Known and novel post-transcriptional regulatory sequences are conserved across plant families

JUSTIN N. VAUGHN,[1] SALLY R. ELLINGSON,[2] FLAVIO MIGNONE,[3] and ALBRECHT VON ARNIM[1,2,4]

[1]Department of Biochemistry, Cellular and Molecular Biology, The University of Tennessee, Knoxville, Tennessee 37996, USA
[2]Graduate School of Genome Science and Technology, The University of Tennessee, Knoxville, Tennessee 37996, USA
[3]Dipartimento di Chimica Strutturale e Stereochimica Inorganica, Università degli Studi di Milano, 20133 Milano, Italy

ABSTRACT

The sequence elements that mediate post-transcriptional gene regulation often reside in the 5′ and 3′ untranslated regions (UTRs) of mRNAs. Using six different families of dicotyledonous plants, we developed a comparative transcriptomics pipeline for the identification and annotation of deeply conserved regulatory sequences in the 5′ and 3′ UTRs. Our approach was robust to confounding effects of poor UTR alignability and rampant paralogy in plants. In the 3′ UTR, motifs resembling PUMILIO-binding sites form a prominent group of conserved motifs. Additionally, Expansins, one of the few plant mRNA families known to be localized to specific subcellular sites, possess a core conserved RCCCGC motif. In the 5′ UTR, one major subset of motifs consists of purine-rich repeats. A distinct and substantial fraction possesses upstream AUG start codons. Half of the AUG containing motifs reveal hidden protein-coding potential in the 5′ UTR, while the other half point to a peptide-independent function related to translation. Among the former, we added four novel peptides to the small catalog of conserved-peptide uORFs. Among the latter, our case studies document patterns of uORF evolution that include gain and loss of uORFs, switches in uORF reading frame, and switches in uORF length and position. In summary, nearly three hundred post-transcriptional elements show evidence of purifying selection across the eudicot branch of flowering plants, indicating a regulatory function spanning at least 70 million years. Some of these sequences have experimental precedent, but many are novel and encourage further exploration.

Keywords: post-transcriptional control; angiosperms; RNA motif; translation reinitiation

INTRODUCTION

In most eukaryotic organisms tested, transcript concentration is only loosely correlated with protein concentration (Baerenfaller et al. 2008; de Sousa Abreu et al. 2009). In mammals and fungi, a substantial fraction of this decoupling results from gene-specific variations in translation efficiency (Ingolia et al. 2009; Vogel et al. 2010). The identification and categorization of elements responsible for this variation would help to assess their biological significance and to more fully understand the pathways in which they act. Thus, we employed a comparative transcriptomics pipeline in order to determine the prevalence and relative proportions of post-transcriptional regulatory sequences within flowering plants.

Upstream start codons (uAUGs) are among the most ubiquitous gene-specific elements affecting an mRNA's protein expression level (Calvo et al. 2009). Because ribosomes scan the mRNA in a 5′ to 3′ direction in search of a start codon, these uAUGs will, with variable frequency, become initiation sites for protein synthesis (Ingolia et al. 2009). Their associated open reading frame (uORF) may either overlap the major ORF (mORF) or terminate upstream of the mORF start codon (Fig. 1A). In either case, uORFs can drastically reduce protein expression (Calvo et al. 2009; Zhou et al. 2010). In some cases, uORFs are conserved at the peptide-level (Franceschetti et al. 2001; Hayden and Jorgensen 2007; Tran et al. 2008; Rahmani et al. 2009)—referred to as "conserved peptide uORFs" (CPuORFs). CPuORFs shared by *Arabidopsis thaliana* and rice fall into 19 homologous groups (Hayden and Jorgensen 2007). The degree to which uORFs are conserved over short or long evolutionary time scales is still not well-delineated, although uAUG triplets in mammals and fungi evolve slower than other 5′ UTR triplets (Churbanov et al. 2005; Neafsey and Galagan 2007).

Many researchers have proposed that groups of specific genes are coregulated at the mRNA level by interactions

[4]**Corresponding author.**
E-mail vonarnim@utk.edu.

Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.031179.111.

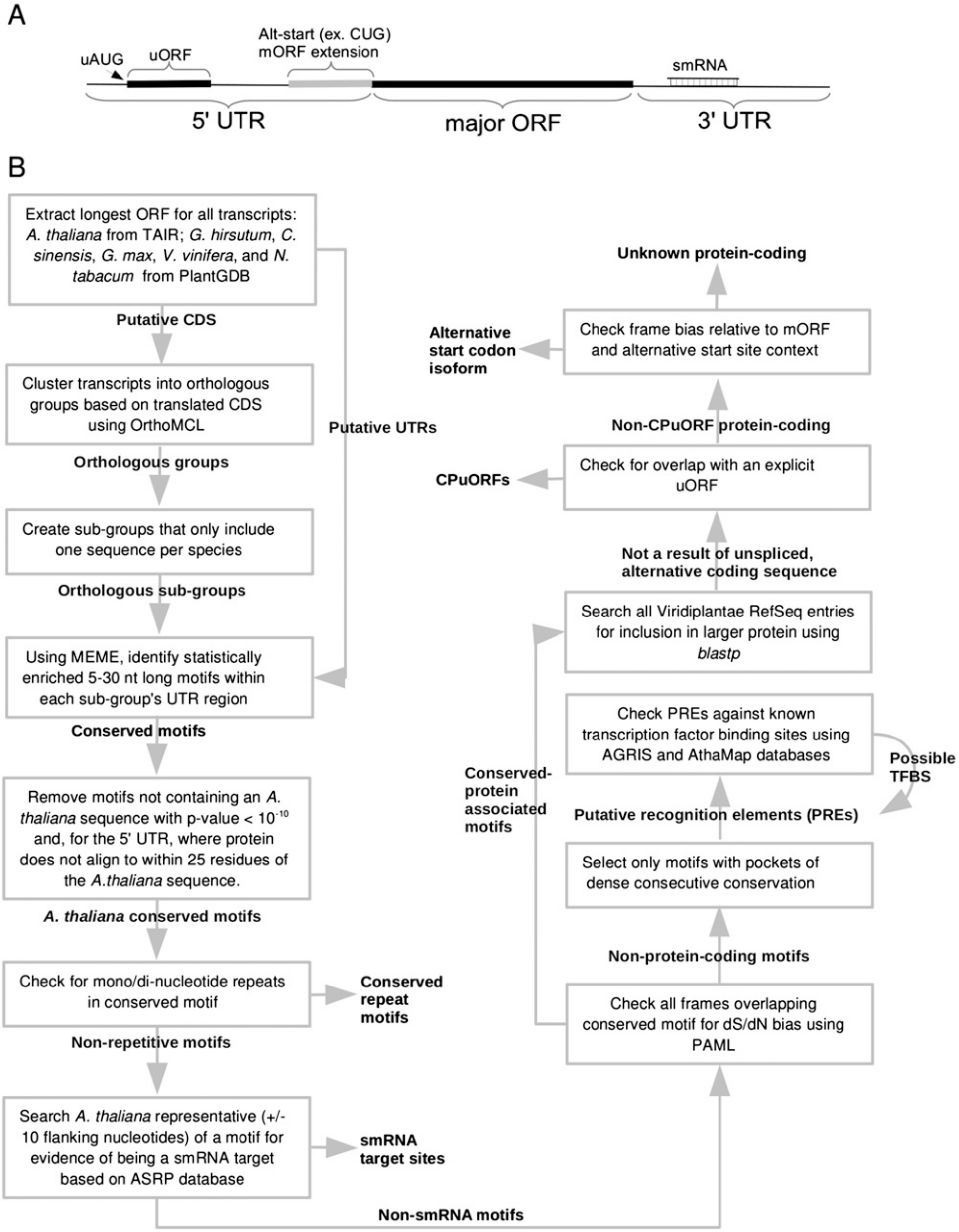
A
uAUG
uORF
Alt-start (ex. CUG)
mORF extension
smRNA
5' UTR
major ORF
3' UTR
B
Extract longest ORF for all transcripts: *A. thaliana* from TAIR; *G. hirsutum*, *C. sinensis*, *G. max*, *V. vinifera*, and *N. tabacum* from PlantGDB
Putative CDS
Cluster transcripts into orthologous groups based on translated CDS using OrthoMCL
Putative UTRs
Orthologous groups
Create sub-groups that only include one sequence per species
Orthologous sub-groups
Using MEME, identify statistically enriched 5-30 nt long motifs within each sub-group's UTR region
Conserved motifs
Remove motifs not containing an *A. thaliana* sequence with p-value < 10^-10 and, for the 5' UTR, where protein does not align to within 25 residues of the *A.thaliana* sequence.
A. thaliana conserved motifs
Check for mono/di-nucleotide repeats in conserved motif
Conserved repeat motifs
Non-repetitive motifs
Search *A. thaliana* representative (+/- 10 flanking nucleotides) of a motif for evidence of being a smRNA target based on ASRP database
smRNA target sites
Non-smRNA motifs
Check all frames overlapping conserved motif for dS/dN bias using PAML
Non-protein-coding motifs
Select only motifs with pockets of dense consecutive conservation
Putative recognition elements (PREs)
Check PREs against known transcription factor binding sites using AGRIS and AthaMap databases
Possible TFBS
Conserved-protein associated motifs
Search all Viridiplantae RefSeq entries for inclusion in larger protein using *blastp*
Not a result of unspliced, alternative coding sequence
Check for overlap with an explicit uORF
CPuORFs
Non-CPuORF protein-coding
Check frame bias relative to mORF and alternative start site context
Alternative start codon isoform
Unknown protein-coding

FIGURE 1. (*A*) Cartoon of a generic plant mRNA with specific sequence regions and known post-transcriptional elements indicated. (*B*) Schematic of the computational pipeline used for conserved sequence identification and categorization. Tasks are described in each box. Outputs are unboxed and overlap an arrow if that output then becomes the input for a downstream processing step.

between a common RNA binding factor and its cognate RNA sequence element (Gerber et al. 2004; Keene 2007). This ''regulon'' concept is supported by experiments on the PUF family of RNA binding proteins within complex eukaryotes (Gerber et al. 2006; Francischini and Quaggio 2009) and on 40 RNA-binding proteins in yeast (Hogan et al. 2008). A striking example is the PUMILIO protein (a PUF-family protein) in *Drosophila*, which interacts with mRNAs for a majority of subunits of the vacuolar ATPase (Gerber et al. 2006). The sequence specificity of PUF-family proteins appears to stay constant across major taxonomic divisions, while the function of their target genes can vary drastically (Gerber et al. 2006). Such variation appears to carry over into plants (Francischini and Quaggio 2009), but its full extent has yet to be determined.

The regulatory sequences described above reside in either the 5′ or 3′ UTR. Because of assorted codon usage constraints, it is difficult to assess peptide-independent nucleotide conservation within individual mORFs, particularly with regard to short motifs (Chen and Blanchette 2007). Therefore, we have focused on the UTRs. Respectively, repression by small RNAs (smRNAs) is another major regulatory process known to act post-transcriptionally on specific mRNAs, and, as opposed to metazoans, most putative smRNA target sites in *A. thaliana* are located in the mORF of mRNAs (Backman et al. 2008). Still, a few smRNA target sites do appear in UTRs, and these can be used to assess trends in target-site conservation without the confounding effects of conservation related to protein encoding.

Although post-transcriptional processes are sometimes mediated by RNA secondary structure, most mRNA-specific interactions require a primary nucleotide sequence component (Rabani et al. 2008; Serganov and Patel 2008; Gilligan et al. 2011). The conservation of such primary sequence elements is typically easier to interpret because similarity in secondary structure is often a result of nucleotide composition (Rivas and Eddy 2000) and, even when constrained structure can be identified via covariance, these structures are difficult to generalize across the genes that harbor them (McGuire and Galagan 2008; Rabani et al. 2008). Since most intergenic or untranslated DNA evolves at a neutral rate, primary sequences separated by a large evolutionary distance cannot be aligned with confidence. Additionally, at such distances, conserved motif identification is confounded by site turnover (Doniger and Fay 2007). The "window of useful divergence" to identify conserved noncoding sequences by alignment is ~10–40 million years for plants (Freeling and Subramaniam 2009). To overcome these difficulties for the eudicot branch of flowering plants, which diverged as early as 70 million years ago, we assessed the statistical enrichment of motifs that are linked to orthologous coding sequences—a technique that is alignment-free and somewhat robust to site turnover or displacement (Blanchette and Tompa 2002). Our approach, based on the MEME algorithm (Bailey and Elkan 1994), is most effective when all sequences compared have undergone effectively randomizing divergence. Since we are looking for deep conservation, this prerequisite is satisfied by default (Galagan et al. 2005; Chen and Rajewsky 2006). Yet, when creating orthologous groups from distantly related plant species, in-paralogs resulting from local duplication or polyploidy can produce false positives caused by short divergence times (Blanchette and Tompa 2002). As described below, we solved this problem by reducing groups of orthologous sequences to combinatorial subgroups.

The identification of sequence conservation across orthologous intergenic or promoter regions is typically interpreted as functional constraint on transcription factor binding sites, while conservation of amino acids in a protein is interpreted as constraint relating to protein function. Because, as described above, variable forms of regulation can act via the UTR, it is easy to misinterpret conservation within the UTR. For example, smRNA binding sites can pose as CPuORFs if synonymous to nonsynonymous mutation bias is not assessed. Therefore, we have categorized conserved motifs into plausible functional groups based on whether or not they (1) have dinucleotide repeats that resemble microsatellites, (2) are target sites of known or predicted microRNAs, (3) code for a peptide that is evolutionarily constrained at the amino acid level, (4) qualify as putative recognition elements (PREs) based on pockets of dense conservation, and (5) match known transcription factor binding sites (Fig. 1B).

The motif identification approach was applied to six eudicot lineages, including tobacco, grapevine, soybean, orange, cotton, and *Arabidopsis*. These six species were chosen not only because each provides substantial transcript data but also because each marks a distinct, highly diverged, branch of the eudicots (Fig. 2A). Moreover, they represent many of the different life histories and characteristics of the eudicots: growth habit (tree, vine, perennial shrub, annual herbaceous plants), geographic origin and latitude, history of domestication, level of polyploidy, and genome size. We found that at least 3% of orthologous groups have one or more deeply conserved UTR motifs. It stands to reason that the majority of these motifs act at the post-transcriptional level. In spite of a similar nucleotide composition, the 5′ and 3′ UTRs have distinct complements of conserved motifs, as predicted by canonical models of eukaryotic translation. Elements that engage the ribosome are found preferentially in the 5′ UTR, while elements resembling protein binding sites are more prominent in the 3′ UTR. We find some evidence for the notion that groups of mRNAs form RNA regulons. However, such groups are typically from the same gene family. Our data suggest that many of the conserved sequence motifs regulate individual genes, rather than gene families, and many of the target genes are regulators of transcription, signaling, or protein turnover.

RESULTS

Three percent of orthologous groups contain a conserved motif in the UTR

UTRs are difficult to predict computationally from genomic DNA (Brown et al. 2005); thus, experimentally confirmed mRNA sequences are needed for valid UTR comparisons. To this end, *A. thaliana* transcripts were acquired from TAIR (version 9). Putative transcript data for five informant species with >60,000 putative transcript entries were downloaded from PlantGDB (Fig. 2A; Supplemental Table S1). The PlantGDB transcripts were assembled from all available expressed sequence tag (EST) and cDNA sequence data in GenBank. The longest continuous reading frame containing an in-frame AUG for each transcript was considered the major ORF. mORFs were translated and clustered into orthologous groups using reciprocal *blastp* and OrthoMCL (Supplemental File 1). The sequence upstream of a mORF was considered its 5′ UTR. Likewise, the sequence downstream was considered its 3′ UTR.

Of the 11,887 groups containing an *A. thaliana* sequence, 10,122 (85.1%) contained at least one informant species

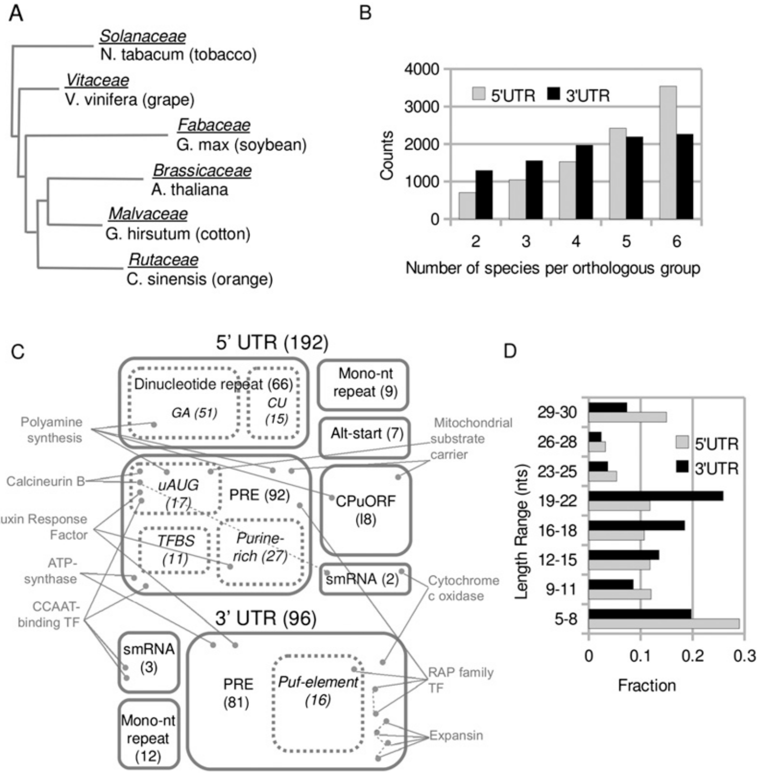

FIGURE 2. The landscape of post-transcriptional regulatory sequences that are conserved in flowering plants. (*A*) Tree representing descent and relative divergence of the species in this analysis. Modified from Bausher et al. (2006) and based on chloroplast genomes. The family name is given *above* its representative species. (*B*) Distributions for the number of species per orthologous group in 5′ and 3′ UTR comparisons. (*C*) A categorical map of conserved motif annotations. Box size roughly indicates the proportion of annotations that we assigned to each category. Exact values are given in parentheses. A box with dashed lines indicates that the category is a subset of the larger, solid-lined box. A dot indicates a single orthologous group. Only a small selection of orthologous groups are shown, and these are placed in the annotation box associated with their conserved motif. Gray labels describe the larger gene families to which these specific orthologous groups belong. A dashed line *between* orthologous groups indicates that the motifs are similiar (RAP family, Expansin). A dashed line added to Calcineurin B indicates multiple annotations of its conserved motif. (TFBS) transcription factor binding site. (*D*) Histograms, calculated independently, of PRE lengths in both UTRs. Related to Supplemental Files 2, 3, and 4.

as well (Supplemental Table S1). The remaining 14.9% of *A. thaliana* groups consist solely of in-paralogs, where no orthology across lineages could be inferred. This number, 10,122, is comparable to the number of orthologs, 10,381, shared by *A. thaliana* and *Ricinus communis*, both of which have sequenced genomes (Chen et al. 2006). This number is necessarily smaller than the total number of genes per genome because many plant genes are closely related in-paralogs derived from recent genome duplications. Therefore, the number of orthologous groups identified here approaches the theoretical maximum. For both UTR regions, the majority of comparisons have good representation, i.e., four to six species (Fig. 2B).

The 5′ and 3′ UTRs of an mRNA generally evolve faster than its main protein coding sequence (CDS). Given the divergence between the species in this study, neutrally evolving portions of the UTR will have a nominal number of consecutive bases conserved as a result of relatedness alone. We, therefore, used the MEME algorithm to search for enriched elements within these UTRs. MEME assumes a random background model and calculates an E-value based on the number of times a given motif is expected to be present by chance in a given set of sequences. Importantly, intra-species paralogs (in-paralogs), resulting from post-speciation duplication events, could potentially have undergone very short divergence times, undermining our assumption of effective randomization and disrupting the identification of conserved motifs. Also, many orthologous groups contain two or more alternative transcripts of the same gene, which are operationally indistinguishable from in-paralogs. To address these issues, all orthologous groups were subdivided combinatorially such that each comparison involved only one sequence from each species in the orthologous group (see Materials and Methods). This approach has two additional benefits: (1) In-paralogs, which may have undergone neo/subfunctionalization at the regulatory level by losing an element (Lockton and Gaut 2005), do not add noise to the identification process; and (2) our false discovery rate can be estimated by simply randomizing orthologous groups, as opposed to simulating mutations.

Our statistical criteria, defined in Materials and Methods, resulted in 194 and 96 conserved motifs for the 5′ and 3′ UTRs, respectively (Fig. 2C). Thus, 3% of orthologous groups have a deeply conserved primary sequence element. Based on randomization of orthologous groups, our false discovery rate was 6.1% (12/194) for the 5′ UTR and 3.1% (3/96) for the 3′ UTR (i.e., for the 3′ UTR, we expect ~3 false positives per 100 positive results). We further characterized motifs identified by MEME based on their composition and their patterns of conservation.

[AG] and [CU] repeats are enriched in the 5′ UTR

Microsatellites (mono/dinucleotide repeats) are common in plants, have a regional bias in the genome (Lawson and Zhang 2006), and are conserved between orthologs and across paralogs (Zhang et al. 2006). The consensus sequence associated with each motif was checked for >5 consecutive mononucleotide repeats or >3 consecutive dinucleotide repeats. We found a dramatic enrichment of conserved dinucleotide repeats in the 5′ UTR relative to the 3′ UTR: 51-fold enrichment for $[AG]_n$ and 15-fold for $[CU]_n$ (Table 1). Mononucleotide repeats showed no such 5′/3′ bias. Randomization of orthologous groups (“random” data sets in Table 1) indicates that only a small proportion of conserved dinucleotide motifs can be attributed to the general enrichment of [AG] or [CU] repeats within 5′ UTRs. GA repeats are enriched around and downstream from the transcription start site in plant genes, and the motif is more often found in TATA-less genes, which have a broad cluster of transcription initiation sites (Yamamoto et al. 2009). Concerning the CU repeats, pyrimidine-rich motifs (Y-patches) are common in the 5′ UTR, but their function remains unclear (Molina and Grotewold 2005; Yamamoto et al. 2007, 2009).

One-fourth of known *A. thaliana* smRNA target sites in the UTR are conserved, and their conservation profiles support the “seed” hypothesis

In striking contrast to metazoans, there are only 238 known small RNA target sites in the *Arabidopsis* small RNA project database (ASRP) (Fahlgren et al. 2007; Backman et al. 2008). Of these, only 15 target sites appear to be located in the 3′ UTR. Of the balance, eight occur in the 5′ UTR, while the remaining ones lie in the CDS of targeted mRNAs.

After removal of repeat motifs, we checked the *A. thaliana* representative of each of our motifs (with +/− 10 flanking nucleotides) against all putative smRNA target sites. Two of the eight known target sites in the 5′ UTR were conserved and seven of the 15 in the 3′ UTR (Table 2—only three entries are shown for the 3′ UTR because the remainder are found within the same orthologous groups). Motifs were also checked against more liberal predictions of smRNA targets and recently discovered miRNAs (Alves et al. 2009; Breakfield et al. 2011), with the same result. The “seed” hypothesis predicts that complementarity to the 5′ end of the mature smRNA is critical for mediating silencing (Brodersen and Voinnet 2009). To varying degrees, all five motifs support this hypothesis. All but three of the 15 genes with a known *Arabidopsis* 3′ UTR target site had the requisite sequence coverage for a valid comparison. In addition, all miRNA families listed in Table 2 are known to be conserved across their entire length in at least two dicots (Willmann and Poethig 2007). Therefore, the 3′ degeneracy of their target sites cannot be accounted for by covariation in the targeting smRNA. Together, these results suggest that many miRNA target sites appear to be lineage-specific.

At least 18 conserved uORFs function at the peptide level

If any region in the UTR codes for a conserved protein, then that region will appear as a highly significant motif in our analysis. A conserved protein could take the form of a conserved-peptide uORF (CPuORF), a segment of main ORF downstream from a non-AUG start codon, or a segment of main ORF that was misannotated as 5′ UTR due to an ambiguous splicing pattern. We, therefore, checked all motifs for their coding potential. The longest continuous reading frame (CRF) for each of the three possible reading frames associated with a motif was aligned as protein. Each alignment was tested for coding potential based on the likelihood of nonsynonymous mutations being under negative selection (see Materials and Methods). After eliminating artifacts stemming from alternative splicing events, motifs were considered CPuORFs when significant protein-coding potential overlapped an explicit uORF (see Fig. 3).

We identified four CPuORFs with little or no precedent in the literature—indicated by “n” in Table 3. These novel CPuORFs show similar conservation profiles and spatial patterns to known CPuORFs (Fig. 3A–C), although CPuORF-24n is particularly unusual in that it begins near the cap and extends the length of the entire 5′ UTR (Fig. 3B). CPuORFs 24n and 25n exhibit extensive amino acid conservation across their entire length, while the others show a 3′ bias in their degree of conservation (Fig. 3D). We also recapitulated 13 of the 19 CPuORFs found previously in an *A. thaliana* (dicot) and *Oryza sativa* (monocot) comparison (Table 3; Hayden and Jorgensen 2007). One might have expected to find a substantial number of dicot-specific CPuORFs, which would not have been detected in prior *Arabidopsis*-monocot comparisons. However, this was not the case. Among the four novel CPuORFs, three were found in one or more monocot lineages (Supplemental Table S2). Only one CPuORF, 27n, could not be found in any of three monocot lineages examined, even though extensive sequence data exists for the 5′ UTR of CIPK6 homologs. Therefore, it appears

TABLE 1. Number of conserved repeat motifs in 5′ and 3′ UTRs

	Repeat[b]									
Data set[a]	$[AC]_n$	$[AG]_n$	$[AU]_n$	$[CG]_n$	$[CU]_n$	$[GU]_n$	A_n	C_n	U_n	G_n
5′ UTR	0	51	0	0	15	0	7	0	1	1
5′ UTR-random	0	5	0	0	1	0	1	0	0	0
3′ UTR	0	1	0	0	1	0	3	5	3	1
3′ UTR-random	0	0	0	0	0	0	0	0	0	1

All genes harboring mono/dinucleotide repeat motifs are listed in Supplemental Table S3.
[a]“Random” refers to the control analysis involving randomization of orthologous groups (see Materials and Methods).
[b]All overlapping dinucleotide repeats, such as $[AG]_n$ and $[GA]_n$, are pooled.

TABLE 2. Motifs implicated in experimentally confirmed smRNA-mediated degradation pathways

mRNA region	*A. thaliana* accession	*A. thaliana* annotation	miRNA family	Mature smRNA consensus[a,c]	Motif (reduced IUPAC consensus[a])
5′ UTR	AT3G15640	cytochrome c oxidase family protein	miR398	UGUGUUCUCAGGUCRCCCCUn	nCnnnnnGGnGnGACCUGAGA (21)
5′ UTR	AT1G08830[b]	Cu/Zn superoxide dismutase (CSD1)	miR398	UGUGUUCUCAGGUCRCCCCUn	AAGGGGUnYYCUGAGAUCACAnAn (24)
3′ UTR	AT1G72830	CCAAT-binding transcription factor	miR169	nAGCCAAGGAUGRCUUGCCGR	GGnAnnUCAUCCUUGGCUn (19)
3′ UTR	AT5G12840	CCAAT-binding transcription factor	miR169	nAGCCAAGGAUGRCUUGCCGR	nGCnAAUCAUUCUUGGCU (18)
3′ UTR	AT1G31280	PAZ/piwi domain-containing protein	miR403	UUAGAUUCACGCACRRAYUCn	AAGnnnnUnnnGCGUnnAnCU (21)

[a] Total length is in parentheses. A motif letter is written as: (1) the actual letter if present at >84% in a motif position; (2) ''R'' or ''Y'' if position composition is G+A >84% or C+U >84%, respectively; (3) ''n'' if otherwise.
[b] Not in ASRP database but from Bonnet et al. (2004).
[c] Based on miRNA alignments of at least two species in our analysis, one of which must be *A. thaliana*.

that CPuORFs are rarely lineage-specific with regard to the angiosperm phylogeny.

In addition to CPuORFs, seven protein-coding motifs were categorized as conserved "non-AUG starts" of the mORF, because they were consistently in-frame with the mORF and not explained as alternative/incomplete splicing artifacts (Table 3). The beginning of each of these putative coding regions contains a variant of the pattern "A(A|C)N(G|U|C)UGG," where "N" indicates any nucleotide and brackets indicate possible residues (Supplemental Fig. S1). This pattern, excepting the variable adenine of the canonical translation start codon, matches the strongest Kozak context for translation initiation in plants (Lukaszewicz et al. 2000). In some prior cases, such sites are used to produce multiple protein isoforms for the same mRNA molecule, which can then be targeted to various organelles (Wamboldt et al. 2009). Also, for seemingly different reasons, a key regulator of flowering time in plants, FCA, consistently uses CUG as an alternative translation initiation site (Simpson et al. 2010). Our data confirm that non-AUG initiation requires a strong context. Moreover, they document that non-AUG initiation is not a fluke restricted to a specific lineage but, given its deep conservation, must have provided a selective advantage during much of the eudicot evolutionary history.

In the 3′ UTR, after comparable filtering criteria, we were left with five potential protein-coding motifs. None of the five were consistently frame-biased in terms of protein-coding potential, and so these likely do not represent conserved read-through or programmed frame-shift events. They may represent exons from rare isoforms or, given their short length (<10 codons), may have coding potential by chance. Given that plant viruses routinely make use of stop codon read-through to expand their protein-coding potential (Skuzeski et al. 1991; Urban et al. 1996), the absence of any evidence for conserved read-through events in these dicot lineages is noteworthy.

Putative recognition elements

Short sequences exhibiting strong conservation in MEME alignments can have equivalent or lower E-values than longer, weakly conserved, sequences. Though both may be important for plant function, one of our aims was to identify potential recognition elements, which experiments suggest are between five and 20 nt long (Farley et al. 2008; Hogan et al. 2008; Pagano et al. 2009). Motifs not belonging to any of the categories discussed above were scanned for regions of dense conservation (see Fig. 1). Within these motifs, the largest window of average consensus-letter frequency greater than 0.92 (see Materials and Methods) was defined as a putative recognition element. See Supplemental File 2 for a complete list and Supplemental Files 3 and 4 for local MEME alignments.

We identified 92 PREs in the 5′ UTR and 81 in the 3′ UTR (Fig. 2C). Motifs in the 5′ UTR are generally shorter than motifs in the 3′ UTR (Fig. 2D). In the 3′ UTR, PREs peak at ~20 nt in length. Again, these conserved motifs have already been filtered to remove likely smRNA target sites; hence, though this is the size expected for smRNA target site conservation (as in Table 2), we consider these to represent non-smRNA binding elements. In both regions, PREs make up the majority of conserved motifs (Fig. 2C). In a gene ontology (GO) term analysis, genes with a 3′ UTR PRE are significantly overrepresented in two specific categories: proteins targeted to cell periphery (20 genes, 3.0-fold enrichment, p-value = 4.44×10^{-3}) and signal transduction proteins (12 genes, 4.2-fold enrichment, p-value = 1.80×10^{-2}). Genes with 5′ UTR PREs are not distinctly enriched in any category.

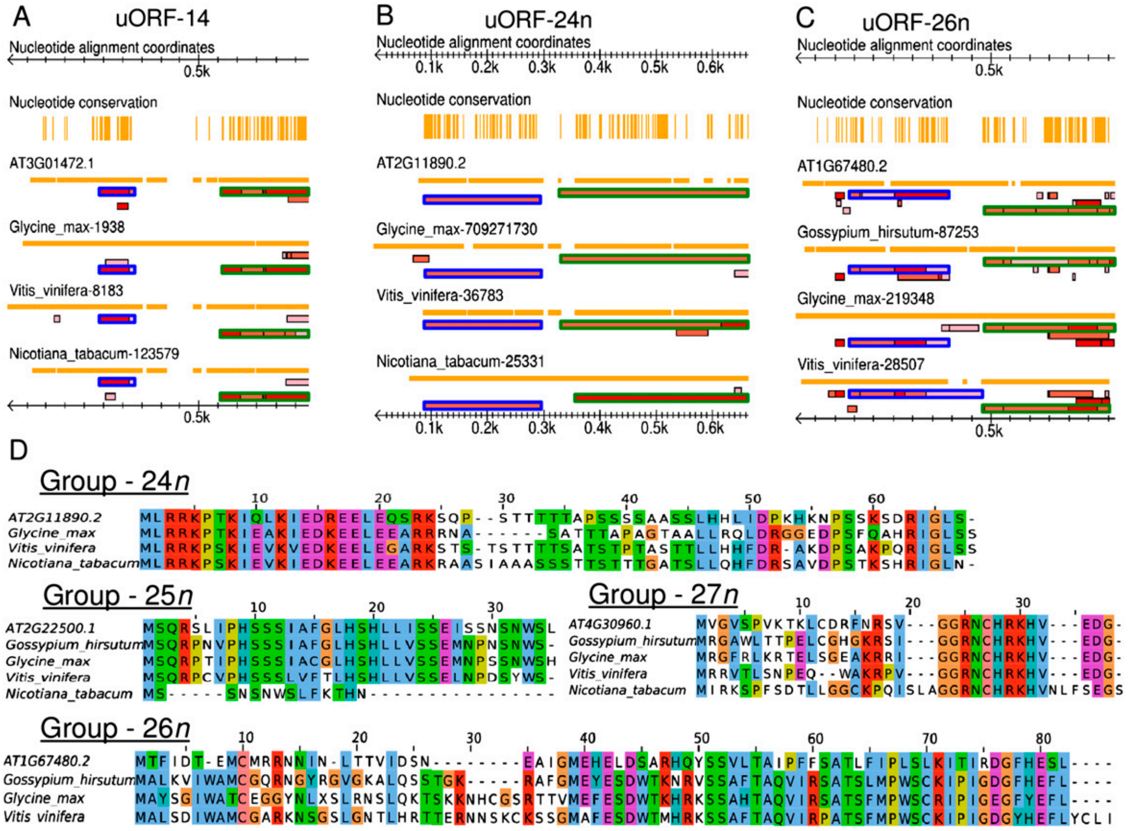
A
uORF-14
Nucleotide alignment coordinates
0.5k
Nucleotide conservation
AT3G01472.1
Glycine_max-1938
Vitis_vinifera-8183
Nicotiana_tabacum-123579
0.5k
B
uORF-24n
Nucleotide alignment coordinates
0.1k 0.2k 0.3k 0.4k 0.5k 0.6k
Nucleotide conservation
AT2G11890.2
Glycine_max-709271730
Vitis_vinifera-36783
Nicotiana_tabacum-25331
0.1k 0.2k 0.3k 0.4k 0.5k 0.6k
C
uORF-26n
Nucleotide alignment coordinates
0.5k
Nucleotide conservation
AT1G67480.2
Gossypium_hirsutum-87253
Glycine_max-219348
Vitis_vinifera-28507
0.5k
D
Group - 24n
10 20 30 40 50 60
AT2G11890.2 MLRRKPTKIQLKIEDREELEQSRKSQP--STTTTTAPSSSSAASSLHHLIDPKHKNPSSKSDRIGLS-
Glycine_max MLRRKPTKIEAKIEDKEELEEARRRNA------SATTTAPAGTAALLRQLDRGGEDPSFQAHRIGLSS
Vitis_vinifera MLRRKPSKIEVKVEDKEELEGARKSTS-TSTTTTSATSTPTASTTLLHHFDR-AKDPSAKPQRIGLSS
Nicotiana_tabacum MLRRKPSKIEVKIEDKEELEEARKRAASIAAASSSTTSTTTGATSLLQHFDRSAVDPSTKSHRIGLN-
Group - 25n
10 20 30
AT2G22500.1 MSQRSLIPHSSSIAFGLHSHLLISSEISSNSNWSL
Gossypium_hirsutum MSQRPNVPHSSSIAFGLHSHLLVSSEMNPNSNWS-
Glycine_max MSQRPTIPHSSSIACGLHSHLLVSSEMNPSSNWSH
Vitis_vinifera MSQRPCVPHSSSLVFTLHSHLLVSSELNPDSYWS-
Nicotiana_tabacum MS-----SNSNWSLFKTHN-----------------
Group - 27n
10 20 30
AT4G30960.1 MVGVSPVKTKLCDRFNRSV---GGRNCHRKHV---EDG-
Gossypium_hirsutum MRGAWLTTPELCGHGKRSI---GGRNCHRKHV---EDG-
Glycine_max MRGFRLKRTELSGEAKRRI---GGRNCHRKHV---EDG-
Vitis_vinifera MRRVTLSNPEQ--WAKRPV---GGRNCHRKHV---EDG-
Nicotiana_tabacum MIRKSPFSDTLLGGCKPQISLAGGRNCHRKHVNLFSEGS
Group - 26n
10 20 30 40 50 60 70 80
AT1G67480.2 MTFIDT-EMCMRRNNIN-LTTVIDSN---------EAIGMEHELDSARHQYSSVLTAIPFFSATLFIPLSLKITIRDGFHESL----
Gossypium_hirsutum MALKVIWAMCGQRNGYRGVGKALQSSTGK-----RAFGMEYESDWTKNRVSSAFTAQVIRSATSLMPWSCKIPIGDGFHEFL----
Glycine_max MAYSGIWATCEGGYNLXSLRNSLQKTSKKNHCGSRTTVMEFESDWTKHRKSSAHTAQVIRSATSFMPWSCRIPIGEGFYEFL----
Vitis_vinifera MALSDIWAMCGARKNSGSLGNTLHRTTERNNSKCKSSGMAFESDWTMHRKSSAFTAQVIRPATSFMPWSCKIPIGDGYHEFLYCLI

FIGURE 3. Novel CPuORFs. (*A–C*) Transcript alignments are shown in miniature followed by all possible ORFs, where the darkness of red indicates the context strength of the AUG. CPuORFs are bordered in blue; the beginning of each associated mORF is bordered in green. Orange vertical lines ("Nucleotide conservation" lane) indicate that all residues are identical in that column. Note gaps in sequence identity *between* uORF and mORF. The orange horizontal line *below* each accession indicates sequence coverage or indels relative to the alignment. In the scale bar, 0.5 k = 500 nt. Sequences are ordered, *top* to *bottom*, relative to their phylogenetic distance from *A. thaliana*. (*A*) Previously confirmed CPuORF-14 (Hayden and Jorgensen 2007). (*B*) CPuORF-24n. (*C*) CPuORF-26n. (*D*) Protein alignments of novel CPuORFs. Color scheme is ClustalX default. Within alignments, sequences are ordered relative to their distance from *A. thaliana*. Related to Table 3.

PUF-binding motifs, possible Expansin localization signals, and other novel elements are conserved in the 3′ UTR

PREs are derived from conserved motifs associated with an individual orthologous group. Yet, it is likely that certain PRE patterns are shared between multiple orthologous groups. To generalize PREs based on their shared sequence content, we assessed the degree to which 4-mer words were enriched among all PREs. We created random expectation distributions (n = 10,000) for each 4-mer such that the length distribution of the PRE data sets was recapitulated. We then assessed the p-value of the actual enrichment based on its position within this distribution. The 3′ UTR and 5′ UTR regions were treated independently.

Nearly all PUF-binding elements in metazoans, plants, and yeast contain a core UGUA sequence (Gerber et al. 2004, 2006; Francischini and Quaggio 2009). Indeed, UGUA was one of the most significantly enriched 4-mers in the PREs of the 3′ UTR and, notably, not in PREs of the 5′ UTR (Table 4). Sixteen of the ~80 genes have UGUA-containing PREs in the 3′ UTR (Table 5). These are most likely authentic PUF binding sites for the following two reasons. First, one of these, *LHCa4*, was one of five mRNAs found to interact with the PUF protein, APUM2, in a yeast three-hybrid screen (Francischini and Quaggio 2009). Another protein identified in the same screen, a DNAJ, also has a conserved UGUA motif but with an E-value (6.7×10^{-5}) that did not pass our conservative statistical cutoff. Of the remaining three APUM2 targets, all have orthologs in at least four species with 3′ UTR coverage, but the UGUA element is not conserved. Second, the mRNA of CLAVATA1 (At1g75820), which is critical for meristem maintenance, was also shown to interact with APUM2 (Francischini and Quaggio 2009). Indeed, the CLAVATA1 3′ UTR contains four UGUA tetramers, although each individual motif had a high E-value in our analysis (data not shown). However, At4g20270.1 is BAM3, an in-paralog of CLAVATA1, and its longer UGUA-PRE is highly conserved (Table 5). These

TABLE 3. All motifs associated with protein-coding potential and their annotations based on the accession for the *A. thaliana* representative of the orthologous group

A. thaliana accession	*A. thaliana* mORF annotation	CRF *p*-value	uORF *p*-value[a]	Predicted cause of coding potential[b]
AT2G11890	Adenylate cyclase	1.64×10^{-19}	1.19×10^{-15}	uORF(24n)
AT4G36990	Heat shock transcription factor 4 (HSTF4)	1.28×10^{-17}	7.33×10^{-15}	uORF(18)
AT3G12012	Mic-1 homolog	1.67×10^{-17}	2.07×10^{-15}	uORF(8)
AT3G62420	bZIP transcription factor (AtbZip53)	5.25×10^{-16}	1.42×10^{-24}	uORF(1)[c,d]
AT3G01470	HD-ZIP 1 transcription factor (ATHB-1)	9.45×10^{-16}	6.19×10^{-18}	uORF(14)
AT1G29950	bHLH transcription factor	5.57×10^{-15}	4.99×10^{-11}	uORF(15)
AT3G25570	S-adenosylmethionine decarboxylase (SAMDC)	3.36×10^{-14}	1.47×10^{-15}	uORF(3)
AT4G25690	Expressed protein	7.01×10^{-12}	2.86×10^{-15}	uORF(4)
AT5G07840	Ankyrin repeat family protein	4.85×10^{-11}	5.18×10^{-11}	uORF(5)
AT1G23150	Expressed protein	3.86×10^{-10}	2.94×10^{-8}	uORF(12)
AT2G43020	Polyamine oxidase family (PAO2, PAO3)	1.11×10^{-8}	6.48×10^{-11}	uORF(6)
AT2G22500	Mitochondrial substrate carrier family protein	4.79×10^{-8}	3.16×10^{-8}	uORF(25n)
AT5G01710	Expressed protein	2.55×10^{-7}	7.77×10^{-5}	uORF(17)
AT1G67480	Kelch repeat-containing F-box family protein	3.54×10^{-7}	1.12×10^{-7}	uORF(26n)
AT4G30960	CBL-interacting protein kinase 6 (CIPK6)	3.77×10^{-7}	8.98×10^{-7}	uORF(27n)[d]
AT1G48600	Methyltransferase	4.42×10^{-7}	1.39×10^{-7}	uORF(13)
AT4G34590	bZIP transcription factor (AtbZip11)	5.56×10^{-7}	8.10×10^{-8}	uORF(1)[c]
AT4G30960	CBL-interacting protein kinase 6 (CIPK6)	1.88×10^{-4}	8.98×10^{-7}	uORF(27n)[d]
AT3G62420	bZIP transcription factor (AtbZip53)	7.80×10^{-4}	1.42×10^{-24}	uORF(1)[d]
AT1G36730	Eukaryotic translation initiation factor 5, putative	1.24×10^{-3}	4.05×10^{-5}	uORF(7)
AT1G03260	Expressed protein	5.19×10^{-11}	#N/A	Non-AUG start (CUG)[e]
AT1G32700	Zinc-binding family protein	5.19×10^{-9}	#N/A	Non-AUG start (CUG)[e]
AT2G25110	MIR domain-containing protein	1.86×10^{-7}	#N/A	Non-AUG start (GUG)
AT3G16630	Kinesin motor family protein	7.26×10^{-5}	#N/A	Non-AUG start (UUG)
AT4G16280	Flowering time control protein (FCA)	1.37×10^{-3}	#N/A	Non-AUG start (CUG)[e]
AT5G14500	Aldose 1-epimerase family protein	5.06×10^{-3}	#N/A	Non-AUG start (CUG)[e]
AT1G55760	BTB/POZ domain-containing protein	5.21×10^{-3}	#N/A	Non-AUG start (CUG)[e]
AT4G26850	Expressed protein	2.48×10^{-11}	#N/A	Unknown[d]
AT2G18040	Peptidyl-prolyl *cis*-trans isomerase	9.79×10^{-7}	#N/A	Unknown
AT1G01060	myb family transcription factor (LHY)	1.31×10^{-5}	#N/A	Unknown
AT4G26850	Expressed protein	6.94×10^{-5}	#N/A	Unknown[d]
AT1G57680	Expressed protein	3.94×10^{-3}	#N/A	Unknown

[a] "uORF" is different from continuous reading frame (CRF) in that it contains an in-frame AUG codon; hence, not all CRFs will contain a uORF.
[b] Parenthetical integer next to "uORF" indicates the homology group associated with prior *A. thaliana* and *O. sativa* comparisons (Hayden and Jorgensen 2007), and parenthetical string next to "non-AUG start" indicates the likely start codon based on alignments.
[c] Associated mORF clusters into a separate group in spite of the uORF being in the same homology group.
[d] Coding potential is found in two distinct frames.
[e] Identified in Simpson et al. (2010) and Wamboldt et al. (2009).

results show that our pipeline was able to identify a subset of authentic 3′ PREs that bind PUF-domain RNA binding proteins. Generally, the UGUA-containing PREs appear to be dispersed among functionally unrelated mRNAs, except for the two subunits of the photosystem I light harvesting complex, AT3G47470.1 and AT3G61470.1 (Table 5). Notably, transcripts associated with photosynthesis are known to be enriched around the chloroplast (Marrison et al. 1996), although the mechanism is still unknown.

The UUUG 4-mer is enriched in 3′ UTRs to a similar extent as UGUA (Table 4). These two 4-mers are occasionally found together but not in a consistent arrangement relative to one another (Table 5). They may be functionally independent. Moreover, their co-occurrence does not deviate from what would be expected given their individual distributions among all PREs (p-value = 0.30, χ^2 test), suggesting that UUUG is generally unrelated to PUF-binding.

Expansins are one of the few groups of plant mRNAs that have been shown to be localized to specific subcellular sites (Im et al. 2000). A conserved [AG]CCCGC-containing motif was found in four out of 12 available Expansin 3′ UTRs (Fig. 4A). A region upstream of the [AG]CCCGC motif is also conserved but more specific to each Expansin. The *Zinnia elegans* Exp1 mRNA (gi|7025490|gb|AF230331.1), which was shown to localize to a particular region of the cell periphery (Im et al. 2000), also contains this pair of elements (Fig. 4B). They, too, appear to be position-independent relative to the mORF stop site and relative to one another. Taken together, Expansin mRNAs may form a small RNA regulon, where a common type of motif drives the coregulation of several

TABLE 4. 4-mer words significantly enriched in PREs from 5′ and 3′ UTRs

5′ UTR		3′ UTR	
4-mer	*p*-value	4-mer	*p*-value
AGAA	$<1 \times 10^{-5}$	UGUA	$<1 \times 10^{-5}$
AGAU	$<1 \times 10^{-5}$	UUUG	$<1 \times 10^{-5}$
UUCU	$<1 \times 10^{-5}$	AAGG	0.0051
AGGG	0.0006	AAUA	0.0066
AUGG	0.0015	UGGU	0.0104
AGGA	0.0019	AAGC	0.0246
UUUU	0.0151	GAGG	0.0296
UCUU	0.0154	UGCA	0.0377
AGAG	0.0224	UUCU	0.0447
CCUC	0.0274		
CGAU	0.0374		

distinct mRNAs, but where a secondary motif confers some degree of additional specificity.

5′ UTR PREs are enriched in purine-rich 4-mers

In an attempt to better classify and characterize the 5′ PREs, we examined which 4-mer words were overrepresented. The most significant 4-mers among the 5′ UTR PREs are purine-rich words (e.g., AGAA) (Table 4). These purine-rich PREs are in a continuum with the [AG] repeat motifs that were set aside early in the motif extraction pipeline. We explored whether the motifs identified here function at the transcriptional or post-transcriptional level. First, [AG] repeats can act at the transcriptional level (Santi et al. 2003; Meister et al. 2004; Kooiker et al. 2005). Purine-rich motifs are also present in many 5′ UTRs and are associated with broad clusters of transcription start sites (Yamamoto et al. 2009). However, purine-rich repeats in the 5′ UTR have also been reported to enhance translation of the *ntp303* gene in *N. tabacum* (Hulzink et al. 2002). Using mRNA half-life data from a prior study (Narsai et al. 2007), we found that transcripts containing a purine-rich PRE in the 5′ UTR have a median half-life of 9.6 h, which is more than twice as long as the median half-life of 3.8 h for *Arabidopsis* transcripts in general ($p < 10^{-5}$, sign test). No other PRE category correlated with RNA stability. This result is consistent with a new role of the purine-rich PRE in RNA stability.

To a lesser extent, pyrimidine-rich words—UUCU, UUUU, UCUU—are more common than expected in the 5′ UTR (Table 4). A few of these PREs resemble "Y-Patch" promoter elements, which can extend into the 5′ UTR (Yamamoto et al. 2007), although it is still unclear if Y-patches are acting post-transcriptionally or transcriptionally. In metazoans, the expression of ribosomal protein mRNAs is governed by 5′ terminal oligopyrimidine motifs (Avni et al. 1994), and translation of mRNAs with this feature is favored in wheat cell lysates (Shama and Meyuhas 1996). Yet, we find no ribosomal proteins with conserved CU-rich elements (Supplemental Files 2 and 3).

We examined whether 5′ PREs are likely to represent transcription factor binding sites. The Telobox (AAACCCUA or its reverse complement) was found in six of the 5′ PREs (Supplemental File 2). Because this motif occurs commonly in both orientations around plant transcription start sites (Tremousaygue et al. 2003; Molina and Grotewold 2005), it is a candidate for functioning at the DNA level.

Based on the AGRIS database of *A. thaliana* promoter motifs (Davuluri et al. 2003), only 11 out of 92 5′ UTR PREs have evidence for containing transcription factor binding sites (Supplemental File 2; Fig. 2C). For comparison,

TABLE 5. Possible PUF-binding PREs within the 3′ UTR

A. thaliana accession	Gene annotation	PRE consensus sequence[a]
AT3G09980.1	Expressed protein	UAUAAACAGG**UUUGUA**ACUAA
AT4G01100.1	Adenine nucleotide transporter 1 (ADNT1)	UGCUAUU**UUUGUA**GGC**AAGG**G
AT5G16000.1	Leucine-rich repeat family protein (NIK1)	UGCU**UGUA**UUCAUC**UGUA**AA
AT3G47470.1	Chlorophyll A-B binding protein 4 (LHCa4)	CUUUAA**UGUA**CAGAGGAACU
AT4G20270.1	Leucine-rich repeat transmembrane protein kinase (BAM3)	**UGUA**CAGUAGGAU**UGGU**GGG
AT3G57200.1	Hypothetical protein	AUUACCCAAGCGC**UGGUGUA**
AT4G14900.1	Hydroxyproline-rich glycoprotein family protein (FRIGIDA-like)	G**UUUGUA**AUCACUAACCGUU
AT2G40110.1	Yippee family protein	AAA**UGUA**CAUUCUUUAACC
AT1G07470.1	Transcription factor IIA large subunit, putative	UUGGCCUGU**UGUA**CAUA
AT1G53910.3	AP2 domain-containing protein RAP2.12	**UGUAAAUA**AAGCUACAU
AT3G11660.1	Harpin-induced family protein	UGAAU**UGUA**CAU**UUUG**C
AT3G18820.1	Ras-related GTP-binding protein, putative	U**UGUA**CAUUAGUG**UUUG**
AT3G61470.1	Chlorophyll A-B binding protein (LHCa2)	**UGUA**CA**AAUA**CC**UUUG**U
AT2G42670.2	Expressed protein	**UGUA**CAUAUU**AAUA**UA
AT1G32400.1	Senescence-associated family protein	GAG**UUUGUGUA**
AT1G08420.1	Kelch repeat-containing protein	**UGUA**U

The accession and associated annotation for the *A. thaliana* representative of the orthologous group is given.
[a]Common 4-mers with *p*-value ≤ 0.01 (Table 4) are in bold font.

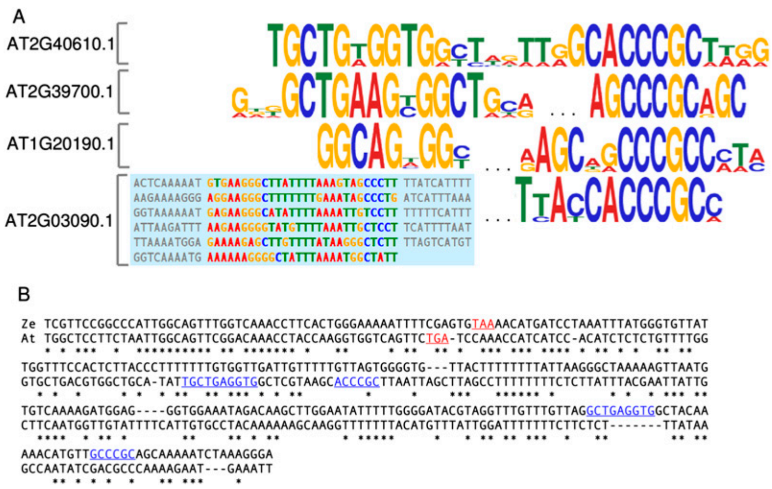
A
AT2G40610.1
AT2G39700.1
AT1G20190.1
AT2G03090.1
ACTCAAAAAT GTGAAGGGCTTATTTTAAAGTAGCCCTT TTATCATTTT
AAGAAAAGGG AGGAAGGGCTTTTTTTGAAATAGCCCTG ATCATTTAAA
GGTAAAAAAT GAGAAGGGCATATTTTAAAATTGTCCTT TTTTTCATTT
ATTAAGATTT AAGAAGGGGTATGTTTTAAATTGCTCCT TCATTTTAAT
TTAAAATGGA GAAAAGAGCTTGTTTTATAAGGGCTCTT TTAGTCATGT
GGTCAAAATG AAAAAAGGGGCTATTTAAAATGGCTATT
B
Ze TCGTTCCGGCCCATTGGCAGTTTGGTCAAACCTTCACTGGGAAAAATTTTCGAGTGTAAAACATGATCCTAAATTTATGGGTGTTAT
At TGGCTCCTTCTAATTGGCAGTTCGGACAAACCTACCAAGGTGGTCAGTTCTGA-TCCAAACCATCATCC-ACATCTCTCTGTTTTGG
TGGTTTCCACTCTTACCCTTTTTTTGTGGTTGATTGTTTTTGTTAGTGGGGTG---TTACTTTTTTTTATTAAGGGCTAAAAAGTTAATG
GTGCTGACGTGGCTGCA-TATTGCTGAGGTGGCTCGTAAGCACCCGCTTAATTAGCTTAGCCTTTTTTTTCTCTTATTTACGAATTATTG
TGTCAAAAGATGGAG----GGTGGAAATAGACAAGCTTGGAATATTTTTGGGGATACGTAGGTTTGTTTGTTAGGCTGAGGTGGCTACAA
CTTCAATGGTTGTATTTTCATTGTGCCTACAAAAAAGCAAGGTTTTTTTACATGTTTATTGGATTTTTTTCTTCTCT-------TTATAA
AAACATGTTGCCCGCAGCAAAAATCTAAAGGGA
GCCAATATCGACGCCCAAAAGAAT---GAAATT

FIGURE 4. Expansin 3′ UTRs contain a combination of conserved sequences, which are also present in the 3′ UTR of the localized *Zinnia elegans* Exp1 mRNA. (*A*) Sequence LOGO plots are generated by MEME; information content of a position is represented by stack height, which is multiplied by letter frequency at that position to give letter height. Each Expansin mRNA has one upstream variable region and one downstream RCCCGC-core motif. The entire MEME-derived alignment is given for AT2G03090.1 group's variable region. (*B*) Alignment of the *Zinnia elegans* ExpansinA1 mRNA (Ze)—gi|7025490|gb|AF230331.1—which is targeted to specific subcellular sites, with its *A. thaliana* ortholog (At)—AT2G40610.1. mORF stop codons are in red, and conserved elements are in blue. Asterisks indicate that the column letters are identical. Note: (*U*)racils are shown as (*T*)hymines.

3′ UTRs contain few confirmed promoter motifs, yet eight out of 96 PREs would be annotated as transcription factor binding sites (Supplemental File 2). A similar result was obtained from the AthaMap project (Bülow et al. 2010) (data not shown). Careful inspection of the 11 cases revealed that the most conserved core of the PRE either lies beside the transcription factor motif or extends beyond the motif (data not shown). Therefore, it seems likely that few of our 5′ UTR PREs are acting as transcription factor binding sites.

Conservation and variation of 5′ PREs reveal patterns of uORF evolution

The 4-mer associated with a strong-context start codon, AUGG, is significantly enriched among PREs in the 5′ UTR (Table 4). In fact, many AUGs appear within conserved motifs, in both weak and strong contexts (Table 6). Since sequences with extensive conserved protein-coding potential were already removed prior to this portion of the analysis, these 4-mers are conserved, at least in part, for peptide-independent effects relating to uAUG initiation/reinitiation.

Because the effect of uAUGs on gene expression is undoubtedly via translation, we focused on AUG-containing 5′ PREs to glean a few trends governing the evolutionary changes in 5′ PREs and their associated uORFs (Fig. 5; Supplemental Fig. S2). As the first trend, when a uORF overlaps the major ORF, their relative frame arrangement is usually conserved (AS1, PRR2/TOC2, smRNP At3g14080). In contrast, uORF clus-

TABLE 6. uAUG-containing PREs within the 5′ UTR

A. thaliana accession	Gene annotation[a]	PRE consensus sequence[b]
AT5G06510.1[c]	CCAAT-binding transcription factor	GUACCGAC**AUG**GCUCCUAACUA**AUG**GGGU
AT4G26570.1	Calcineurin B-like protein 3 (CBL3)	GAA**AUG**GUUAAAAGGU**AUG**GAGUGUUUUG
AT4G18020.1	Pseudo-response regulator 2 (APRR2) (TOC2)	GAGAAAGG**AUG**CCAAACCAG
AT3G48210.1	Expressed protein	AAGUAAAA**AUG**GCGGGCUAA
AT1G71980.1	Zinc finger (C3HC4-type RING finger) protein	**AUG**GAAGCUG**AUG**UUUCCAU
AT1G19330.1	Expressed protein	UCAGCA**AUG**C**AUG**AUCUUCA
AT5G62000.2[c]	Transcription factor B3 protein (ARF2)	CAG**AUG**AGAGAUCUGAGC
AT3G54020.1	Phosphatidic acid phosphatase-related (IPCS1)	UGAAGUAAU**AUG**GAAGUG
AT3G63200.1[b]	Patatin-related	CCAUUA**AUG**CCUCUCAGC
AT5G47100.1	Calcineurin B-like protein 9 (CBL9) - **miR847**	AAG**AUG**GUUUUG**AUG**A
AT2G02710.3	PAC motif-containing protein	CAC**AUG**GGAUUGGG
AT1G18660.1[c]	Zinc finger (C3HC4-type RING finger) protein	UGGUCCGUGU**AUG**
AT1G72820.1	Mitochondrial substrate carrier family protein	CGACG**AUG**GUCG
AT3G14080.2[c]	Small nuclear ribonucleoprotein, putative - **miR159**	CCA**AUG**CCAUU
AT4G03415.1[c]	Protein phosphatase 2C family protein	AUCAG**AUG**U
AT5G17640.1	Expressed protein	CA**AUG**GGG
AT2G22430.1	Homeobox-leucine zipper protein 6 (HB-6)	G**AUG**G
AT2G37630.1	myb family transcription factor (MYB91)	**AUG**GG

[a] If the *A. thaliana* representative of the motif is a possible microRNA binding site as predicted by psRNATarget (see Materials and Methods), the microRNA family is given in bold next to the gene annotation.
[b] AUG's are in bold font.
[c] AUG has been lost in the *Arabidopsis* lineage but is present in all others.

ters that do not overlap the mORF lack a consistent frame arrangement with the mORF (e.g., S6K, CBL9, NF-YA10, HD-ZIP-6, At1g19330, ADC). This pattern is quite striking in the case of smRNP, where all three overlap-uORFs occupy the −1 frame with respect to the mORF. In contrast, four nonoverlap uORFs occupy two different frames.

Second, in cases where the PRE harbors multiple AUGs, their relative reading frame arrangement is generally highly conserved. Examples include calcineurin CBL3, myb protein AS1 uORFs 2 and 3, RING At1g71980, CCAAT enhancer binding factor NF-YA10, and expressed protein At1g19330. In contrast, uORFs that are not initiated in the PRE tend to have a variable frame position with respect to the PRE's uORFs (PRR2 uORF2, RING At1g71980 uORF3). Although we selected for situations like this by focusing on conserved PREs, the trend is, nevertheless, noteworthy. Because the initiation at a given AUG is strongly affected by an overlapping uORF (Hanfrey et al. 2005; Roy et al. 2010), we suspect that most of these uORFs modulate the strength of initiation among each other and at the major ORF.

Third, in the case of PREs that do not specify overlap-uORFs, the relative position of the PRE in the 5′ UTR appears to be somewhat restrained (RING, CBL9, HD-ZIP-6, At1g19330, ADC), even though the remaining 5′ UTR sequences have usually diverged to the point that they cannot be aligned.

Fourth, the uORF pattern in *Arabidopsis* and other Brassicaceae often differs from the consensus in the other families. Examples can be seen in S6K, PRR2/TOC2, the smRNP, the RING At1g71980, NF-YA10, and HD-ZIP-6.

Fifth, in-paralogs often differ in their PRE and uORF pattern. Examples include *Arabidopsis* S6K, calcineurins CBL9 and CBL1, NF-YA10 and 2, HD-ZIP-6 and 16, expressed protein At1g19330, ADC, and CTR1.

Together, these case studies show that uORFs associated with PREs tend to be highly conserved, whereas the uORFs that reside outside of PREs tend to be quite variable. However, even uORFs associated with PREs have been subject to significant variation during evolution of the Brassicaceae, and even otherwise highly conserved uORFs tend to vary between in-paralogs. All of these observations suggest that uORFs are one of the dials of molecular evolution that alter gene expression levels such that they are fine-tuned to be adaptive in specific lineages.

DISCUSSION

Patterns of element enrichment in the 5′ versus 3′ UTR reflect the canonical model of eukaryotic translation

Judging by the distribution of PREs, the 5′ and 3′ UTRs mediate distinct forms of post-transcriptional regulation. This is expected given the distinct molecular events occurring within each region during translation (Jackson et al. 2010). Scanning of the small ribosomal subunit through the 5′ UTR is likely to displace transient interaction between RNA and *trans*-acting factors whereas no such restriction applies to the 3′ UTR, which is thought to be free of ribosome traffic (Gu et al. 2009). The 5′ UTRs retain sequences that engage the ribosome itself, i.e., CPuORFs, non-CPuORFs and non-AUG initiation sites. The 5′ UTRs also harbor purine-rich and pyrimidine-rich repeat sequences (Fig. 2C). The remaining PREs in the 5′ UTR tend to be short (Fig. 2D). In contrast, the 3′ UTR largely lacked

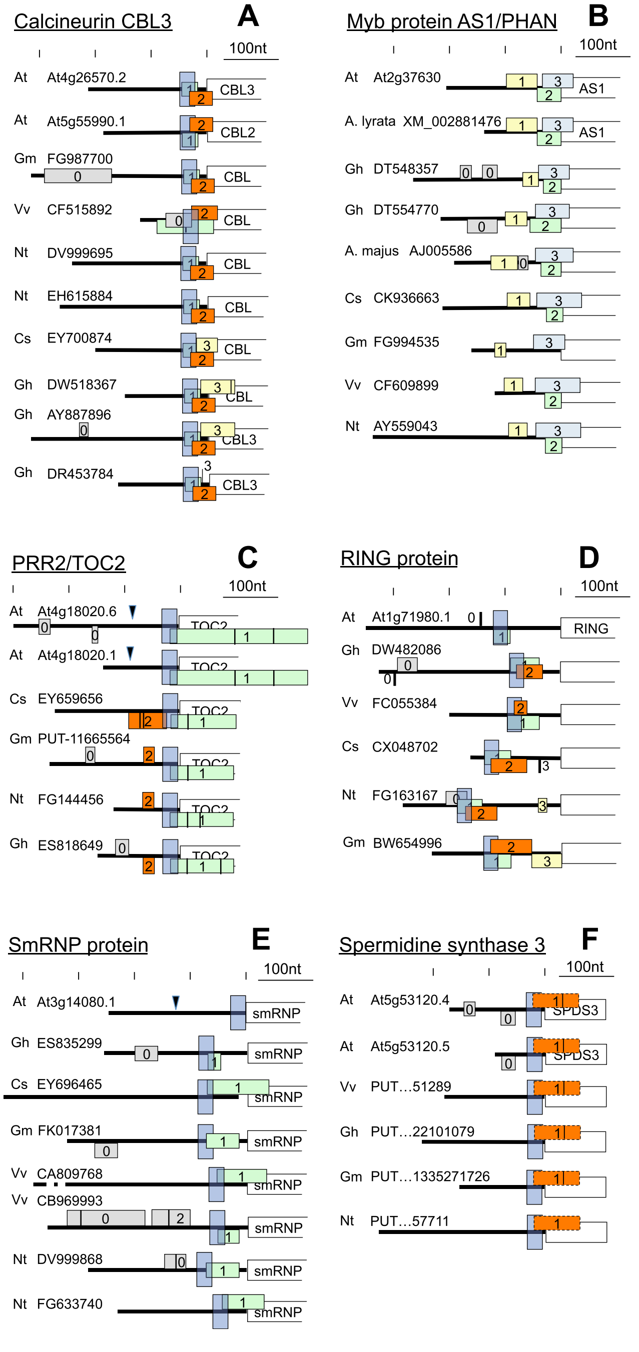

FIGURE 5. (Legend on next page)

conserved peptides and AUG-related motifs, and low-complexity repeat motifs were uncommon. Instead, the 3′ UTR is where more complex PREs are found (Fig. 2C). Among these, putative binding sites for PUMILIO-type RNA binding proteins were detected only in the 3′ UTR. In the case of Expansin mRNAs, which must harbor a determinant responsible for their subcellular targeting, a candidate PRE was discovered in the 3′ UTR (Fig. 4).

Conserved target sites for smRNA are an exception to this trend of regional bias (Table 2). This is in striking contrast to metazoans, where 3′ UTR sites are much more common. Experiments on select genes in *A. thaliana* indicate that the position of a smRNA target site—3′ UTR vs. 5′ UTR vs. CDS—does not correlate with the degree to which an affected mRNA is degraded and/or translationally silenced (Brodersen et al. 2008), so perhaps a lack of regional bias is to be expected.

One extension of the RNA regulon concept is that, analogous to transcriptional modules, some mRNAs will be under combinatorial control by a suite of factors. Overall, only 8% ($n = 7$) of 5′ UTRs harbor legitimate examples of multiple conserved motifs (Supplemental File 2). In contrast, 16% ($n = 13$) of PRE-containing 3′ UTRs have significant conservation outside of the most conserved PRE, and generally this conservation is longer and more informative (Supplemental File 2). This pattern supports a model in which the 3′ UTR, because it is free of ribosomes, is a more appropriate platform for combinatorial regulation and signal integration than the 5′ UTR.

FIGURE 5. Conservation and variation in the 5′ UTR of mRNAs that harbor uAUG containing PREs and/or uORFs. Each subpanel shows the uORF pattern (colored rectangles) and, where present, the PRE (tall blue box) of one orthologous group; some orthologous groups were supplemented manually with members from additional species. The Gene Identifiers come from TAIR (At#g#####), PlantGDB (PUT...), or GenBank (all others). The start of the major ORF is symbolized by a white box. The 5′ upstream region is indicated by a thick line. Vertical lines *inside* the uORF stand for internal in-frame AUG codons. Unless indicated, identical color or number does not imply sequence similarity. uORFs *centered* on the UTR are in the same frame as the major ORF, while those *above* the line and *below* the line are in the −1 and +1 frames, respectively. Selected exon-exon junctions are indicated by black arrowheads. Species abbreviations are as follows: (At) *Arabidopsis thaliana*, (Cs) *Citrus sinensis*, (Gh) *Gossypium hirsutum*, (Gm) *Glycine max*, (Nt) *Nicotiana tabacum*, (Vv) *Vitis vinifera*. (*A*) Calcineurin CBL3 and CBL2. This 5′ UTR contains a highly conserved PRE with two AUGs in a conserved frame arrangement. The downstream uAUG leads into a uORF that overlaps the main ORF (orange). A third uORF is present in some species (yellow). Its uAUG overlaps the stop of uORF1. (*B*) ASSYMMETRIC1/PHANTASTICA. This 5′ UTR lacks a recognizable PRE, yet its uORF pattern is highly conserved. Invariable features include one overlap uORF (blue) and one upstream uORF (yellow). An additional uORF ends at the start codon of the main ORF (green); it is present in all species except soybean. The consistent pattern notwithstanding, the uORF nucleotide and peptide sequences are not highly conserved. (*A. majus*) *Antirrhinum majus*, (*A. lyrata*) *Arabidopsis lyrata*. (*C*) PSEUDO-RESPONSE REGULATOR2/TOC2. This 5′ UTR harbors a conserved PRE with a conserved uAUG that leads into a long overlap-uORF (green). An additional uORF is usually present upstream, except in *Arabidopsis*. (*D*) RING protein At1g71980. Its PRE invariably harbors the start codon of a short uORF (green). A second uAUG is usually present in the PRE (orange), except in *Arabidopsis*. Additional short uORFs are common. (*E*) smRNP protein At3g14080. A PRE with a conserved uAUG leads into a uORF of variable length (green), which may overlap the major ORF. Upstream uORFs may be present. In *Arabidopsis*, the uAUG and uORF have been lost from the PRE. (*F*) Spermidine synthase. The PRE contains a putative non-AUG start codon. The non-AUG uORF downstream (orange, stippled) is highly conserved in length and amino acid sequence (data not shown). Related to Supplemental Figure S2.

Identification of novel CPuORFs

We identified four novel CPuORFs, at least three of which are conserved in dicots and monocots, i.e., pan-angiosperm. The extensive length and conservation of the newly identified CPuORFs 26n and 24n (Fig. 3B–D) begs the question of whether CPuORFs are, in fact, regulatory elements or, alternatively, whether their associated mRNAs should be considered bicistronic transcripts. Strikingly, CPuORF-24n is identical to GenBank accession CDC26, a small subunit of the anaphase promoting complex (APC) (Dong et al. 2007; de F Lima et al. 2010). CPuORF-24n/CDC26 appears as a distinct transcript in metazoans and fungi, but in plants it is consistently found as a uORF upstream of a mORF that is annotated as adenylate cyclase. Notwithstanding that the existence of adenylate cyclase and cyclic-AMP in plants is controversial, as it stands, the associated mRNA is a good candidate for a multicistronic eukaryotic transcript. Yet, the conserved synteny, proximity, and cotranscription of CPuORF-24n/CDC26 and the putative adenylate cyclase suggest that, in plants, these two proteins require cotranslation.

Of the remaining three CPuORF loci identified in this study, little is known. CPuORF-27n is linked to a protein kinase involved in salt stress tolerance, and CPuORF-26n is upstream of an uncharacterized F-box protein. CPuORF-25n is associated with a dicarboxylic acid transporter in mitochondria.

What do conserved RNA sequence motifs reveal about the architecture of post-transcriptional regulation?

Certain genes and gene families are tightly associated with specific types of RNA sequence elements; the CPuORFs in the bZip transcription factor family are well-known. The entire polyamine synthesis pathway is a “hotspot” for translational control that incorporates CPuORFs, non-CPuORFs, non-AUG uORFs, and ribosomal frameshifting (Ivanov et al. 2010). In plants, thermospermine synthase is regulated by a uORF-dependent transcription factor (Imai et al. 2006), and arginine decarboxylase contains a conserved uORF (Chang et al. 2000), as do several catabolic polyamine oxidases (Hayden and Jorgensen 2007; Table 3; Supplemental Fig. S2). Aside from the well-investigated CPuORF in S-adenosylmethionine decarboxylase (Table 3) that reduces translation through polyamine-induced ribosome stalling, we have identified a PRE in arginine decarboxylase, a rate-limiting enzyme in the plant polyamine synthesis pathway (Supplemental Fig. S2), and a PRE in plant spermidine synthase (AT5G53120) (Supplemental File 3; Fig. 5F). Interestingly, this PRE contains the best match of all 5′ PREs to the noncanonical start site, AgaCUGG, possibly representing a uORF with a non-AUG start codon.

A second example for a focal point of RNA-level control that arose from this study is the CCAAT enhancer binding proteins. Two CCAAT-binding transcription factors have conserved microRNA target sites in the 3′ UTR (Table 2). Additionally, two other family members have PREs located in the 5′ UTR (Table 6; Supplemental File 2 and 3). Of these, the PRE of At3g05690 may possess a peptide-independent uORF. Notably, mammalian CCAAT-binding transcription factors also possess uORFs, which, in this case, govern CDS start-site selection (Calkhoven et al. 1994; Lincoln et al. 1998). This finding suggests that CCAAT-binding protein joins a small number of other genes, including S-adenosylmethionine decarboxylase and bZip transcription factors, which are regulated by uORFs across the plant-metazoan divide.

Also of note are the calcineurins. The subtype comprising *Arabidopsis* CBL1 and CBL9 has a 5′ PRE with a variable uORF, and the PRE in CBL9 can be recognized by the seed region of miRNA847 (data not shown); whereas the subtype comprising CBL2 and CBL3 has a 5′ PRE with a highly conserved uORF cluster. Finally, among the cytochrome c oxidases two family members have seemingly different modes of post-transcriptional control; AT3G15640.1 has a conserved microRNA binding site in the 5′ UTR (Table 2), while AT2G07687.1 has an extensive but clearly unrelated PRE in the 3′ UTR (Table 5). Such elements could be the basis for differential regulation.

Concerning the concept of RNA regulons, one expectation is that multiple, functionally related mRNAs share the same type of PRE. Again, the common sucrose-regulated CPuORF found in several bZip transcription factor mRNAs is a valid example. Here, we took note of a common PRE found in four Expansin mRNAs, including Expansins known to be targeted to specific subcellular sites.

It has been previously reported that RNA-binding proteins, among them multiple PUF-family proteins, associate with functionally related sets of genes (Hogan et al. 2008). Indeed, in our analysis, we find that LHCa2 and LHCa4, closely related genes, both contain a UGUA-PRE, a putative PUF-binding site.

Overall, however, we find very few instances of conserved coregulation of functionally related genes by the same motif and no examples to match the coregulation of the vacuolar ATPase subunits in *Drosophila* (Gerber et al. 2006). For example, genes with conserved UGUA-containing PREs were not significantly enriched in any GO category (data not shown). Undoubtedly, we have false negatives resulting from a lack of sequence coverage in the 3′ UTR, but genome comparisons within the fungal genus *Aspergillus* revealed only 48 such conserved PUF-binding sites, and the associated genes also do not have a GO bias (Galagan et al. 2005). This situation begs the question whether binding sites discovered via functional assays show evidence for selection in comparative genome sequence analyses. Many legitimate, sequence-specific mRNA-protein interactions may have little bearing on the fitness of the organism in which they occur. Differentiating those interactions that are neutral happenstance from those that are not will be critical in understanding not just how post-transcriptional control works but why it is used at all. Additionally, novel UTR sequences are fast emerging as a byproduct of extensive plant transcriptome sequencing. We anticipate that the techniques and motifs described herein will serve as an entrée into those data.

MATERIALS AND METHODS

Sequence acquisition and preparation

See Figure 1 for a general guide to the following computation pipeline. Transcript data for *A. thaliana* were downloaded from http://www.arabidopsis.org/help/helppages/BLAST_help.jsp#datasets on 24 September 2009 (Version 9). Putative transcripts for all other plant species were downloaded from http://www.plantgdb.org/prj/ESTCluster/progress.php on 7 October 2009 (Duvick et al. 2008). The PlantGDB versions of each species are as follows: *Gossypium hirsutum*, PUT-165a; *Citrus sinensis*, PUT-167a; *Glycine max*, PUT-169a; *Vitis vinifera*, PUT-173a; *Nicotiana tabacum*, PUT-169a. None of sequence sets have undergone significant additions. The longest ORF from each putative transcript was extracted using a custom Perl module (uORF.pm; all scripts are available at http://web.utk.edu/~jvaughn7/code/) based on the criteria that an AUG be followed in-frame by a UAA, UGA, or UAG or that an ORF extend to the end of the putative transcript. These ORFs were translated to peptide sequences and reciprocal BLAST searched, using *blastp* (version 2.2.16), in species-wise fashion with an E-value cutoff of $<10^{-30}$. In order to diminish confounding effects of lineage specific gene loss or incomplete sequence data on orthology assessment, accessions were then clustered based on their BLAST scores using OrthoMCL (version 1. 4) (Li et al. 2003), with default parameters. 5′ or 3′ UTRs shorter than 8 nt were excluded from analysis. Sequences completely overlapped by a larger, identical sequence were subsumed into that sequence.

Motif identification

We used MEME (version 4.3.0) (Bailey and Elkan 1994) to search for overrepresented sequences across the UTRs associated with orthologous groups of coding sequences (Wels et al. 2006). Because in-paralogs/alternative-transcripts within a group confound direct interpretation of E-values produced by MEME, each orthologous group was divided into all possible subgroups, where each subgroup contains only one sequence per species. For tractability reasons, if there were more than 30 possible subgroups per orthologous groups, only 30 randomly selected subgroups were processed further. For example, if an orthologous group contained four *A. thaliana,* three *G. max*, and six *V. vinifera* sequences, there would be 4 × 3 × 6, or 72 subgroup combinations; only 30 of these would be randomly selected for further evaluation. Only 11% of orthologous groups were large enough to require such reduction. Again, because of incomplete sequence information or lineage specific loss, we used the MEME option that detects zero or one motif in all sequences ("zoops"), where motifs could be 6–30 nt long. A second-order Markov model based on all *A. thaliana* 5′ UTRs (or 3′ UTRs, depending on the region being searched) was used as a background model in all MEME searches (Fan et al. 2009). To correct for multiple tests, we divided an E-value cutoff of 0.05 by the number of orthologous subgroups compared: 91,331 for 5′ UTR and 73,893 for 3′ UTR. Only the lowest scoring subgroup of an orthologous group was processed further. A motif was excluded from further analysis if the *A. thaliana* representative had a *p*-value of $>10^{-9}$ of belonging to the motif by chance. Also, motifs were excluded from 5′ UTR comparisons unless the mORF proteins for >70% of informant species aligned, using ClustalW with default parameters, to within 25 amino-acids of the 5′ terminus of the *A. thaliana* representative. All analyses were also carried out on 5′ and 3′ UTR control data sets, which were generated by randomizing all orthologous groups such that species composition per orthologous group was maintained: all sequences were pooled based on the species to which they belong, and, for every orthologous group, the actual sequence was replaced with another sequence, drawn at random with replacement, from the same species pool. These data sets were used to determine our false discovery rate.

Though not reported in detail, we tested numerous alternative motif-identification algorithms. Generally word-based approaches such as Weeder (Pavesi et al. 2004) and FootPrinter (Blanchette and Tompa 2003) were less sensitive than MEME, presumably because they are less tolerant of site degeneracy. For example, at a comparable false discovery rate, Weeder found 12 enriched motifs in the 5′ UTR compared to 194 by MEME. Additionally, algorithms that took phylogeny into account, PhyloGibbs (Siddharthan et al. 2005) and PhyloCon (Wang and Stormo 2003), had equivalent or lower sensitivity than MEME, which is expected given that the lineages under analysis are highly diverged (Storms et al. 2010).

Motif categorization

The sequence of the most frequent letter at each site (consensus sequence) associated with each motif was checked for >5 consecutive mononucleotide repeats or >3 consecutive dinucleotide repeats (Supplemental Table S3). These motifs were removed from downstream processing. Next, the *A. thaliana* representative with +/−10 flanking nucleotides from each significant motif was searched against known and predicted smRNA target sites using the *Arabidopsis* Small RNA Project (ASRP) database (2 June 2010) (Backman et al. 2008) and data from Alves et al. (2009).

All possible open reading frames associated with a motif were checked for protein-coding potential. Each continuous reading frame from the *A. thaliana* representative was translated to protein and aligned using pairwise BLAST (*b2seq*) to every other sequence (also translated) of the appropriate frame in the MEME alignment. Only *b2seq* alignments with an E-value of <0.01 were considered homologous. All sequences passing this criteria were then aligned together using ClustalW (version 1.82), back-translated, and assessed for purifying selection against nonsynonymous mutations using PAML (version 3.14) (Yang 1997) according to the protocol in Nekrutenko et al. (2002) and a tree topology based on Figure 2A (Bausher et al. 2006). In brief, a likelihood ratio test was evaluated for a model in which the nonsynonymous (dN) to synonymous substitution (dS) ratio was allowed to vary and another model in which it was fixed at 1. Motifs with resulting *p*-values, from the χ^2 distribution, of <0.01 were considered protein-coding motifs. Additionally, the motif was removed as an artifact of alternative splicing if the *A. thaliana* peptide associated with the coding potential had a *blastp* match (E-value cutoff of 10^{-5}) to any Viridiplantae protein in the GenBank Refseq database that was 1.5 times longer than itself. The same analysis was done with explicit uORFs that overlap the motif. These were considered CPuORFs if, by manual curation (as in Fig. 3), their nucleotide conservation was exclusively associated with the uORF.

To identify PREs, we further differentiated remaining motifs based on the largest window in a positional weight matrix for which the average of all highest scoring letters for each column in the window was >0.92 (searched from left to right). The positional weight matrix was supplied by MEME; each letter in the matrix represents the frequency of that letter in the site. Windows that were >4 nt long, that were not known smRNA targets, did not show coding potential, and did not contain mono- or dinucleotide repeats were annotated as PREs.

All possible 4-mer word frequencies were measured separately for the combined sets of 5′ and 3′ PREs. The most frequent word was removed, leaving a gap to prevent artifactual fusions, and then the search was rerun until no more 4-mers existed. This prevents confounding effects of 4-mer overlap—UGGA overlapping GGAA, for example. Null distributions for each 4-mer were then generated by (1) randomly simulating a sequence data set with per-element length intact, (2) assessing the frequency of 4-mers, and (3) repeating the process 10,000 times. Because our initial inference of conservation accounts for higher-order statistical properties of the UTR, each letter in these randomizations was considered equally probable. An actual 4-mer frequency was placed within its null distribution, and the number of null distribution values greater than the actual 4-mer frequency was divided by 10,000 to get the *p*-value.

To further differentiate between PREs acting at the transcriptional versus post-transcriptional level, we searched the AGRIS database (Davuluri et al. 2003) in the manner described above for smRNA-related motifs. The IUPAC form used for AGRIS elements was converted to a regular expression prior to searching. Any *A. thaliana* regions associated with a PRE, as with the smRNA search above, were also checked again for imperfect matches to mature smRNAs using psRNATarget web server (Dai et al. 2010) with default parameters. Significant matching smRNA families are reported in Tables 5 and 6.

GO term enrichment

GO term enrichment was performed using the Amigo term enrichment web service with default parameters and all *A. thaliana* genes with a GO annotation as the null model ("TAIR" background filter). All *A. thaliana* accessions associated with the 5′ UTR PREs were pooled and searched; likewise for 3′ UTR PREs. Only one member per orthologous group was included.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

This work was supported by the Division of Chemical Sciences, Geosciences, and Biosciences, Office of Basic Energy Sciences of the US Department of Energy through grant DE-FG02-96ER20223 (most figures), as well as the NSF Arabidopsis 2010 project DBI-0820047 (Table 2; Fig. 5; Supplemental Fig. S2). We thank Timothy Bailey, Charles Grant, and Jon Duvick for generous advice.

Received October 31, 2011; accepted November 21, 2011.

REFERENCES

Alves L, Niemeier S, Hauenschild A, Rehmsmeier M, Merkle T. 2009. Comprehensive prediction of novel microRNA targets in *Arabidopsis thaliana*. *Nucleic Acids Res* **37:** 4010–4021.

Avni D, Shama S, Loreni F, Meyuhas O. 1994. Vertebrate mRNAs with a 5′-terminal pyrimidine tract are candidates for translational repression in quiescent cells: Characterization of the translational *cis*-regulatory element. *Mol Cell Biol* **14:** 3822–3833.

Backman TWH, Sullivan CM, Cumbie JS, Miller ZA, Chapman EJ, Fahlgren N, Givan SA, Carrington JC, Kasschau KD. 2008. Update of ASRP: The *Arabidopsis* Small RNA Project database. *Nucleic Acids Res* **36:** D982–D985.

Baerenfaller K, Grossmann J, Grobei MA, Hull R, Hirsch-Hoffmann M, Yalovsky S, Zimmermann P, Grossniklaus U, Gruissem W, Baginsky S. 2008. Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. *Science* **320:** 938–941.

Bailey TL, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2:** 28–36.

Bausher MG, Singh ND, Lee S-B, Jansen RK, Daniell H. 2006. The complete chloroplast genome sequence of *Citrus sinensis* (L.) Osbeck var "Ridge Pineapple": Organization and phylogenetic relationships to other angiosperms. *BMC Plant Biol* **6:** 21. doi: 10.1186/1471-2229-6-21.

Blanchette M, Tompa M. 2002. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res* **12:** 739–748.

Blanchette M, Tompa M. 2003. FootPrinter: A program designed for phylogenetic footprinting. *Nucleic Acids Res* **31:** 3840–3842.

Bonnet E, Wuyts J, Rouzé P, de Peer YV. 2004. Detection of 91 potential conserved plant microRNAs in *Arabidopsis thaliana* and *Oryza sativa* identifies important target genes. *Proc Natl Acad Sci* **101:** 11511–11516.

Breakfield NW, Corcoran DL, Petricka JJ, Shen J, Sae-Seaw J, Rubio-Somoza I, Weigel D, Ohler U, Benfey PN. 2011. High-resolution experimental and computational profiling of tissue-specific known and novel miRNAs in *Arabidopsis*. *Genome Res* doi: 10.1101/gr.123547.111.

Brodersen P, Voinnet O. 2009. Revisiting the principles of microRNA target recognition and mode of action. *Nat Rev Mol Cell Biol* **10:** 141–148.

Brodersen P, Sakvarelidze-Achard L, Bruun-Rasmussen M, Dunoyer P, Yamamoto YY, Sieburth L, Voinnet O. 2008. Widespread translational inhibition by plant miRNAs and siRNAs. *Science* **320:** 1185–1190.

Brown RH, Gross SS, Brent MR. 2005. Begin at the beginning: Predicting genes with 5′ UTRs. *Genome Res* **15:** 742–747.

Bülow L, Brill Y, Hehl R. 2010. AthaMap-assisted transcription factor target gene identification in *Arabidopsis thaliana*. *Database* **2010:** baq034. doi: 10.1093/database/baq034.

Calkhoven CF, Bouwman PR, Snippe L, Ab G. 1994. Translation start site multiplicity of the CCAAT/enhancer binding protein alpha mRNA is dictated by a small 5′ open reading frame. *Nucleic Acids Res* **22:** 5540–5547.

Calvo SE, Pagliarini DJ, Mootha VK. 2009. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc Natl Acad Sci* **106:** 7507–7512.

Chang KS, Lee SH, Hwang SB, Park KY. 2000. Characterization and translational regulation of the arginine decarboxylase gene in carnation (*Dianthus caryophyllus* L.). *Plant J* **24:** 45–56.

Chen H, Blanchette M. 2007. Detecting non-coding selective pressure in coding regions. *BMC Evol Biol* (Suppl 1) **7:** S9. doi: 10.1186/1471-2148-7-S1-S9.

Chen K, Rajewsky N. 2006. Deep conservation of microRNA-target relationships and 3′UTR motifs in vertebrates, flies, and nematodes. *Cold Spring Harb Symp Quant Biol* **71:** 149–156.

Chen F, Mackey AJ, Stoeckert CJ, Roos DS. 2006. OrthoMCL-DB: Querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res* **34:** D363–D368.

Churbanov A, Rogozin IB, Babenko VN, Ali H, Koonin EV. 2005. Evolutionary conservation suggests a regulatory function of AUG triplets in 5′-UTRs of eukaryotic genes. *Nucleic Acids Res* **33:** 5512–5520.

Dai X, Zhuang Z, Zhao PX. 2010. Computational analysis of miRNA targets in plants: Current status and challenges. *Brief Bioinform* **12:** 115–121.

Davuluri RV, Sun H, Palaniswamy SK, Matthews N, Molina C, Kurtz M, Grotewold E. 2003. AGRIS: *Arabidopsis* gene regulatory information server, an information resource of *Arabidopsis cis*-regulatory elements and transcription factors. *BMC Bioinformatics* **4:** 25. doi: 10.1186/1471-2105-4-25.

de F Lima M, Eloy NB, Pegoraro C, Sagit R, Rojas C, Bretz T, Vargas L, Elofsson A, de Oliveira AC, Hemerly AS, Ferreira PC. 2010. Genomic evolution and complexity of the Anaphase-promoting Complex (APC) in land plants. *BMC Plant Biol* **10:** 254. doi: 10.1186/1471-2229-10-254.

de Sousa Abreu R, Penalva LO, Marcotte EM, Vogel C. 2009. Global signatures of protein and mRNA expression levels. *Mol Biosyst* **5:** 1512–1526.

Dong Y, Bogdanova A, Habermann B, Zachariae W, Ahringer J. 2007. Identification of the *C. elegans* anaphase promoting complex subunit Cdc26 by phenotypic profiling and functional rescue in yeast. *BMC Dev Biol* **7:** 19. doi: 10.1186/1471-213X-7-19.

Doniger SW, Fay JC. 2007. Frequent gain and loss of functional transcription factor binding sites. *PLoS Comput Biol* **3:** e99. doi: 10.1371/annotation/363b6074-caec-4238-b88f-acbf45de498f.

Duvick J, Fu A, Muppirala U, Sabharwal M, Wilkerson MD, Lawrence CJ, Lushbough C, Brendel V. 2008. PlantGDB: A resource for comparative plant genomics. *Nucleic Acids Res* **36:** D959–D965.

Fahlgren N, Howell MD, Kasschau KD, Chapman EJ, Sullivan CM, Cumbie JS, Givan SA, Law TF, Grant SR, Dangl JL, et al. 2007. High-throughput sequencing of *Arabidopsis* microRNAs: Evidence for frequent birth and death of miRNA genes. *PLoS ONE* **2:** e219. doi: 10.1371/journal.pone.0000219.

Fan D, Bitterman PB, Larsson O. 2009. Regulatory element identification in subsets of transcripts: Comparison and integration of current computational methods. *RNA* **15:** 1469–1482.

Farley BM, Pagano JM, Ryder SP. 2008. RNA target specificity of the embryonic cell fate determinant POS-1. *RNA* **14:** 2685–2697.

Franceschetti M, Hanfrey C, Scaramagli S, Torrigiani P, Bagni N, Burtin D, Michael AJ. 2001. 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* **353:** 403–409.

Francischini CW, Quaggio RB. 2009. Molecular characterization of *Arabidopsis thaliana* PUF proteins–binding specificity and target candidates. *FEBS J* **276:** 5456–5470.

Freeling M, Subramaniam S. 2009. Conserved noncoding sequences (CNSs) in higher plants. *Curr Opin Plant Biol* **12:** 126–132.

Galagan JE, Calvo SE, Cuomo C, Ma L-J, Wortman JR, Batzoglou S, Lee S-I, Batürkmen M, Spevak CC, Clutterbuck J, et al. 2005. Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. *Nature* **438:** 1105–1115.

Gerber AP, Herschlag D, Brown PO. 2004. Extensive association of functionally and cytotopically related mRNAs with Puf family RNA-binding proteins in yeast. *PLoS Biol* **2:** E79. doi: 10.1371/journal.pbio.0020079.

Gerber AP, Luschnig S, Krasnow MA, Brown PO, Herschlag D. 2006. Genome-wide identification of mRNAs associated with the translational regulator PUMILIO in *Drosophila melanogaster*. *Proc Natl Acad Sci* **103:** 4487–4492.

Gilligan PC, Kumari P, Lim S, Cheong A, Chang A, Sampath K. 2011. Conservation defines functional motifs in the squint/nodal-related 1 RNA dorsal localization element. *Nucleic Acids Res* **39:** 3340–3349.

Gu S, Jin L, Zhang F, Sarnow P, Kay MA. 2009. Biological basis for restriction of microRNA targets to the 3′ untranslated region in mammalian mRNAs. *Nat Struct Mol Biol* **16:** 144–150.

Hanfrey C, Elliott KA, Franceschetti M, Mayer MJ, Illingworth C, Michael AJ. 2005. A dual upstream open reading frame-based autoregulatory circuit controlling polyamine-responsive translation. *J Biol Chem* **280:** 39229–39237.

Hayden C, Jorgensen R. 2007. 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* **5:** 32. doi: 10.1186/1741-7007-5-32.

Hogan DJ, Riordan DP, Gerber AP, Herschlag D, Brown PO. 2008. Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. *PLoS Biol* **6:** e255. doi: 10.1371/journal.pbio.0060255.

Hulzink RJM, de Groot PFM, Croes AF, Quaedvlieg W, Twell D, Wullems GJ, Herpen MMAV. 2002. The 5′-untranslated region of the ntp303 gene strongly enhances translation during pollen tube growth, but not during pollen maturation. *Plant Physiol* **129:** 342–353.

Im KH, Cosgrove DJ, Jones AM. 2000. Subcellular localization of expansin mRNA in xylem cells. *Plant Physiol* **123:** 463–470.

Imai A, Hanzawa Y, Komura M, Yamamoto KT, Komeda Y, Takahashi T. 2006. The dwarf phenotype of the *Arabidopsis acl5* mutant is suppressed by a mutation in an upstream ORF of a bHLH gene. *Development* **133:** 3575–3585.

Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324:** 218–223.

Ivanov IP, Atkins JF, Michael AJ. 2010. A profusion of upstream open reading frame mechanisms in polyamine-responsive translational regulation. *Nucleic Acids Res* **38:** 353–359.

Jackson RJ, Hellen CUT, Pestova TV. 2010. The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat Rev Mol Cell Biol* **11:** 113–127.

Keene JD. 2007. RNA regulons: Coordination of post-transcriptional events. *Nat Rev Genet* **8:** 533–543.

Kooiker M, Airoldi CA, Losa A, Manzotti PS, Finzi L, Kater MM, Colombo L. 2005. BASIC PENTACYSTEINE1, a GA binding protein that induces conformational changes in the regulatory region of the homeotic *Arabidopsis* gene SEEDSTICK. *Plant Cell* **17:** 722–729.

Lawson MJ, Zhang L. 2006. Distinct patterns of SSR distribution in the *Arabidopsis thaliana* and rice genomes. *Genome Biol* **7:** R14. doi: 10.1186/gb-2006-7-2-r14.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res* **13:** 2178–2189.

Lincoln AJ, Monczak Y, Williams SC, Johnson PF. 1998. Inhibition of CCAAT/enhancer-binding protein alpha and beta translation by upstream open reading frames. *J Biol Chem* **273:** 9552–9560.

Lockton S, Gaut BS. 2005. Plant conserved non-coding sequences and paralogue evolution. *Trends Genet* **21:** 60–65.

Lukaszewicz M, Feuermann M, Jerouville B, Stas A, Boutry M. 2000. In vivo evaluation of the context sequence of the translation initiation codon in plants. *Plant Sci* **154:** 89–98.

Marrison JL, Schunmann PHD, Ougham HJ, Leech RM. 1996. Subcellular visualization of gene transcripts encoding key proteins of the chlorophyll accumulation process in developing chloroplasts. *Plant Physiol* **110:** 1089–1096.

McGuire AM, Galagan JE. 2008. Conserved secondary structures in *Aspergillus*. *PLoS ONE* **3:** e2812. doi: 10.1371/journal.pone.0002812.

Meister RJ, Williams LA, Monfared MM, Gallagher TL, Kraft EA, Nelson CG, Gasser CS. 2004. Definition and interactions of a positive regulatory element of the *Arabidopsis INNER NO OUTER* promoter. *Plant J* **37:** 426–438.

Molina C, Grotewold E. 2005. Genome wide analysis of *Arabidopsis* core promoters. *BMC Genomics* **6:** 25. doi: 10.1186/1471-2164-6-25.

Narsai R, Howell KA, Millar AH, O'Toole N, Small I, Whelan J. 2007. Genome-wide analysis of mRNA decay rates and their determinants in *Arabidopsis thaliana*. *Plant Cell* **19:** 3418–3436.

Neafsey DE, Galagan JE. 2007. Dual modes of natural selection on upstream open reading frames. *Mol Biol Evol* **24:** 1744–1751.

Nekrutenko A, Makova KD, Li W-H. 2002. The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: An empirical and simulation study. *Genome Res* **12:** 198–202.

Pagano JM, Farley BM, Essien KI, Ryder SP. 2009. RNA recognition by the embryonic cell fate determinant and germline totipotency factor MEX-3. *Proc Natl Acad Sci* **106:** 20252–20257.

Pavesi G, Mauri G, Pesole G. 2004. In silico representation and discovery of transcription factor binding sites. *Brief Bioinform* **5:** 217–236.

Rabani M, Kertesz M, Segal E. 2008. Computational prediction of RNA structural motifs involved in posttranscriptional regulatory processes. *Proc Natl Acad Sci* **105:** 14885–14890.

Rahmani F, Hummel M, Schuurmans J, Wiese-Klinkenberg A, Smeekens S, Hanson J. 2009. Sucrose control of translation mediated by an upstream open reading frame-encoded peptide. *Plant Physiol* **150:** 1356–1367.

Rivas E, Eddy SR. 2000. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics* **16:** 583–605.

Roy B, Vaughn JN, Zhou F, Gilchrist MA, von Arnim AG. 2010. The h subunit of eIF3 promotes reinitiation competence during translation of mRNAs harboring upstream open reading frames. *RNA* **16:** 748–761.

Santi L, Wang Y, Stile MR, Berendzen K, Wanke D, Roig C, Pozzi C, Müller K, Müller J, Rohde W, et al. 2003. The GA octodinucleotide repeat binding factor BBR participates in the transcriptional regulation of the homeobox gene Bkn3. *Plant J* **34:** 813–826.

Serganov A, Patel DJ. 2008. Towards deciphering the principles underlying an mRNA recognition code. *Curr Opin Struct Biol* **18:** 120–129.

Shama S, Meyuhas O. 1996. The translational *cis*-regulatory element of mammalian ribosomal protein mRNAs is recognized by the plant translational apparatus. *Eur J Biochem* **236:** 383–388.

Siddharthan R, Siggia ED, van Nimwegen E. 2005. PhyloGibbs: A Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biol* **1:** e67. doi: 10.1371/journal.pcbi.0010067.

Simpson GG, Laurie RE, Dijkwel PP, Quesada V, Stockwell PA, Dean C, Macknight RC. 2010. Noncanonical translation initiation of the *Arabidopsis* flowering time and alternative polyadenylation regulator FCA. *Plant Cell* **22:** 3764–3777.

Skuzeski JM, Nichols LM, Gesteland RF, Atkins JF. 1991. The signal for a leaky UAG stop codon in several plant viruses includes the two downstream codons. *J Mol Biol* **218:** 365–373.

Storms V, Claeys M, Sanchez A, Moor BD, Verstuyf A, Marchal K. 2010. The effect of orthology and coregulation on detecting regulatory motifs. *PLoS ONE* **5:** e8938. doi: 10.1371/journal.pone.0008938.

Tran MK, Schultz CJ, Baumann U. 2008. Conserved upstream open reading frames in higher plants. *BMC Genomics* **9:** 361. doi: 10.1186/1471-2164-9-361.

Tremousaygue D, Garnier L, Bardet C, Dabos P, Herve C, Lescure B. 2003. Internal telomeric repeats and "TCP domain" protein-binding sites co-operate to regulate gene expression in *Arabidopsis thaliana* cycling cells. *Plant J* **33:** 957–966.

Urban C, Zerfass K, Fingerhut C, Beier H. 1996. UGA suppression by tRNAC^{m}CA^{Trp} occurs in diverse virus RNAs due to a limited influence of the codon context. *Nucleic Acids Res* **24:** 3424–3430.

Vogel C, de Sousa Abreu R, Ko D, Le S-Y, Shapiro BA, Burns SC, Sandhu D, Boutz DR, Marcotte EM, Penalva LO. 2010. Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol Syst Biol* **6:** 400. doi: 10.1038/msb.2010.59.

Wamboldt Y, Mohammed S, Elowsky C, Wittgren C, de Paula WBM, Mackenzie SA. 2009. Participation of leaky ribosome scanning in protein dual targeting by alternative translation initiation in higher plants. *Plant Cell* **21:** 157–167.

Wang T, Stormo GD. 2003. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics* **19:** 2369–2380.

Wels M, Francke C, Kerkhoven R, Kleerebezem M, Siezen RJ. 2006. Predicting *cis*-acting elements of *Lactobacillus plantarum* by comparative genomics with different taxonomic subgroups. *Nucleic Acids Res* **34:** 1947–1958.

Willmann MR, Poethig RS. 2007. Conservation and evolution of miRNA regulatory programs in plant development. *Curr Opin Plant Biol* **10:** 503–511.

Yamamoto YY, Ichida H, Matsui M, Obokata J, Sakurai T, Satou M, Seki M, Shinozaki K, Abe T. 2007. Identification of plant promoter constituents by analysis of local distribution of short sequences. *BMC Genomics* **8:** 67. doi: 10.1186/1471-2164-8-67.

Yamamoto YY, Yoshitsugu T, Sakurai T, Seki M, Shinozaki K, Obokata J. 2009. Heterogeneity of *Arabidopsis* core promoters revealed by high-density TSS analysis. *Plant J* **60:** 350–362.

Yang Z. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13:** 555–556.

Zhang L, Zuo K, Zhang F, Cao Y, Wang J, Zhang Y, Sun X, Tang K. 2006. Conservation of noncoding microsatellites in plants: Implication for gene regulation. *BMC Genomics* **7:** 323. doi: 10.1186/1471-2164-7-323.

Zhou F, Roy B, von Arnim AG. 2010. Translation reinitiation and development are compromised in similar ways by mutations in translation initiation factor eIF3h and the ribosomal protein RPL24. *BMC Plant Biol* **10:** 193. doi: 10.1186/1471-2229-10-193.