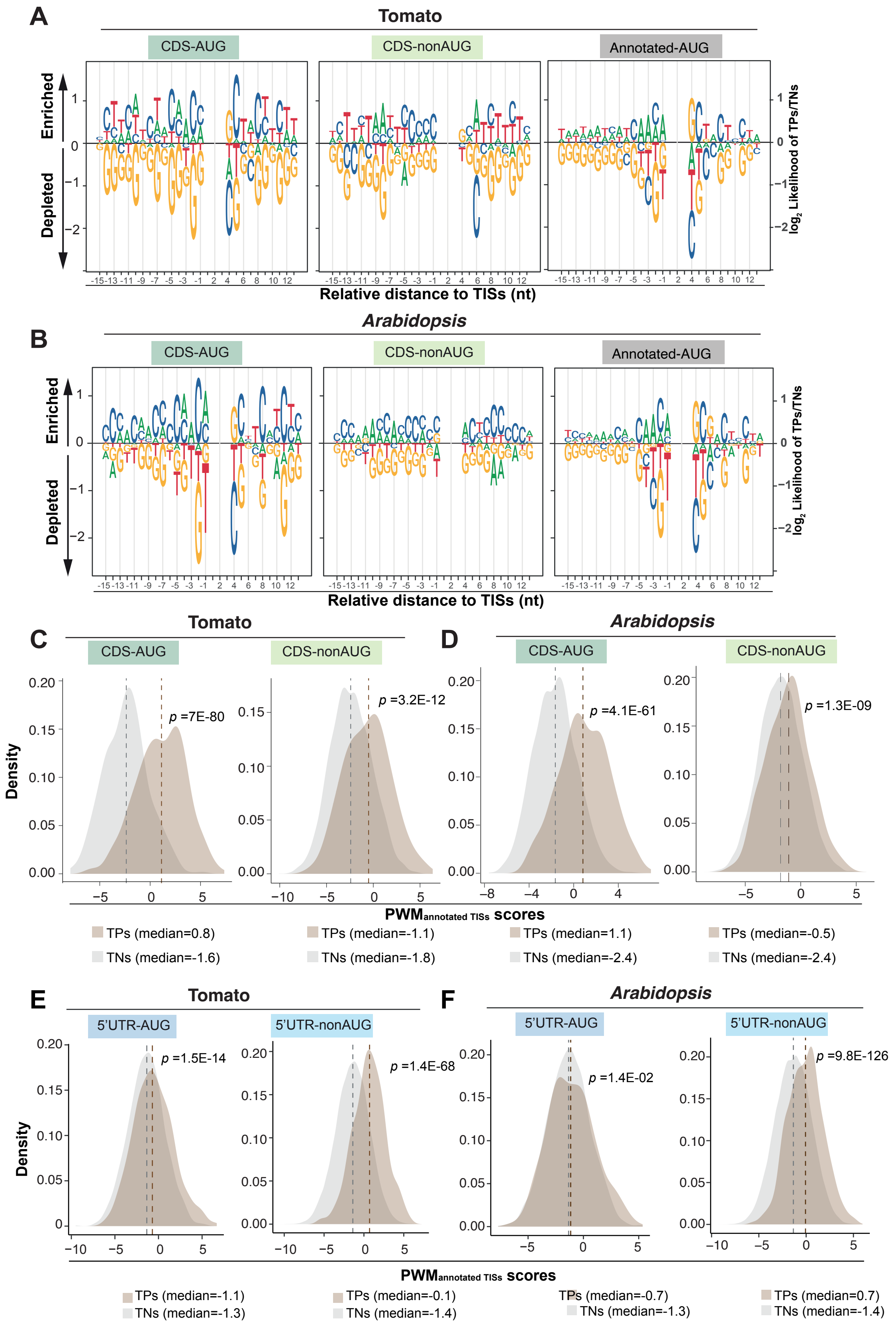
(A) The AUC-ROC curves for the best models (indicated by the arrows in Fig. 1 and Supplemental Fig. S1F) employing all features to predict the 5′ UTR -AUG, 5′ UTR -nonAUG, CDS -AUG and CDS -nonAUG TIS groups in tomato and *Arabidopsis*. (B) As described in (A), but the cross-species prediction performance of the models indicated in Fig. 1F. Tomato (light blue:) using the best model built in Arabidopsis to predict the TISs in tomato. *Arabidopsis* (orange:) using the best model built in tomato to predict the TISs in Arabidopsis. (C) As described in (A), but for the best models (indicated by the arrows in Fig. 4) employing putative *cis*-elements to predict TISs.

Supplemental_Figure_S3



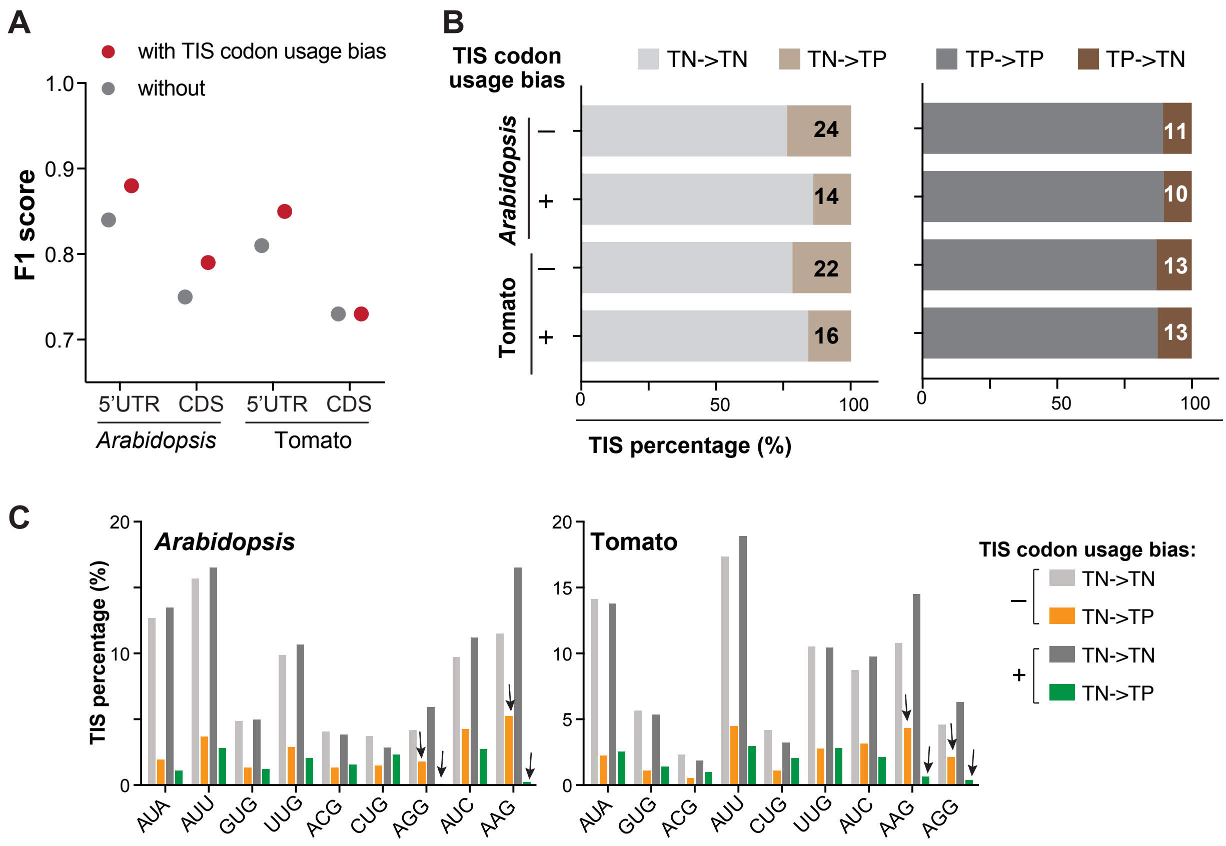
Supplemental Figure S3. The features that were most informative for predicting TISs in tomato and *Arabidopsis*. (A,B) Feature Elimination analyses of selecting the top 10 features with the highest importance to the performance of the best model. Boxplots present the F1 scores generated from 10 balanced datasets (black dots) randomly chosen from the four TIS groups in tomato (A) and *Arabidopsis* (B); the median values, the first and third quartiles, and whiskers of maximum and minimum values are shown. (C) Comparison of the importance scores derived from the best model and the statistical significance of differences (-log$_{10}$(FDR), determined by Wilcoxon signed-rank test) between tomato CDS-AUG TPs and TNs for the features used in the best model. Rho: Spearman's rank correlation coefficient. The black line indicates the fitted linear regression line, and the gray area indicates the 95% confidence level interval. (D) The means of the feature values in the tomato CDS-AUG TP and TN datasets (left) and the frequency of the Feature Elimination-determined top10 features (ranked using their importance revealed in (A)) identified in 10 randomly balanced datasets (left). The rank and frequency indicate the importance of a given feature in the prediction model and their robustness using 10 randomly balanced datasets. The features with frequency >7 within the top 10 are shown. Orange indicates the TIS group with the higher feature value. (E-J) As described in (c,d), but for the tomato CDS-nonAUG TIS group (E,F), the *Arabidopsis* CDS-AUG TIS group (G,HS) and *Arabidopsis* CDS-nonAUG TIS group (I,J).

Supplemental_Figure_S4



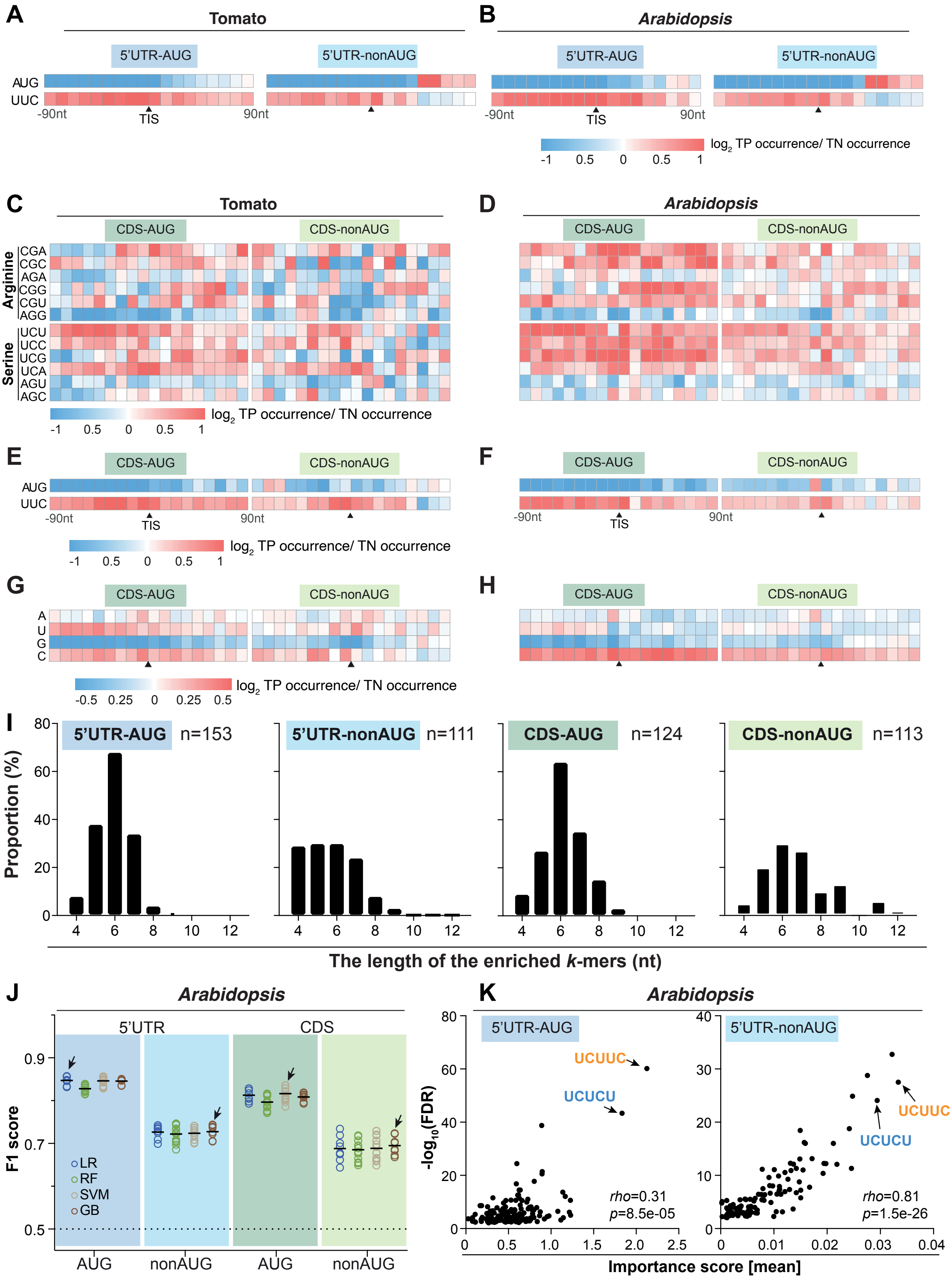
Supplemental Figure S4. The C/U nucleotide compositions and the flanking sequences of CDS TISs. (A,B) Sequence logo plots showing the differential enrichment of A/U/C/G nucleotides between TPs and TNs in the region 15-bp upstream and 13-bp downstream of the CDS-AUG and CDS-nonAUG TISs, represented as the log$_2$ ratio of the site frequencies between TPs and TNs, in tomato (A) and *Arabidopsis* (B). (C,D) Position-weight matrix (PWM) scores showing the sequence similarity of the TIS-flanking regions between TPs (brown)/TNs (gray) and annotated TISs for the tomato (C) and *Arabidopsis* CDS TISs (D). PWMannotated TISs matrix was determined based on the annotated TISs with in vivo translation initiation activities (See Supplemental Methods). P-values derived from Mann–Whitney U test were used to evaluate the significance of differences of PWM scores between the TP and TN datasets. Dashed line indicates the median value. (E,F) As indicated in (C,D), but for the 5'UTR-AUG and 5'UTR-nonAUG TISs.

Supplemental_Figure_S5

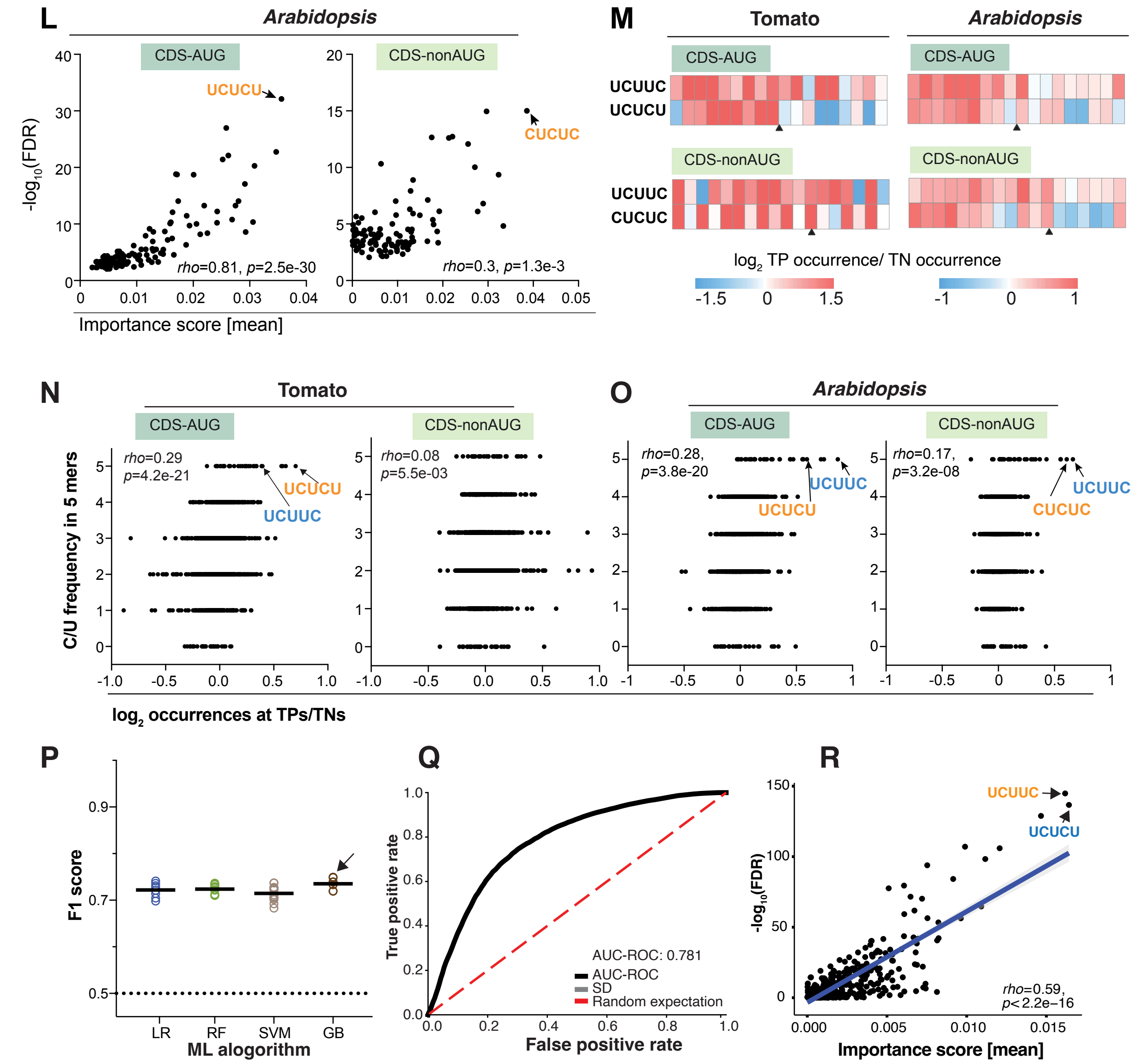


Supplemental Figure S5. The inclusion of TIS codon usage bias as a feature improves prediction of nonAUG TISs in tomato and *Arabidopsis*.
(A) Comparison of the prediction performances with (red) and without (gray) TIS codon usage bias as a feature for the 5′ UTR-nonAUG and CDS-nonAUG TIS groups in *Arabidopsis* and tomato. As described in Fig. 1E, but for the median F1 scores of the ML algorithms with the highest performance. (B) Box plots showing how the TIS codon preference feature affects the proportion of the mis-predicted TISs in the TNs (TN→TP; light brown, left) and TPs (TP→TN; dark brown, right) of the 5′ UTR-nonAUG TIS groups. (C) As described in (B), but for individual near-cognate codons. Arrows: the top two codons with the highest degree of difference in mis-prediction with/without using the feature of TIS codon usage bias (orange vs. green).

Supplemental_Figure_S6



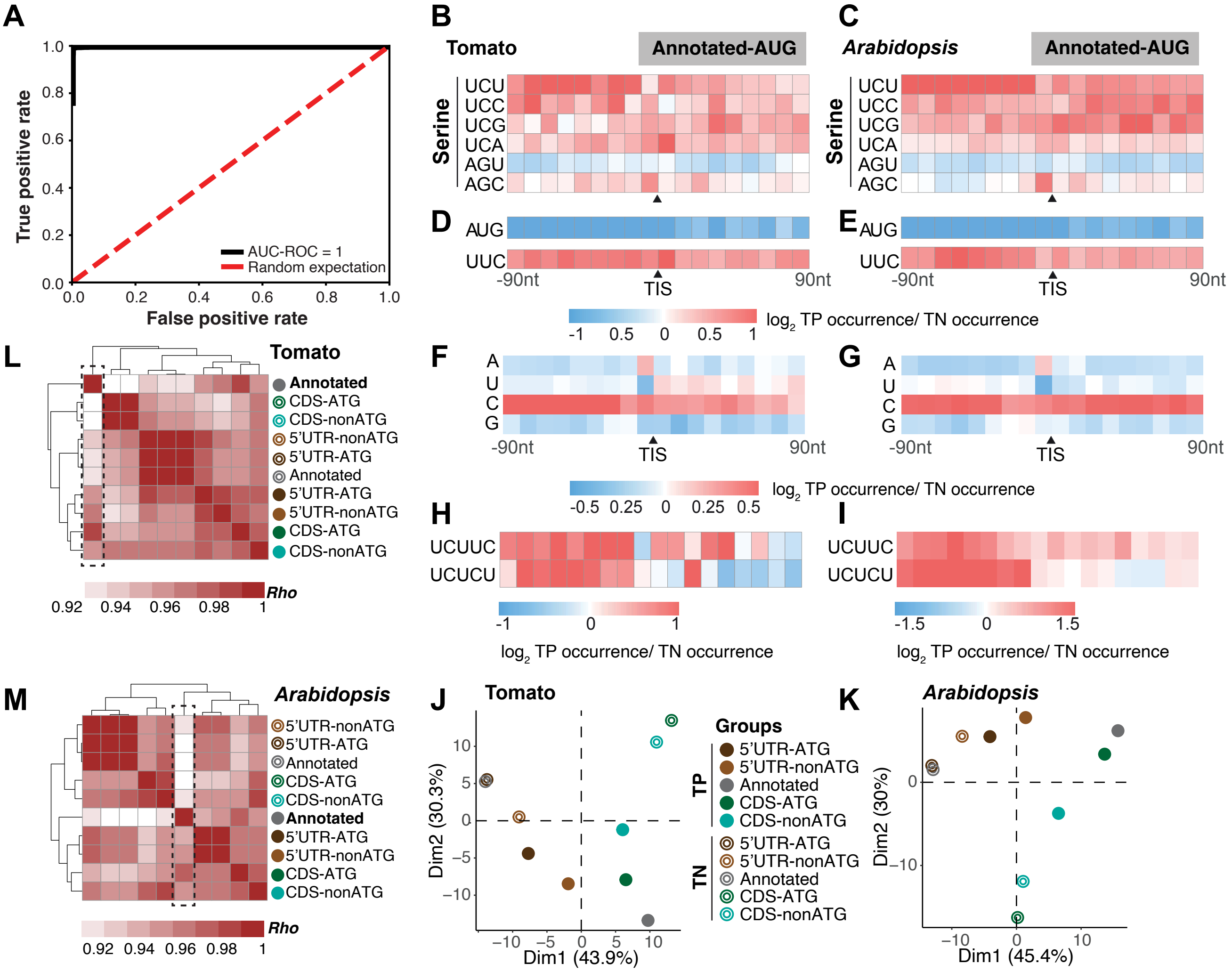
Supplemental_Figure_S6 (continued)



Supplemental Figure S6. The C/U nucleotide enrichments in the flanking sequences of tomato and *Arabidopsis* TISs and the prediction performance of enriched k-mers in all *Arabidopsis* TISs.

(A,B) Enrichment of sites with the indicated 3-mer sequences in the 5'UTR-AUG and 5'UTR-nonAUG TIS groups, represented as the $\log_2$ ratio of the site frequencies between TPs and TNs in the 180-bp region centered on TISs with a 10-bp window, in tomato (A) and *Arabidopsis* (B). (C-H, M) As described in (A,B), but for the CDS-AUG and CDS-nonAUG groups (C,D,E,F), the A, U, C, and G mononucleotides (G,H) and the "UCUUC" and "UCUCU" sequences (M). (I) The length distribution of the putative RNA cis-elements identified from the *k*-mer enrichment searching pipeline in four TIS groups in *Arabidopsis*. (J) As described in Fig. 1E, but for the model employing the identified putative RNA *cis*-elements to predict the four *Arabidopsis* TIS groups. Arrows: the best model. (K,L) As described in Fig. 2A, but for the putative RNA *cis*-elements in predicting the *Arabidopsis* 5′ UTR-AUG and 5′ UTR-nonAUG TIS groups (K) as well as CDS-AUG and CDS-nonAUG groups (L). (N,O) Relationships between the number of C/Us (y-axis) and the enrichment of sequence occurrence (x-axis) for all the 5-mers in the tomato (N) and *Arabidopsis* (O) CDS-AUG and CDS-nonAUG TIS groups. The enrichment is represented as the $\log_2$ ratio of the median of the sequence occurrence between TPs and TNs in the 200-bp regions centered on TIS sites. Rho: Spearman's rank correlation coefficient. Arrows: the putative *cis*-elements discussed in Fig. 4. (P) As indicated in Fig. 1E, but for the F1 scores showing the performances of four different ML algorithms based on the mixed set of putative *cis*-regulatory RNA elements from 4 TIS groups (n=444) in predicting the mixed set of the four *Arabidopsis* TIS groups. (Q) The AUC-ROC curve for the the performances of the best model in (P). (R) As indicted in Fig. 2A, but for the comparison of the importance scores derived from the model and the statistical significance of differences for features employed in (P).

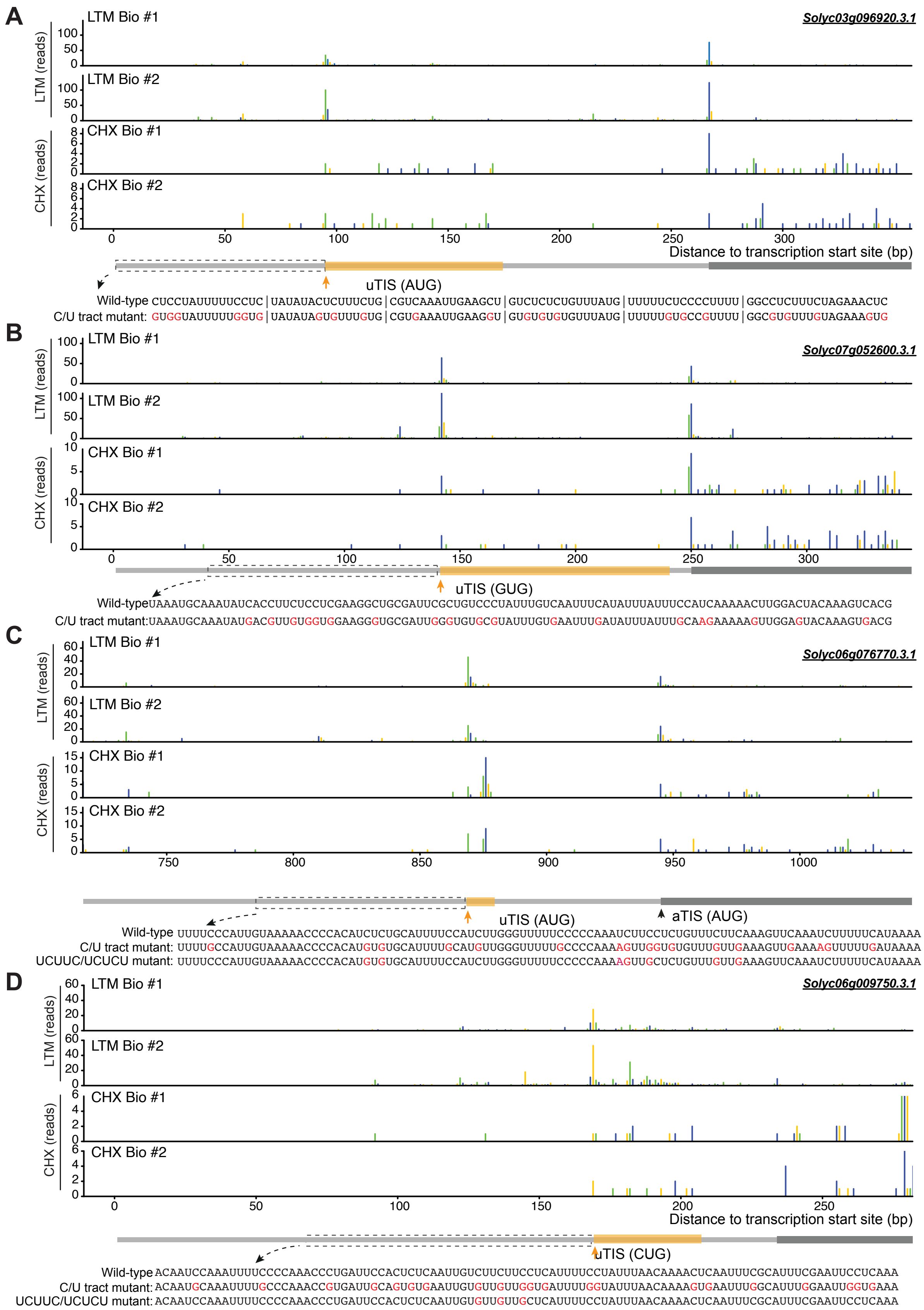
Supplemental_Figure_S7



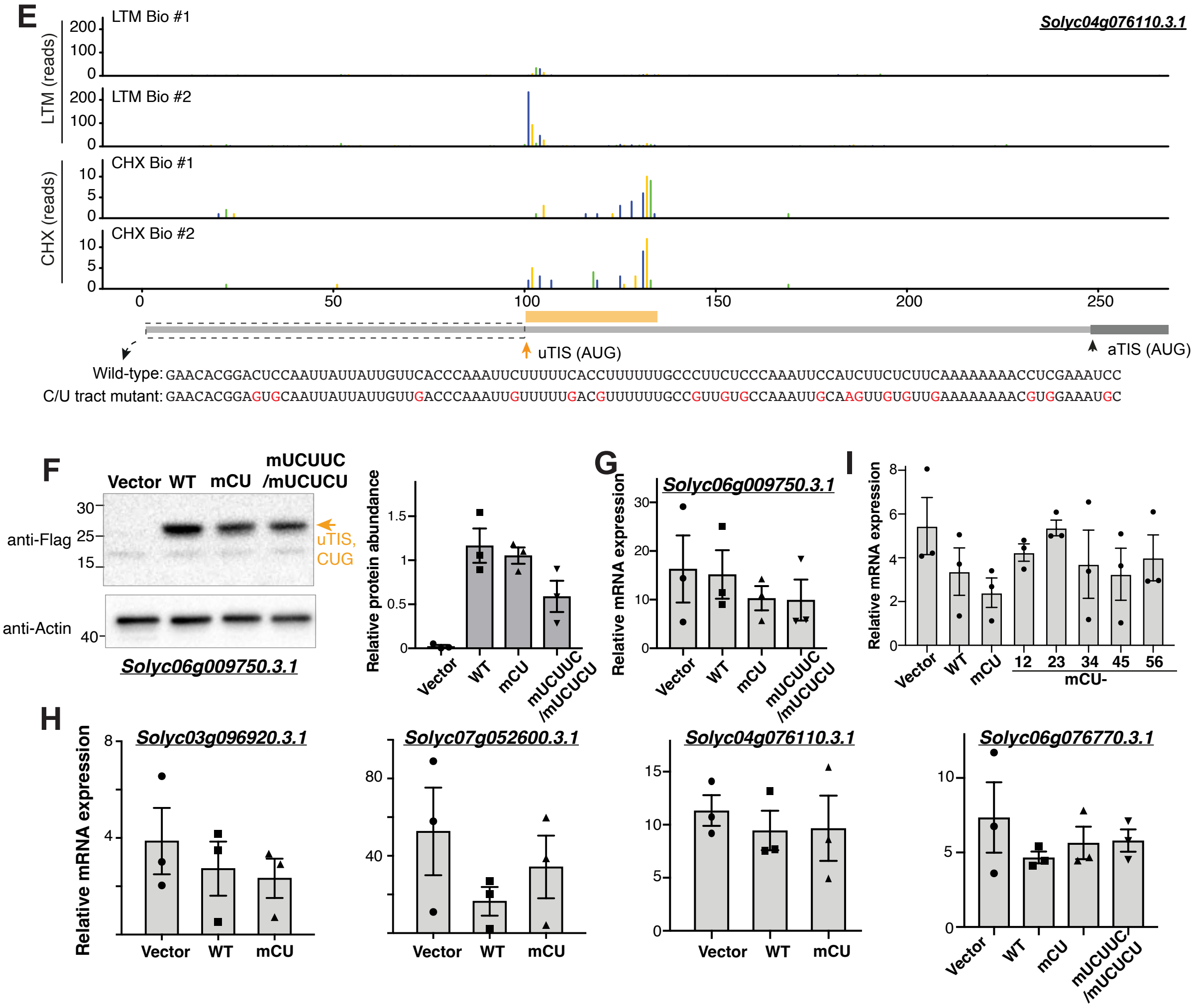
Supplemental Figure S7. The C/U nucleotide compositions and the flanking sequences of toamto and *Arabidopsis* annotated TISs.

(I) The AUC-ROC curve showed the performance of models emplying all features to classify annotated TP AUG TISs against AUG TNs located in 5'UTRs. (B-I) As described in Supplemental Fig. S6 but focusing on the enrichment of the indicated 3-mer sequences (B), the sequences "AUG" and "UUC" (D), the mononucleotides A, U, C, and G (F) and the sequences "UCUUC" and "UCUCU" (H) for the annotated TP TISs (i.e., the annotated TISs with *in vivo* initiation signals) in tomato. (C,E,G,I,) As described in (B,D,F,H), but for the annotated TP TISs identified in *Arabidopsis*. (J,K) The Principal Component Analyses for the TP and TN groups of 5'UTR-AUG, 5'UTR-nonAUG, CDS-AUG and CDS-nonAUG TIS and annotated AUG TISs based on the enriched features identided via ML pipelines in Fig. 1 in tomato (J) and *Arabidopsis* (K). (L,M) The pariwise comparison of the Spearman correlation for feature values of enriched features (indictaed in (J,K)) among the 5 TIS TP and TN groups. Black dashed box: the correlation between the annotated TP TIS (bold) and the rest of TIS TP and TN groups.

Supplemental_Figure_S8

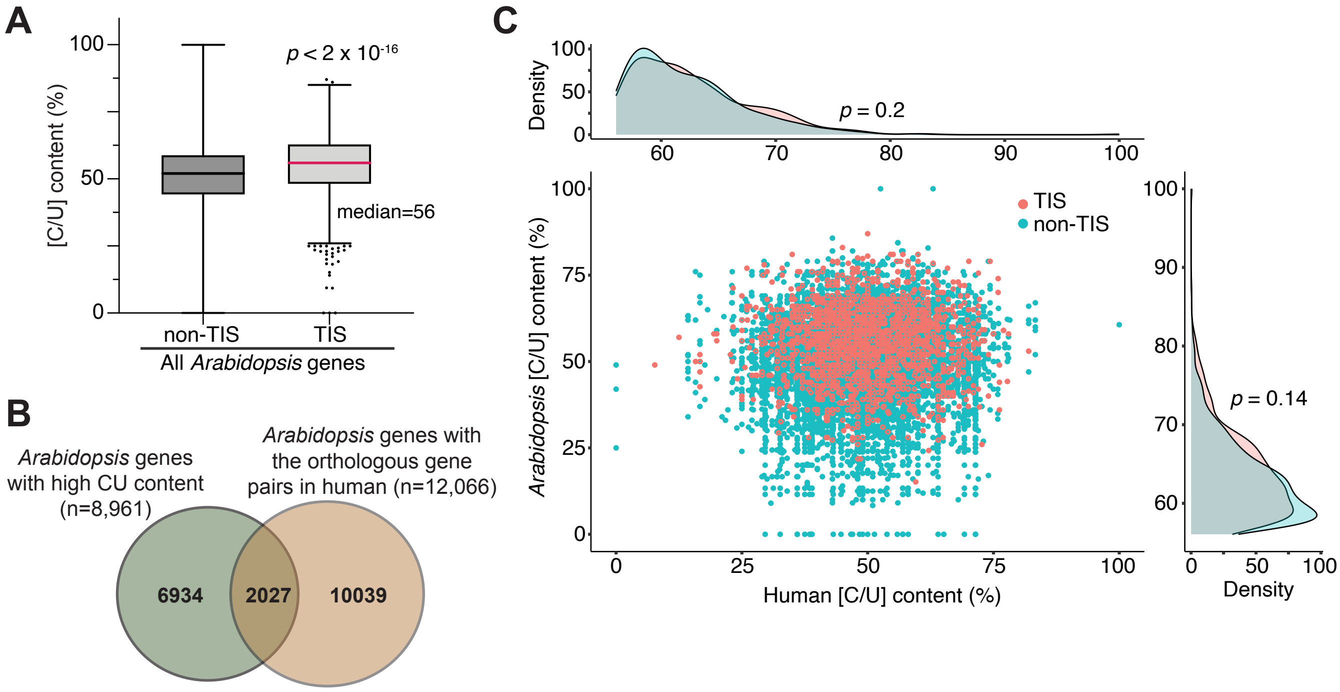


Supplemental_Figure_S8 (continued)



Supplemental Figure S8. *In vivo* initiation activities of misclassified TISs revealed by ribosome profiling
(A) Plots showing the LTM read counts of the indicated genic regions in two biological replicates for Solyc03g096920.3.1, which has a 5′ UTR-AUG TIS (uTIS, AUG; orange arrow) validated previously(Li and Liu 2020). The gene model (bottom) with the UTRs (light gray boxes), annotated CDSs (dark gray boxes), introns (thin lines), annotated TIS (aTIS, black arrow), and the upstream 100-nt wild-type (WT) sequence or sequence with mutations of CT tracts (red) are shown. The upstream 100-nt region was divided into 6 subregions whereas the leftmost one is the 1st subregion as indicated in Fig. 5D. (B) As described in (A), but for Solyc07g052600.3.1, which has a 5′UTR-AUG uTIS site (uTIS, GUG; black arrow) validated previously (Li and Liu 2020). (C) As described in (A), but for Solyc06g76770.3.1, which has a mis-classified 5′ UTR -AUG site (uTIS, AUG; orange arrow). There were statistically significant TIS signals for this TIS in replicate #1 but not in replicate #2 because of the low read counts in replicate #2. Although the reads did not pass the detection threshold, the prediction score of 0.94 passed the prediction threshold of 0.41. (D) As described in (A), but for Solyc06g009750.3.1, which has a mis-classified 5′ UTR-CUG site (uTIS, CUG; orange arrow). The signals for this TIS were statistically significant in replicate #2 but not in replicate #1, because of a low read count in replicate #1, which did not reach the detection threshold, but the prediction score of 0.81 passed the prediction threshold of 0.44. (E) As described in (A), but for Solyc04g76110.3.1 with the indicated mis-classified 5′ UTR-AUG site (uTIS, AUG; orange arrow). The TIS signals were significant in replicate #2 but not in replicate #1, probably because of the low precision of RPF read mapping at the uTIS, but the prediction score of 0.99 passed the prediction threshold of 0.41. (F,G) As indicated in Fig. 5A, but for the protein expression diven by 5′ UTR CUG TIS of Solyc06g009750.3.1 (F) and mRNA abundance of the TIS-containing transcripts (G). (H,I) As indicated in Fig. 5A, but for the mRNA abundance of the TIS-containing transcripts relative to the UBQ3 in transformed plants in Fig. 5A (H) and Fig. 5D (I) and determined by quantitative RT-PCR analyses.

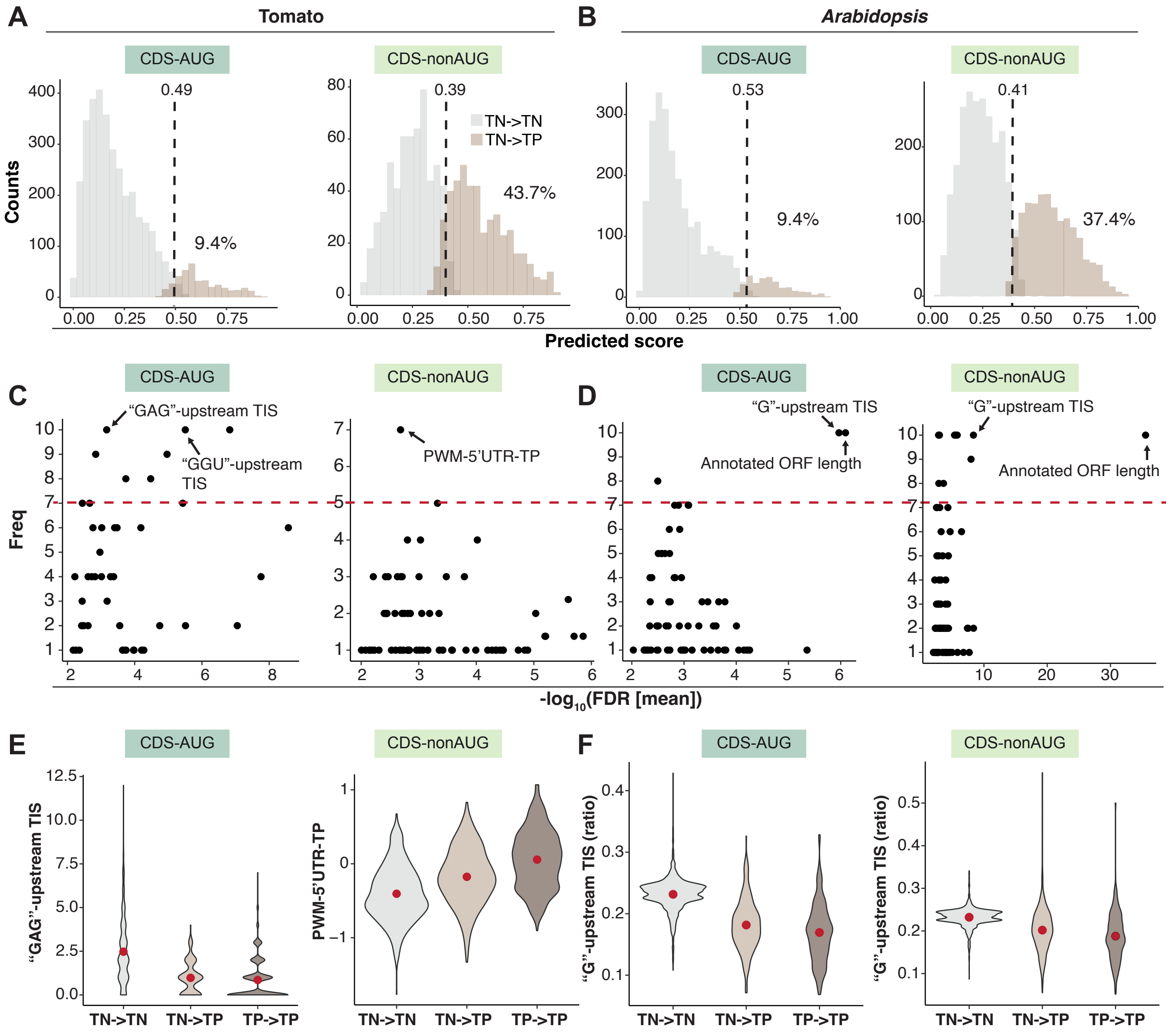
Supplemental_Figure_S9



Supplemental Figure S9. The correlation of CU-rich content around annotated TISs between human and *Arabidopsis*.

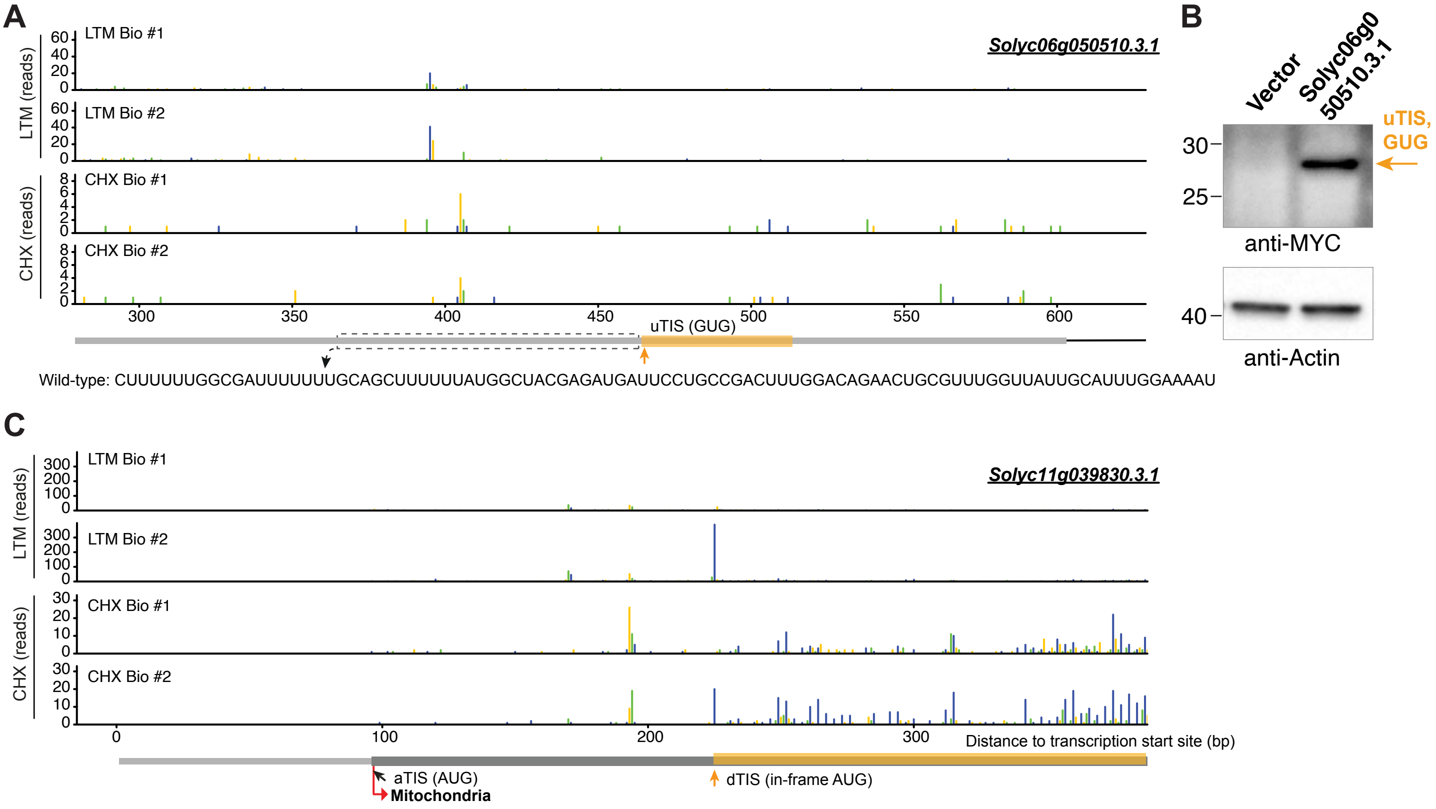
(A) To assesse the relationship between CU-rich content of annotated TISs and the orthologous gene (OG) pairs between *Arabidopsis* and human, we first, by examining the CU-rich contents of the annotated TP TISs (i.e., the annotated TISs with in vivo initiation signals; indicated in black in Supplemental Fig. S1A). We used their median (56%) as the threshold of “CU-rich” content and identified the CU-rich *Arabidopsis* genes. Box plots showed the CU contents (%) of the upstream 100-nt regions of the *Arabidopsis* annotated TISs with (TIS, median=56, n=3,093) and without *in vivo* initiation signals (non-TIS, median=52, n=20,099).(B) With the threshold of CU enrichmeneindicated in (A), we found 2027 out of 8961 CU-rich *Arabidopsis* genes has the OGs in human. The OG list was retrived from the ‘Orthologous Matrix’ database (https://omabrowser.org/oma/home/). (C) Scatter plots and distribution for the CU contents of the OG pairs in *Arabidopsis* (y-axis) and human (x-axis). Shown for the OGs with (TIS, pink, n=1,764) and without *in vivo* initiation signals (non-TIS, pale blue, n=8,616) in *Arabidopsis*. P-values derived from Mann–Whitney U test were used to evaluate the significance of differences. We observed that the CU contents of the annotated TP TISsof *Arabidopsis* OGs were higher than those without initiation signals (pink vs. pale blue; the y-axis). Intriguingly, we further found that the annotated TP TISs of their human OGs also showed marginally higher CT-contents compared to those without TIS activity (x-axis), although the difference was not significant. These results indicate that CU-tract might be a conserved feature for the regulation of translation in both species and the further experimental investigation of the TIS activities on OGs will facilitate to reveal the mechanisms of TIS selections and translation control across different species.

Supplemental_Figure_S10



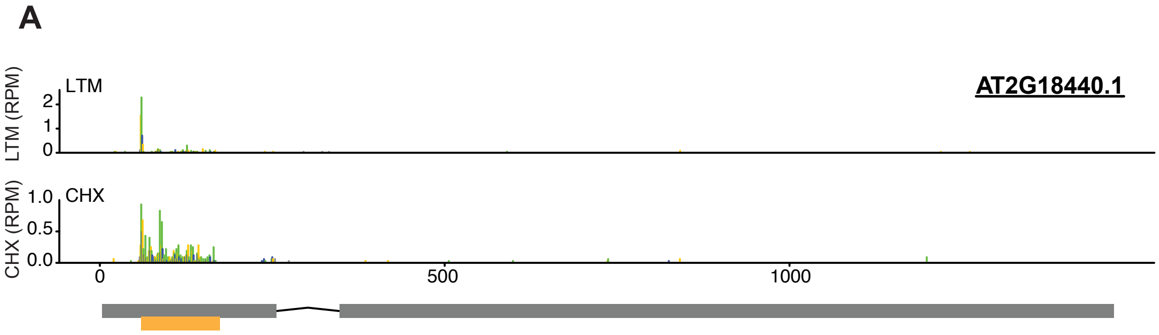
Supplemental Figure S10. The feature characteristics of the misclassified TN TISs
(A,B) Prediction score distribution of the CDS-AUG and CDS-nonAUG TIS groups in tomato (A) and in *Arabidopsis* (B). The mean threshold (dashed lines) for classifying the TN→TN (the TN TISs predicted as TNs, gray) and TN→TP (the TN TISs predicted as TPs, light brown) groups derived from the models on the10 randomly balanced datasets (as indicated in Fig. 1) are shown. (C,D) Dot plots show the frequency (y-axis) of a given feature used for TIS prediction in 10 randomly balanced datasets and the feature enrichment (FDR, x-axis) between the TN→TN and TN→TP groups for the TIS groups indicated in (A,B). The red line represents the threshold (frequency ≥ 7) of important features as indicated in Fig. 2. (E,F) Violin plots show the feature value distributions for the features that were most enriched in (C,D) for the TN→TN (gray) and TN→TP (light brown) groups, indicated in (A,B), and the TP→TP (the TP TISs predicted as TPs, dark brown) group. The red dot represents the median value.

Supplemental_Figure_S11



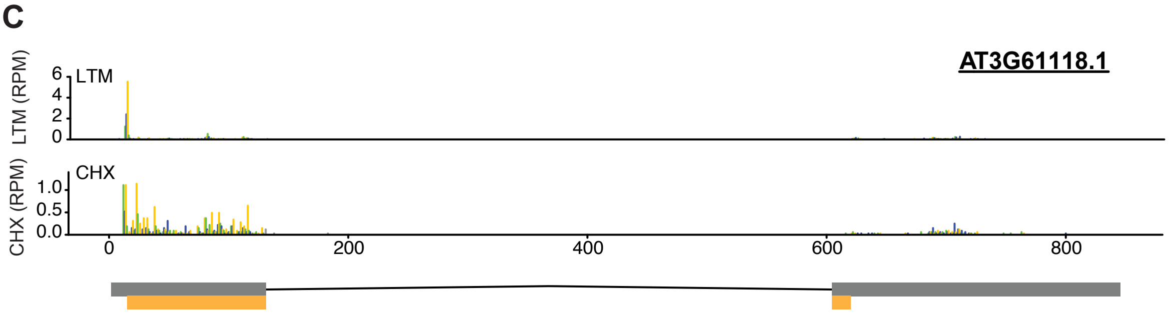
Supplemental Figure S11. *In vivo* initiation activities of misclassified TN TISs
(A) As described in Supplemental Fig. S8, but for Solyc05g050510.3.1, which has the indicated mis-classified 5′ UTR-nonAUG TIS (uTIS, GUG; orange arrow). There were no significant LTM signals for this TIS in both replicates, but the prediction score of 0.61 passed the prediction threshold of 0.44. (B) As described in Fig. 5A, but for the Immunoblot analyses of proteins translated from the misclassified TISs in (A). (C) As described in (A), but for Solyc11g039830.3.1, which has the indicated misclassified CDS-AUG TIS site (dTIS, in-frame AUG; orange arrow). The signals for this TIS were statistically significant in replicate #2 but not in replicate #1 because of the low read count in replicate #1, which did not pass the detection threshold, but the prediction score of 0.66 passed the prediction threshold of 0.49. The annotated AUG TIS-encoded protein isoform, but not the AUG dTIS-encoded one, has a predicted mitochondria-targeting signal (red). uTIS and dTIS denote upstream and downstream TIS, respectively.

Supplemental_Figure_S12



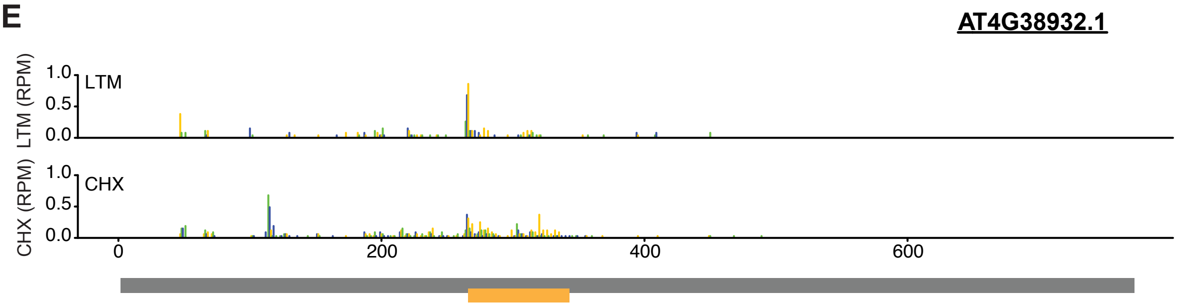
B

UUGAGAUCUUAUCUAACCGUUUAAUCGAGACCGUCCAUCUUAUCUCAUCACCGCCUCCAUAUUCUUUCGCGUCUUUUGCUUCCUCCCUCUUUUUCGCCAA**<u>AUG</u>**



D

UUUUAAUAGAAAAAUUAAGAUAAUGGACUCUAUAAGUACUACUUUCUUCGGACAGAUGAUCUGAAUUCUCUUACUCGUUCUUUUGUUCUUUUUCUUACAA**<u>AUG</u>**

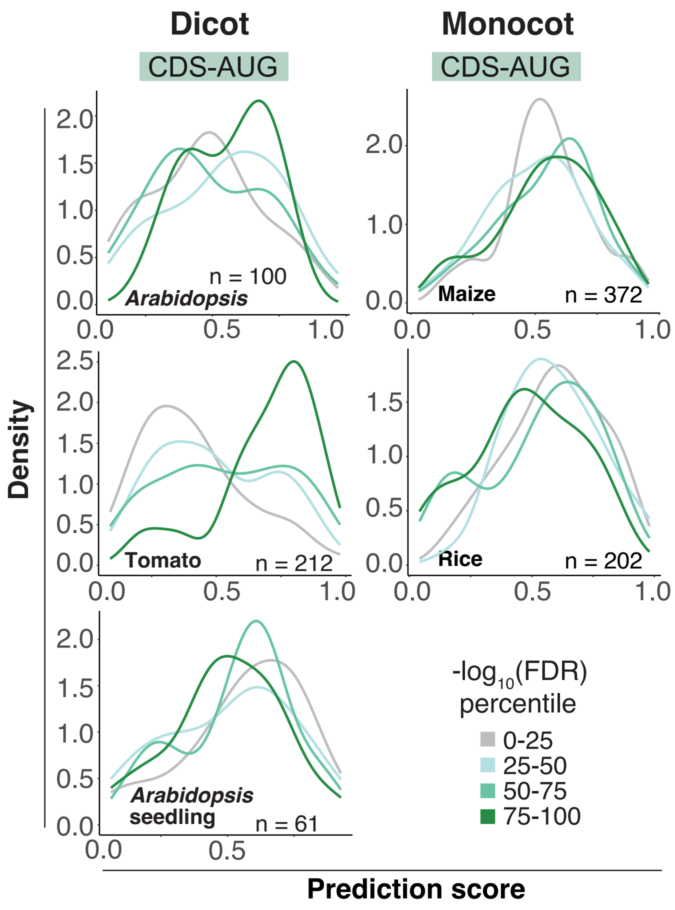


F

AUAAAUUUUCGUUUUAACGGAUAAAGCUUCCGGCGUUGUCCGGUUUAUCUUUCCGAUUAGAUCGGAUUUUUUGUUUGUGUUUUAGUUUUUCUUUUUCCAC**<u>AUG</u>**

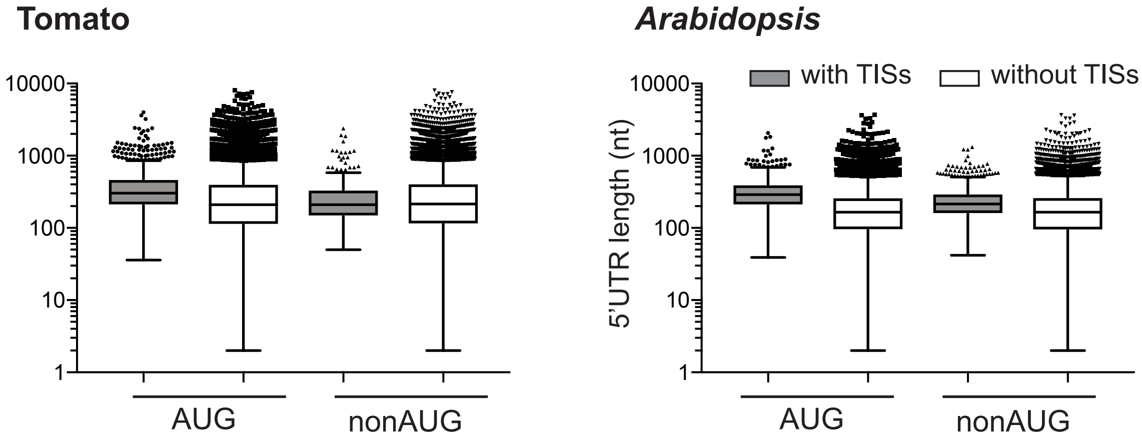
Supplemental Figure S12. The novel TIS-initiated ORFs in the non-coding RNA genes
(A) Plots showing the LTM and CHX read counts (RPM, read per million mapped reads) of the indicated genic regions of AT2G18440.1, which has a novel TIS and a hidden small open-reading frame (ORF) reported previously (Hsu et al., 2016). The gene model (bottom) with the exon (gray boxes), introns (thin lines), and novel ORFs (orange box). Blue, green and orange bars indicate the reads whose assigned P-sites map to the codon positions 0, 1, and 2 (i.e. phases 1–3, respectively) relative to the 5′ end of the transcripts. (B)The upstream 100-nt sequence with the CU-rich regions (light gray box) and the identified novel TIS (bold and underlined) are shown. (C,D) As indicated in (A,B), but for the novel AUG-initiated ORF in AT3G1118.1. (E,F) As indicated in (A,B), but for the novel AUG-initiated ORF in AT4G38920.1.

Supplemental_Figure_S13



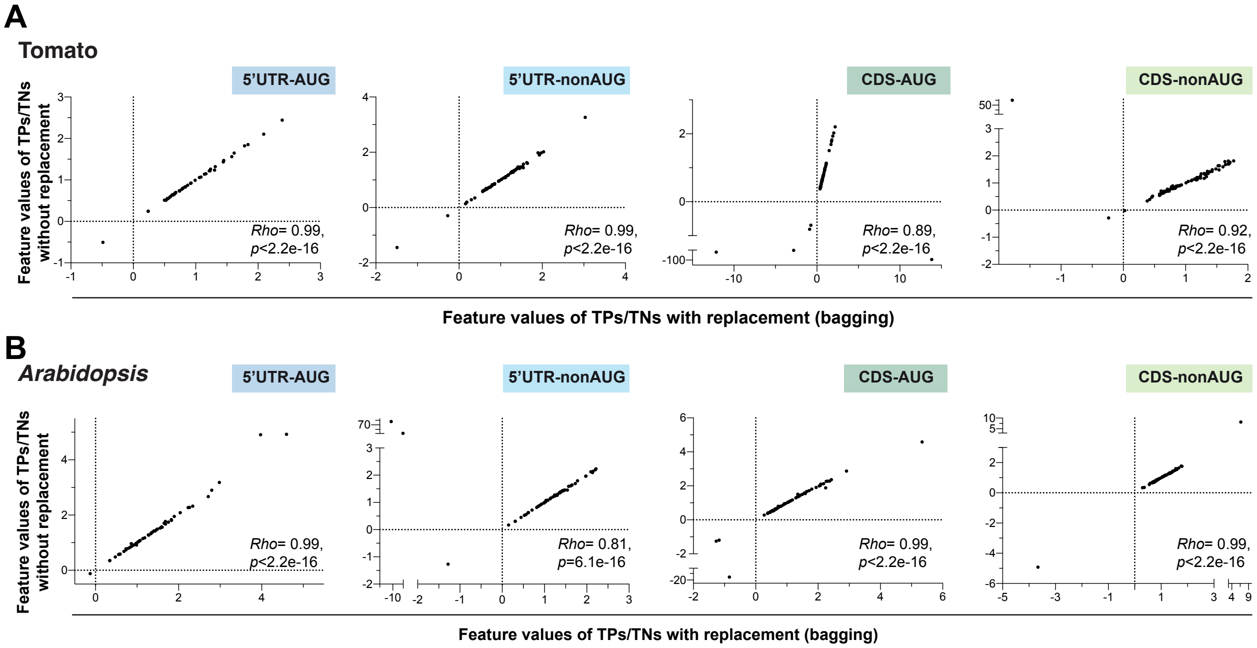
Supplemental Figure S13. Prediction of monocot and dicot TISs using transfer learning. Distribution of the TIS prediction scores generated by the tomato best models for the CDS-AUG and CDS-nonAUG TIS groups identified by the RiboTISH algorism and with RiboTISH-reported FDRs (FDR percentile in which 0-25 category includes TISs with lowest FDR values) using ribosome profiling datasets generated from dicot plants including *Arabidopsis* (suspension cells), tomato (leaves) and *Arabidopsis* seedling and monocot plants including maize and rice.

Supplemental_Figure_S14



Supplemental Figure S14. The distribution of 5'UTR lengths for genes with and without 5'UTR AUG and nonAUG TISs in tomato and *Arabidopsis*.

Supplemental_Figure_S15



Supplemental Figure S15. The correlation of the feature values between two different strategies of random sampling.

(A,B) The correlation of the feature values for TNs randomly undersampled without replacement (y-axis) and with replacement (bagging; x-axis) is shown for four TIS groups in tomato (A) and *Arabidopsis* (B). The features enriched in at least one of 10 randomly balanced datasets generated by each undersampling approach were included. The graph respresents the mean fold-changes of the feature values between TP and TN TIS sets across 10 randomly balanced datasets. Rho: Spearman's rank correlation coefficient.

Supplementary Table S1. The mean F1 measure, Matthews correlation coefficient (MCC; indicated in column) and Area Under The Curve Receiver Operating Characteristics scores (AUROC, indicated in column) scores showing the performances of models predicting translation initiation sites (TISs) based on the indicated features.

TIS type: AUG, 5UTR, *Arabidopsis*

Models	Known features	ORF features	Contextual features	All features	All features (MCC/AUROC)	Putative cis-elements	Putative cis-elements (MCC/AUROC)
GB	0.61	0.77	0.83	0.85	0.73/0.94	0.85	0.71/0.92
LogReg	0.62	0.73	0.83	0.86	0.73/0.93	0.85	0.69/0.92
RF	0.60	0.76	0.83	0.83	0.70/0.92	0.83	0.65/0.90
SVM	0.62	0.72	0.83	0.86	0.72/0.93	0.85	0.70/0.92

TIS type: nonAUG, 5UTR, *Arabidopsis*

Models	Known features	ORF features	Contextual features	All features	All features (MCC/AUROC)	With TIS codon preference	Putative cis-elements	Putative cis-elements (MCC/AUROC)
GB	0.72	0.73	0.80	0.82	0.67/0.90	0.88	0.73	0.41/0.76
LogReg	0.72	0.71	0.80	0.84	0.66/0.90	0.85	0.73	0.39/0.76
RF	0.70	0.72	0.79	0.82	0.65/0.89	0.86	0.72	0.40/0.74
SVM	0.72	0.71	0.80	0.83	0.66/0.90	0.85	0.72	0.39/0.76

TIS type: AUG, CDS, *Arabidopsis*

Models	Known features	ORF features	Contextual features	All features	All features (MCC/AUROC)	Putative cis-elements	Putative cis-elements (MCC/AUROC)
GB	0.72	0.86	0.83	0.85	0.82/0.96	0.81	0.63/0.88
LogReg	0.72	0.84	0.84	0.87	0.76/0.95	0.81	0.64/0.88
RF	0.71	0.85	0.82	0.88	0.80/0.95	0.80	0.64/0.87
SVM	0.72	0.83	0.83	0.87	0.76/0.95	0.82	0.67/0.88

TIS type: nonAUG, CDS, *Arabidopsis*

Models	Known features	ORF features	Contextual features	All features	All features (MCC/AUROC)	With TIS codon preference	Putative cis-elements	Putative cis-elements (MCC/AUROC)
GB	0.51	0.76	0.65	0.61	0.49/0.82	0.79	0.70	0.40/0.77
LogReg	0.59	0.74	0.64	0.72	0.42/0.79	0.74	0.69	0.37/0.76
RF	0.54	0.75	0.64	0.75	0.51/0.81	0.78	0.69	0.38/0.75
SVM	0.58	0.74	0.65	0.72	0.42/0.79	0.74	0.69	0.38/0.77

TIS type: AUG, 5UTR, Tomato

Models	All features	All features (MCC/AUROC)
GB	0.78	0.65/0.91
LogReg	0.82	0.64/0.90
RF	0.81	0.64/0.89
SVM	0.82	0.64/0.90

TIS type: nonAUG, 5UTR, Tomato

Models	All features	All features (MCC/AUROC)	With TIS codon preference
GB	0.74	0.69/0.91	0.85
LogReg	0.81	0.68/0.93	0.85

RF	0.80	0.65/0.91	0.83
SVM	0.81	0.65/0.91	0.85

TIS type: AUG, CDS, Tomato

Models	All features	All features (MCC/AUROC)
GB	0.87	0.77/0.95
LogReg	0.88	0.74/0.93
RF	0.88	0.76/0.94
SVM	0.88	0.75/0.93

TIS type: nonAUG, CDS, Tomato

Models	All features	All features (MCC/AUROC)	With TIS codon preference
GB	0.45	0.5/0.76	0.73
LogReg	0.73	0.43/0.78	0.72
RF	0.70	0.43/0.75	0.71
SVM	0.72	0.41/0.78	0.71

Supplemental Table S2. List of primers used in this study.

Primer	5’ to 3’ sequence
Solyc03g096920_WT_F	TTTGTACAAAAAAGCAGGCTCCGCTCCTATTTTTCCTCTATATACTCTTTCTGCGTCAAATTGAAGCTGTCTCTCTGTTTATGTTTTTCTCC
Solyc03g096920_WT_R	CTTTGTACAAGAAAGCTGGGTCGAAAAACATGAGTTTCTAGAAAGAGGCCAAAAGGGGAGAAAAACATAAACAGAGAGACAGCTTC
Solyc03g096920_CUmutation_F	TTTGTACAAAAAAGCAGGCTCCGGTGGTATTTTTGGTGTATATAGTGTTTGTGCGTGAAATTGAAGGTGTGTGTGTGTTTATGTTTTTGTG
Solyc03g096920_CUmutation _R	CTTTGTACAAGAAAGCTGGGTCGAAAAACATCACTTTCTACAAACACGCCAAAACGGCACAAAAACATAAACACACACACACCTTC
Solyc07g052600_WT_F	TTTGTACAAAAAAGCAGGCTCCGTAAATGCAAATATCACCTTCTCCTCGAAGGCTGCGATTCGCTGTCCCTATTTGTCAATTTCATATTTATTTCC
Solyc07g052600_WT_R	CTTTGTACAAGAAAGCTGGGTCGGTCTCCACCGTGACTTTGTAGTCCAAGTTTTTGATGGAAATAAATATGAAATTGACAAATAGGGACAG
Solyc07g052600_CUmutation _F	TTTGTACAAAAAAGCAGGCTCCGTAAATGCAAATATGACGTTGTGGTGGAAGGGTGCGATTGGGTGTGCGTATTTGTGAATTTGATATTTATTTG
Solyc07g052600_CUmutation _R	CTTTGTACAAGAAAGCTGGGTCGGTCTCCACCGTCACTTTGTACTCCAACTTTTTCTTGCAAATAAATATCAAATTCACAAATACGCACAC
Solyc06g050510_WT_F	CTTTTTTGGCGATTTTTTTGCAGC
Solyc06g050510_WT_R	TAAAAACACGGTTAATTTTCCAAATGCAATAAC
Solyc06g009750_WT_F	ACAATCCAAATTTTCCCCAAACCC
Solyc06g009750_WT_R	AAAATTCAGTTTGAGGAATTCGAAATGCG
Solyc06g009750_CUmutation _F	TTTGTACAAAAAAGCAGGCTCCGAATTCGCCCTTACAATGCAAATTTTGCCCAAACCGTGATTGCAGTGTGAATTGTGTTGTTGGTGATTTTGGTATTTAACAAAAG
Solyc06g009750_CUmutation _R	CTTTGTACAAGAAAGCTGGGTCGAATTCGCCCTTAAAATTCAGTTTCACCAATTCCAAATGCCAAATTCACTTTTGTTAAATACCAAAATCACCAACAACAC
Solyc06g009750_CUmotif_mutation_F	TTTGTACAAAAAAGCAGGCTCCGAATTCGCCCTTACAATCCAAATTTTCCCCAAACCCTGATTCCACTCTCAATTGTGTTGTTGCTCATTTTCCTATTTAAC
Solyc06g009750_CUmotif_mutation _R	CTTTGTACAAGAAAGCTGGGTCGAATTCGCCCTTAAAATTCAGTTTGAGGAATTCGAAATGCGAAATTGAGTTTTGTTAAATAGGAAAATGAGCAACAACACAATTG
Solyc06g076770_WT _F	TTTTCCCATTGTAAAAACCCCACATC
Solyc06g076770_WT _R	AGACAGCATTTTTATGAAAAAGATTTGAACTTTG
Solyc06g076770_CUmutation _F	TTTGTACAAAAAAGCAGGCTCCGAATTCGCCCTTTTTTGCCATTGTAAAAACCCCACATGTGTGCATTTTGCATGTTGGGTTTTTGCCCCAAAAGTTGG

Solyc06g076770_CUmutation _R	CTTTGTACAAGAAAGCTGGGTCGAATTCGCCCTTAGACAGCATTTTTATCAAAAACtTTTCAACTTTCAACAAACACACCAACTTTTGGGGCAAAAACCC
Solyc06g076770_CUmotif_mutation _F	TTTGTACAAAAAAGCAGGCTCCGAATTCGCCCTTTTTTCCCATTGTAAAAACCCCACATGTGTGCATTTTCCATCTTGGGTTTTTCCCCCAAAAGTTGC
Solyc06g076770_CUmotif_mutation _R	CTTTGTACAAGAAAGCTGGGTCGAATTCGCCCTTAGACAGCATTTTTATGAAAAAGATTTGAACTTTCAACAAACAGAGCAACTTTTGGGGGAAAAACCC
Solyc04g076110_WT _F	GAACACGGACTCCAATTATTATTGTTCAC
Solyc04g076110_WT_R	AACCTCCATGGATTTCGAGGT
eGFP_F	CAAGGGCGAGGAGCTGTTCAC
eGFP_R	GGTCAGGGTGGTCACGAGG
ubi3_F	GCCGACTACAACATCCAGAAGG
ubi3_R	TGCAACACAGCGAGCTTAACC
nLUCF_F	GAACAGGGAGGTGTGTCCAG
nLUCR_R	CGCTCAGACCTTCATACGGG
Solyc03g096920_mCU-12_F	TTTGTACAAAAAAGCAGGCTCCGGTGGTATTTTTGGTGTATATAGTGTTTGTGCGTCAAATTGAAGCTGTCTCTCTGTTTATGTTTTTCTCC
Solyc03g096920_mCU-23_F	TTTGTACAAAAAAGCAGGCTCCGCTCCTATTTTTCCTCTATATAGTGTTTGTGCGTGAAATTGAAGGTGTCTCTCTGTTTATGTTTTTCTCC
Solyc03g096920_mCU-23_R	CTTTGTACAAGAAAGCTGGGTCGAAAAACATGAGTTTCTAGAAAGAGGCCAAAAGGGGAGAAAAACATAAACAGAGAGACAcCTTC
Solyc03g096920_mCU-34_F	TTTGTACAAAAAAGCAGGCTCCGCTCCTATTTTTCCTCTATATACTCTTTCTGCGTGAAATTGAAGGTGTGTGTGTGTTTATGTTTTTCTCC
Solyc03g096920_mCU-34_R	CTTTGTACAAGAAAGCTGGGTCGAAAAACATGAGTTTCTAGAAAGAGGCCAAAAGGGGAGAAAAACATAAACACACACACACCTTC
Solyc03g096920_mCU-45_F	TTTGTACAAAAAAGCAGGCTCCGCTCCTATTTTTCCTCTATATACTCTTTCTGCGTCAAATTGAAGCTGTGTGTGTGTTTATGTTTTTGTG
Solyc03g096920_mCU-45_R	CTTTGTACAAGAAAGCTGGGTCGAAAAACATGAGTTTCTAGAAAGAGGCCAAAACGGCACAAAAACATAAACACACACACAGCTTC
Solyc03g096920_mCU-56_F	TTTGTACAAAAAAGCAGGCTCCGCTCCTATTTTTCCTCTATATACTCTTTCTGCGTCAAATTGAAGCTGTCTCTCTGTTTATGTTTTTGTG
Solyc03g096920_mCU-56_R	CTTTGTACAAGAAAGCTGGGTCGAAAAACATCAATTTCTACAAACACGCCAAAACGGCACAAAAACATAAACAGAGAGACAGCTTC
Solyc03g096920_mCU_R2	CTTTGTACAAGAAAGCTGGGTCGAAAAACATCAATTTCTACAAACACGCCAAAACGGCACAAAAACATAAACACACACACACCTTC