University of Tennessee, Knoxville
Trace: Tennessee Research and Creative Exchange

Doctoral Dissertations

Graduate School

8-2011

The Evolution and Mechanics of Translational Control in Plants

Justin N. Vaughn
jvaughn7@utk.edu

Recommended Citation

Vaughn, Justin N., "The Evolution and Mechanics of Translational Control in Plants. " PhD diss., University of Tennessee, 2011.
https://trace.tennessee.edu/utk_graddiss/1136

To the Graduate Council:

I am submitting herewith a dissertation written by Justin N. Vaughn entitled "The Evolution and Mechanics of Translational Control in Plants." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Biochemistry and Cellular and Molecular Biology.

Albrecht G. von Arnim, Major Professor

We have read this dissertation and recommend its acceptance:

Daniel M. Roberts, Michael A. Gilchrist, Igor B. Jouline, Feng Chen

Accepted for the Council:
Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

The Evolution and Mechanics of Translational Control in Plants

A Dissertation Presented for the
Doctor of Philosophy Degree
The University of Tennessee, Knoxville

Justin N. Vaughn
August 2011

Acknowledgments

I would like to thank Albrecht von Arnim for his guidance and support of this research over the last five years. It is rare to find a person so adept at observing and criticizing data yet so patient with the people who generated it. His biological intuition, careful scholarship, and effective adoption of emerging techniques are qualities that we would all like to have in such quantity. I would also like to thank my other committee members - Michael Gilchrist, Dan Roberts, Igor Zhulin, and Feng Chen - for their time and their direction. I would particularly like to thank Dr. Gilchrist for much hard-fought mathematical advise. Though not on my committee, Andreas Nebenfuer has also been an excellent academic instructor and professional advisor. The BCMB department was kind enough to send me to Italy for a bioinformatics sabbatical in 2007. I would like to thank my hosts there: Flavio Mignone, for his valuable instruction, and his grandmother, for graciously renting me her apartment. There are many past and present students/postdocs within the department to thank, but I will specifically acknowledge Bijoyita Roy, Fujun Zhou, Byung-Hoon Kim, Krzysztof Bobik, Evan Riddick, and Ju Guan for letting me talk about what I didn't understand until I understood it. Lastly, this dissertation is for Nunally and Waylon, who are the stars that I sail by, and my parents, who are . . . my parents.

Abstract

The expression of numerous plant mRNAs is attenuated by RNA sequence elements located in the 5' and 3' untranslated regions (UTRs). For example, in plants and many higher eukaryotes, roughly 35% of genes encode mRNAs that contain one or more upstream open reading frames (uORFs) in the 5' UTR. For this dissertation I have analyzed the pattern of conservation of such mRNA sequence elements. In the first set of studies, I have taken a comparative transcriptomics approach to address which RNA sequence elements are conserved between various families of angiosperm plants. Such conservation indicates an element's fundamental importance to plant biology, points to pathways for which it is most vital, and suggests the mechanism by which it acts. Conserved motifs were detected in 3% of genes. These include di-purine repeat motifs, uORF-associated motifs, putative binding sites for PUMILIO-like RNA binding proteins, small RNA targets, and a wide range of other sequence motifs. Due to the scanning process that precedes translation initiation, uORFs are often translated, thereby repressing initiation at the an mRNA's main ORF. As one might predict, I found a clear bias against the AUG start codon within the 5' untranslated region (5' UTR) among all plants examined. Further supporting this finding, comparative analysis indicates that, for ~42% of genes, AUGs and their resultant uORFs reduce carrier fitness. Interestingly, for at least 5% of genes, uORFs are not only tolerated, but enriched. The remaining uORFs appear to be neutral. Because of their tangible impact on plant biology, it is critical to differentiate how uORFs affect translation and how, in many cases, their inhibitory effects are neutralized. In pursuit of this aim, I developed a computational model of the initiation process that uses five parameters to account for uORF presence. *In vivo* translation efficiency data from uORF-containing reporter constructs were used to estimate the model's parameters in wild type Arabidopsis. In addition, the model was applied to identify salient defects associated with a mutation in the subunit h of eukaryotic initiation factor 3 (eIF3h). The model indicates that eIF3h, by supporting re-initation during uORF elongation, facilitates uORF tolerance.

Table of Contents

Index of Tables

Index of Figures

List of Attachments

Chapter 1: Introduction

Proteins are encoded by DNA. An intermediate transcript molecule, mRNA, is used to relay this coding information from the genome to the ribosomes. At ribosomes, amino acid polymerization and, hence, protein production takes place. Under this simple model, one would expect a large positive correlation between mRNA and protein concentrations, but for many genes in plants, this is not the case [1]. Roughly 50% of this decoupling is the result of processes that affect translation, primarily through initiation [2]. Moreover, the translation of an mRNA is intimately related to its degradation and cellular localization [3]. Thus, translation control is central to many developmental, metabolic, and pathological processes [4]. Additionally, a critical step in the life cycle of all viruses is to co-opt the host ribosome for the production of their requisite proteins. Viruses do this through a seemingly unending variety of mechanisms, and these are typically understood only to the degree that we understand the translation apparatus being hijacked [5]. Finally, understanding the mechanics and distribution of both global and mRNA-specific regulatory mechanisms will help to complete our understanding of gene expression and to broaden the palette of tools for engineering specific expression levels [6].

With this dissertation, I have attempted to identify which modes of translation control are most common in plants and which plant genes are specifically regulated. Of all candiate modes of regulation, my main focus has been on small upstream open reading frames (uORFs) found in many mature mRNAs. Additionally, I have developed a series of computational models of translation initiation in the presense of uORFs, both with regard to wild-type plants and initiation factor mutants. These biological questions have led to the development of software that may have general utility in mutliple fields, from comparative sequence analysis in plants to the computational modeling of molecular processes.

Section 1: The life and death of an eukaryotic mRNA

From its initial transcription to its final degradation, an mRNA is continuously partnered with proteins [3]. Unbound, 'naked' mRNA is rapidly degraded. The composition of interacting proteins dictates an mRNA's stage in life - transcription, splicing, export, transport, translation, or degradation. Whether the protein encoded by a gene is expressed at continuous levels or whether expression varies over time and space depends not only on the transcription rate but on the relative kinetics of these downstream events. In turn, these kinetics can be dramatically influenced by *cis* and *trans*-acting factors. Some of these pathways have global effects, such as the phosphorylation state of a translation initiation factor [7], while other act in mRNA-specific ways, such the interaction between a short mRNA sequence and a Pumilio-like RNA-binding protein [8].

Transcription of mRNA is followed by immediate and continual interaction with a diverse set of proteins

Short sequences in the genome attract the transcriptional machinery to a position upstream of the DNA encoding a protein. mRNA is copied from the DNA antisense strand in the 3' to 5' direction, creating a letter for letter transcript (Figure 1.1). In mRNA, a uridine (U) supplants thymine (T). In eukaryotes, proteins are encoded in fragments, called 'exons', that are stitched together in the nucleus by the splicesome. Exon junction complexes (EJCs) interact with the splicing machinery and cooperatively bind mRNA, in a sequence-independent manner, 20-25 nucleotides (nts) upstream of the splice site. The position of an EJC relative to the stop-codon of the protein coding frame can have dramatic effects on the stability of a transcript. While still in the nucleus the spliced mRNA is matured by adding a 5'-methylguanosine cap and a poly-Adenine (poly-A) tail. Bound nuclear export factors allow the mRNA passage through the nuclear pore and into the cytoplasm. These factors are removed in transit, leaving exposed landing sites for the translation pre-initation complex.

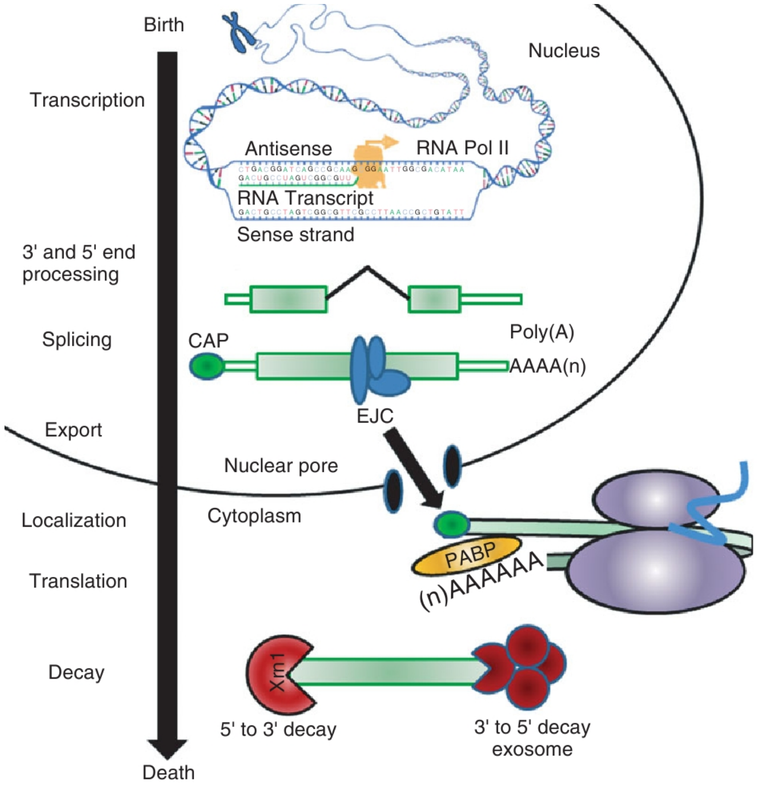


Figure 1.1: Metabolic cycle of mRNA.
mRNAs are transcribed by RNA polymerase II. Introns are removed via splicing and an exon junction complex (EJC) remains bound at the site of ligation. Capped and polyadenylated mRNAs are exported. Poly-A binding proteins interact with the cap to facilitate initiation. After multiple rounds of translation mRNAs are degraded. From [9].

The pre-initiation complex binds near the mRNA 5' cap and begins scanning

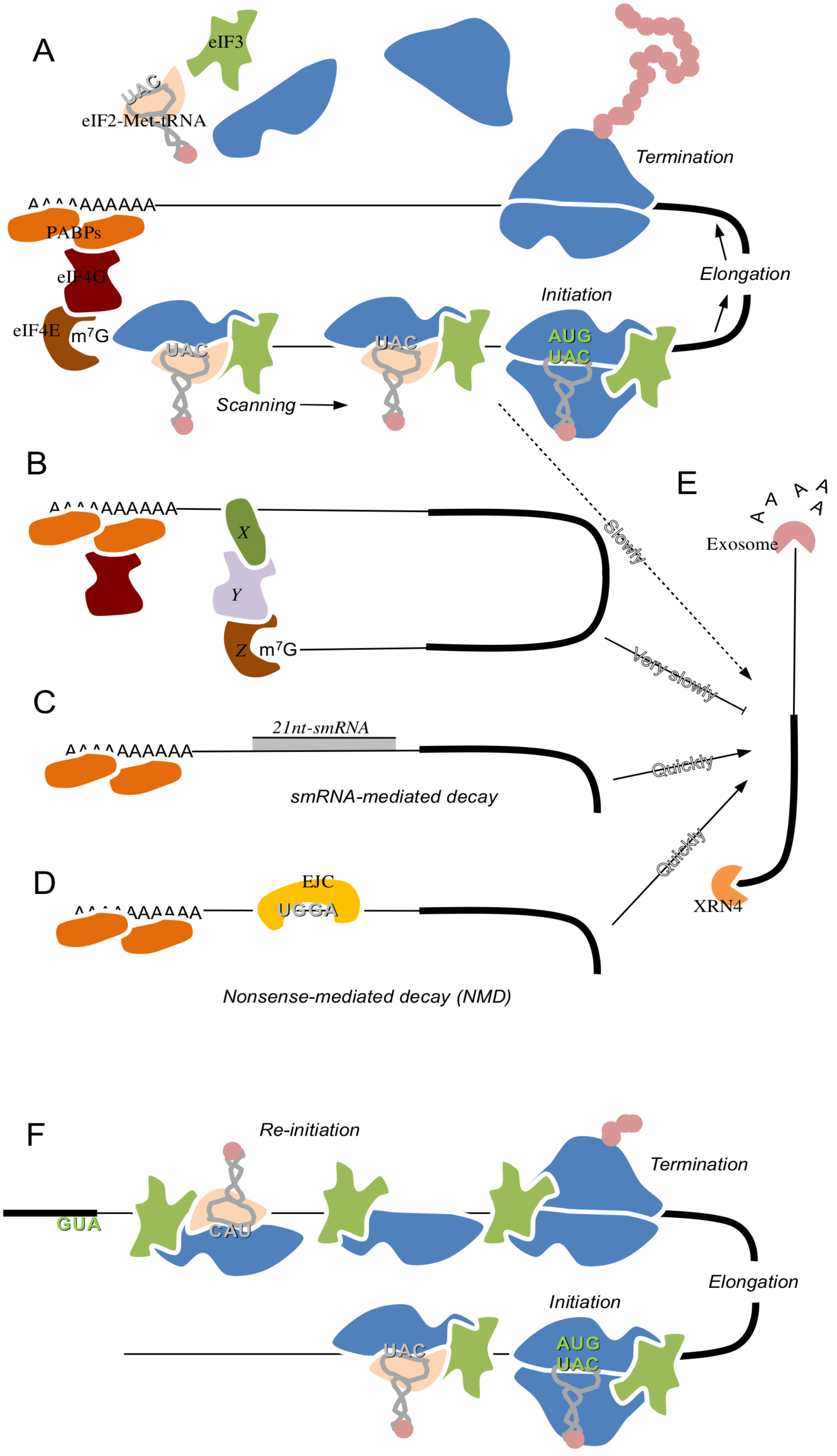
Although the exact sequence of binding events is unclear, the charged methionyl-tRNA, the small-ribosomal subunit, and a suite of eukaryotic initation factors (eIFs) - 2, 3, 4, 5 - converge at the 5' cap of the mRNA (Figure 1.2A). Additionally, eIF4G interacts with poly-A binding proteins to circularize the mRNA. This closed-loop form is thought to facilitate ribosome recycling and/or to stabilize the interactions between the mRNA and initiation factors that would otherwise be lost to the cytosol after translation initation [7].

Once established, the pre-initation complex moves in the 5' direction via a hypothetical Brownian ratchet [10]. The first AUG encountered by the pre-intiation complex should stimulate initation, although the effectiveness of an AUG is never 100% and can vary by 3-fold based on the surrounding sequence [11,12].

A pioneering round of translation occurs during which residual proteins are displaced by the processing ribosome. A degradation pathway, called 'nonsense-mediated decay' (NMD), can be triggered when a ribosome encounters a stop-codon upstream of an EJC (Figure 1.2D). The 'nonsense' in NMD refers to polymerase errors that introduce inappropriate stop codons; hence, this pathway is thought to be an error-control check-point for transcripts. Once the pioneering round is finished, the mRNA forms a relatively stable circular polyribosome and conventional protein production proceeds.

Figure 1.2 (next page): General models of initiation, active mRNA translation, and mRNA degradation.

A) Cap-binding, scanning, initiation, elongation, and termination occur in sequential order. 5' and 3' ends of mRNA are indirectly associated via the Poly-A binding protein (orange), eIF4G (crimson), and eIF4E (brown). eIF4E attracts a complex containing the small-ribosomal subunit (blue), eIF3 (green), eIF2 (light pink), and the Met-tRNA (gray with anticodon), referred to as the 43S complex. Upon AUG recognition eIF2 leaves and the large ribosomal subunit (blue, larger) binds. Polypeptide (dark pink) is polymerized. Deadenylation eventually results in mRNA degradation. B) XYZ model of mRNA specific protein interaction and translation repression. While maintaining mRNA stability, the XYZ complex inhibits translation because the XY proteins outcompete the eFI4g for eIF4e and prevent the 43S-complex from binding to the 5' end. C) Small-RNA (21-24nts) interact with the mRNA and target it for degradation. D) When an exon junction occurs downstream of an elongation/termination reaction, the mRNA is directed into the NMD pathway. E) Degradation occurs after the poly-A tail has been removed. Decapping follows (not shown), and the 3'-to-5' and 5'-to-3' exoribonuclease activity of the Exosome and XRN4, respectively, catabolizes the mRNA. F) 'Re-scanning' and, thus, re-initiation can occur after elongation and termination, but the eIF2-Met-tRNA must be reaquired in order to decode 3' AUGs.


A
eIF3
UAC
eIF2-Met-tRNA
Termination
AAAAAAAAA
PABPs
eIF4G
eIF4E
m7G
UAC
UAC
Initiation
AUG
UAC
Elongation
Scanning
B
AAAAAAAAA
X
Y
Z
m7G
Slowly
Very slowly
E
A
A
A
A
A
Exosome
C
21nt-smRNA
AAAAAAAAA
smRNA-mediated decay
Quickly
D
EJC
UGGA
AAAAAAAA
Quickly
XRN4
Nonsense-mediated decay (NMD)
F
Re-initiation
Termination
GUA
CAU
Elongation
Initiation
UAC
AUG
UAC

Initiation is the rate-limiting step of translation

During translation, mRNAs typically form a polyribosome ('polysome', for short). As the name implies, multiple ribosomes simultaneously translate the mRNA. Interestingly, during log-phase growth in yeast, the ribosome density across an mRNA is far lower than the maximum expected by the spatial dimensions of the ribosome [13] (Figure 1.3A). This may be explained, in part, by conformational demands on the polysome [14], but most research indicates that initiation is rate-limiting relative to elongation and termination [7]. ORF length is also thought to have a non-linear effect on translation rates (Figure 1.3B) [15,16], although there is currently conflicting data as to the nature of this relationship - some data indicates that shorter transcripts have more efficient initiation [17], while other data suggest that the relationship relates to the elongation phase [16].

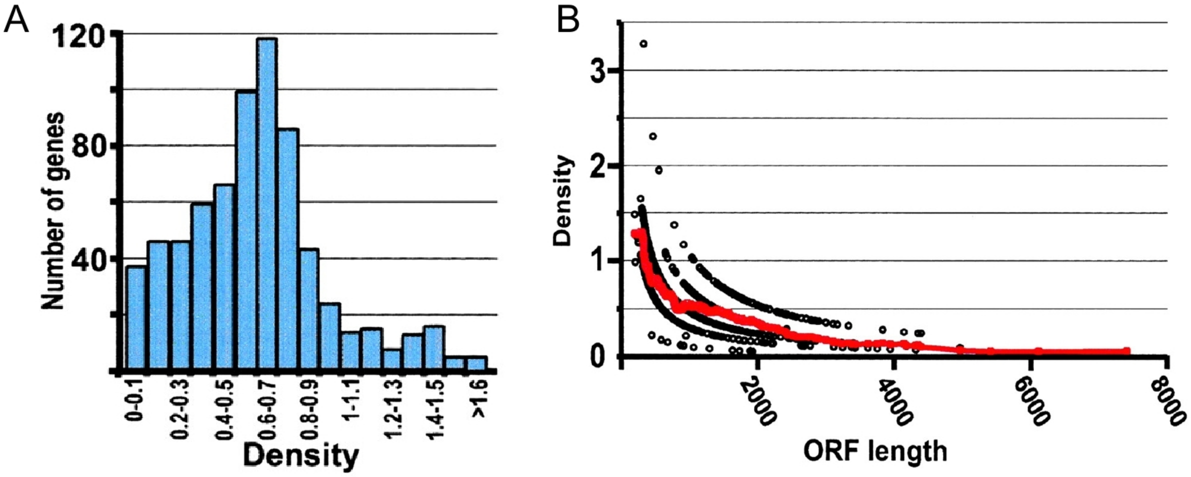


Figure 1.3: Genome-wide ribosome density in yeast
A) Ribosome density is, on average, ~5-fold lower than theoretical maximum. Plot shows the number of genes that fall within a particular range of ribosome-density (ribosomes per 100 nts).
B) As ORF length increases, ribosome-density declines. Density is plotted against ORF length. Multiple curves in the data result from the division of total ribosome number (integer) by ORF length to get ribosome-density. Red line is moving average (50 nt window) Modified from [13].

mRNA half-life is highly variable and is loosely coupled to translation

By stopping transcription and monitoring mRNA content over time with microarrays, researchers can observe the degradation rates of nearly all constitutively expressed mRNAs in

Arabidopsis [18]. Based on those studies, the median half-life of an mRNA in Arabidopsis is ~4 hrs. The absence of an intron and the presence of an small-RNA binding site positively correlate with the speed of decay. In plants, the lack of an intron reduces transcript half-life to ~2hrs. Energy metabolism and protein synthesis genes produce very stable mRNAs with a median half-life of ~8hrs. By comparing unstable and stable transcripts, the authors attempted to define 6-letter sequences that mediate decay rate. Unstable transcripts tended to have uridine-rich elements in their 3' UTRs and, to a lesser extent, 5' UTRs. Otherwise, there were no obvious trends.

Translation and mRNA decay rates are almost certainly related, yet molecular explanations linking these processes are still speculative. As discussed below, the poly-A tail is a critical point for integrating these two processes, making it hard to untangle the exact causal chain. At the very least, stalled elongating ribosomes are thought to trigger mRNA decay, or 'no-go' decay [19], a process that directly links abnormal translation with decay. Two cellular structures - P-bodies and stress granules - are thought to be sites of translational senescence and mRNA degradation. Because these structures appear to be devoid of large ribosomal subunits, the contained mRNAs are not in an elongating phase of translation [19]. Interestingly, mRNAs can be resurrected from these structures, particularly stress granules, and returned to the general pool of translating mRNA.

5' and 3' untranslated regions (UTRs) are operationally distinct

An open reading frame (ORF) is a sequence of codons that starts with a start codon (AUG) and ends with an in-frame stop codon (UAA, UAG, or UGA). Conventionally, a coding sequence (CDS) is an ORF with some evidence of actually encoding a protein - experimental confirmation, mutation bias, presence in an mRNA, or homology to another known CDS. Some authors use 'ORF' to more generally describe a sequence of codons that is devoid of stop codons, regardless of whether or not it possesses a start codon. Instead, the term 'continuous reading frame' should be used for such situations. ('Unidentified reading frame' has also been used synonymously but is less clear.)

Based on the previous subsections, we can divide a mRNA into regions in which distinct events are occuring. The 5' UTR, defined as the region of the mRNA from the 5' cap (or *transcriptional* start site) to the *translational* start site of the mORF, is continually scanned by the pre-initation complex. Such a process should in theory displace bound proteins and disrupt low-energy secondary structures [12]. Alternatively, the region of the mRNA from the stop of the longest open-reading frame to start of the poly-A tail, the 3' UTR, is relatively free of such ribosomal traffic. In mammalian systems, as predicted, the 3' UTR appears to house the bulk of small-RNA targets and other *cis*-acting elements.

General decoupling of mRNA and protein concentrations

Proteins are usually the operational manifestation of a gene. mRNAs are an intermediate in gene expression. Thus, many microarray experiments operate under the assumption that transcriptional change inevitably leads to a change in protein concentration. More narrowly defined, the assumption is often that the concentration of mRNA is linearly proportional to the concentration of protein. Based on proteomic quantification techniques using mammalian cell lines, it appears that this is not the case, and that mRNA concentration only accounts for one-third of the variation observed in protein concentration [2]. Moreover, of this third, it is still unclear whether transcription or mRNA stability (either related to translation or direct degradation) is more important in predicting mRNA concentration. Comparable results have emerged from assessing translation rates in yeast wherein differential translation rates account for 25% of variation in protein abundance, while mRNA abundance accounts for only 17% [16]. A revised model of gene expression may be emerging in which promoter modules are responsible for whether a gene is expressed or not, but precise dosage of a gene product is mediated by translation, mRNA degradation, and protein degradation.

Section 2: Mechanisms of translational control

Measuring translation efficiency

In order to assess translational control, we must be able to identify cases where translation efficiency is reduced or enhanced in response to a signal or mutation. This, in turn, requires a valid way to assess the translation state of an mRNA, while controlling for possible changes in other steps of gene expression, such as transcription, mRNA degradation, and protein degradation.

Ideally, one could watch a single mRNA *in vivo* and count the number of initiation events that occur per second. Though single molecule studies are approaching this goal [6,20], such resolution is still intractable, particularly for large-scale characterization of translation. As an alternative approach, mRNAs can be isolated from cell extracts based on the number of ribosomes with which they are associated. A shift in this number suggests some change in the translation rate of an mRNA [17]. Moreover, by counting the number of mRNAs in each ribosomal fraction and assuming that elongation and termination rates are fairly constant, one can estimate the initiation rate for that mRNA based on the ribosomal density across the mRNA - the lower the density, the lower the initiation rate.

Alternatively, the concentration of a mRNA that encodes a protein, typically a light-emitting reporter, and the concentration of that protein can be quantitated and divided by one another [20]. This then gives a protein per mRNA value. Assuming protein degradation is not affected, this value can be used to calculate a relative change in translation rates across conditions.

Another recently developed approach to estimating translation efficiency involves sequencing fragments of mRNA that are protected by ribosomes [16]. Of the techniques described, nucleotide-specific ribosome mapping is the only one to allow researchers to interrogate the *in vivo* kinetics of translation [21], although this requires numerous assumptions including *a priori* prediction of elongation times based on tRNA abundance.

Initiation factor availability and activity has a global impact on mRNA translation

eIF2 delivers the initiation tRNA (Met-tRNA) to small-ribosomal subunit prior to scanning. Phosphorylation of eIF2 reduces the activity of eIF2B, which is responsible for exchanging guanosine di-phosphate (GDP) for a GTP in order to create an active eIF2. In the absense of the eIF2-Met-tRNA complex, nearly all genes are repressed. This pathway appears to be present in all eukaryotic kingdoms. Interestingly, as discussed below, some genes in fungi and mammals that are already repressed through other mechanisms are de-repressed when eIF2 is phosphorylated. It is not clear, how common such de-repression is in plants.

Any change in expression involving key translation factors will have an impact on global translation efficency. During plant stress such as drought and oxygen-deprivation, the translation state of most mRNAs is reduced [22,23]. Whether this response is mediated solely by the eIF2-kinase pathway is unknown. Many eIFs are phosphorylated but the functional significance of such modification has not been determined. Nearly all related mutational analysis in mammals suggest that, excluding eIF2, phosphorylation state is not a major factor in the global translation response [7].

Generic sequence features of eukaryotic mRNAs affect both translation and mRNA degradation

Nearly all eukaryotic mRNAs have a methylated cap and a poly-A tail, both of which are critical for mRNA stability. Additionally, the poly-A tail is the recognition site for poly-A binding proteins (PABPs) that interact with initiation factors to enhance translation. In this regard, the poly-A/PABP interaction is one of the few RNA-protein interactions thought to enhance translation. As the poly-A tail gets degraded in the cystosol, the mRNA becomes more susceptible to degradation. This shortening typically coincides with reduced translation efficiency, but not always [19].

Upstream open-reading-frames allow for mRNA specific de-repression

As mentioned above, some mRNAs avoid the repressive effects of phosphorylated-eIF2. More exactly, they increase in concentration during stress conditions relative to their non-stressed concentrations. This de-repression is mediated by upstream open reading frames (uORFs). These elements and their molecular effects will be discussed in detail at many points throughout this thesis. In short, the 5' UTR can contain short ORFs that, because of 5' to 3' scanning, get translated as short peptides. Surprisingly, after termination the small ribosomal subunit continues scanning, but, because the eIF2-Met-tRNA is lost at initiation, this translation event temporarily reduces the small ribosome's ability to recognize downstream start sites (Figure 1.2F and 1.3). If post-termination scanning continues long enough, eIF2-Met-tRNA can be regained. In terms of de-repression, mRNAs can have multiple uORFs that are positioned in such a way that, at typical levels of active eIF2 , one uORF is highly suppressive, but, when eIF2 has low activity (phosphorylated state), the repressive uORF is skipped. Two of the best known examples are GCN4 in yeast and ATF4 and ATF5 in mammals (Figure 1.4), all of which are transcription factors that reside at the head of key metabolic pathways.

De-repression mechanisms aside, most uORFs simply repress translation [24]. Moreover, this repression (or de-repression) is rarely dependent on the encoded peptide, although there are very interesting cases where uORF peptide sequences are functionally constrained [25,26]. In plants, two functional uORF peptides have received experimental attention. bZip11 is a short transcription factor. The bZip11 mRNA harbors a uORF that encodes a sucrose-responsive peptide [27]. The nascent peptide possibly interacts with sucrose and stalls the translating ribosome, thus repressing the translation of the downstream transcription factor. uORFs in AdoMetDC are thought to respond to polyamines in an analogous way [28,29].

small-RNAs target mRNA for translation repression and degradation

Small-RNA-dependent (smRNA) repression ("silencing") involves the processing of a double-stranded RNA precursor into a short, single-stranded RNA molecule that then targets complementary RNAs for degradation (Figure 1.2C). The pathway is likely to have evolved as an immune response to viruses and transposable elements, not for the regulation of endogenous transcripts [30,31], and the number of such transcripts thought to be under smRNA control in plants is actually quite small [32]. Still, smRNA target sites of particular mRNAs appear to be conserved and, by inference, important to the carrier's survival and reproduction in its native environment [33].

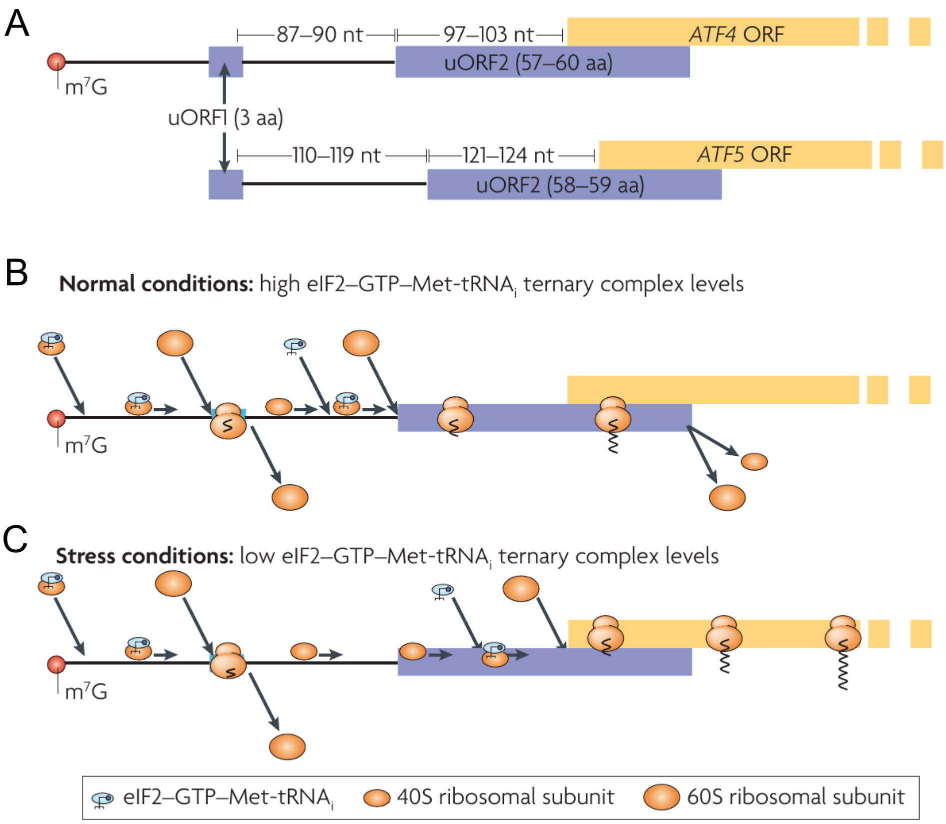


Figure 1.4: The mechanism of regulation of ATF4 and ATF5 mRNA translation. A) Diagram showing the sizes and spacing disposition of the two uORFs in activating transcription factor 4 (ATF4) mRNAs and ATF5 mRNAs. B) The pattern of translation in unstressed conditions, when eIF2–GTP–Met- tRNA ternary complexes (eIF2-TCs) are abundant. Small (40S) ribosomal subunits, with associated eIF2-TCs (blue), scan the mRNA in the direction shown. If eIF2-TCs are abundant, most of the 40S subunits that resume scanning after uORF1 translation will acquire a new eIF2-TC in time to initiate translation of uORF2, and ribosomes that translate this second uORF will be unable to initiate at the ATF4 or ATF5 AUG because uORF2 is too long to allow rescanning, and because it would require backwards scanning, which doesn't seem to occur over long distances. C) Pattern of translation in stressed conditions, when eIF2-TC availability is low owing to eIF2 phosphorylation. Consequently, most of the 40S subunits that resume scanning after translating uORF1 acquire a new eIF2-TC only after they have migrated past the uORF2 initiation codon, but in time to initiate at the next AUG, which is at the start of the ATF ORF in both cases. From [7].

The structural basis for RNA recogniton

While the mechanism of direct interaction between an smRNA-target and the mature smRNA is fairly intuitive, RNA-protein interactions are much more various. In terms of RNA-specific protein interactions known to affect translation in eukaryotes, structural studies indicate that nearly all such interactions require some primary sequence element, although these are often supplemented by local secondary structure [34]. The primary sequence element can be as short as four nucleotides and, even at this length, can contain degenerate sites. Though we are focused on mRNA-specific interactions, there are many other RNA-protein interactions that do not require sequence or structural specificity. Moreover, RNA and protein, because of their often opposing charges and their conformational variability, form interactions of moderate affinity by default [35]. It is doubtful that many of these interactions have any functional significance. With regard to large-scale interaction studies, this complication highlights the need for assessing sequence constraint through evolutionary analysis in order to differentiate functionally significant interactions from those that are effectively inert.

Other *cis*-acting elements affect the stability, translation, and localization of specific plant mRNAs

Deriving most of their inspiration from research in *Drosophila*, Jackson, *et. al* [7] have proposed the XYZ paradigm for mRNA-specific translation repression by *trans*-acting proteins (Figure 1.2B). Under this paradigm, protein X interacts with a mRNA's 3' UTR, via structures described above. Another region of Protein X interacts with an adapter, Protein Y, that in turn interacts with a cap-binding Protein Z. This series of interaction creates an inhibitory, closed-loop structure. It is somewhat paradoxical that the closed-loop structure formed by the poly-A tail binding is excitatory while the XYZ structure is inhibitory, but poly-A binding itself can be thought of as an XYZ process, where PABP is the X protein and Y is the pre-initiation complex. The critical difference is that Y in the case of poly-A binding goes on to initiate translation while

Y in the inhibitory case remains locked to the cap. Based on this generic model, stoichiometric competition for the 5' UTR cap should have a substantial impact on translation rate.

In the animal kingdom, translation repression and mRNA localization are often coordinated processes. XYZ complexes, which regulate Nanos, Caudal, Oskar, and Hunchback proteins, are vital for establishing morphogen gradients in the zygote. Precise regulation and localization of these proteins at this stage is required because the specific sectors of a single cell must be partitioned into smaller, more differentiated cells. Because they have fewer body axises, plants may not need such a precise combinatorial pattern of morphogens. That said, the first division of a plant zygote is asymetrical and can be modified via mutations in the kinase, *SHORT SUSPENSOR* [36]. Like the Drosophila proteins discussed above, this protein has a parent-of-origin effect and is translationally repressed.

Many mRNAs in somatic cells also need specific localization and translation. Perhaps, best known of these is the β-Actin transcript that is targeted to the leading edge of motile cells [37]. This transport is coupled with inhibition of initiation [38]. In fact, localization appears to be dependent on translation repression in all transcripts known to be actively localized [39]. In plants, examples of mRNA localization that have had this degree of experimental attention are rare. mRNAs associated with seed-storage storage proteins are targeted to particular subdomains of the ER. Importantly, mis-targeting of the mRNA disrupts resultant protein localization [40]; in the absence of such evidence, unknown nascent peptide signals could explain mRNA targeting. In turn, many transcripts whose encoded proteins function in the chloroplast are known to be enriched around the chloroplast - subunits of the Mg chelatase, protochlorophyllide oxidoreductase, and chlorophyll a/b binding protein - but the molecular mechanism has yet to be specified [41].

Translational control has been mapped to specific regions of the mRNA for a handful of plant genes (Figure 1.5). These elements are typically located within the untranslated region of the mRNA, although the 5' terminus of the protein coding region (CDS) is often involved in 5' UTR-mediated control. Coding constraints and codon usage bias within the protein coding region of the mRNA clearly limit its ability to harbor *cis*-elements, but there are exceptions. Surprisingly, plant smRNA-binding sites appear to have no bias toward any particular mRNA region (see Chapter 2).

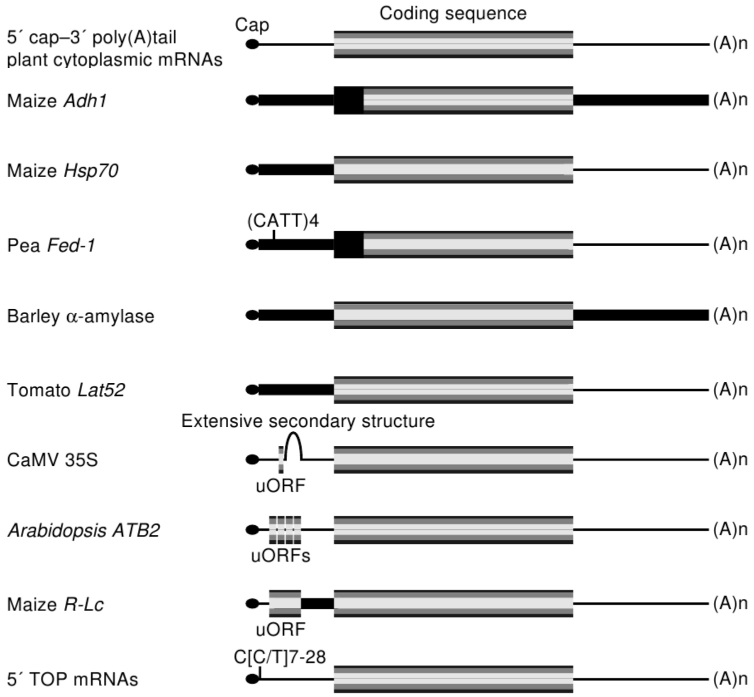


Figure 1.5: The mRNA cis-acting elements involved in translation of representative plant mRNAs.
The 5'-7mGpppN-cap structure is indicated with a filled circle. Regions of mRNAs with cis-acting sequences that regulate translation are indicated as black boxes. Abbreviations: Adh1, alcohol dehydrogenase-1; Hsp70, heat shock protein 70; Fed-1, ferredoxin-1; Lat52, tomato pollen specific mRNA; CaMV, cauliflower mosaic virus; ATB2, bZip mRNA with four uORFs; R-Lc, myc-like transcription factor with one uORF; 5' TOP mRNAs, 5' terminal polypyrimidine tract mRNAs of animals. Regions of mRNAs with cis-acting sequences that regulate translation are indicated as black boxes. Modified from [42]

Section 3: Aims of the dissertation

It should be clear from the introduction that most of our knowledge of translational control is derived from mammalian and fungal systems. While the biochemistry of these processes is homologous and mechanically similar, the genes under such control appear to be very different in plants. Thus the first aim of this dissertation was to use comparative sequence analysis in order to determine which translational control elements are conserved across plants, and to identify the pathways in which they are most important. This aim is addressed in **Chapter 2**.

uORFs appear to be surprisingly common across plant transcriptomes. Prior work in mammals and fungi indicates that the associated uAUGs are conserved more often than any other triplet in the 5'UTR, and this thesis bears out that conclusion in plants. Such conservation has lead the community to propose many functional explanations, which may or may not be mutually exclusive. These explanations make predictions that can be tested using comparative sequence analysis. In **Chapter 3**, the uORF content of plants and its functional implications are explored in more detail.

The elements described in **Chapters 2 and 3** are likely to have some bearing on the actual translation state of the genes identified. In **Chapter 4**, I describe work in which both single-gene experiments and whole-transcriptome translation profiling were used to assess that prediction. Additionally, specialized models of uORF-mediated repression was developed that makes quantitative estimates of critical initiation parameters and of the effect of initiation factor mutation on those estimates.

Many pre-existing computational tools were used to derive the biological insights emerging from this work, but, over the course of this research, it was necessary to develop software that was either missing from the set of available tools or simply lacked the extensibility to perform the tasks required. The major libraries of computer code developed to stem these gaps are described in the applicable sections as well as appendices following those sections.

In **Chapter 5**, we discuss general conclusions and future directions.

Chapter 2: The evolutionary conservation of sequences that mediate post-transcriptional control

This chapter has been submitted for publication as:

Vaughn JN, Ellingson SR, Mignone F, and von Arnim AG. Known and novel post-transcriptional elements are conserved across plant families.

Abstract

The expression of numerous plant mRNAs is attenuated by upstream open reading frames (uORFs), RNA-binding proteins, and small RNAs (smRNA). Many of these regulatory regimes are mediated by primary sequence elements. Conservation of such elements across multiple plant lineages would indicate their fundamental importance to plant biology, point to pathways for which they are most vital, and suggest the mechanism by which they act. To assess the degree of element conservation, we identified orthologous groups of mRNAs using all available EST/cDNA data from six different families of dicotyledonous plants. We then used an alignment-free technique to search for conserved motifs within associated untranslated regions (UTRs) and developed a pipeline for categorizing these motifs. Conserved motifs were detected in 3% of orthologous groups. In the 3' UTR, motifs resembling Pumilio binding elements are the most prominent group of putative recognition elements. Additionally, Expansins, one of the few plant gene families with actively localized mRNAs, possess a conserved RCCCGC motif with a more variable yet conserved upstream region. In the 5' UTR, we discovered four novel conserved-peptide upstream open reading frames (CPuORFs), three of which have monocot homologs. uORFs also appear to be conserved for their peptide-independent functions. We also found seven cases of conserved non-canonical translation initiation sites. In addition, purine-rich elements are highly enriched in the 5' UTR, as previously reported, but have a strand bias as well, suggesting that they participate in common and fundamental post-transcriptional processes in dicots. Several major forms of post-transcriptional regulation are deeply conserved in plants. Many protein families also appear to rely on a variety of these forms of regulation. Though we find some evidence for conserved co-regulation of mRNAs in a single pathway, or "RNA regulons", it remains to be seen how applicable this model is to plants.

Introduction

For many genes in plants, transcript concentration is not correlated with protein concentration [1]. This decoupling is the result of translation regulation as well as protein degradation [43]. Variation in translational efficiency, subcellular localization, and degradation rate must be encoded in the mRNA sequence. Although not all of these processes are solely mediated by primary sequence elements, the identification of conserved mRNA elements could help to assess their biological significance and to understand the pathways in which they act. We employed phylogenetic footprinting (see *Addendum*) to determine the prevalence and relative proportions of these post-transcriptional regulatory elements in plants, where, relative to mammals and fungi, such features have received less attention.

Recognition elements for RNA-binding proteins typically occur in the 3' untranslated region (UTR) assumably because ribosomes that scan the 5' UTR - the portion of an mRNA between its 5' cap and its major protein-coding region - displace any bound protein [7]. In addition, mRNAs are known to possess other sequence-specific features that change expression. One of the more ubiquitous of these features, upstream start codons (uATGs), are found in the 5' UTR. (Note that, to maintain consistency with figures, we use 'T' for 'thymine' as opposed to 'U' for 'uracil' throughout this chapter.) Because ribosomes scan the mRNA in a 5' to 3' direction in search of a start codon, these uATGs will, with variable frequency, become initiation sites for protein synthesis [16]. Their associated open-reading-frame (uORF) may either overlap the major ORF (mORF) or terminate upstream of the mORF start codon. In either case, uATGs can drastically and often detrimentally reduce protein expression [24,44]. In some cases, the uORFs resulting from uATGs are known to be conserved at the peptide-level [25,27,28] - referred to as 'conserved peptide uORFs' (CPuORFs). Additionally, cells express uORFs in quantity [45] and some of these uORF-peptides mediate responses to small molecules [27]. CPuORFs shared by *A. thaliana* and rice fall into 19 homologous groups [25]. It is still unclear how the prevalence of uORFs changes with regard to phylogenetic scope or the degree to which the 5' UTR harbors other forms of protein-coding potential.

Many researchers have proposed that groups of specific genes are co-regulated at the mRNA level by interactions between a common RNA binding factor and its cognate RNA sequence elements [46,47]. This proposition is supported by experiments on the PUF family of RNA binding proteins within complex eukaryotes [48,49] and on 40 RNA-binding proteins in yeast [35]. A striking example is the PUMILIO protein in *Drosophila*, which interacts with mRNAs for a majority of subunits of the vacuolar ATPase [48]. An extension of the regulon model predicts that, similar to elaborate transcriptional modules, some mRNAs will be under combinatorial control. Like many inferences of regulatory function, it is unclear which interactions are significant for organismal fitness and which are true positives that are operationally inert and, hence, evolutionarily neutral [50]. It is clear from work on PUF-family proteins that, while binding elements appear to stay constant, the function of their target genes can vary drastically across major taxonomic divisions [48]. Such variation appears to carry over into plants [49], 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 [51]. Therefore, we have focused on the UTRs. Respectively, small-RNA-induced repression is another major regulatory process known to act post-transcriptionally on specific mRNAs, and, as opposed to metazoans, most putative small-RNA (smRNA) target sites in *A. thaliana* are located in the mORF of mRNAs [32]. Still, 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.

With regard to comparative sequence analysis, it has been argued that highly-diverged sequences fall outside the 'window of useful divergence' (~20 to ~50 million years since speciation, for plants) [52]. This appears to be true with regard to alignment-dependent identification of transcription-factor binding sites because of site turnover or high insertion-deletion activity around orthologous sites [53]. Regardless of positional constraint or lack thereof, distant evolutionary relationships between intergenic or untranslated regions typically preclude alignment. Instead, we assessed the statistical enrichment of motifs across highly diverged UTRs that are linked to orthologous coding sequences - a technique that is alignment-free and somewhat robust to site turnover or displacement [54]. Our approach, based on the MEME algorithm [55], requires effectively randomizing divergence between all sequences compared. Since we are looking for deep conservation, this prerequisite is satisfied by default [56]. Yet, when creating orthologous groups from distantly related plant species, in-paralogs, either resulting from local duplication or polyploidy, may result in false positives caused by small divergence times [54]. As described below, we addressed this problem by reducing groups of orthologous coding 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; similarly, 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 non-synonymous mutation bias is not assessed [57]. Hence, we have attempted to categorize conserved motifs into plausible functional groups based on whether or not they 1) overlap microsatellites, 2) are target sites of known/predicted microRNAs, 3) code for a peptide that is constrained at the amino-acid level, 4) can be considered recognition elements based on pockets of dense conservation longer than four nucleotides, and 5) are likely transcription factor binding sites or unknown smRNA targets (Figure 2.1).

By applying our motif identification approach to six eudicot lineages with substantial transcript data, we found that at least 3% of orthologous groups have one or more conserved UTR motifs. The majority of these motifs appear to be acting 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. With regard to the RNA regulon model, we find few examples where conserved post-transcriptional co-regulation is occuring. Our data suggest that, in plants, it is more appropriate to consider post-transcriptional regulation as acting on individual, perhaps 'keystone', genes and gene families, which are under multiple forms of UTR-mediated control.

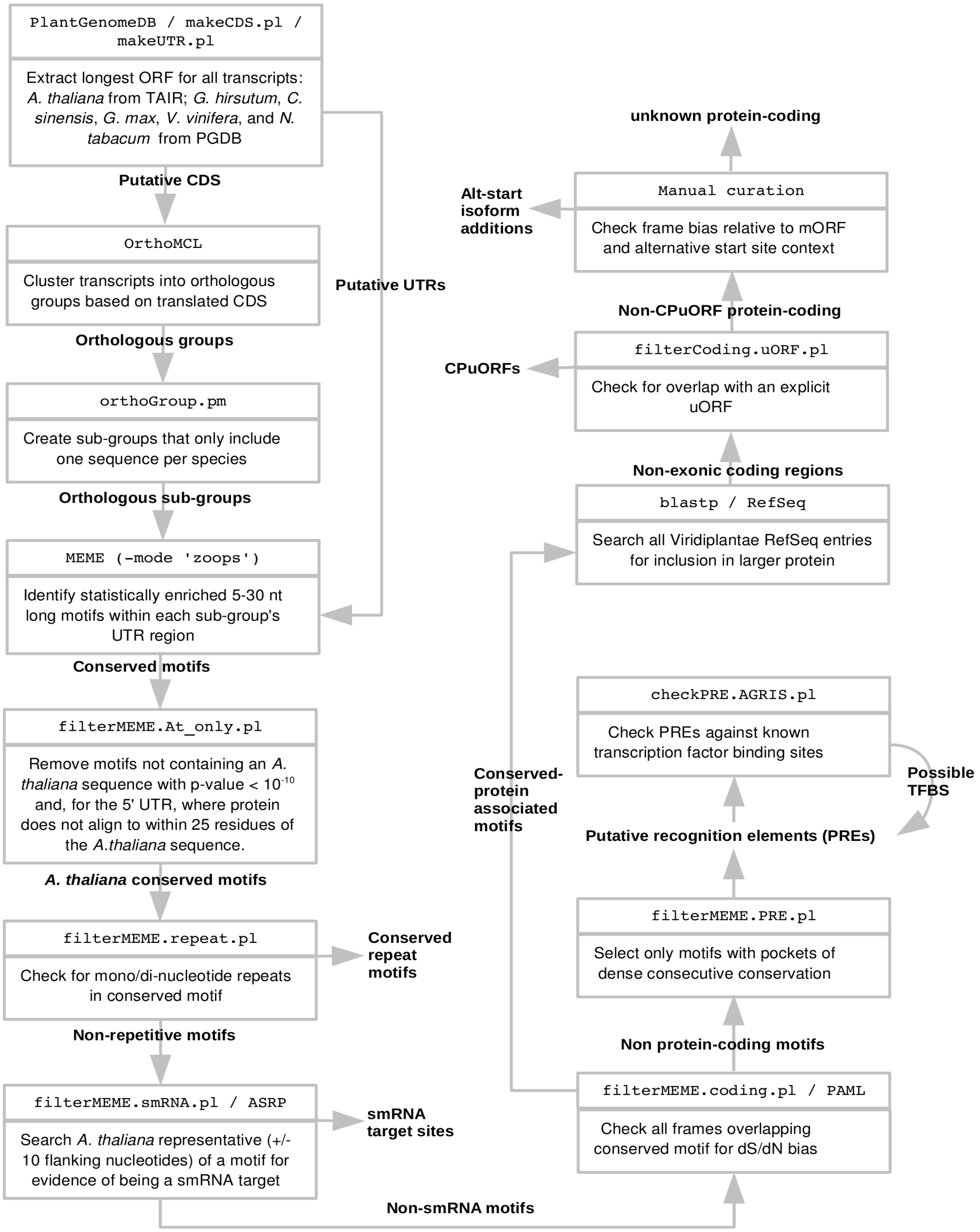

PlantGenomeDB / makeCDS.pl / makeUTR.pl
Extract longest ORF for all transcripts: A. thaliana from TAIR; G. hirsutum, C. sinensis, G. max, V. vinifera, and N. tabacum from PGDB
Putative CDS
Putative UTRs
OrthoMCL
Cluster transcripts into orthologous groups based on translated CDS
Orthologous groups
orthoGroup.pm
Create sub-groups that only include one sequence per species
Orthologous sub-groups
MEME (-mode 'zoops')
Identify statistically enriched 5-30 nt long motifs within each sub-group's UTR region
Conserved motifs
filterMEME.At_only.pl
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
filterMEME.repeat.pl
Check for mono/di-nucleotide repeats in conserved motif
Conserved repeat motifs
Non-repetitive motifs
filterMEME.smRNA.pl / ASRP
Search A. thaliana representative (+/- 10 flanking nucleotides) of a motif for evidence of being a smRNA target
smRNA target sites
Non-smRNA motifs
filterMEME.coding.pl / PAML
Check all frames overlapping conserved motif for dS/dN bias
Non protein-coding motifs
Conserved-protein associated motifs
filterMEME.PRE.pl
Select only motifs with pockets of dense consecutive conservation
Putative recognition elements (PREs)
checkPRE.AGRIS.pl
Check PREs against known transcription factor binding sites
Possible TFBS
blastp / RefSeq
Search all Viridiplantae RefSeq entries for inclusion in larger protein
Non-exonic coding regions
filterCoding.uORF.pl
Check for overlap with an explicit uORF
CPuORFs
Non-CPuORF protein-coding
Manual curation
Check frame bias relative to mORF and alternative start site context
Alt-start isoform additions
unknown protein-coding

Figure 2.1: Schematic of the computational pipeline used for conserved sequence identification and categorization.

Tasks are described in the bottom section of each box. The program(s) that performed each task are given in the top section of each box.

Results

EST/cDNA sequence coverage of the eudicot proteome approaches the theoretical maximum predicted by the fully sequenced *Ricinus communis* genome

UTRs are difficult to predict computationally from genomic DNA [58]; hence,experimentally confirmed mRNA sequences are required for valid UTR comparisons. *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 (Table 2.1 and Figure 2.2D). The PlantGDB transcripts were assembled from all available expressed sequence tag (EST) and cDNA sequence data in GenBank. The species were chosen because of their degree of divergence from one another and the extent of their sequence coverage. The longest continuous reading frame containing an in-frame ATG for each transcript was considered the major ORF (mORF). mORFs were translated and clustered into orthologous groups using reciprocal *blastp* and OrthoMCL. The sequence upstream of an mORF was considered its 5' UTR. Likewise, the sequence downstream was considered its 3' UTR.

Table 2.1: A numerical account of the clustering of dicot mRNA sequence data into orthologous groups.

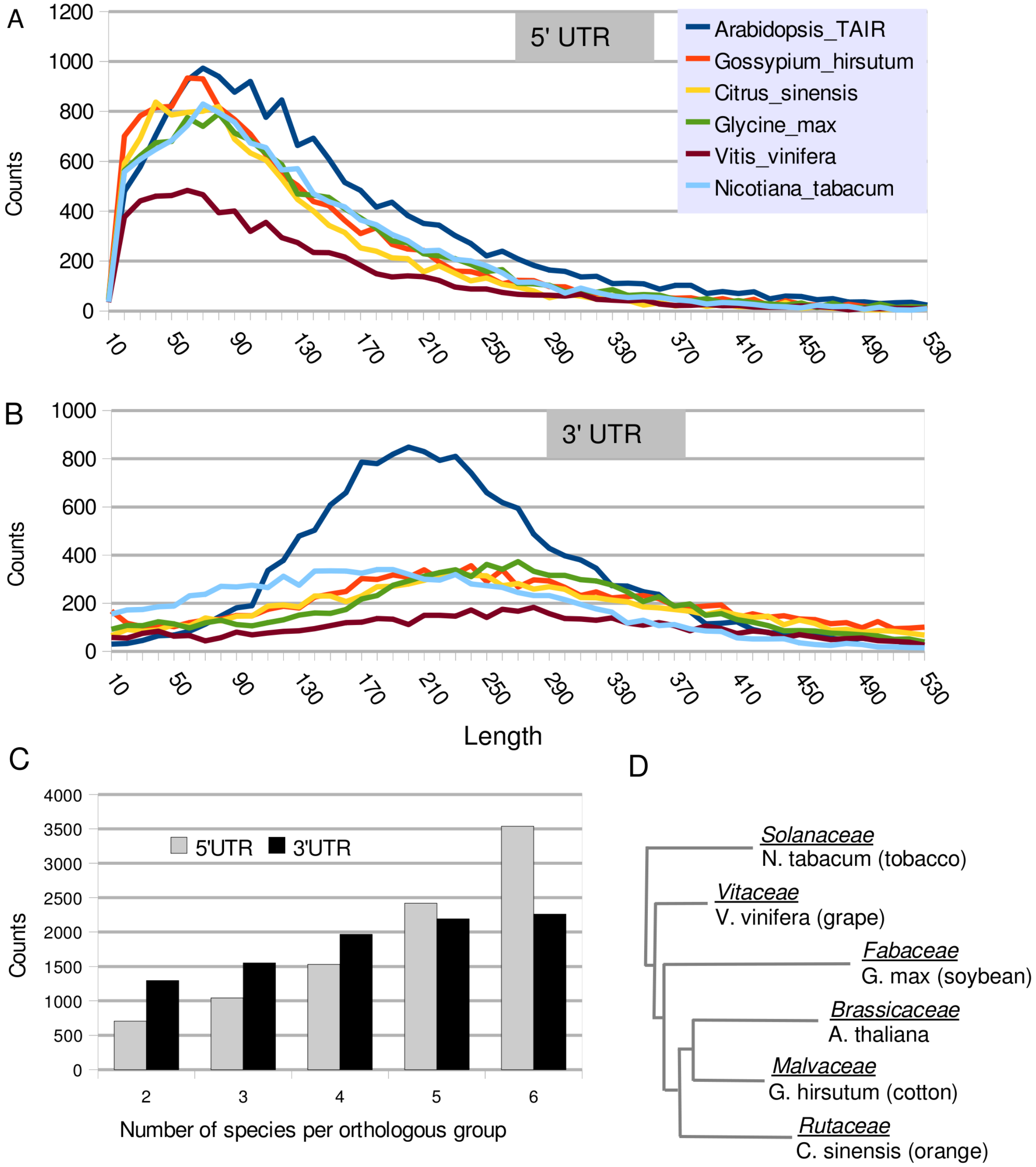
Species	Putative transcripts	Number of sequences after clustering	Number of orthologous groups with the species present	Number of orthologous groups containing the given species and an *A. thaliana* sequence
Total	*698,941*	*158,613*	*31,909*	*10,122*
Arabidopsis thaliana[a]	39,640	29,823	11,887	10,122
Gossypium hirsutum	98,420	28,385	14,541	8,016
Citrus sinensis	105,294	34,063	15,168	7,187
Glycine max	258,849	26,192	14,124	8,216
Vitis vinifera	64,796	14,091	9,763	6,380
Nicotiana tabacum	131,942	26,059	13,995	8,060

[a] From TAIR9 release, as opposed to the informant species from PlantGDB.

Of the 11,887 groups containing an *A. thaliana* sequence, 10,122 (85.1%) contained at least one informant species as well (Table 2.1). (The remaining 14.9% of *A. thaliana*-groups consist solely of in-paralogs, were no orthology across lineages could be inferred.) 10,122 is comparable to the number of orthologs, 10,381, shared by *A. thaliana* and *Ricinus communis*, both of which have sequenced genomes [59]. As expected, a bulk of putative transcripts represented fragmentary and unspliced mRNAs and were eliminated by orthologous clustering (Table 2.1, Columns 2 and 3). Many of these fragments persist as groups with a single taxa; hence, though there are only 10,122 groups with an *A. thaliana* and an informant species representative, there are 31,909 groups total. Some of these groups represent lineage-specific genes, but many groups are likely to result from the clustering of multiple versions of a sequence fragment that were too short to cluster with their appropriate full-length transcripts (not shown).

In terms of sequence coverage, there is a clear bias toward coverage across the 5' UTR (Figure 2.2). While both 5' and 3' UTRs show a median length that is comparable to *A. thaliana* (Figure 2.2A and B), informant species have fewer 3' UTRs present, and these 3' UTRs, when present, are often longer or shorter than *A. thaliana*. Whether this length variation is a result of less thorough sequencing or a genuine trend among the dicots sampled is unclear. For both regions, the majority of comparisons have good representation, i. e. four to six species (Figure 2.2C). As expected [60], the 3' UTR median length in all species is longer than the 5' UTR (Figure 2.2A-B). The distinction between length distributions of 5' and 3' UTRs is thought to result from countervailing effects in the 5' UTR involving the migration of *transcriptional* start site and acquisition of potentially lethal pre-mature start codons - processes that do not come to bear on the 3' UTR [61].

Figure 2.2 (next page): Both 5' and 3' UTRs have multiple informant species per comparison, but the 3' UTRs in informant species are more variable in length than A. thaliana.
A) and B) 5' UTR and 3' UTR length distributions, respectively, per species for all putative transcripts that could be clustered into orthologous groups. The width of each x-axis bin is 10 nucleotides. C) Distributions for the number of species per orthologous group in 5' and 3' UTR comparisons. Total counts for each region will be slightly lower than the number of orthologous groups containing an A. thaliana representative and at least one other species (Table 2.1) because some orthologous groups lack UTR data. D) Tree representing descent and relative divergence of the species in this analysis. Modified from - and based on chloroplast genomes. The family name is given above its representative species.


A
5' UTR
Arabidopsis_TAIR
Gossypium_hirsutum
Citrus_sinensis
Glycine_max
Vitis_vinifera
Nicotiana_tabacum
Counts
1200
1000
800
600
400
200
0
10
50
90
130
170
210
250
290
330
370
410
450
490
530
B
3' UTR
Counts
1000
800
600
400
200
0
10
50
90
130
170
210
250
290
330
370
410
450
490
530
Length
C
Counts
4000
3500
3000
2500
2000
1500
1000
500
0
5'UTR
3'UTR
2
3
4
5
6
Number of species per orthologous group
D
Solanaceae
N. tabacum (tobacco)
Vitaceae
V. vinifera (grape)
Fabaceae
G. max (soybean)
Brassicaceae
A. thaliana
Malvaceae
G. hirsutum (cotton)
Rutaceae
C. sinensis (orange)

3% of orthologous groups contain a conserved motif in the UTR

5' and 3' UTRs generally evolve faster than the CDS. Given the divergence between the species in this study, we assumed that neutrally-evolving portions of the UTR will have a nominal number of consecutive bases conserved as a result of relatedness alone (see *Results* below). We therefore used the MEME algorithm to search for enriched elements within these UTRs. MEME assumes a random background model and calculates the number of times a given motif is expected to be present by chance alone in a given set of sequences (E-value). Importantly, in-paralogs, resulting from post-speciation duplication events, could potentially have undergone very short divergence times, undermining our assumption of effective randomization and disrupting conserved motif identification. Also, our dataset contains many unknown 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 *Methods*). This approach has two additional benefits. One, in-paralogs, which may have undergone neo/subfunctionalization at the regulatory level by losing an element, do not add noise to the identification process [62]. Two, our false discovery rate can be estimated by simply randomizing orthologous groups, as opposed to simulating mutations.

Our statistical criteria, defined in *Methods*, resulted in 194 and 96 conserved motifs for the 5' and 3' UTRs, respectively (see Figure 2.5B below). Only one orthologous group was found to contain both a 5' and 3' UTR motif; hence, ~3% [(194 + 96 - 1) / 10,122] of genes in our study have a conserved UTR motif. 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 ~2 false positives per 100 positive results). We further characterized motifs based on their composition and their patterns of conservation.

[AG]|[CT] repeats are enriched in the 5' UTR and have a strand bias

Microsatellites (mono/dinucleotide repeats) are common in plants, have a regional bias in the genome [63], and are conserved between orthologs and across paralogs [64]. The consensus sequence associated with each motif was checked for >5 consecutive mononucleotide repeats or >3 consecutive dinucleotide repeats. As seen previously [63], we found a dramatic enrichment in dinucleotide repeats in the 5' UTR relative to the 3' UTR (Table 2.2). Mononucleotide repeats show no such 5'/3' bias suggesting that differential sequence coverage between 5' and 3' UTRs (Figure 2.2) is not a major factor, particularly to the extent that it would explain such an extreme difference: 51-fold enrichment for [AG]$_n$ and 15-fold for [CT]$_n$ (Table 2.2). When orthologous groups were randomized and significant motifs reassessed (see *Methods*), the few recovered motifs were primarily [AG]|[CT] repeats ('random' datasets in Table 2.2). This indicates that, in terms of repeats, a proportion of our positive results are caused by general enrichment of [AG]|[CT] repeats within the 5' UTR, although the actual dataset still has significantly and substantially more conserved [AG] and [CT] repeats: p-value $<10^{-33}$ and $<10^{-13}$, respectively, based on a binomial distribution where p is calculated from [random_dataset_repeats / number_of_orthologous_groups_with_Arabidopsis] (Table 2.1 and 2.2).

Table 2.2: Number of repeat motifs in 5' and 3' UTRs.

Dataset[a]	**Repeat**[b]									
	[AC]$_n$	[AG]$_n$	[AT]$_n$	[CG]$_n$	[CT]$_n$	[GT]$_n$	A$_n$	C$_n$	T$_n$	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

[a] 'random' refers to the control analysis involving randomization of orthologous groups (see *Methods*). [b] All overlapping dinucleotide repeats, such as [AC]$_n$ and [CA]$_n$, are pooled.

Interestingly, we found that [AG] repeats appear to have a strand bias in the 5' UTR (p-value $<10^{-5}$; binomial distribution, $p = 0.5$). While [AG] repeats are known to act at the transcriptional level [65,66], the responsible transcription factors appear to be orientation insensitive [67]. Our identification of strand bias indicates that [AG]-repeats are acting at the mRNA level as well. Based on a 1st-order Markov model, none of these findings could be explained by differences in background nucleotide composition, which are comparable across 5' UTR and 3' UTR as well as the forward and reverse strand (data not shown).

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

Numerous whole-transcriptome sequencing experiments have been used to generate the ASRP database of smRNAs and their mRNA targets [32,68-70]. In striking contrast to metazoans, there are only 238 smRNA targets in the ASRP database and, of these, only 15 target sites appear to be located in the 3' UTR. Of the balance, 8 occur in the 5' UTR while the remainder lie in the CDS of targeted mRNAs.

After removal of repeats motifs, we checked the *A. thaliana* representative of each of our motifs (with +/- 10 flanking nucleotides) against all putative target sites. Of these, we identified two in the 5' UTR and three in the 3' UTR (Table 2.3). Motifs were also checked against more liberal predictions of smRNA targets [71], with the same result. All miRNA families listed in Table 2.3 are known to be conserved in at least 2 dicots. Interestingly, *G. max* is represented in the conserved miR403 target site, but, while both miR398 and miR169 genes were identified in prior studies of *G. max* [72,73], the miR403 gene was not.

The 'seed' hypothesis predicts that complementarity to the 5' end of the mature smRNA is critical for mediating silencing [74]. To varying degrees, all five motifs support this hypothesis, although it appears that certain genes, AT1G08830 (with respect to AT3G15640) and AT1G72830 (with respect to AT5G12840), also have substantial complementarity to the 3' end of the mature smRNA across dicots. In turn, these appear to be more conserved across their entire length. Though it still is biased toward the 'seed' end, AT1G31280 has more sporadic conservation and appears to require a short pocket of 3' complementarity as well.

Table 2.3: Motifs implicated in experimentally confirmed smRNA-mediated degradation pathways.

mRNA region	***A. thaliana* accession**	***A. thaliana* annotation**	**miRNA family**	**reduced IUPAC consensus sequence**[a]
5' UTR	AT3G15640	cytochrome c oxidase family protein	miR398	nCnnnnnGGnGnGACCTGAGA(21)
5' UTR	AT1G08830[b]	Cu/Zn superoxide dismutase (CSD1)	miR398	AAGGGGTnYYCTGAGATCACAnAn(24)
3' UTR	AT1G72830	CCAAT-binding transcription factor	miR169	GGnAnnTCATCCTTGGCTn(19)
3' UTR	AT5G12840	CCAAT-binding transcription factor	miR169	nGCnAATCATTCTTGGCT(18)
3' UTR	AT1G31280	PAZ/piwi domain-containing protein	miR403	AAGnnnnTnnnGCGTnnAnCT(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% AND C+T >84%, respectively, 3) 'n' if otherwise. [b] Not in ASRP database but from [75].

At least 18 uORFs function at the peptide level

If any region in the UTR codes for a conserved protein, either as a result of inaccurate sequence data or a legitimate biological process, then that region will appear as a highly significant motif in our analysis. We therefore checked all motifs for their coding potential. The longest continuous reading frame (CRF) among each of the three possible reading frames from each species associated with a motif was aligned as protein. Each alignment was tested for coding potential based on a likelihood ratio test between a model in which non-synonymous and synonymous mutations rates are equivalent (null hypothesis) and a model in which this ratio is allowed to vary (alternative hypothesis). A similar analysis was done for the longest uORF overlapping a motif. Motifs were annotated as protein coding if the p-value result from this test was <0.01 (see *Methods*). By then searching sequences with significant coding potential against all known Viridiplantae proteins, we could determine whether a conserved motif is an artifact of alternative splicing events. Such motifs were removed from further consideration. uORFs were considered conserved peptide uORFs, or CPuORFs, if they passed the above criteria.

Figure 2.3(next page): Novel uORFs show positional sequence conservation patterns similar to known uORFs, but have little phylogenetic bias.

(A-C) Transcript alignments are shown in miniature followed by all possible ORFs, where the darkness of red indicates the context strength of the ATG. CPuORFs are bordered in blue; the beginning of each associated mORF is bordered in green. Orange vertical lines in 'Nucleotide conservation' lane indicate that all residues are identical in that column. The orange horizontal line below each accession indicates sequence coverage or indels relative to the alignment. Sequences are ordered, top to bottom, relative to their phylogenetic distance from A. thaliana. (A) Previously confirmed CPuORF-14. (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. E) Y-axis indicates the number of representatives for a given species among all CPuORF alignments. The maximal number of CPuORF alignments possible is equal to the A. thaliana value. X-axis is ordered left to right based on the phylogenetic distance of a given species from A. thaliana (not to scale).

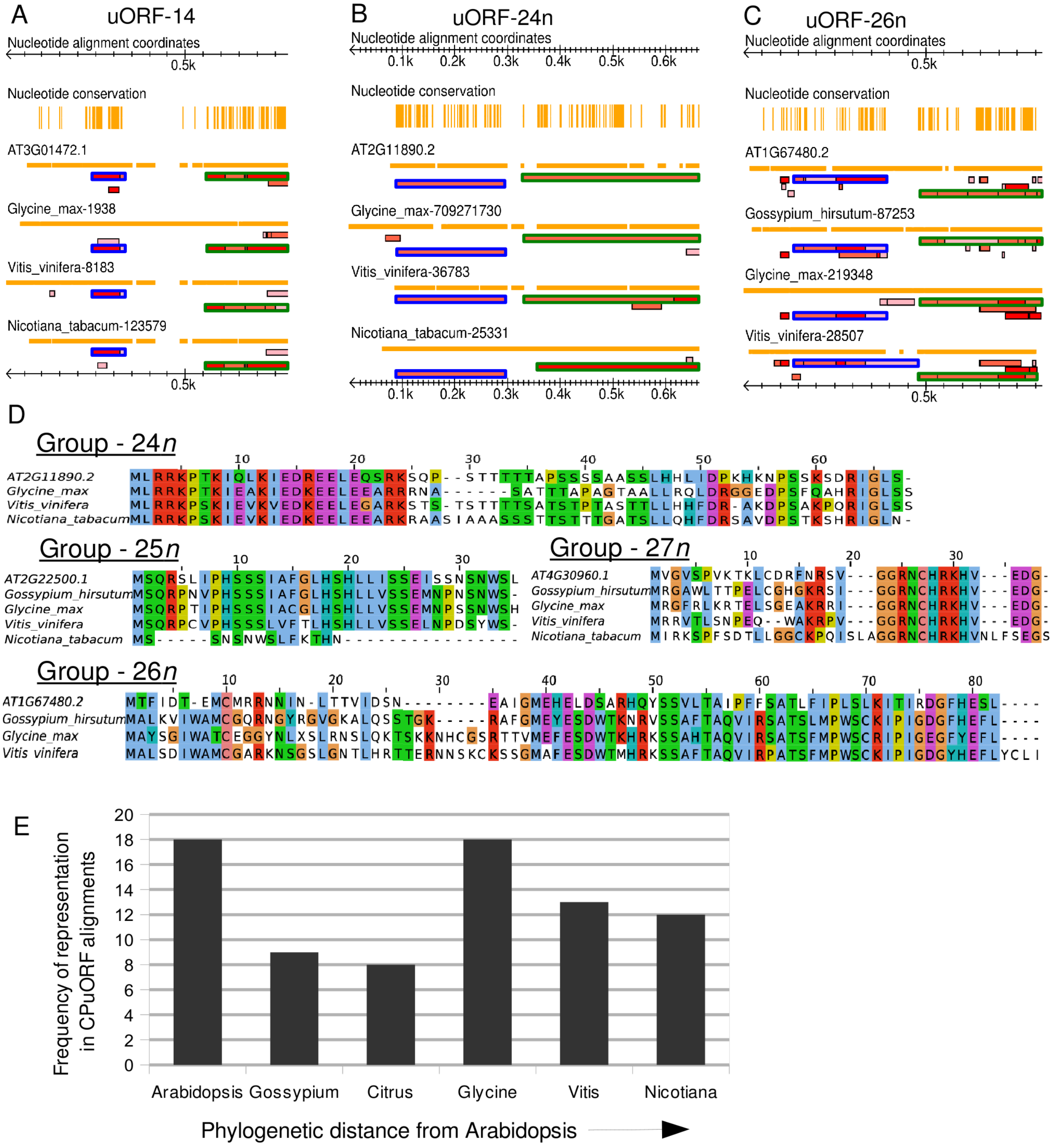

A
uORF-14
Nucleotide alignment coordinates
0.5k
Nucleotide conservation
AT3G01472.1
Glycine_max-1938
Vitis_vinifera-8183
Nicotiana_tabacum-123579
0.5k
B
uORF-24n
Nucleotide alignment coordinates
0.1k 0.2k 0.3k 0.4k 0.5k 0.6k
Nucleotide conservation
AT2G11890.2
Glycine_max-709271730
Vitis_vinifera-36783
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
E
Frequency of representation in CPuORF alignments
0 2 4 6 8 10 12 14 16 18 20
Arabidopsis
Gossypium
Citrus
Glycine
Vitis
Nicotiana
Phylogenetic distance from Arabidopsis

Table 2.4: All motifs associated with protein coding potential.

A. thaliana **gene accession**	***A. thaliana*** **mORF annotation**	**alternative splicing**[a]	**CRF p-value**	**uORF p-value**[b]	**Predicted cause of coding potential**[c]
AT2G11890	adenylate cyclase	none	1.64E−19	1.19E−15	uORF(24n)
AT4G36990	heat shock transcription factor 4 (HSTF4)	none	1.28E−17	7.33E−15	uORF(18)
AT3G12012	Mic-1 homolog	XP_002882761.1	1.67E−17	2.07E−15	uORF(8)
AT3G62420	bZIP transcription factor family protein	none	5.25E−16	1.42E−24	uORF(1)[d,e]
AT3G01472	HD-ZIP 1 transcription factor	none	9.45E−16	6.19E−18	uORF(14)
AT1G29950	bHLH transcription factor	none	5.57E−15	4.99E−11	uORF(15)
AT3G25570	adenosylmethionine decarboxylase family protein	XP_002272179.1	3.36E−14	1.47E−15	uORF(3)
AT4G25690	expressed protein	XP_002513657.1	7.01E−12	2.86E−15	uORF(4)
AT5G07840	ankyrin repeat family protein	none	4.85E−11	5.18E−11	uORF(5)
AT1G23150	expressed protein	none	3.86E−10	2.94E−08	uORF(12)
AT2G43020	amine oxidase family protein	none	1.11E−08	6.48E−11	uORF(6)
AT2G22500	mitochondrial substrate carrier family protein	none	4.79E−08	3.16E−08	uORF(25n)
AT5G01710	expressed protein	none	2.55E−07	7.77E−05	uORF(17)
AT1G67480	kelch repeat-containing F-box family protein	none	3.54E−07	1.12E−07	uORF(26n)
AT4G30960	CBL-interacting protein kinase 6 (CIPK6)	none	3.77E−07	8.98E−07	uORF(27n)[e]
AT1G48600	methyltransferase	none	4.42E−07	1.39E−07	uORF(13)
AT4G34590	bZIP transcription factor family protein	none	5.56E−07	8.10E−08	uORF(1)[d]
AT4G30960	CBL-interacting protein kinase 6 (CIPK6)	none	1.88E−04	8.98E−07	uORF(27n)[e]
AT3G62420	bZIP transcription factor family protein	none	7.80E−04	1.42E−24	uORF(1)[e]
AT1G36730	eukaryotic translation initiation factor 5, putative	none	1.24E−03	4.05E−05	uORF(7)
AT1G03260	expressed protein	none	5.19E−11	#N/A	non-AUG start (CTG)[f]
AT1G32700	zinc-binding family protein	none	5.19E−09	#N/A	non-AUG start (CTG)[f]
AT2G25110	MIR domain-containing protein	none	1.86E−07	#N/A	non-AUG start (GTG)
AT3G16630	kinesin motor family protein	none	7.26E−05	#N/A	non-AUG start (TTG)
AT4G16280	flowering time control protein (FCA)	none	1.37E-03	#N/A	non-AUG start (CTG)[f]
AT5G14500	aldose 1-epimerase family protein	none	5.06E−03	#N/A	non-AUG start (CTG)[f]
AT1G55760	BTB/POZ domain-containing protein	none	5.21E−03	#N/A	non-AUG start (CTG)[f]
AT4G26850	expressed protein	none	2.48E-11	#N/A	unknown
AT2G18040	peptidyl-prolyl cis-trans isomerase (PIN1)	none	9.79E−07	#N/A	unknown
AT1G01060	myb family transcription factor	none	1.31E−05	#N/A	unknown
AT4G26850	expressed protein	none	6.94E−05	#N/A	unknown[e]
AT1G57680	expressed protein	none	3.94E−03	#N/A	unknow

[a]Entries in 'alternative splicing' column indicate the RefSeq Viridiplantae protein matches that were 50% longer than the translated uORF query from *A. thaliana*. [b] 'uORF' is different from continuous reading frame (CRF) in that it contains an in-frame ATG codon; hence, not all CRFs will contain a uORF. [c] Value next to 'uORF' indicates the homology group associated with prior *A. thaliana* and *O. sativa* comparisons [25] and parenthetical string next to 'non-AUG start' indicates the likely start codon based on alignments. [d]Associated mORF clusters into a separate group in spite of the uORF being in the same homology group. [e]Coding potential is found in two distinct frames. [f]Recently identified in [76] as well.

We identified 4 CPuORFs with little or no precedent in the literature - indicated by '*n*' in Table 2.4. These novel CPuORFs show similar conservation profiles and spatial patterns to known CPuORFs (Figure 2.3A-C), although CPuORF-24n is particularly unusual in that it begins near the cap and extends the length of the entire 5' UTR (Figure 2.3B). The *A. thaliana* mORF associated with this CPuORF has deep EST support and there is no neighboring gene within 1kb upstream of the transcriptional start site (see http://gbrowse.arabidopsis.org). CPuORFs 24*n* and 25*n* exhibit extensive amino acid conservation across their entire length, while the others show a 3' bias in their degree of conservation (Figure 2.3D). We also recapitulated 13 of the 19 CPuORFs found previously in an *A. thaliana* (dicot) and *Oryza sativa* (monocot) comparison (Table 2.4) [25]. Of the balance, 1 negative result was due to a lack of identifiable mORF orthologs, and 5 negative results either lacked sufficient 5' UTR sequence data (false negative) or are true negatives that have dicot-monocot homology but are not present in the other dicots under study.

Based on prior analysis comparing *A. thaliana* paralogs resulting from a recent whole-genome duplication, four CPuORFs - 20, 23, 25, 26 - have no monocot homologs [25]. Using *Zea mays*, *Oryza sativa*, and *Sorghum bicolor,* we find three of the four novel CPuORFs in one or more monocot lineages (Table 2.5). Additionally, there was no clear phylogenetic bias among CPuORFs (Figure 2.3E), particularly when considering sequence coverage (Figure 2.2). For example, while *G. hirsutum* has more sequence coverage and is more closely related to *A. thaliana* than *V. vinifera*, it is involved in fewer CPuORF alignments. That being said, there is a reduction in CPuORF content between *G. max* and *N. tabacum,* the most distantly related species, in spite of comparable sequence coverage; CPuORF substitutions and truncations of CPuORF-25n in *N. tabacum* and, to a lesser extent, CPuORF-27n bears this out (Figure 2.3D). One CPuORF, 27n, could not be found in any of the three monocot lineages even though extensive sequence data exists for the 5' UTR of CIPK6 homologs. Though this may be another example of a dicot-specific CPuORF, taken together, it appears that most CPuORFs have monocot homologs. These finding indicate that, generally, CPuORFs are not lineage-specific.

Conserved non-canonical start sites have a strong Kozak context and vary at the +1 position

In addition to explicit uORFs, we identified regions of the 5' UTR that appear to code for sequences constrained at the peptide level but that have no apparent start codon. The frame of coding potential for each such example was checked for whether or not it was consistently in-frame with the mORF. Where all sequences were in-frame, these were further checked for an in-frame stop codon upstream of the coding potential. As with uORFs, these regions were also searched against known full-length plant proteins to guard against alternative/incomplete splicing artifacts. Seven cases satisfied these criteria and were categorized as a 'non-AUG start' (Table 2.4).

Strikingly, the beginning of each of these putative coding regions contains a variant of the pattern 'A[AC]N[GTC]TGG', where 'N' indicates any nucleic acid and brackets indicate possible residues (Figure 2.4A). This pattern, excepting the variable adenine of the canonical site, matches the strongest Kozak context in plants [11]. Thus, it would appear that any nucleotide can effectively replace adenine at the first position. In one case, AT1G32700.1, strong context can potentially overcome two substitutions in the canonical ATG triplet. None of these putative start sites significantly approximates the splice site consensus. Curiously, though the first position of the start codon appears mutable, the alternative nucleotide is almost always conserved.

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. A comparable number of protein coding motifs, which we could not categorize, were found in the 5' UTR and are annotated as 'unknown' in Table 2.4.

Table 2.5: Novel CPuORF monocot homologs.

CPuORF accession*	Protein sequence alignments
24n - implicit *Zea mays, Oryza sativa, Arabidopsis thaliana*	NM_001143483.1_1 --TSNPTRIELTSSDRDELEDHLRAAAAATTTKDPSGYTTPSPVLGPQTSNPLLQFLHPK NM_001068020.2_1 --------IELRSSDRDELEDHLRAAAAAAAS-TPTASSTPTTTPPPSNSNPLLHLLHPP AT2G11890.1_1 MLRRKPTKIQLKIEDREELEQSRKSQPSTTTTTAPSSSSAAS---------SLHHLIDPK *:* .**:***: :: .::::: *:. ::.: .* :::.* NM_001143483.1_1 PGAVPSKSQRIGIGLSTPPAPAPNPRPPHPPHGG NM_001068020.2_1 PGAAPSKSHRIGL----PTNPNPNPKP------- AT2G11890.1_1 HKNPSSKSDRIGLS-------------------- .***.***:
25n - explicit *Saccharum officinarum, Zea mays, Sorghum bicolor, Arabidopsis thaliana*	AY644463.1_1 MSQRASVPHSSCIGCALHSHLLVSSEMCPSSYWQQ NM_001157169.1_1 MSQRASVPHSSCIAFALHSHLLVSSEMCPSSYWQQ XM_002445603.1_1 MSQRASVPHSSCIAFALHSHLLVSSEMCPSSSWQQ AT2G22500.1_1 MSQRSLIPHSSSIAFGLHSHLLISSEISSNSNWSL ****: :****.*. .******:***:...* *.
26n - explicit *Arabidopsis thaliana, Oryza sativa*	AT1G67480.1_2 MTFIDTEMCMRRNNINLTTVIDSNEAIGMEHELDSARHQYSS------VLTAIPFFSATL NM_001053449.1_3 MAIDNCAMCVGGKGFYLNSKETSDPSRKNHSKVSQYRMAFDAPRITKTETSKLKNLISAS *:: : **: :.: *.: *: : . ::.. * :.: : : : :: AT1G67480.1_2 FIPLSL------------------- NM_001053449.1_3 FKPLSLTIPIGDGFHELFPVGHCHL * ****

*'Implicit' indicates that the monocot homolog's start codon is not covered by the current 5'UTR annotation. Species are listed in the order that they appear in the alignment.

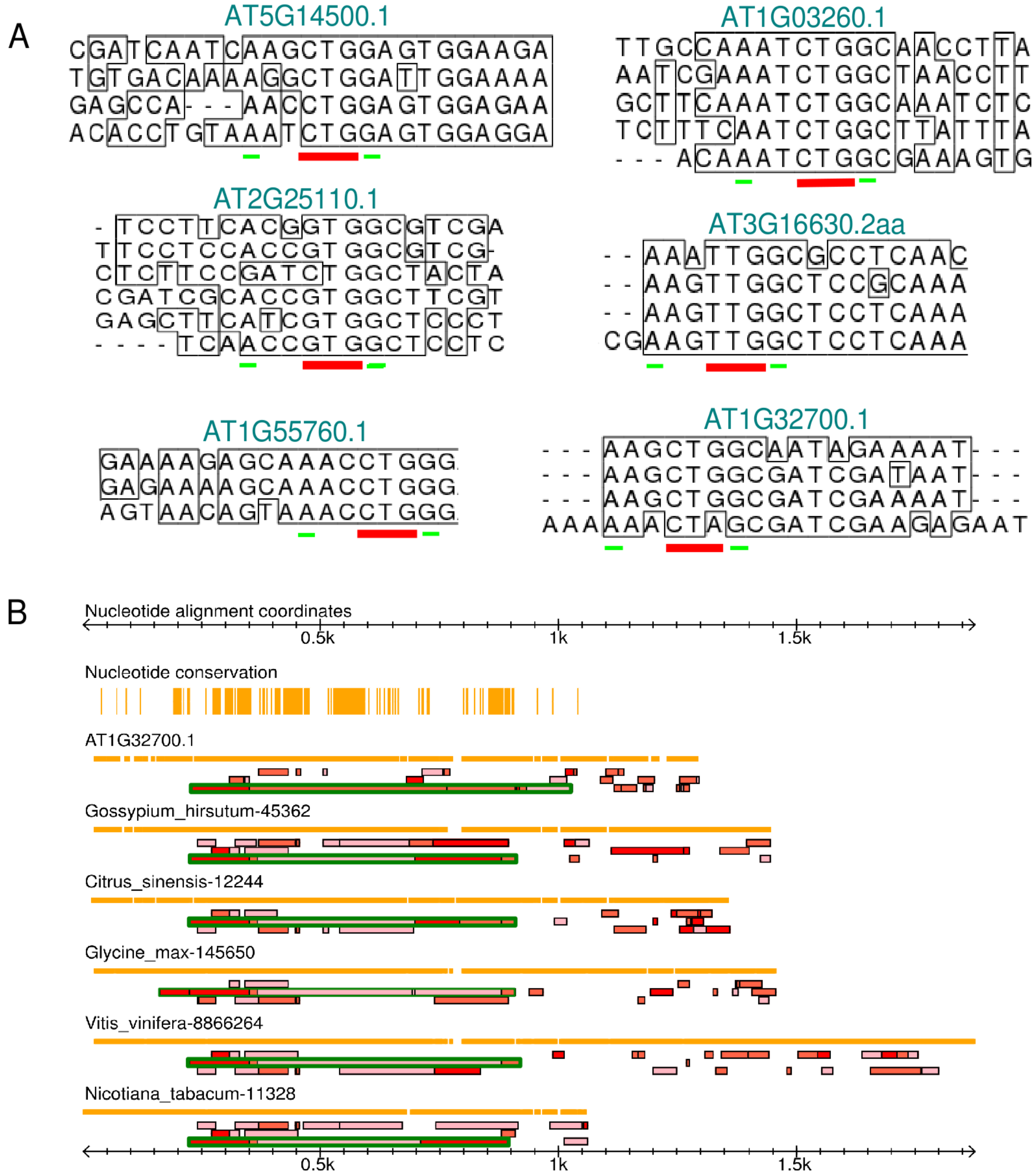


Figure 2.4: Non-canonical start sites have strong context and a variable +1 residue.
A) Each sub-panel is a sample from back-translated protein alignments of translated sequences from the conserved motif. Conserved blocks are demarcated by boxes. Red lines indicate the proposed start codon and green lines indicate the two most critical context sites. The A. thaliana accession from each group appears above the respective alignment. B) The only group in which the non-canonical start has mutated back to the AUG (G. max sequence). See Figure 2.3 legend for explanation of coloring and layout.

A small but distinct fraction of genes with putative 3' UTR recognition elements support the RNA regulon model

Short sequences exhibiting strong conservation in MEME alignments can have equivalent or, often, higher E-values than longer weakly-conserved sequences. Though both may be important for plant function, our aim was to identify potential recognition elements, which experiments suggest are between 5 and 20 nucleotides long [35,77,78]. Motifs not belonging to any of the above categories were scanned for regions of dense conservation. Within these motifs, the largest window of average consensus-letter frequency greater than 0.92 (see *Methods*) was defined as a putative recognition element (PRE). See thesis supplemental File 1, 5UTR_motif.pdf and File 2, 3UTR_motif.pdf for local alignments associated with all PREs.

We identified 92 PREs in the 5'UTR and 81 in the 3' UTR (Figure 2.5B). 29% of all 5' UTR PREs fall within our smallest length category (5-8 nts) (Figure 2.5A). In the 3' UTR, PREs peak at ~20 nts 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 2.2), we consider these to represent non-smRNA binding elements. In both regions, PREs make up the majority of conserved motifs (Figure 2.5B). In a gene ontology (GO) term analysis, two subsets among the 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 = $4.44x10^{-3}$) and signal transduction proteins (12, genes, 4.2-fold enrichment, *p*-value = $1.80x10^{-2}$) - based on GO term-enrichment web service with all *A. thaliana* genes as a null model [79]. Genes with 5' UTR PREs are not distinctly enriched in any category.

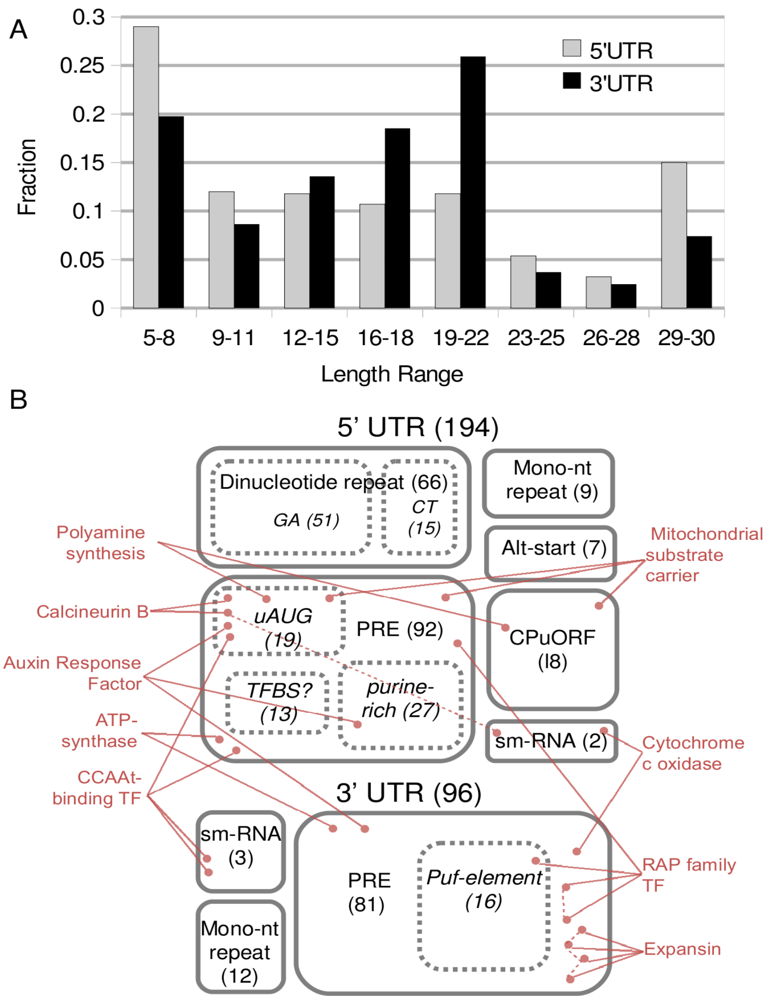


Figure 2.5: Though their length profiles differ with regard to region, PREs account for the bulk of conserved motifs within the both 3' and 5' UTRs.
A) Histograms, calculated independently, of PRE lengths in both UTRs. B) Treemap of conserved motif annotations. Box size roughly indicates the proportion of annotations that we assigned to conserved motifs. Exact values are given in parentheses next to the appropriate category. A box with dashed lines indicates that the category is a subset of the larger, solid-lined box. A red 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. Red labels describe the larger gene families or functional categories into which these specific orthologous groups can be grouped. A dashed line between orthologous groups indicates that the motifs are similiar. A dashed line added to a single orthologous group (such a one Calcineurin B group) represents multiple or overlapping annotations of that group's conserved motif. TF - transcription factor; TFBS - TF binding site.

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

Nearly all PUF-binding elements (see *Introduction*) known to metazoans, plants, and yeast contain a core TGTA sequence [46,48,49]. We assessed the degree to which this 4-mer word was enriched among PREs and if there were other 4-mers that were likewise enriched. We created random expectation distributions (n=10,000) for each 4-mer such that the length distribution of the PREs dataset was recapitulated (see *Methods*). We then assessed the *p*-value of the actual enrichment based on its position within this distribution. 3' UTR and 5' UTR regions were treated independently.

Table 2.6: 4-mer words significantly enriched in PREs from 5' and 3' UTRs.

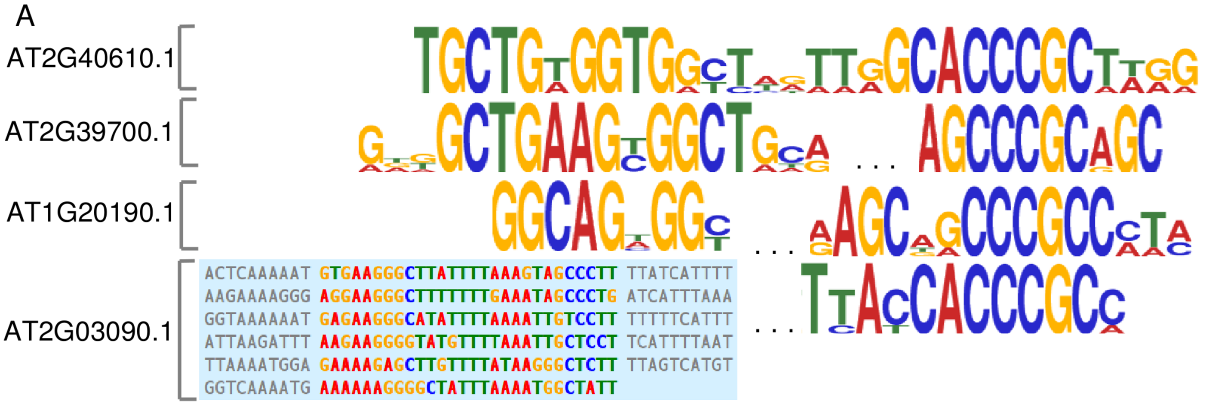
5' UTR		3' UTR	
4-mer	**p-value**	**4-mer**	**p-value**
AGAA	<1E-05	TGTA	<1E-05
AGAT	<1E-05	TTTG	<1E-05
TTCT	<1E-05	AAGG	0.0051
AGGG	0.0006	AATA	0.0066
ATGG	0.0015	TGGT	0.0104
AGGA	0.0019	AAGC	0.0246
TTTT	0.0151	GAGG	0.0296
TCTT	0.0154	TGCA	0.0377
AGAG	0.0224	TTCT	0.0447
CCTC	0.0274		
CGAT	0.0374		

TGTA was one of the most significantly enriched 4-mers in the 3' UTR and, notably, not in the 5' UTR (Table 2.6). Sixteen genes have TGTA-containing PRE in the 3' UTR. The mRNA from CLAVATA1, which is critical for meristem maintenance, binds to APUM2 (a PUF-domain-containing protein) with roughly half the affinity of a Nanos Responsive Element [49]. AT4G20270.1 is an in-paralog of CLAVATA1. Based on our results, its TGTA-element is conserved (Table 2.7). Our pipeline identified a TGTA-motif in the 3'UTR of the actual

CLAVATA1 in-paralog assayed previously (AT1G75820) [49], but it had a substantially higher E-value than AT4G20270. Interestingly, it did contain four TGTA tetramers (7% chance of four or more based on 3'UTR mononucleotide composition and length). The remaining TGTA-containing PREs appear to be dispersed among functionally unrelated genes, except for two subunits of the photosystem-I light-harvesting complex, AT3G47470.1 and AT3G61470.1. It is noteworthy that their mRNAs are known to be enriched around the chloroplast [80].

The TTTG 4-mer has an equivalent p-value to TGTA. These two 4-mers are occasionally found together, but not in a consistent arrangement relative to one another (Table 2.7). Moreover, their co-occurrence does not deviate from what would be expected given their individual distributions among all PREs (p-value = 0.30, Chi-squared test; observed: TGTA/TTTG, 6; TGTA/-, 10; -/TTTG, 9; -/-, 57), suggesting that TTTG is generally unrelated to PUF-binding.

Expansins are one of the few groups of mRNAs that have been shown to be localized to specific subcellular sites in plants [81]. An [AG]CCCGC-containing motif was found in four Expansin 3' UTRs (Figure 2.6A). A region upstream of the [AG]CCCGC motif is also conserved but more specific to each Expansin. As the alignment of AT2G03090.1's variable region indicates, this conservation may be a result of secondary structure or a gap-tolerant binding partner. The *Zinnia elegans* Exp1 (gi|7025490|gb|AF230331.1) mRNA, which was shown to localize to a particular region of the cell periphery, also contains this pair of elements although they too appear to be position-independent relative to the mORF stop site and relative to one another (Figure 2.6B). Though we identified 12 Expansin orthologous groups with substantial sequence coverage, only four contained conserved motifs. None of these motifs are the result of mis-annotation or overlap with regions of extensive conservation as indicated by whole transcript alignments (not shown).



```
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 2.6: 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. Left motifs for each group represent the 3' variable region. Right motifs contain the RCCCGC-core and are found downstream. Motifs are so proximal in the AT2G40610.1 group that they are identified together. The the entire MEME-derived alignment is given for AT2G03090.1 group's variable region. B) Alignment of the localizing Zinnia elegans ExpansinA1 mRNA (Ze) - gi|7025490|gb|AF230331.1 - with its A. thaliana ortholog (At) - AT2G40610.1. mORF stop codons and conserved elements are colored red and blue, respectively. Asterisks indicate that the column letters are identical.

Table 2.7: Select PREs within the 3' UTR.

Hypothetical function of PRE	***A. thaliana* accession**	**Gene annotation**[a]	**PRE sequence**[b]
PUF-element	AT3G09980.1	expressed protein	TATAAACAGG**TTTGTA**ACTAA
	AT4G01100.1	adenine nucleotide transporter 1	TGCTATT**TTTGTA**GGC**AAGG**G
	AT5G16000.1	leucine-rich repeat family protein	TGCT**TGTA**TTCATC**TGTA**AA
	AT3G47470.1	chlorophyll A-B binding protein 4 (LHCa4)	CTTTAA**TGTA**CAGAGGAACT
	AT4G20270.1	leucine-rich repeat transmembrane protein kinase, (CLAVATA1-like)	**TGTA**CAGTAGGAT**TGGT**GGG
	AT3G57200.1	hypothetical protein	ATTACCCAAGCGC**TGGTGTA**
	AT4G14900.1	hydroxyproline-rich glycoprotein family protein	G**TTTGTA**ATCACTAACCGTT
	AT2G40110.1	yippee family protein	AAA**TGTA**CATTCTTTAACC
	AT1G07470.1	transcription factor IIA large subunit, putative	TTGGCCTGT**TGTA**CATA
	AT1G53910.3	AP2 domain-containing protein RAP2.12 (RAP2.12)	**TGTAAATA**AAGCTACAT
	AT3G11660.1	harpin-induced family protein	TGAAT**TGTA**CAT**TTTG**C
	AT3G18820.1	Ras-related GtP-binding protein, putative	T**TGTA**CATTAGTG**TTTG**
	AT3G61470.1	chlorophyll A-B binding protein (LHCa2)	**TGTA**CA**AATA**CC**TTTG**T
	AT2G42670.2	expressed protein	**TGTA**CATATT**AATA**TA
	AT1G32400.1	senescence-associated family protein	GAG**TTTGTGTA**
	AT1G08420.1	kelch repeat-containing protein	**TGTA**T
unknown	AT2G07687.1	cytochrome c oxidase subunit 3	ATGAAAGCTCGAAGACAAAGAGAACCGGG
	AT1G76160.1	multi-copper oxidase type I family protein	GACCTCAACTCGAGGTCTCATTCTTT
	AT2G07733.1	similar to NADH dehydrogenase subunit 2	CGGCGGCAGCGGCGCGAGGAGTTAACGAC
	AT1G78080.1	AP2 domain-containing transcription factor RAP2.4	TGCAATGGAGTT**TTTG**GCAATTGCA
	AT4G39780.1	AP2 domain-containing transcription factor, putative	TCTGCAATGAAATT**TTTG**ACA**TTTG**
	AT4G00730.1	anthocyaninless2 (ANL2)	GAGTCAAGAACGAACCGCGCGTG
	AT1G72230.1	plastocyanin-like domain-containing protein - ***miR408***	GGCCAGGATAGAGGCAGTGC
	AT3G54180.1	cell division control protein 2 homolog B (CDC2B)	TGTCATCATCT**TGGT**GATTT
	AT4G26210.1	mitochondrial AtP synthase g subunit family protein	AAGCTGAG
	AT1G33470.1	RNA recognition motif (RRM)-containing protein	ACCA**TGGT**
	AT5G20160.1	ribosomal protein L7Ae	GGGCCTC
	AT1G59750.2	auxin-responsive factor (ARF1)	ACATG

[a] If the *A. thaliana* representative of the motif is a possible microRNA binding site as predicted by psRNATarget (see *Methods*), the microRNA family is given in bold italics next to the gene annotation. [b] 4-mers with a p-value of greater than or equal to 0.01 are in large, bold font.

5'UTR PREs are enriched in purine-rich 4-mers, and uATGs are conserved in a non-CPuORF capacity

Based on the AGRIS database of *A. thaliana* promoter motifs [82], only 13 out of 92 5' UTR PREs have evidence for being transcription factor binding sites (File 1, 5UTR_motifs.pdf and Figure 2.5B). Additionally, no 4-mer and its reverse compliment had significant equivalent enrichment among PREs (Table 2.6), suggesting that PREs are either orientation specific transcriptional elements or that they act at the mRNA level.

The most significant 4-mers within the 5' UTR region appear to represent purine-rich motifs (Table 2.6). Purine-rich repeats in the 5' UTR have been reported to enhance translation of the *ntp303* gene in *N. tabacum* [83]. While this gene has a 3-species orthologous group in our analysis, we do not observe conservation of the respective motif in *A. thaliana* or *V. vinifera*. Yet, GAA was found in one of the most significantly enriched tetramers among 5'UTR PREs (AGAA in Table 2.6). To a lesser extent pyrimidine-rich elements are also frequently enriched (TTCT, TTTT, TCTT). The pyrimidine-rich Y-Patch is a common promoter element in plants. It peaks in frequency at the -13 nt position relative to the transcriptional start site but often extends deep into the 5' UTR [84]. While this may be the cause of CT-rich element conservation, these elements are thought to be orientation specific, and so this does not explain purine enrichment. If CT-rich sequences are acting as core promoter elements, then they will be positionally biased toward the transcriptional start site. We tested positional bias of CT-rich and AG-rich PREs - 80% pyrimidine or purine content, respectively - using all other PREs as a null model. Though CT-rich tetramers are enriched among the PREs, entire elements that have 80% CT content were actually quite rare (8%). When present, their median position relative to transcription start site (TSS) was 36 nts (12-48nts - 95% confidence intervals based on resampling with replacement; n = 7). AG-elements were positioned farther from the TSS, 45 nts (39-89nts; n = 27), but still closer than the null model 109 nts (89-129nts; n = 58). In summary, though some of these CT-rich 4-mers may be residual Y-Patch signals, most are found in motifs with a balanced nucleotide composition and without a 5' positional bias. Alternatively, 29% of all 5'UTR PREs are AG-rich. As with the repetitive elements, this suggests that AG-rich elements are acting at the post-transcriptional level.

Interestingly, the 4-mer associated with a strong-context start codon, ATGG, is significantly enriched among PREs in the 5' UTR. Since sequences with conserved protein coding potential were already removed from this portion of the analysis, these 4-mers, if they do function as start codons, are conserved for peptide-independent effects relating to uATG initiation/reinitiation. This finding corroborates with low substitution rates of uATG triplets in mammals and fungi [85,86]. In fact, many ATGs appear within conserved motifs, in both weak and strong contexts (Table 2.8).   In most of the ATG-containing motifs conservation extends beyond the -3 and +4 positions. For example, AT4G26570.1 and AT5G06510.1 show extensive downstream conservation. While too short to be identified as 'protein-coding', these may be functionally constrained at the peptide-level. Alternatively, both of these as well as AT5G47100.1, show downstream conserved ATGs without a frame bias, suggesting that premature initation as opposed to protein-coding may be the more critical function of these uATGs. Respectively, some mRNAs of CCAAT-binding proteins are regulated in a peptide-independent manner by uORFs in metazoans [87]   Likewise, though not an ortholog, AT5G06510.1 has a lengthy, conserved motif that harbors multiple ATGs (Table 2.8).

Conspicuously absent from the 5' and 3' enrichment lists is the AGGT tetramer representing canonical splice sites.   For comparison, UTR exon boundaries are recognizably conserved in *Cryptococcus* [88]. Because we do detect 4-5 nucleotide pockets of high conservation, particularly in the 5' UTR where coverage is more extensive, conservation of such sites appears to be weak across the more distantly related lineages in this study.

Table 2.8: uATG-containing PREs within the 5' UTR.

***A. thaliana* accession**[a]	**Gene annotation**	**PRE 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	protease-associated zinc finger (C3HC4-type RING finger)	**AUG**GAAGCUG**AUG**UUUCCAU
AT1G19330.1	expressed protein	UCAGCA**AUG**C**AUG**AUCUUCA
AT5G62000.2[c]	transcriptional factor B3 family protein (Auxin Response Factor 1)	CAG**AUG**AGAGAUCUGAGC
AT3G54020.1	phosphatidic acid phosphatase-related	UGAAGUAAU**AUG**GAAGUG
AT3G63200.1[c]	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) family 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 italics 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.

Discussion

Can putative transcript coverage reconstitute the eudicot proteome?

We used putative transcripts assembled from all relevant EST and cDNA entries in GenBank and made available through PlantGDB. Of the ~30,000 *A. thaliana* genes, at least 85.1% have one or more orthologs among plant species separated by more than 70 million years (Table 2.1, Column 5). Like *A. thaliana*, *Ricinus communis* has a sequenced genome, and it is represented in 10,381 groups containing an *A. thaliana* gene when clustered using OrthoMCL [59]. *G. max* is as distantly related to *A. thaliana* as *R. communis* but is only present in 8,216 such groups in our analysis. This suggests that transcript coverage is still incomplete for *G. max*, in spite of its many putative transcripts (Table 2.1, Column 2).

Do patterns of element enrichment in the 5' versus 3' UTR corroborate with canonical models of eukaryotic translation?

The 5' and 3' UTRs appear to mediate distinct forms of post-transcriptional regulation. This is expected given the distinct molecular events occurring within each region during translation [7]. Small ribosomal scanning 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. Based on this work, 5' UTRs across dicots retain uORF-related and purine-rich sequences, whereas the 3' UTR is more likely to retain probable PUF-binding sites and other elements with experimental precedent for mediating mRNA targeting (Figure 2.5B). Though our sample size is small, the exception to this trend is smRNA binding, which has little conservation bias for either region of the mRNA. As discussed above, 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 target site - 3' UTR vs. 5' UTR vs. CDS - does not correlate with the degree to which an affected mRNA is degraded by Dicer1 and/or translationally silenced [89], so perhaps a lack of regional bias is to be expected.

Are CPuORF-containing mRNAs polycistronic?

The extensive length and conservation of the newly identified CPuORFs 26n and 24n (Figure 2.3B-D) begs the question of whether CPuORFs are in fact regulatory elements or, alternatively, whether their associated mRNAs should be considered multi-cistronic transcripts. The identification of CPuORF-24n, linked to adenine cyclase, implicates CPuORF-mediated regulation in the cyclic-AMP pathway, although the physiological role of cyclic-AMPs in plants is still debated. Strikingly, CPuORF-24n is identical to GenBank accession, AAN10198, an *A.*

thaliana gene annotated by the depositor as CDC26 - a small subunit of the anaphase promoting complex (APC). The peptide aligns to all eukaryotic CDC26s [90], but the conservation is found mainly in the N-terminal region, where it is also most conserved in plants (Figure 2.3D). CPuORF-24n appears as a distinct transcript in much more distantly related eukaryotes. Moreover, though the conservation between CP-uORF-24n and the metazoan CDC26 is quite weak, this CPuORF appears to be a legitimate component of the APC [91]. Thus, as it stands, the associated mRNA is a good candidate for a multi-cistronic eukaryotic transcript. Yet, the conserved synteny, proximity, and co-transcription of CPuORF-24n and the putative adenylate cyclase suggests that, in plants, these two proteins require co-translation.

Of the remaining three loci identified in this study, little is known. CPuORF-27n is linked to a protein kinase involved in salt stress tolerance. The At1g67480 loci associated with CPuORF-26n codes for two major isoforms, the sole difference being an alternative 5' UTR, but this difference has no bearing on uORF-26n presence, its peptide sequence, or the introduction of other uORFs in the 5' UTR. CPuORF-25n is associated with a dicarboxylic acid transporter in mitochondia. The associated mORF of CPuORF-25n has a paralog that has the same biochemical activity and basal expression pattern, but lacks CPuORF-25n. Interestingly, the mRNA expression levels of these two paralogs diverge only under various stress conditions [92].

At least two known plant CPuORFs - those linked to bZip transcription factors [27] and to adenosylmethionine decarboxylase [29] - are known to reduce translation of the linked mORF in response to small-molecules. Thus, as opposed to being merely co-translated peptides, these CPuORFs appear to be actively regulating their downstream protein. As seen in bacteria, there are many short evolutionary paths to small-molecule-induced translational repression via the 5' end of an elongating protein [93], a situation analogous to CPuORF-mediated repression. Interestingly, as indicated by our work and prior research, few CPuORFs are genetically linked outside of a specific protein family (Table 2.4). For example, the sucrose responsive CPuORF-1 [27] is found only in mRNAs that code for bZip transcription factors, not for sucrose metabolizing enzymes. As formulated previously [25], this suggests that initial signal transduction is mediated by translation repression and the response cascade is transcriptional.

Why would non-ATG start sites be conserved?

We have shown that translation initiation is likely to occur at certain non-ATG start sites, and, because these sites are deeply conserved, that the non-ATG nature of these sites appear to be important for plant fitness. This non-ATG initiation requires a strong Kozak context. A comparable conservation profile has been seen for the mammalian eIF4g2 gene [94]. Interestingly, translation can initiate at various non-ATG sites, such as, among others, ACG for AGAMOUS in *A. thaliana* and ATT for the AZI1 antizyme in mammals [95]. Although we were unable to identify AGAMOUS orthologs suitable for comparison in this study, we do not observe other ACG sites. Thus, non-ATG sites that are conserved only appear to vary in the +1 position. An alternative explanation is simply that ATT/ACG start sites may be less dependent on context and the downstream coding region is under much less sequence constraint.

Strikingly, we only find evidence for mutation back to a canonical ATG in one case - the *G. max* ortholog of aldose 1-epimerase (Figure 2.4B). Recent work identified a mutation of the non-canonical CTG-start of floral regulator FCA back to ATG in two out of six lineages, but, as with our data, this is rare for the other non-canonical CTG-starts identified [76]. One hypothesis for a conserved non-ATG start is that it is used to generate an N-terminal extension, which then targets the protein to a different subcellular location. At least for CTGs, this does not appear to be the case [76]. TargetP [96] does not consistently predict the N-terminal additions in this analysis to re-target the protein. If the AZI1 gene discussed above is any guide, it may be that non-ATG start sites serve as inducible translation initiation sites, producing N-terminal extensions that are themselves responsive to the same inducing signal - in the case of AZI1, polyamine concentration [95]. Thus, a non-ATG start site followed by a conserved N-terminal addition may result a very sensitive form of inducible translation repression.

Does PRE identification reveal trends concerning RNA regulons in plants?

It has been previously reported that RNA-binding proteins, among them mulitple PUF-family proteins, associate with functionally related sets of genes [35]. Indeed, in our analysis, we find that LHCa2 and LHCa4, closely related genes that are likely co-regulated, both contain a putative PUF-binding PREs. However, TGTA-containing PREs generally do not appear to be significantly enriched in any Gene Ontology (GO) category based on all other *A. thaliana* genes as a background model and using GO term-enrichment web service (data not shown) [79]. It is likely that, even in the case of PUF-binding elements, specificity is achieved or enhanced through a combination of primary and secondary mRNA structures as indicated by atomic models of protein-mRNA interaction [34]. Additionally, the cellular /developmental context of PUMILIO has been shown to influence the cohort of mRNAs to which it binds [48]. PUF-elements aside, between our GO analysis and manual curation, we find very few instances of conserved co-regulation of functionally related genes by the same motif, and essentially none to match that seen for the vacuolar ATPase in *Drosophila* (see *Introduction*) [48]. Undoubtedly, we have false negatives resulting from a lack of sequence coverage in the 3' UTR, but a yeast three-hybrid screen against the Arabidopsis PUF-family protein, APUM-2, revealed interaction with the 3'UTRs from only five genes, which appear to be unrelated. Of those we found a conserved motif in the chlorophyll binding protein, as discussed above. Also, a DNAJ protein identified in the screen has a motif with a low E-value ($6.7x10^{-5}$) that did not pass our conservative statistical cutoff. Of the remaining three, all have orthologs in least four species with 3' UTR coverage, but the element is not conserved. In summary, either co-regulation at the mRNA level is rarely conserved at the levels of divergence used in our analysis, or it is uncommon in dicots. Better ways to assess such small sequence elements within narrow phylogenetic scopes and the requisite sequence data to do so will help to differentiate these possibilities.

One extension of the RNA regulon model is that, analagous to transcriptional modules, some mRNAs will be under combinatorial control by a suite of factors. Only 7% (n=6) of 5'

UTRs harbor legitimate examples of multiple-conserved motifs (File 3, PRE_categories.xls). In contrast, 17% (n=14) of PRE-containing 3'UTRs have significant conservation outside of the most conserved PRE (File 3, PRE_categories.xls). Such a difference has a *p*-value of $<10^{-2}$ of being a result of sampling error if, under the null-hypothesis, both regions have an equivalent number of multi-site-containing groups, calculated from their combined frequency of 11.4%. Although the difference is not extreme, it does support a model in which the 3' UTR, because it is free of ribosomes, is a more appropriate platform for combinatorial regulation, or signal integration, than the 5' UTR.

It is still unclear whether or not particular types of post-transcriptional regulation are specific to a protein's function. We've found that many mRNAs, either from single genes or genes that share a function or domain, appear to be under multiple forms of post-transcriptional regulation (Figure 2.5B).

Two CCAAt-binding transcription factors have conserved microRNA target sites (Table 2.3). Additionally, two other family members have PREs located in the 5' UTR (Table 2.8 and File 1, 5UTR_motif.pdf). Neither of these PREs bears any resemblance to the microRNA target site nor are they detected as other microRNA target sites. As discussed above, one PRE, linked to AT5G06510.1, may be the result of the uATG initiation, while the other mRNA, AT3G53340.1, appears to have an extensive but uncharacterized PRE.

Two components of cytochrome c oxidase have seemingly different modes of post-transcriptional control; AT3G15640.1 has a conserve microRNA binding site (Table 2.3) in the 5' UTR while AT2G07687.1 (Table 2.7 has an extensive but clearly unrelated PRE in the 3' UTR. Such elements could be the basis for differential regulation.

Calcineurin B-like proteins are calcium responsive kinase regulators. AT4G26570.1 and AT5G47100.1 represent the two major branches of this protein family and both harbor a multiple-ATG-containing element within their 5' UTR, suggesting that premature initiation maybe a conserved mechanism of translation repression for these genes. Additionally, based on complementarity to miR847, AT5G47100.1 may overlap a conserved smRNA-binding site.

The RAP family members contain an AP2-transcription-factor domain, and RAP2.4 (AT178080.1) is thought to serve as a mutual control point for ethylene and light responses [97]. AP2 is known to be under smRNA regulation (mORF target-site), but, in spite of their length, the RAP PREs do not appear to be result of such regulation. Notably RAP2.4 is the only gene we have identified that possesses a conserved element in both its 3' and 5' UTRs. A paralog (AT4G00730.1) has a similar 3' UTR PRE but lacks the 5' UTR PRE. The RAP2.12 (AT1G53910.3) PRE matches the canonical PUF-binding site, which is missing from an alternative isoform that varies only in its 3' UTR.

As discussed elsewhere in this paper, polyamine synthesis in known to be a 'hotspot' for translational control. In all eukaryotes, a CPuORF is thought to reduce AdoMetDC (Table 2.4) levels through polyamine-induced ribosome stalling, although through somewhat distinct peptide sequences [95] . In another polyamine pathway, mammalian spermidine synthase is alternatively spliced, in response to polyamine depletion, to include a premature stop site which triggers the NMD pathway [98]. Also, mammalian *spermine* synthase also contains an uORF that overlaps the mORF. We have identified a spermidine synthase homolog (AT5G53120.5) with an extensive PRE (File 1, 5UTR_motif.pdf and Figure 2.5B). This gene codes for five different alternative transcripts, which vary solely in their 5' UTR. However, the PRE is present in all isoforms, while the variation lies upstream. Interestingly, the PRE contains a possible non-canonical start site, Aga**CTG**G. As discussed above, it is the upstream region of the FCA mRNA that controls non-canonical initiation of that protein [76]. Also, alternative start sites are employed in other mammalian genes associated with spermidine synthesis. Thus, a reasonable untested hypothesis is that spermidine synthase expression in plants is controlled in part by inducible premature initiation.

Taken together, it appears that particular protein families are disposed to regulation by *cis*-elements within their mRNAs but that the mechanism of regulation is surprisingly variable. Moreover, unless our knowledge of smRNA targets is dramatically incomplete, many of these mRNAs fall outside the jurisdiction of smRNA surveillance, suggesting that, in plants, transcript-specific protein interactions and premature translation initiation may be more important than smRNA-mediated-repression in explaining post-transcriptional variation in gene expression [2].

Methods

Sequence acquisition and preparation

See Figure 2.1 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 [99]. None of sequence sets have undergone significant addition since that date. The longest ORF from each putative transcript was extracted using a custom Perl module (uORF.pm available at http://web.utk.edu/~jvaughn7/code/bio/) based on the criteria that an ATG be followed in-frame by a TAA, TGA, or TAG 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 <1E-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) [100], with default parameters. 5' or 3' UTRs shorter than 8 nucleotides 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) [55] to search for overrepresented sequences across the UTRs associated with orthologous groups of coding sequences [101]. 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 4 *A. thaliana,* 3 *G. max*, and 6 *V. vinifera* sequences, there would be 4*3*6 or 72 subgroup combinations; only 30 of these would be randomly selected for further evaluation. 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 between 6 and 30 nucleotides long. A 2nd-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 [102]. To correct for multiple tests, we divided an E-value cut-off 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 greater than $1x10^{-9}$ of belonging to the motif by chance. Also, motifs were excluded from 5' UTR comparisons unless the mORF proteins for more than 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 datasets, which were generated by randomizing all orthologous groups, such that species composition per orthologous group was maintained: all sequences were put into a pool 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 datasets were used to determine our false discovery rate. Though not reported in detail, we tested numerous alternative algorithms. Generally word-based approaches such as Weeder [103] and FootPrinter [104] were less sensitive than MEME; assumably 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 [105] and PhyloCon [106], had equivalent or lower sensitivity than MEME, which is expected given that the lineages under analysis are highly diverged [107].

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. These motifs were removed from downstream processing.

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) [32] and data from [71].

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 non-synonymous mutations using PAML (version 3.14) [108] according to the protocol in [109] and a tree topology based on Figure 2.2D [110]. In brief, a likelihood ratio test was evaluated for a model in which the non-synonymous to synonymous substitution ratio was allowed to vary and another model in which it was fixed at 1. Motifs with resulting p-values, from the Chi-squared distribution, of <0.01 were considered protein-coding motifs. Additionally, the motif was removed if the *A. thaliana* peptide associated with the coding potential had a *blastp* match (e-value cut-off of 10^{-5}) to any Viridaeplantae 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 Figure 2.3), 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 nucleotides long, that were not known smRNA targets, did not show coding potential, and did not contain mono- or di- nucleotide repeats were annotated as PREs.

All possible 4-mer word frequencies were assessed for combined sets of PREs - 5' and 3' processed separately. 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 4-mer overlap - TGGA overlapping GGAA, for example. Null distributions for each *4*-mer were then simulated by randomizing the dataset 10,000 times with per-element length intact. Because our initial inference of conservation accounts for higher-order statistical properties of the UTR, each letter in these randomizations was considered equally probable. The same 4-mer frequency assessment was perform for each randomized dataset. Actual frequencies were placed within the simulated distributions and the number of distribution values for the 4-mer in question 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 [82] 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* region associated with a PRE, as with smRNA search above, were also checked again for imperfect matches to mature smRNAs using psRNATarget web server [111] with default parameters. Significant matching smRNA families are reported in Table 2.6 and 2.7.

Addendum: Comparative transcriptomics in plants with orthoGroup.pm and Anchored-MEME

The sequence similarity between two proteins suggests their structural similarity (although the reverse is not always true). In turn, structural similarity suggests that the two proteins have similar biochemistry, be it enzymatic activitiy, protein-protein interaction, or another molecular function. Beyond mere similarity, the determination that two proteins are orthologous allows for a more powerful inference of similar function. If two proteins are orthologous then they were the same sequence in the last shared ancestor. It is inferred that these two proteins will continue to serve the same role in the extant lineages that they were serving in the ancestor [112].

As opposed to two lineages in a pairwise comparison, many lineages can be used to assemble orthologous group. Like the pairwise approach, the interpretation remains the same: all of the genes in the ortholgous groups were the same gene in the last shared ancestor of all the species considered. Because of local and/or global duplication events within a particular lineage being compared, orthologous groups may contain more than one gene from a single species; such a sequence is called 'in-paralogous' relative to the other genes in respective species and 'co-orthologous' or simply 'orthologous' to genes in the other species being compared.

Manipulating orthologous group data with orthoGroup.pm

The following set of Perl libraries uses BioPerl and original code to bundle many common tasks one might desire to perform on/with an orthologous group into a uniform object interface. As with all associated code in this dissertaion, these Perl classes can be downloaded from http://web.utk.edu/~jvaughn7/code. The `orthoGroup` object is acquired from an input/output object, `orthoGroupIO`, which takes many input formats - an OrthoMCL file [100], a directory of FASTA files, a directory of alignments, or a BLAST report - and links them with a raw sequence database (Bio::DB::Fasta). An `orthoGroupIO` can then be looped through, producing a series of `orthoGroup` objects. `orthoGroup` is designed such that certain filtering procedures can be done on the object and only those sequences passing the filtering procedure get acted on during future method calls (Table 2.9). This can dramatically clarify the scripts that uses these objects.

Table 2.9: Methods supplied by the `orthoGroup` class

`Method_Name` **Function**	**Argument**	**Return**
`getAccNamesBySpecies` see Argument/Return	String: Species name	Ref to list: All accessions in `orthoGroup` belonging to species
`getAccNamesByPattern` see Argument/Return	String: Regular expression pattern	Ref to list: All accesssion in `orthoGroup` matching regex
`seqObjArray` see Argument/Return	none	Ref to list: All Bio::Seq objects
`filterBasedOnLength` see Argument/Return	Scalar: Minimum raw sequence length required	1 or 0, filters internal state
`trimRawSeqsTo` Trims the raw sequence to the length argument. If no start position is given, starts from the begining. If 'fromEnd' is given only the last nucleotides specified by (1) are retained.	(1) Scalar: length <2> Scalar: start position - can be integer or 'fromEnd'	1 or 0, filters internal state
`removeAcc` see Argument/Return	String: Name of the accession to remove	1 or 0, filters internal state
`next` for cycling through each sequence in the group; come also use seqObjArray	none	Bio::Seq object
`makeFasta` Makes a fasta file from the `orthoGroup`; if [2] is set to '1' then a 'temp.fasta' file is created (or overwritten)	(1) String: location, where to put the file <2> or 0, make a temporary file	1 or 0
`geneCount` see Argument/Return	none	Integer: number of genes in `orthoGroup`
`taxaCount` see Argument/Return	none	Integer: number of taxa in `orthoGroup`
`printOrthoMCLForm` see Argument/Return	none	String: string representation of `orthoGroup` in OrthoMCL format
`abbreviateNames` see Argument/Return	String: regex pattern to remove from matching accessions	1 or 0, filters internal state
`doesItContain` does the list of current accessions match a given regular expression pattern	String: regex pattern	1 or 0: match or not
`speciesWiseSubsets` Reduces `orthoGroup` to subsets of `orthoGroups` such that each only contains one sequence from every species in the original `orthoGroup` and no two subsets have an identical sequence composition.	Hash: <'maximumSubsets' => Integer, 'useRandIfTooBig' => 0 or 1>	`orthoGroup`::`orthoGroupIO` object; 0 if subset number > 'maximumSubsets'
`filterSpecies` Only accessions belonging to the species in the existence hash argument are kept	Ref to Hash: {'species1' => 1, 'species2 => 1}	1 or 0, filters internal state
`filterInclusiveSeqs` Raw sequences smaller than and identical to a larger sequence are removed	none	1 or 0, filters internal state

Extending MEME with anchored-MEME preprocessor

In this chapter, we confronted a major hurdle with regard to comparative sequence analysis in plants - rampant paralogy. This feature of plant genomes is a result of numerous genome duplication events. To avoid many of the complications associated with paralogy in Section 1, we adopted a brute-force approach, looking at all possible subgroups in an orthologous group (sub-group method). A more sophisticated approach would incorporate all data from an orthologous group concurrently, but would be robust to the possibility of element loss in an in-paralog. We have developed software to address this need.

Phylogenetic footprinting is a technique for identifying short regulatory regions in the genome. The technique requires that a region be under a slower rate of substitution than its surrounding genomic context [113]. As originally proposed, phylogentic footprinting requires an estimate of nucleotide substitution rates. This estimate, because it is based on a common register of sites, requires neutrally evolving regions to be aligned. These alignments are often the source of errors and discrepancies in the field of *cis*-regulatory element (*cis*-element) identification [114]. Additionally, both experimental and theoretical data is accumulating to suggest that *cis*-elements are commonly turned over. In brief, temporary redundancy resulting from *de novo cis*-element creation allows the 'native' *cis*-element to be replaced without a lapse in carrier fitness [53,115]. Thus, ideally, *cis*-element identification would not require alignments. Alignment-free techniques have existed for more than a decade and have been used to find motifs among co-regulated genes and homologous proteins . They have also been exploited for *cis*-element identification in orthologous promoters. Furthermore, researchers have incorporated phylogenetic information into alignment-free approaches, although such techniques begin to strain the definition of phylogenetic footprinting, in that they are taking an indirect measure of the hypothetical element substitution rate relative to its surroundings. As an example, FootPrinter [104], uses a tree to calculate substrings of a particular length with the least number of substitutions, given the underlying phylogenetic tree. Though it has many insightful features and optimizations, the algorithm and other comparable methods, has significant drawbacks: 1) The tree used to relate the sequences must be manually generated. Optionally, it can be estimated from the sequences under study, but, in order to avoid dampening the signal of a long stretch of conservation, the tree should come from neighboring genomic space not involved in the search. As proposed by the authors, the closest protein-coding region would be suitable to the task. 2) By using the phylogenetic tree as a guide, the algorithm assumes that *de novo* binding sites, which are not related by decent, do not occur with a substantial frequency. Again, experimental and theoretical work indicates otherwise. 3) Lineage-specific motif loss is penalized in a fairly *ad hoc* manner. Though this may be innocuous, or even advantageous, for single-gene investigations, it becomes problematic for large-scale analysis and, to an even greater extent, for handling in-paralogous sequences. 4) The algorithm does not produce an absolute statistical criteria and, so, requires comparison of a motif's score with a simulated null distribution of scores. This distribution is derived from the computationally-intensive simulation of random sequences based on neutral distances between species on the tree.

We favor a simpler approach in which sequence data is divided into protein-coding space (the 'anchor' in 'Anchored-MEME') and proximal regulatory space. The coding space is used to assess the substitution rate between sequences and establish a correction factor by which to weigh regulatory sequences. Underlying our approach, we are assuming that, as long as it is not a result of random expectation or sequence relatedness, sequence similarity and genomic context are sufficient to infer the functional significance of a motif. In the end, lineage bias in regulatory element composition and evolution is of utmost interest, but it should be an *a posteri* consideration. We envision a framework wherein results from Anchored-MEME are then fed back into a pipeline as implemented in [53].

Design and Implementation

The user is asked to supply a codon alignment of orthologous CDSs along with an unaligned set of hypothesized regulatory regions - promoters, UTRs, exons, or combinations of each.

If a neutral site has undergone a substitution, by definition, we assumed the substitution to have been a stochastic event. If all sites have undergone substitutions, then the two sequences are effectively randomized. Thus, to quantitate the degree of randomization, we estimated the substitutions per site (*K*) by aligning translated coding sequences (CDSs) from each orthologous group and then back-translating each sequence relative to the alignment, such that the codon frame was maintained. In pairwise fashion, each sequence within a codon alignment was compared with every other sequence, such that the percent identity across all fourfold degenerate sites shared by the two sequences was evaluated. Any comparison with fewer than 10 available sites was not processed further; the rationale for selecting this value is described below. This percent identity was then used to estimate the number of fourfold degenerate site substitutions (K_{4d}) between the two sequences. In turn, a binomial distribution was used to estimate the fraction of sites that have not undergone a substitution. In calculating the chance that no substitutions have occurred, the binomial function reduces to:

$$Pr(K_{4\mathrm{d}})=(1-\frac{1}{s})^{sK_{4\mathrm{d}}}$$

where,

s = total sites compared, K_{4d} = the substitutions per site between the two sequences
As the number of sites, s, becomes greater than 10, this function further reduces to:

$$Pr(K_{4\mathrm{d}})=(\frac{1}{e})^{K_{4\mathrm{d}}}$$

For the MEME analysis described below, each sequence was weighed such that sequences that have undergone effectively randomizing divergence should weigh ~1 relative to one another, whereas identical sequences should weigh 0.5 relative to one another. Three identical sequences should each be weighed 0.3334, etc. Hence, we used the following equation to weigh each sequence with regard to its possible contribution to a motif score:

$$\frac{1}{1+\sum_{i=1}^{n}(Pr(K_{4d}^{i}))^{m}}$$

where,

i = a pairwise comparison, n = total pairwise comparisons for that species, m = minimum motif size

Results and Extensions

As a proof-of-principle, we ran Anchored-MEME on the 5' UTR dicot dataset used above. In 9,647 orthologous groups, we identified substantially more elements using the same correction for multiple-tests used for the sub-groups method (Figure 2.7). Because there are fewer tests, the threshold is higher. Interestingly, there are groups that were not identified using anchored-MEME approach but were found only under the sub-group approach. These may reflect groups where element loss in a paralog dilutes the signal, suggesting extensive subfunctionalization [62]. Alternatively, these may reflect the inclusion of coding sequences resulting from sequencing errors, an artifact to which the sub-group method is more susceptible.

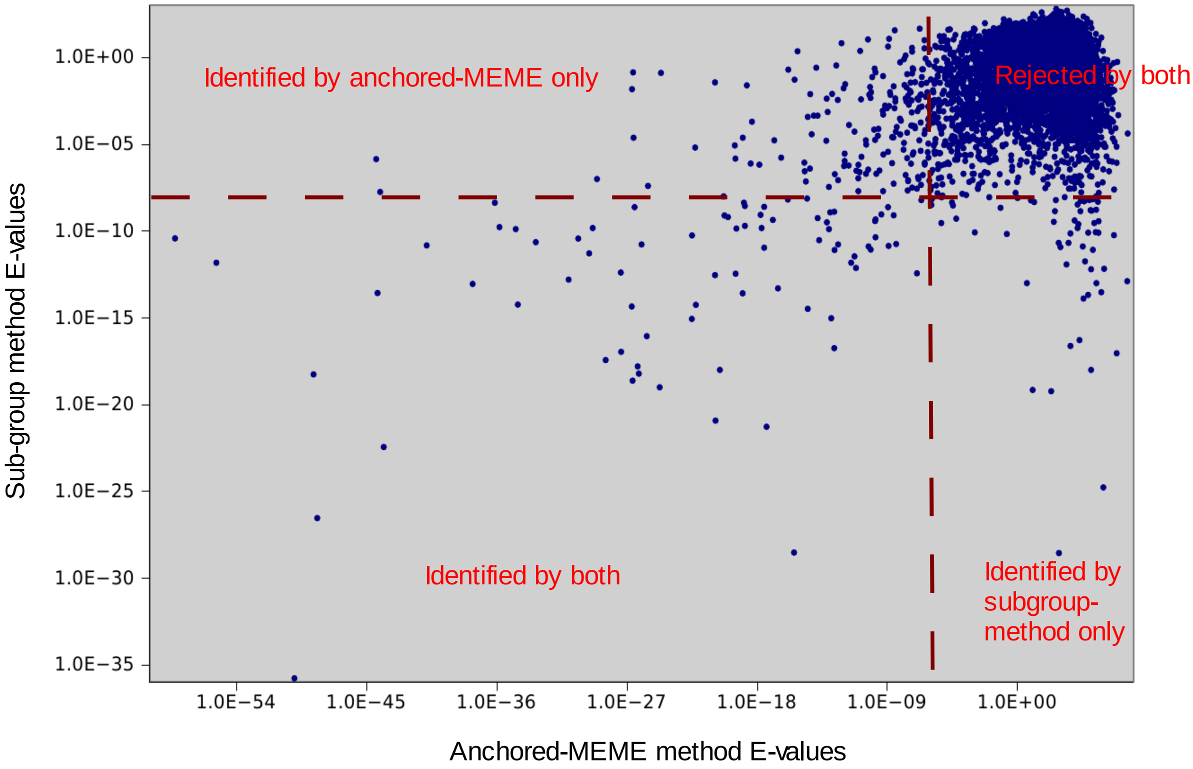


Figure 2.7: E-values of Anchored-MEME method versus E-values of sub-group method. Each point represents lowest E-value of a motif identified for an individual orthologous group.

Critically absent from our algorithm is a proper accounting of the tree topology relating the sequences within an orthologous group. This level of detail has little impact when a sequence is highly diverged ($K = 1$) from all other sequences or when it is very similiar to another sequence ($K = 0.01$), but becomes importent in terms of middle distances ($K = 0.1$). Unfortunately, many in-paralogous sequences are located at this distance; thus, the number of elements identified in Figure 2.7 is likely to be inflated. In short, the algorithm overestimates the weight a sequence should have because we do not consider correlated mutations among the sequences to which it is compared. Conceptually, this failing should be easily corrected: a tree topology can be generated based on the mORF alignment and, for each pairwise relationship, evolutionary distances between nodes would then be divided by the number of leaves associated with that node (see Figure 2.8). The algorithm would then proceed as described above.

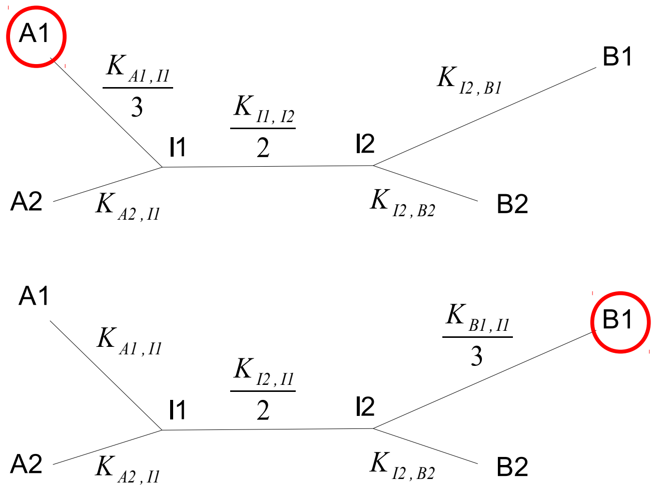


Figure 2.8: Correcting for tree topology in sequence weighing.
Red circles indicate the sequence being weighed. In this scenerio, A1 is in-paralogous to A2 and B1 is in-paralogous to B2. I1 and I2 indicates inferred nodes in the tree. $K_{x,y}$ *indicates the evolutionary distance between* x *and y as inferred from the CDS.*

Chapter 3: Upstream start sites of translation - a plant transcriptomics perspective

Large portions of this *Introduction* are soon to be published as part of:
von Arnim AG and Vaughn JN. uORF-mediated translational control in eukaryotes. In: The Encyclopedia of Systems Biology, Springer, expected 2011.

Small portions of the methods and results have been published as part of:
Kim, BH, Cai, X, Vaughn, JN, and von Arnim, AG. On the functions of the h subunit of eukaryotic initiation factor 3 in late stages of translation initiation. *Genome Biol,* **2007***, 8*, R60

Though I had an advisory role, the algorithm used to calculate uAUG enrichment scores was implemented into software by Sagar Utturkar.

Abstract

Numerous sequence elements within the 5' untranslated region (5'UTR) have been shown to inhibit and, to a lesser extent, enhance translation. The most ubiquitous of these elements, upstream start sites of translation (uAUGs), are found in abundance (>30%) in many eukaryotic transcripts. We examined the distribution and conservation of these elements across eleven angiosperm species with extensive transcriptome coverage. Between 30% and 45% of genes in each species give rise to uAUG-containing transcripts. Additionally, AUG is consistently the most depleted triplet in the 5' UTR. We employed pairwise comparisons of *Arabidopsis thaliana / Brassica napus* and *Solanum tuberosum / Solanum lycopersicum* transcriptomes to assess the conservation of these sequence elements in the 5' UTR. As found previously in mammals and fungi, AUGs within orthologous 5' UTRs show significantly higher conservation than any other triplet. We observed only a weak functional bias in the extent of uAUG conservation among different functional groups of genes, the extreme groups being involved in stress response and development. Extending our analysis in hopes of finding more precise molecular functions of uAUGs, we used the angiosperm dataset of eleven species and scored the uORF enrichment of a subset of 2,094 orthologous groups. Based on these results, we estimated

that 5% of orthologous groups fall into an uAUG-enriched category, 42% are uAUG-depleted, and 53% are uAUG-neutral. Among genes that are uAUG-depleted, there is a substantial bias for genes related to translation. Using groups with precedent for uAUG-enrichment from Chapter 2, we examined gene-specific patterns of uAUG conservation across dicot and monocot 5' UTRs. In summary, uAUGs appear to be functionally significant across both narrow and wide evolutionary distances in plants.

Introduction

In Chapter 2, we identified examples of local conservation in the 5' UTR associated with the AUG triplet that signals initiation to the scanning ribosome. These upstream AUGs (uAUGs) result in upstream open-reading frames (uORFs). An uORF is best described as a short ORF that begins 5' to the longest ORF in an eukaryotic mRNA. We refer to the portion of an mRNA found upstream of the longest ORF as the "5' untranslated region" (5' UTR), in spite of the fact that uORFs, which are by definition found within this region, are often translated. Because of the mechanics of eukaryotic translation, downstream ORFs, which exist in the 3' UTR, are not translated. Therefore, they have little impact on the biology of a cell. The term "upstream ORF" is not particularly applicable to prokaryotes, where mRNA transcripts commonly contain multiple protein-coding regions and initiate translation in a different way.

According to the scanning model of eukaryotic translation initiation, the small (40S) ribosomal subunit, in association with a bound $tRNA^{Met}$ and assorted initiation factors, scans from the 5' terminus (5' cap) of an mRNA toward its 3' end. A start codon is recognized primarily through codon-anticodon pairing with $tRNA^{Met}$. At this point, the large (60S) ribosomal subunit joins the complex, and translation elongation begins. An uORF poses a barrier to the scanning 40S ribosome because, upon recognition of the uORF start codon, the uORF peptide must be translated and terminated. Upon translation of an uORF, the translation machinery must perform a reinitiation event (see Figure 1.2 and 1.4). Reinitiation differs from standard 5' cap-dependent initiation in ways that are not yet fully understood. A ribosome whose 40S subunit dissociates from the mRNA after termination can be regarded as having suffered a permanent loss of reinitiation competence. Conversely, a 40S ribosome that resumes scanning downstream of the uORF displays a conditional loss of reinitiation competence because it lacks, among other factors, a tRNA^{Met}. These factors must be reacquired while the ribosome is scanning in order to successfully recognize the start codon of the main ORF further downstream. uORF-mediated translation repression is primarily dictated by a) the probability of initiation at the start codon of an uORF (based on sequence context and upstream events), b) the uORF's length, and c) the distance between an uORF stop and the next downstream start codon (see Chapter 4, Section 1).

An uORF can be created or removed by a single base change. Such mutations may have dramatic consequences for gene expression and phenotype of the organism, on par with mutations that alter mRNA splice sites or the active sites of enzymes and other proteins [24]. About one third of eukaryotic genes give rise to uORF-containing mRNAs (ranging from ~13% in yeast to ~47% in mammals). Prior attempts to address uAUG and uORF conservation have focused on general trends across multiple genomes. Analysis of human and mouse as well as yeast lineages focused on the frequency of triplet conservation within alignable regions of the 5'UTR [116]. uAUGs in these lineages show higher conservation than controls regardless of the frame of translation relative to the main ORF. Examination of four closely related fungal species within the *Cryptococcus* genus revealed that at least one-third of uORFs are conserved for their effect on translation [86]. This conservation does not appear to be related to constraint at the peptide level. All of the aforementioned studies focused on genome-wide uAUG/uORF conservation. Attempts to identify gene-specific examples of functional uORFs have also been undertaken in yeast, where researchers developed an expert system to search for constrained patterns of uORFs across 8 lineages separated by 100 million years of evolution [117]. The expert system was trained on the GCN4 mRNA's 5' UTR and, so, is biased in that regard. Still researchers claim to have identified 252 uORFs in the baker's yeast genome that could be functional and 32 that are very likely to be functional. As in the mammalian and *Cryptococcus* studies, yeast uORFs are typically not constrained at the peptide level.

While experimental evidence for effects of uORFs on fitness is sparse and largely indirect, at least 14 uORF-altering single nucleotide polymorphisms (SNPs) have been linked to human disease phenotypes [24]. Moreover, uORFs are particularly abundant in mRNAs that code for regulatory proteins, such as transcription factors and protein kinases [118].

The inference of functional constraint on uAUGs has led to a number of different hypotheses concerning their biological mode of action:

a) uORFs produce short peptides. A peptide-dependent uORF is one whose amino acid sequence is critical for function. With rare exceptions the peptides function in *cis*, i.e. one peptide molecule affects the expression of the mRNA molecule from which it was translated. Nascent uORF peptides are thought to slow progression of the ribosome by stalling it or by preventing termination and thus reinitiation [119]. As expected, peptide-dependent uORFs tend to be conserved at the amino acid level, and have a codon substitution bias favoring synonymous over nonsynonymous changes. Though clearly necessary for the regulation of some genes, they represent a minority of uORFs in all species examined [25,26].

b) Peptide-independent uORFs regulate translation via initiation-reinitiation regardless of their encoded peptide. Thus, uORFs may be a quick evolutionary route to reduce protein expression without disrupting more sensitive spatio-temporal gene regulation at the transcriptional level.

c) In addition to their direct impact on initiation efficiency, some uORFs indirectly reduce the stability of the mRNA via nonsense-mediated decay (NMD). In plants, this pathway is thought to require uORF longer than 90 nts [120].

d) A uORF may increase gene expression at the major ORF if, for example, initiation at one upstream uAUG prevents initiation at a second, highly inhibitory, downstream uAUG (see

Chapter 1, Section 2). Examples include the uORFs in the yeast bZip transcription factor, GCN4, and related mammalian bZips [121]. Note, in such cases, gene expression will most likely still be reduced relative to a comparable uORF-less mRNA.

e) An uORF can affect the choice of start codons for the post-transcriptional regulation major ORF, allowing multiple N-terminal isoforms to be produced from a single mRNA molecule [122].

f) Low-level transcription is stochastic and thus contributes substantially to the molecular noise of gene expression. Translational repression in conjunction with high-level transcription is a strategy for controlling noise in gene expression [123]. However, this hypothesis has yet to be tested with regard to uORFs.

As it stands, there is little data on which genes are under which forms of uORF regulation. In what follows, we extend many of the described findings from yeast and mammals to plants. We also identify gene-specific forms uAUG enrichment and analyze the uAUG/uORF structure of these groups with regard to the aforementioned hypotheses.

Results

Angiosperm orthologous groups

Using an approach described in Chapter 2 (see *Methods*), we identified orthologous groups across the following lineages (Figure 3.1):

Dicots - *Arabidopsis thaliana, Brassica napus (rapeseed), Citrus sinesis (orange), Glycine max (soybean), Gossypium hirsutum (cotton),* and *Nicotiana tabacum (tobacco).*

Monocots - *Hordeum vulgare (barely), Oryza sativa (rice), Panicum virgatum (switch-grass), Saccharum officinarum (sugarcane),* and *Zea mays (corn).*

The split between these two major branches of the angiosperm tree occurred ~150 million years ago [124]. While the monocot lineages are all from the same plant family, Poaceae, most are usefully diverged, such that long stretches of identical sequences are not solely a result of relatedness [52].

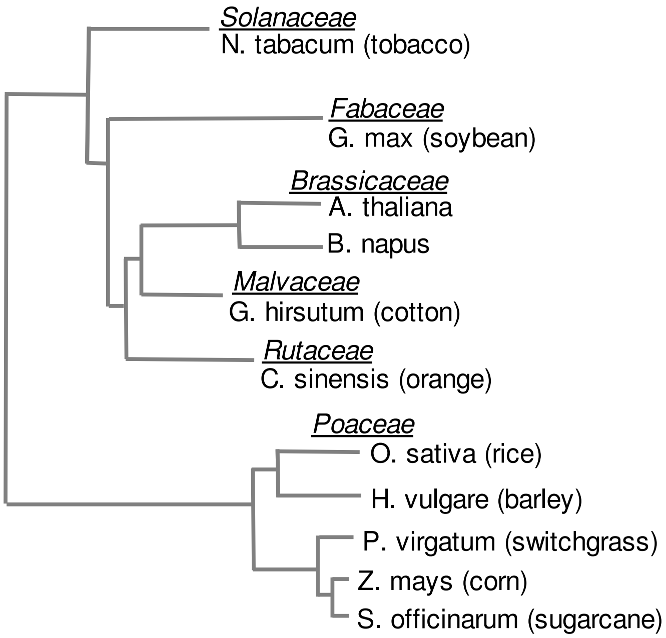


Figure 3.1: Tree representing descent and relative divergence of the species in this analysis. Modified from [110] and based on chloroplast genomes. The family name is given above representative species.

uAUGs are present in ~38% of angiosperm transcripts

Between 30% and 45% of genes in each angiosperm species gives rise to uAUG-containing transcripts (Figure 3.2A). General inclusion of protein coding sequence would bias poorly curated transciptomes toward the higher uAUG frequencies. Since well-curated *Arabidopsis* and *Oryza* are both near average, such artifacts appear to be minimal; thus, the range of values seen likely represents natural variation. Additionally, there is no relationship between genome size and percent of transcripts with an uAUG - *Zea*, a ~2.7Gbp genome, and *Arabidopsis*, a 157 Mbp genome, are roughly equivalent.

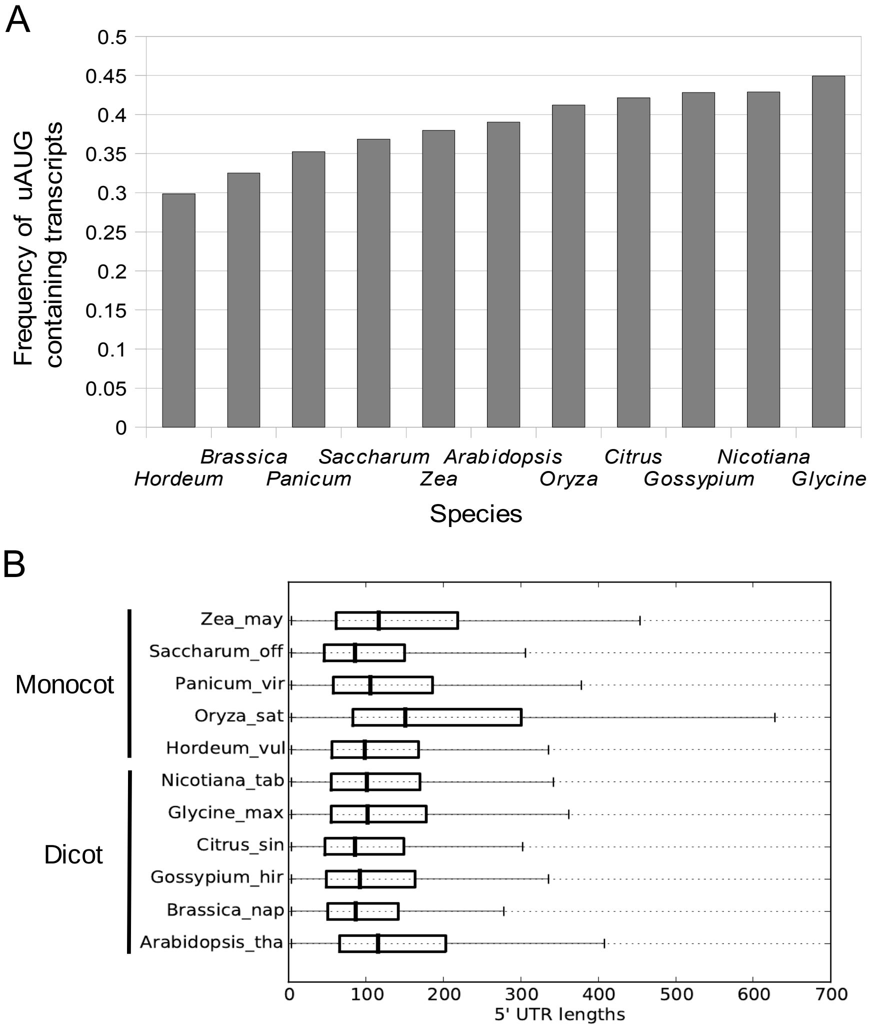


Figure 3.2: uAUGs are found in ~38% of mRNAs in angiosperm transcriptomes.
A) Bar graph of the percentage of known transcripts for the given angiosperms that contain at least one uAUG. Note, alternative transcripts of the same gene are included in the analysis. Species are sorted lowest, left, to highest, right, based on their frequency. B) 5' UTR length distributions for each species represented as a box plot. The whiskers of box plots represent the minimum and maximum values in the distribution. Boxes represent 25% below the median (thick, center line) and 25% above the median.

uAUGs are depleted in the 5'UTR of plants

Even though many transcripts contain uAUGs, because of their often inhibitory effect, it is expected that the 5' UTR will generally be depleted in AUGs across the transcriptome. This assumes that there is a substantial fraction of genes for which uAUG introduction is selected against. Using putative 5' UTR sequences from the 11 angiosperm lineages, we found that uAUGs show consistently lower than expected frequencies in the 5' UTR (Figure 3.3). This bias does not appear in the 3' UTR (Figure 3.4). For ATG, the bias does not appear to be related to the degree of sequence coverage or sequence curation: *Oryza* and *Aradidopsis* have extensive full-length-cDNA support but very different degrees of bias. Other triplets also show significant bias, some negative, like AUG, and some positive. Interestingly, the direction of the bias appears to be highly conserved among all eleven species. Of note, triplets with same base in position 1 and 3 typically show a positive bias, whereas triplets with three different bases show a negative bias. As a possible explanation, one might consider that 5' UTRs are enriched in repeat motifs (AGAGAG, CTCTCT, etc), which could be a cause for this bias. Yet, the bias doesn't appear to correlate with the absolute number of repeats within the 5' UTR - CT|TC and AG|GA repeats account for ~90% of the dinucleotide repeats in the 5' UTR (not shown). Also, removal of these repeats (5 or more dinucleotides in a row) had a discernible but minor effect on the general profile of triplet bias (Figure 3.5). One could further enhance the stringency of the repeat filter, but, as the threshold begins to approach the size of the word being analyzed, the analysis becomes circular. Though much weaker, some of these triplet bias patterns do appear in the 3' UTR. In summary, mutational mechanisms such as replication slippage may be biasing the frequency of short repeats in both UTR regions, but, based on their differential enrichment in the 5' UTR, these short repeats appear to have some function, either in transcription or translation.

Importantly, while ATGs are the most biased triplet, the bias is surprisingly weak: there are roughly 25% fewer than expected ATGs in the 5' UTR. Thus, either a large fraction of proteins are tolerant of uAUGs or most uAUGs have only a mild effect on expression levels.

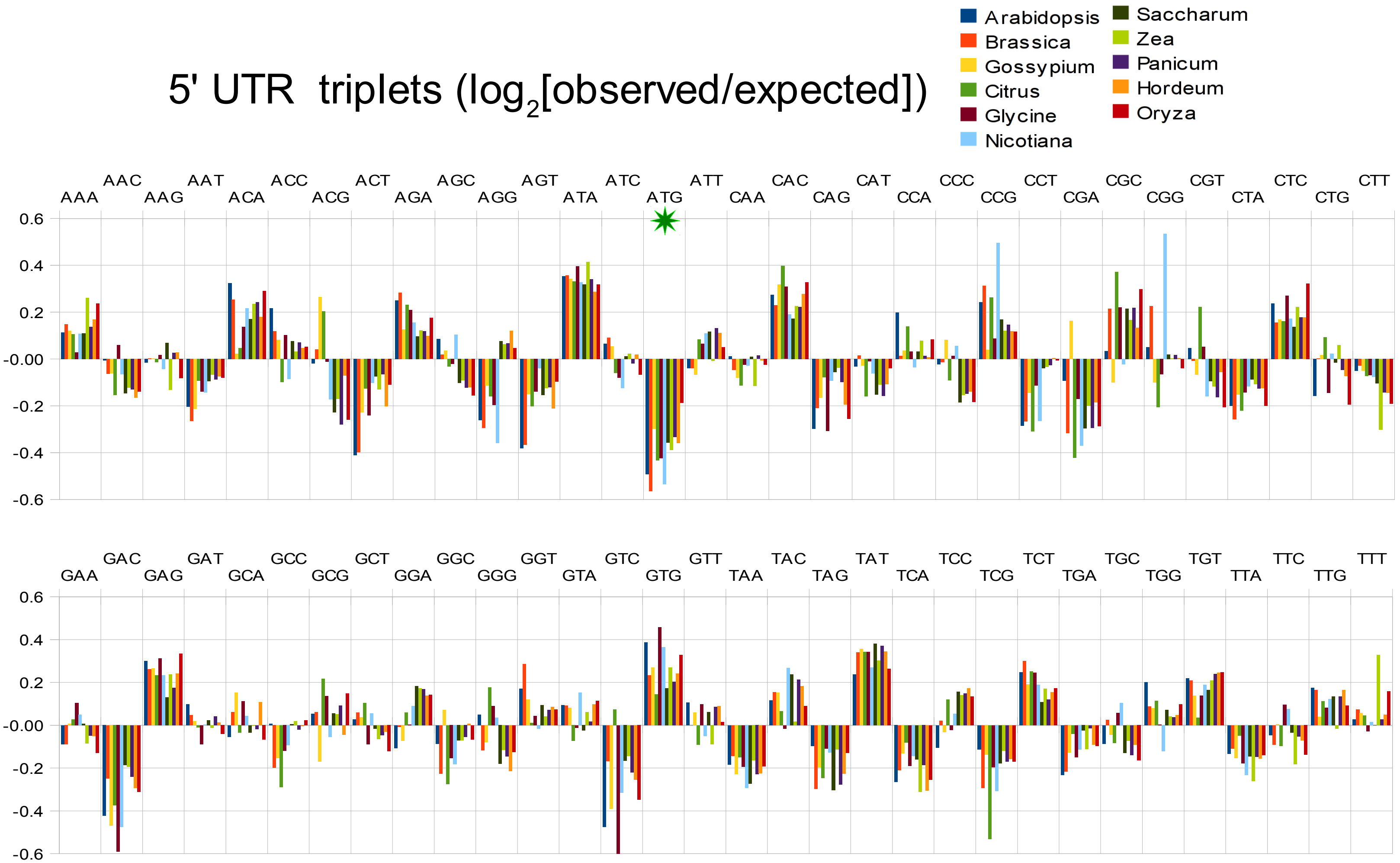


Figure 3.3: uAUGs in the 5' UTR are observed at consistently lower frequencies than expected.
Triplets starting with A and C are shown in the top panel. Triple starting G and T are shown in the bottom panel. Y-axis shows the $\log_2$*(observed/expected), where expected is calculated from the dinucleotide frequency (see Methods).*

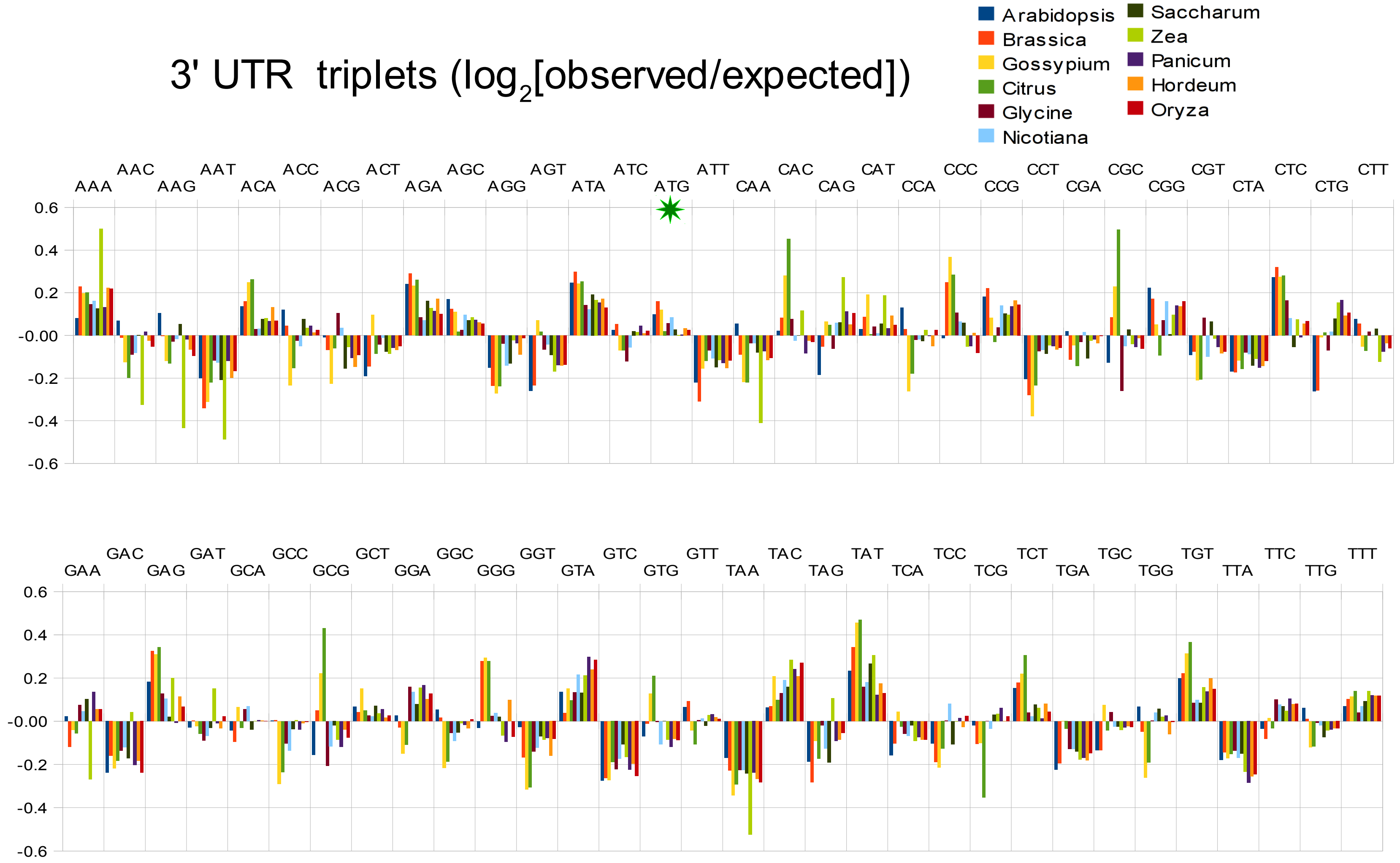


Figure 3.4: AUGs in the 3' UTR have no frequency bias.
See Figure 3.3 for description.

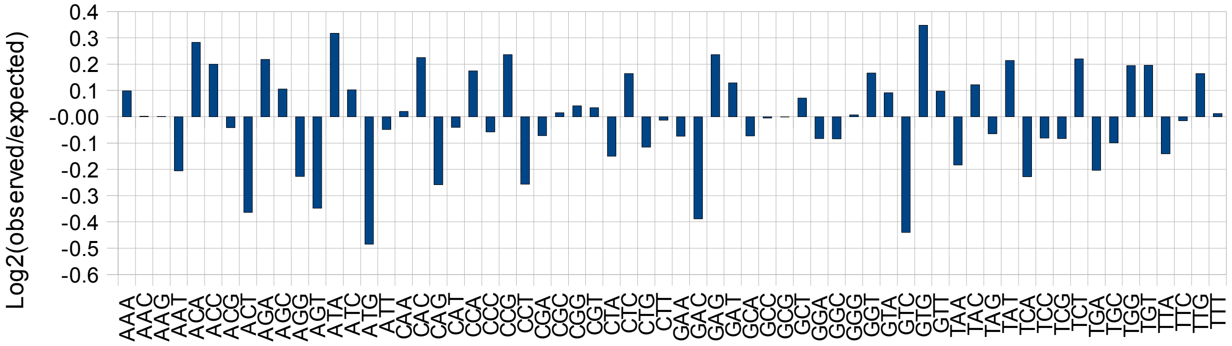


Figure 3.5: Removal of 5' UTRs with dinucleotide repeats does not substantially change triplet bias profile.
Only calculations from Arabidopsis *5' UTRs* are *shown as a representative profile. Repeats were considered those with 5 or more dinucleotides in a row.*

uAUGs are under purifying selection in flowering plants

As described in *Introduction*, fungal and mammalian 5' UTRs possess conserved uAUGs. Using pairs of species from two separate branches of the dicot tree, we tested if such conservation existed in plants. The 5' UTRs of orthologous proteins between *Arabidopsis thaliana / Brassica napus* (rape seed) and *Solanum lycopersicum* (tomato) */ Solanum tuberosum* (potato) were aligned with one another in pairwise fashion using the BLAST pairwise alignment program, *b2seq*. Additionally, human/mouse alignments were used as input to our computational pipeline in order to recapitulate results from an earlier study [116] using the software we had developed. Since the plant 5' UTRs had already been clustered based on relatedness, our alignability criteria were less stringent than the analysis in mammals. Regions were considered alignable if their match length was greater than 50 nts and if the comparison had an e-value of less than 0.01. The conservation score for each triplet was calculated as the number of perfectly conserved triplets in the alignment divided by the total number of triplets within all sequences in the alignment.

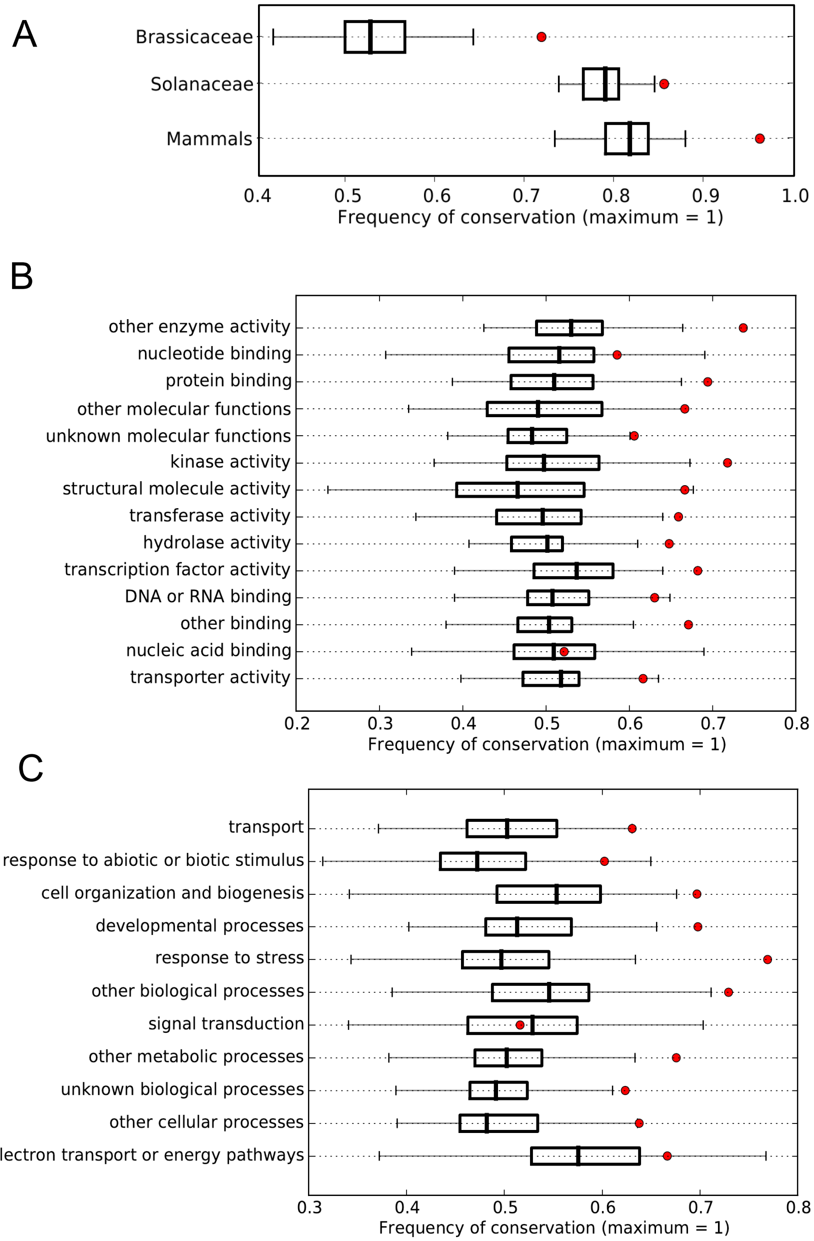


Figure 3.6: uAUGs are conserved in a higher percentage of applicable alignment columns than any other triplet in the 5' UTR.

Dots indicate the AUG conservation frequency. Boxes indicate the distribution of conservation frequencies for all other non-AUG triplets. The whiskers of box plots represent the minimum and maximum values in the distribution. Boxes represent 25% below the median (thick, center line) and 25% above the median. A) Whole-transcriptome pairwise comparisons: Brassicaceae - Arabidopsis thaliana and Brassica napus, Solanaceae - tomato and potato, Mammals - human and mouse. B-C) Brassicaceae uAUG conservation within high-level GO categories - molecular function categories (B) and biological process categories (C).

uAUG conservation within the 5' UTR appears to extend to plants (Figure 3.6A). As expected, triplets are more conserved between two species from a single genus, *Solanum*, than two species from different genera within the same family, *Brassicaceae*. Again, the extensive conservation across mammals reflects the more stringent alignability criteria used previously.

Functional groups appear to be equivalently enriched in conserved uAUGs

Purifying selection appears to be acting on mutations that disrupt AUGs in the 5' UTR. In the *Introduction*, we discussed many hypothetical functional explanations for uAUGs. As a preliminary approach toward understanding uAUG function, we examined uAUG conservation with regard to functional categorization. Each pairwise alignment in the *Brassicaceae* comparison used above, was grouped into a Gene Ontology (GO) Slim category. In essence, these categories represent root nodes of the three main GO hierarchical categorization schemes.

Few GO categories deviate from the general uAUG conservation profile seen in whole transcriptome comparisons (Figure 3.6B-C). Of all categories, 'Response to stress' and 'Signal transduction' show the strongest and weakest uAUG conservation, respectively. These also happen to be two of the most narrowly defined of the GO categories used, suggesting that more resolution in classification may be critical to differentiating functional bias among genes with conserved uAUGs. Ideally, gene specific examples of such conservation could be identified, and we pursued this aim in the next section.

Very few orthologous groups show uAUG enrichment across angiosperms

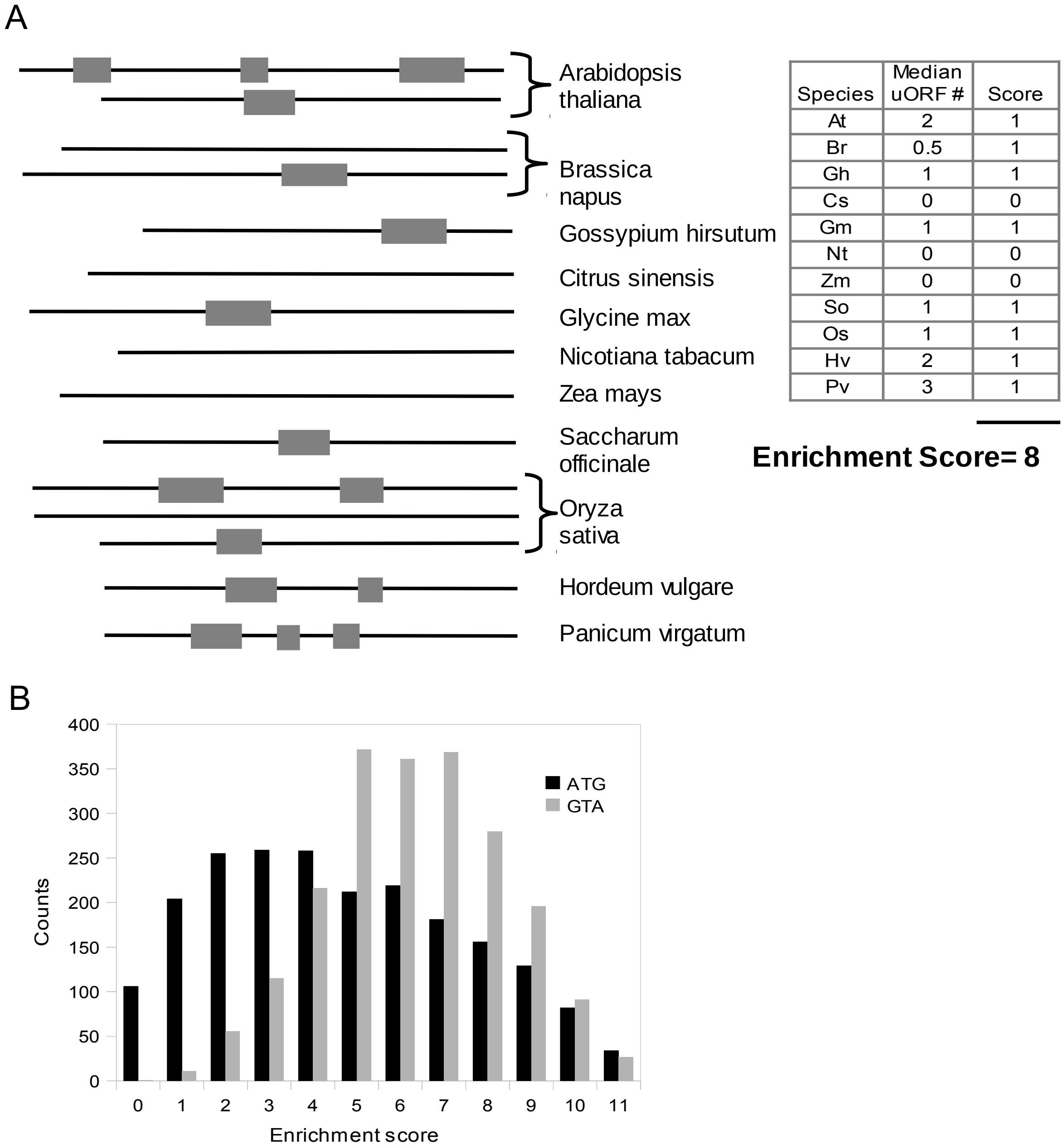
The techniques used in the previous section required that the 5' UTRs of compared species be aligned. Because an AUG is only three letters long, in order to arrive at gene-specific resolution, such a method would require tens to hundreds of transcriptomes at the family level of phylogenetic scope. Even with current sequencing advances, such coverage is unlikely in the near future. If portions of a sequence are constrained, large evolutionary distances between compared sequences increases the signal-to-noise ratio in a comparative sequence analysis. Thus, a comparison of all angiosperms would broaden our available sequence pool and enhance signals if they do in fact exist. Orthologous groups generated from 11 angiosperm species were used to extend our analysis of uAUG conservation.

As stated, because of extensive divergence across the angiosperm tree, 5' UTRs used in this analysis could no longer be aligned with any appreciable confidence. Instead, we focused on uAUG enrichment or depletion. By this we mean that the 5' UTRs associated with an orthologous group possess more or less AUGs relative to a suitable control, GUA [86]. GUA is the reverse of AUG and, since the scanning complex reads the start codon 5' to 3', the GUA is not expected to have any bearing on translation. Additionally, it has the same mononucleotide composition, is replicated identically, and does not share a dinucleotide overlap with AUG. Moreover, GUA shows no discernible bias in the 5' UTR of any angiosperm analyzed (Figure 3.3).

Our hypothesis was that there are genes for which uAUGs are not tolerated, genes for which uAUGs are neutral, and a small group for which uAUG may be selectively retained. Of these we expected the first group to be the largest because the general depletion of AUGs from the 5' UTR of angiosperms (Figure 3.3). Concerning uAUG enrichment, though we observed uAUG conservation in alignments between plants in the same family (Figure 3.6), we did not know if such purifying selection would span the angiosperm tree.

We returned to the angiosperm orthologous groups for this analysis. Of these groups, only 2,094 had all 11 species represented. Only these 2,094 groups were processed further. We then scored uORF enrichment in the linked 5' UTRs based on an algorithm in which the per-sequence AUG content for all sequences from each species was tabulated using a simple string search (Figure 3.7A). The per-species median uORF number was then calculated. We used the median, as opposed to the mean, in order to guard against possible artifacts resulting from the inclusion of coding sequence where AUG may be very common. If the median was >0.5, species received a score of 1; otherwise, 0. The per-species scores were tabulated and summed to give an enrichment score. It is also conceivable to use the median score directly, and analyze resultant distributions. We found that results were more interpretable using the 0 or 1 approach because it more effectively identified groups where uAUG enrichment or depletion spanned the entire phylogenetic tree. Using the median directly, some groups may have many uAUGs in a UTR within dicots but few in monocots, and would still score as high as groups where single-uAUG enrichment spans the entire tree.

Many clusters have a substantially different uAUG enrichment score relative to the control, GUA (Figure 3.7B). Specifically, uAUG depletion is much more common than depletion of the control. These results are consistent with our hypothesis that mutations that would create an uAUG are selected against in many genes, whereas they are neutral in others. *Vice versa*, we hypothesized that mutations that disrupt uAUGs may be selected against in some genes (uAUG enrichment). However, on first glance, this analysis did not reveal a clear surplus of uAUG-enriched groups (score of 8-11). We reexamined this question after considering that the AUG dataset may be the aggregate of three underlying distributions, depleted, neutral and enriched, whereas the GUA dataset consists entirely of the neutral distribution.



Species	Median uORF #	Score
At	2	1
Br	0.5	1
Gh	1	1
Cs	0	0
Gm	1	1
Nt	0	0
Zm	0	0
So	1	1
Os	1	1
Hv	2	1
Pv	3	1

Figure 3.7: uAUGs are depleted from many orthologous groups.
A) Example of how an enrichment score is derived from 5' UTRs from an orthologous group. Gray boxes represent uORFs in the 5' UTR (black line). Species names beside a sequence are abbreviated to the first letter of the genus and specific epithet in the adjacent table. B) Histogram of enrichment scores for all orthologous groups with 11 species represented.

Using the GUA distribution as representative of a neutral distribution, we modeled enriched and depleted fractions as binomial distributions with 11 trials. Because of variation resulting from species relatedness, this model is unlikely to be the best reflection of distributions resulting from biased enrichment scores, but it models the control distribution well barring a slight underestimate of variance (not shown). Each of these distributions - enriched, depleted, and neutral - was assigned a relative contribution to the actual distribution (Figure 3.7B). The GUA distribution was used directly as the neutral distribution. This results in a five parameters: proportion accounted for by the neutral distribution (f_n), proportion accounted for by the depleted distribution (f_d), *p* for depleted distribution (p_d), proportion accounted for by the enriched distribution (f_e), *p* for enriched distribution (p_e). Since, the total contributions must sum to 1 ($f_n = 1 - (f_d + f_e)$), a four parameter model results. This is the 'NDE-model (w/ enriched)' in Figure 3.8A.

For contrast, we also modeled the AUG distribution, under the assumption that there is no AUG-enriched fraction of gene ('ND-model (no enriched)' in Figure 3.8A), which, for reasons defined above, results in a two-parameter model. Though the two models produce a comparable fit across most of the distribution, the NDE-model with enrichment is, as expected, substantially better at representing the enriched categories (9 through 11). For validation, we checked enrichment scores for previously identified uORFs that appeared to be conserved at the peptide-level across dicots; as many of these are known to be conserved across the angiosperm tree as well, these conserved peptide uORFs would likely exhibit AUG-enrichment. Of these 19 groups (Table 2.4), only five had orthologous groups with all 11 representatives. Of these five, four had an enrichment score of >9. Therefore, our enrichment score does appear to reflect groups in which AUGs are functional, and that the NDE-model, versus the ND-model, is justified.

The NDE-model suggests that 5% of orthologous groups fall into the AUG-enriched category, while the balance is split between the AUG-depleted (42%) and neutral (53%) categories (Figure 3.8B). For Arabidopsis, 5% equates to roughly 1,250 genes, a value that dramatically exceeds characterized cases of uORF function (~70 genes). In spite of illuminating the size of the fraction of orthologous groups that might generally require an AUG, the NDE-model makes it clear that we would, at best, have a 50% chance of picking a group that truly requires AUGs. Alternatively, the AUG-depleted groups are easily identified, and groups with a 0 enrichment score are almost certainly intolerant of uAUGs.

Genes from orthologous groups with uAUG depletion serve fundamental roles in the cell

As expected from general depletion in AUGs across all 5' UTRs in the transcriptome (Figure 3.4), many orthologous groups are depleted in AUGs (Figure 3.7). But, unlike simply examining general features within the transcriptome of a single species, comparative analysis allowed us to identify particular genes in which uAUGs are not tolerated (Figure 3.8B). If uAUGs, at best, lead to a small reduction in protein production, then genes that are depleted in AUGs should be fundamental to central metabolic pathways in the plant cell.

To test this hypothesis, we extracted the *Arabidopsis* accessions from orthologous groups having enrichment scores of 0 (Figure 3.8B). These accessions were then used to check for GO term enrichment relative to the entire *Arabidopsis thaliana* transcriptome using the AMIGO term enrichment web-service. Among enriched terms, translation-related categories are the most significant and ubiquitous (Figure 3.9). As translation is perhaps the most fundamental process in biology, this result coincides with our expectation. Genes related to nucleotide biosynthesis, another central pathway, also appear to be differentially depleted in uAUGs. Auxin synthesis is one of the only plant-specific categories that shows a significant uAUG-depletion bias, although response to heat, salt, and metal ions all have lower but significant enrichment as well (shown as 'Stress Response' in Figure 3.9).

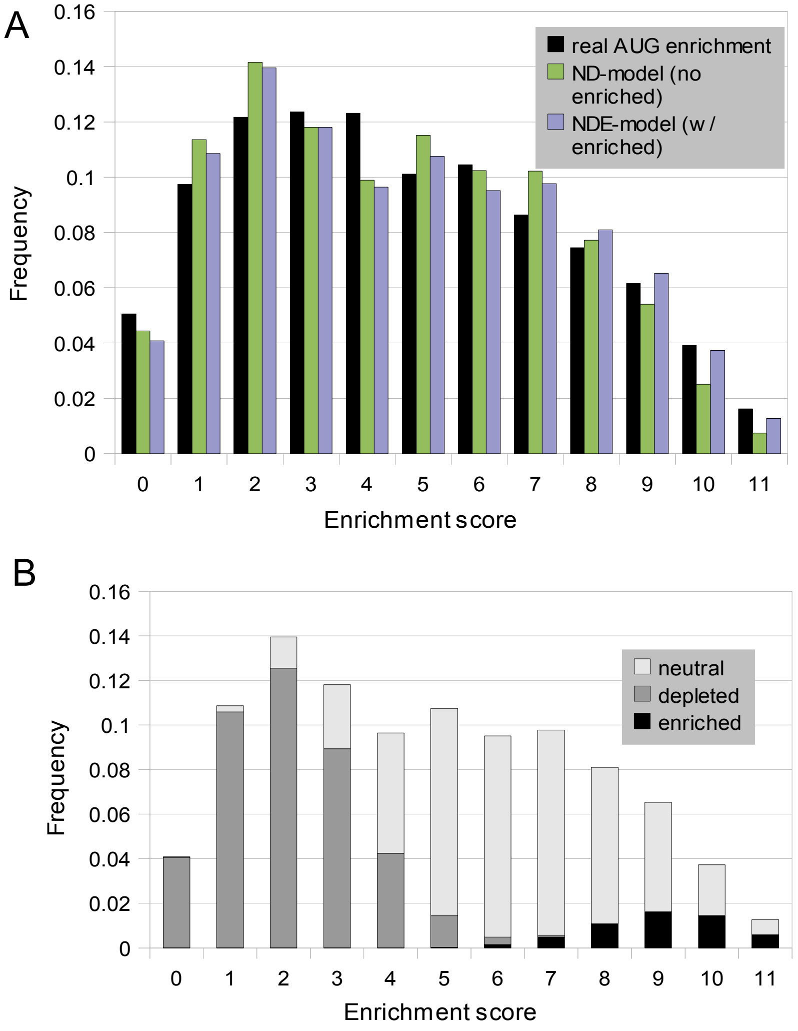


Figure 3.8: NDE model predicts the relative proportion of uAUG-depleted, enriched, and neutral orthologous groups.

A) Black bars indicate the actual distribution AUG enrichments scores (Figure 3.7) converted to frequencies. Green bars represent prediction of the ND-model, which has two parameters and does not account for an enriched component. Blue bars represent the NDE-model, which has four parameters and does account for an enriched component. B) Relative proportions of 'neutral', 'depleted', and 'enriched' categories within the total predicted enrichment distribution.

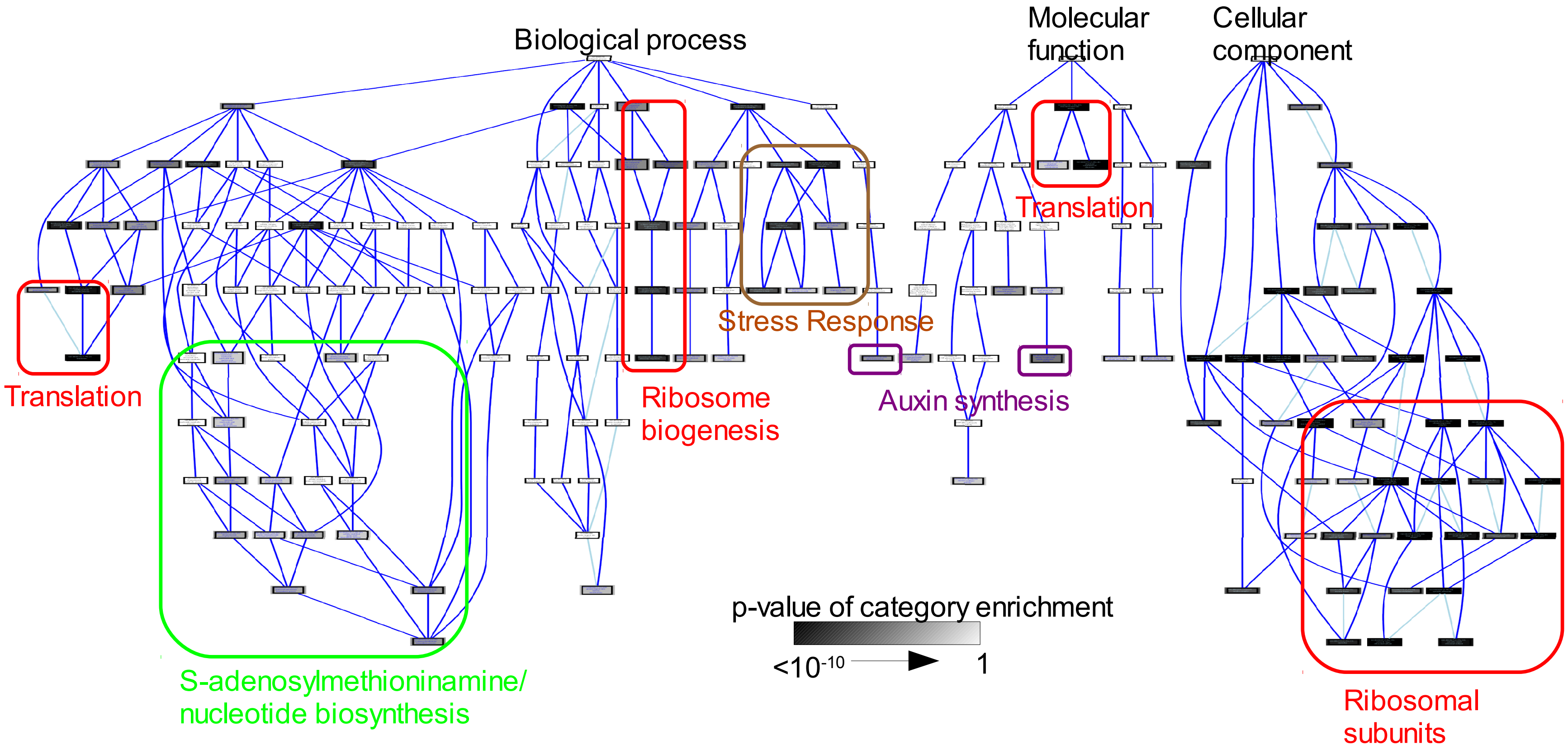


Figure 3.9: uAUGs are depleted in pathways central to all cellular life.
GO term enrichment was assessed for Arabidopsis genes associated with orthologous groups that had an enrichment score of 0. The interconnected graph represents the hierarchy of GO terms. The darker the box-node the smaller the p-value of term enrichment among uAUG-depleted genes relative to the Arabidopsis transcriptome. Size of the box-node is irrelevant

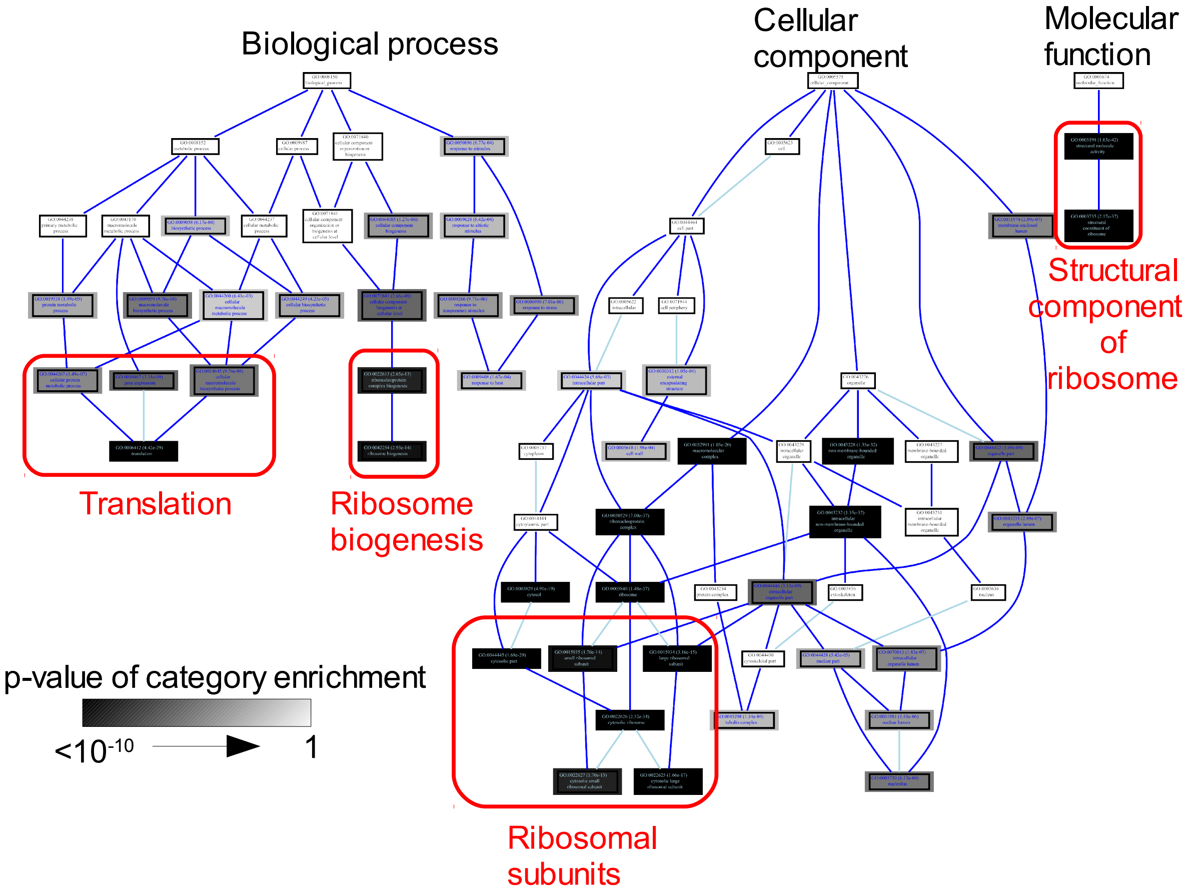


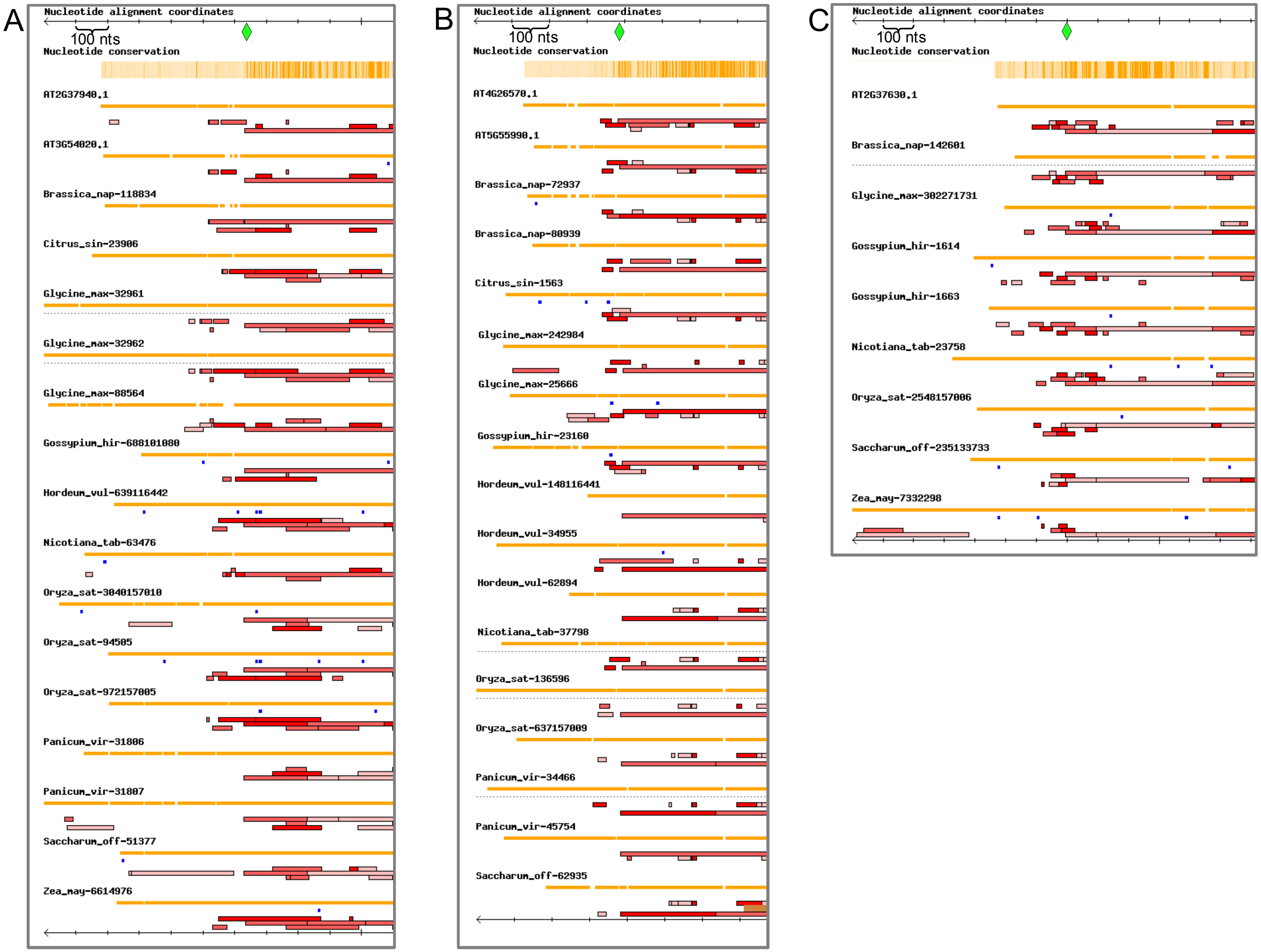
Figure 3.10: Moderate enrichment control.
Similiar to Figure 3.9, except that these data represent enrichment relative to a background of genes with moderate (5-8) enrichment scores (Figure 3.8). See Figure 3.9 for description.

Because we initially limited our analysis to orthologous groups with 11 representatives, we were concerned that we were introducing bias into the GO enrichment analysis: genes shared across 11 species are likely to be genes related to fundamental biological processes. To address this issue, we ran a similar analysis as above, except that instead of the entire *Arabidopsis* transcriptome, *Arabidopsis* accessions from the moderate enrichment class (5 through 8 in Figure 3.8B) were used as a background model. This result also indicates that the introduction of uAUGs into genes related to translation is not tolerated (Figure 3.10). Because we limit the size of our background model by ~10-fold, enrichment categories with moderate significance in Figure 3.9 disappear. Thus, while these results bear out the conclusions above, they are perhaps overly conservative. Genes related to the categories that disappear are not anymore significantly enriched among the 2,094 gene analyzed than many other categories (not shown), and so their presence in Figure 3.9 is not likely to be an artifact of general enrichment in the analysis.

Specific examples of peptide-independent uAUG conservation show uORF stacking

As described above, it is difficult to differentiate true-positives from false-positives in terms of orthologous groups that exhibit AUG-enrichment. Still, because 5% of these groups may require uAUGs (Figure 3.8B), we considered it worthwhile to further investigate uAUG coverage across the monocot-dicot divide. In Chapter 2, we described a set of 19 AUG-containing conserved elements within the 5' UTRs of orthologous dicot genes (Table 2.8). These regions did not appear to overlap known smRNA binding sites or to be constrained at the peptide level. Using this dataset as a starting point, we examined these groups in more detail with regard to the additional monocot data. Note that many of these groups do not appear in the enrichment score analysis (Figure 3.7) because they do not have the requisite 11 species represented. Of the four that did, all four have an enrichment score >8.

Figure 3.11 (next page): Orthologous groups exhibiting uORF stacking.
(A-C) Transcript alignments are shown in miniature followed by all possible ORFs, where the darkness of red indicates the context strength of the ATG. ORFs are positioned relative to the alignment and appear in separate rows based on their frame relative to the end of the transcript. The consensus start of the main ORF is indicated by a green diamond on the scale. The 3' region of main ORF is clipped for viewing purposes. Orange vertical lines in 'Nucleotide conservation' lane indicates the level of identity in that column. The orange horizontal line below each accession indicates sequence coverage for the given accession relative to the alignment. Blue boxes indicate the presence of a splice-site consensus sequence - [A|G]GGA. A) INOSITOL PHOSPHORYLCERAMIDE SYNTHASE 2. B) CALCINEURIN B-LIKE 3 C) ASYMMETRIC LEAVES 1.


A
Nucleotide alignment coordinates
100 nts
Nucleotide conservation
AT2G37940.1
AT3G54020.1
Brassica_nap-118834
Citrus_sin-23906
Glycine_max-32961
Glycine_max-32962
Glycine_max-88564
Gossypium_hir-688101080
Hordeum_vul-639116442
Nicotiana_tab-63476
Oryza_sat-3040157010
Oryza_sat-94505
Oryza_sat-972157005
Panicum_vir-31806
Panicum_vir-31807
Saccharum_off-51377
Zea_may-6614976
B
Nucleotide alignment coordinates
100 nts
Nucleotide conservation
AT4G26570.1
AT5G55990.1
Brassica_nap-72937
Brassica_nap-80939
Citrus_sin-1563
Glycine_max-242984
Glycine_max-25666
Gossypium_hir-23160
Hordeum_vul-148116441
Hordeum_vul-34955
Hordeum_vul-62894
Nicotiana_tab-37798
Oryza_sat-136596
Oryza_sat-637157009
Panicum_vir-34466
Panicum_vir-45754
Saccharum_off-62935
C
Nucleotide alignment coordinates
100 nts
Nucleotide conservation
AT2G37630.1
Brassica_nap-142601
Glycine_max-302271731
Gossypium_hir-1614
Gossypium_hir-1663
Nicotiana_tab-23758
Oryza_sat-2548157006
Saccharum_off-235133733
Zea_may-7332298

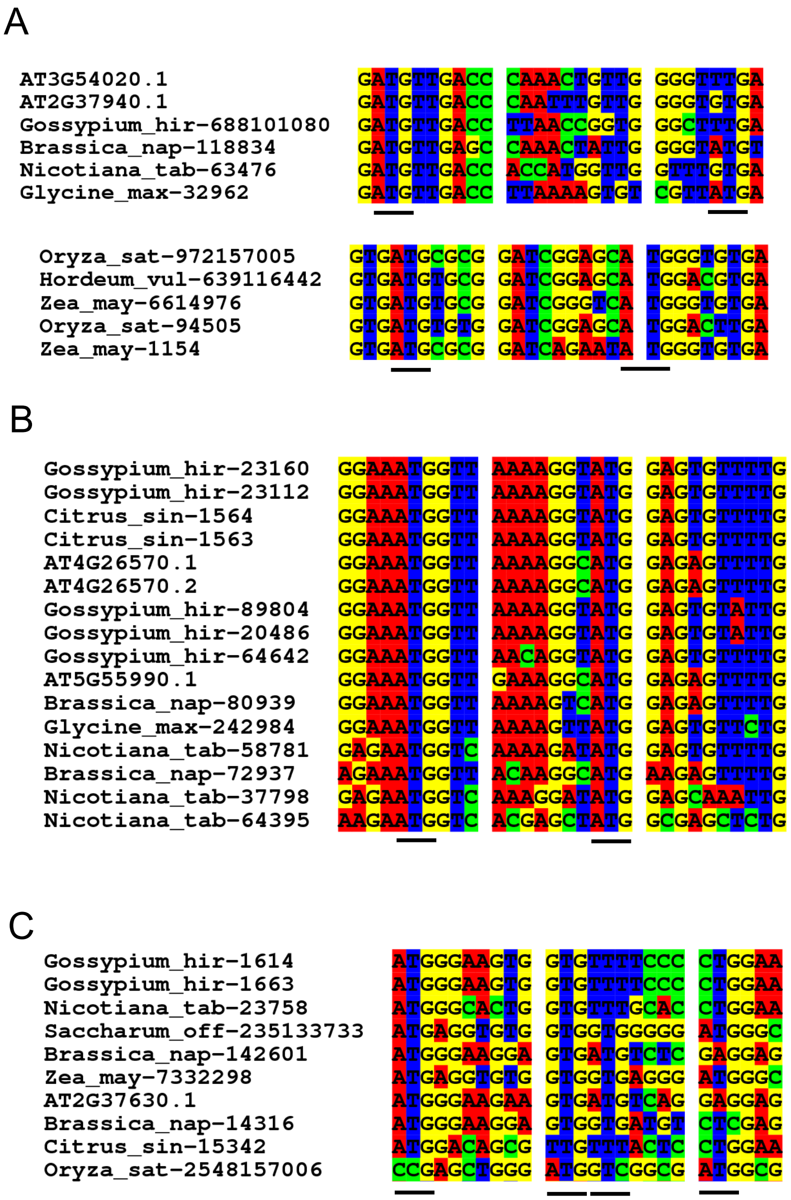

A
AT3G54020.1 GATGTTGACC CAAACTGTTG GGGTTTGA
AT2G37940.1 GATGTTGACC CAATTTGTTG GGGTGTGA
Gossypium_hir-688101080 GATGTTGACC TTAACCGGTG GGCTTTGA
Brassica_nap-118834 GATGTTGAGC CAAACTATTG GGGTATGT
Nicotiana_tab-63476 GATGTTGACC ACCATGGTTG GTTTGTGA
Glycine_max-32962 GATGTTGACC TTAAAAGTGT CGTTATGA
Oryza_sat-972157005 GTGATGCGCG GATCGGAGCA TGGGTGTGA
Hordeum_vul-639116442 GTGATGTGCG GATCGGAGCA TGGACGTGA
Zea_may-6614976 GTGATGTGCG GATCGGGTCA TGGGTGTGA
Oryza_sat-94505 GTGATGTGTG GATCGGAGCA TGGACTTGA
Zea_may-1154 GTGATGCGCG GATCAGAATA TGGGTGTGA
B
Gossypium_hir-23160 GGAAATGGTT AAAAGGTATG GAGTGTTTTG
Gossypium_hir-23112 GGAAATGGTT AAAAGGTATG GAGTGTTTTG
Citrus_sin-1564 GGAAATGGTT AAAAGGTATG GAGTGTTTTG
Citrus_sin-1563 GGAAATGGTT AAAAGGTATG GAGTGTTTTG
AT4G26570.1 GGAAATGGTT AAAAGGCATG GAGAGTTTTG
AT4G26570.2 GGAAATGGTT AAAAGGCATG GAGAGTTTTG
Gossypium_hir-89804 GGAAATGGTT AAAAGGTATG GAGTGTATTG
Gossypium_hir-20486 GGAAATGGTT AAAAGGTATG GAGTGTATTG
Gossypium_hir-64642 GGAAATGGTT AACAGGTATG GAGTGTTTTG
AT5G55990.1 GGAAATGGTT GAAAGGCATG GAGAGTTTTG
Brassica_nap-80939 GGAAATGGTT AAAAGTCATG GAGAGTTTTG
Glycine_max-242984 GGAAATGGTT AAAAGTTATG GAGTGTTCTG
Nicotiana_tab-58781 GAGAATGGTC AAAAGATATG GAGTGTTTTG
Brassica_nap-72937 AGAAATGGTT ACAAGGCATG AAGAGTTTTG
Nicotiana_tab-37798 GAGAATGGTC AAAGGATATG GAGCAAATTG
Nicotiana_tab-64395 AAGAATGGTC ACGAGCTATG GCGAGCTCTG
C
Gossypium_hir-1614 ATGGGAAGTG GTGTTTTCCC CTGGAA
Gossypium_hir-1663 ATGGGAAGTG GTGTTTTCCC CTGGAA
Nicotiana_tab-23758 ATGGGCACTG GTGTTTGCAC CTGGAA
Saccharum_off-235133733 ATGAGGTGTG GTGGTGGGGG ATGGGC
Brassica_nap-142601 ATGGGAAGGA GTGATGTCTC GAGGAG
Zea_may-7332298 ATGAGGTGTG GTGGTGAGGG ATGGGC
AT2G37630.1 ATGGGAAGAA GTGATGTCAG GAGGAG
Brassica_nap-14316 ATGGGAAGGA GTGGTGATGT CTCGAG
Citrus_sin-15342 ATGGACAGCG TTGTTTACTC CTGGAA
Oryza_sat-2548157006 CCGAGCTGGG ATGGTCGGCG ATGGCG

Figure 3.12: Local alignments of regions around the uORF start sites. Heavy black underlines indicate an ATG-containing column. A) INOSITOL PHOSPHORYLCERAMIDE SYNTHASE 2. B) CALCINEURIN B-LIKE 3 C) ASYMMETRIC LEAVES 1.

Of the genes we examined with more than two clearly defined monocot orthologs (7 out of 19), few trends were evident. Using diagnostic graphics like those in Figure 3.11, the only discernible pan-angiosperm pattern was the occurrence of multiple overlapping uORFs in three separate orthologous groups. The three groups depicted in Figure 3.11 are functionally unrelated both in terms of pathways in which they act and molecular functions that they perform. *INOSITOL PHOSPHORYLCERAMIDE SYNTHASE 2* (Figure 3.11A) encodes an enzyme in the sphingolipid biosynthesis pathway. *CALCINEURIN B-LIKE 3* (Figure 3.11B) encodes a calcium-responsive regulator of kinase activity. *ASYMMETRIC LEAVES 1* (Figure 3.11C) encodes a developmental transcription factor. Yet, each of these proteins shares of general pattern of uORF stacking around and overlapping the start site of the main ORF, a pattern that could potentially be highly inhibitory. The general alignability of many of these uORF start sites across very divergent lineages indicates that they have been under purifying selection since the earliest angiosperms (Figure 3.12).

Discussion

uORFs appear to to be present in many transcripts and are functionally constrained across both narrow and wide evolutionary distances in plants (Figure 3.2A, 3.8B, and 3.12). Monocots generally appear to have fewer uAUG-containing transcripts (Figure 3.2A), and this is not related to 5' UTR length (Figure 3.2B). More transcriptomes are needed to confirm this trend, and also to differentiate whether these lower values are unique to grasses. In any case, the bias against AUGs in the 5' UTR is not significantly different between monocot and dicot lineages (Figure 3.3).

Examination of closely related plant species reveals that, like mammals, some mutations that disrupt uAUGs are under purifying selection relative to other 5' UTR triplets (Figure 3.6A). Though this functional constraint appears to span genes associated with all biological processes, uAUGs harbored by stress response genes are dramatically constrained (Figure 3.6C). Interestingly, orthologous groups in the pan-angiosperm analysis that are depleted in uAUGs are also commonly found among genes associated with stress response (Figure 3.9). Thus, stress responsive genes are perhaps the most dramatic example of the NDE-model (Figure 3.8B), in which many genes cannot tolerate uAUGs but those that possess uAUGs require them for proper function. Though solely speculation, it stands to reason that certain genes belonging to these stress response categories - Hsp70 and Hsp101, for example [110] - need to be translated at full capacity when the transcriptional response is initiated. Alternatively, genes that mediate the transcriptional activation of these 'responder' genes may need to be expressed either with reduced yield or reduced noise to prevent premature investment in a major physiological transition. We still lack the resolution to test this prediction, but requisite sequence data may be available in the near future (http://www.onekp.com/index.html).

Using data from Chapter 1, we were able to identify a small sample of orthologous groups in which peptide-independent uORF conservation may span the angiosperm tree. *INOSITOL PHOSPHORYLCERAMIDE SYNTHASE 2* contains uORFs ranging from ~80 to ~200 nts and these are often followed by a canonical splice sequence; thus, this transcript may be a good candidate for conserved uORF-mediated NMD (Figure 3.11A). Interestingly, sequence constraint appears around the uORF start sites in both monocot and dicots but these regions are unalignable and, thus, may not be homologous (Figure 3.12A). Although, given that purifying selection may be acting on a very small region of this 5' UTR, we would not expect homology to be discernible. *CALCINEURIN B-LIKE 3* uORFs are typically shorter and closer to the start of the main ORF (Figure 3.11B). Also, while the sequence region around the two uORFs is highly constrained in dicots (Figure 3.12B), there is no extensive sequence constraint in the 5' UTRs of monocots. The high degree of conservation within this region, at least in dicots, may either result from the requirement of very strong start codon context for both start sites or from overlapping an unknown *cis*-element. Lastly, like the CALCINEURIN B-LIKE, the

ASYMMETRIC LEAVES 1 mRNA has a high concentration of uORFs around the start of the main ORF (Figure 3.11C). These are typically short. Additionally, there is discernible sequence constraint associated with this region acting in both monocots and dicots (Figure 3.12C). Interestingly, in all sequences, there is a large span (~300 nts) of ORF-less sequence space downstream of these uORF clusters. Because of the possibility of re-initiation, this mRNA is perhaps the best candidate as of yet for uORF-mediated N-terminal isoform production [122]. Also, while Oryza has lost its initial uORF ATG (relative to other monocots), it has gained one downstream (Figure 3.12C).

Interestingly, uORF-stacking is not only associated with peptide-independent uAUG conservation, but also with many conserved peptide uORFs. For example, the beginning of the sucrose-responsive bZip11 uORF is overlapped by a shorter uORF in all lineages that we have analyzed (not shown). These uORFs may be anti-inhibitory in the case of bZip11 (see Chapter 4, Section 1), which also appears to be a valid explanation for many of the initial uORFs in Figure 3.11. Why nature would maintain an inhibitory uORF only to diminish its effect with an anit-inhibitory uORF remains somewhat puzzling but speaks to the possibility that these genes require very precisely tuned expression levels.

In this analysis, we have focused on orthologous groups in which all species were represented; thus, we reduce our total pool of orthologous groups from ~9,000 to 2,094. As described above, this was done mainly to accommodate our scoring algorithm and to emphasize uAUG enrichment that spanned the entire angiosperm tree. Ideally, we could use all of these groups. This may be possible by using a Branch Length Scores, which can account for relatedness among lineages under study [125]. In effect, the underlying tree relating the species is used to calculate the score. In our case, the presence/absence score would be used (Figure 3.7A), such that the length of the subtree connecting all 1's would become the score.

Methods

Sequence acquisition and preparation for analysis of 5' UTR triplet composition

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 (see *Results*) were downloaded from http://www.plantgdb.org/prj/ESTCluster/progress.php on 2 February 2011 [99]. The longest ORF from each putative transcript was extracted using a custom Perl module (uORF.pm available at http://web.utk.edu/~jvaughn7/code/bio/) based on the criteria that an ATG be followed in-frame by a TAA, TGA, or TAG 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 2.2.21), in species-wise fashion with an e-value cutoff of <1E-30. All BLAST searches were performed using the Newton supercomputer at the University of Tennessee. 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) with default parameters [100]. Only 5' UTRs linked with those proteins clustered into orthologous groups were used.

Analysis of 5' UTR triplet composition

UTRs having annotated residues other than A, C, G, or T were removed; counts were of all overlapping words containing an A, C, G, or T. UTRs shorter than 3 nts were removed. Percentage of transcripts with an AUG (Figure 3.7A) was found by searching each resultant 5' UTR for one or more AUGs. For triplet expectation (Figure 3.4 and 3.5), dinucleotides and triplet words were counted using a sliding window incremented by 1 residue. Importantly, to avoid artificial fusion triplets, UTR sequences were not concatenated prior to assessment of dinucleotide and triplet frequencies. Dinucleotide frequencies were then used in a conditional probability formula to predict the expected triplet frequency, for example:

$$p(AUG) = p(AU) \times p(UG)/(p(UG) + p(UA) + p(UT) + p(UC))$$

Analysis of AUG substitution rates and their functional bias

For the pairwise uAUG conservation analysis (Figure 3.6), transcript data for Arabidopsis thaliana was downloaded from http://www.arabidopsis.org/help/helppages/BLAST_help.jsp#datasets on 24 September 2009. Putative transcripts for all other plant species were downloaded from http://www.plantgdb.org/prj/ESTCluster/progress.php on 7 October 2009 [99]. Orthologous groups were generated in the manner given above. Transcripts with 5' or 3' UTRs shorter than 8 nuclotides were excluded from analysis. Data for the mammalian lineage comes from prealigned sequences made available at ftp://ftp.ncbi.nih.gov/pub/koonin/uAUG/.

Each 5' UTR from a given species within a particular orthologous group was aligned with all 5' UTRs of the other species in that group. Pairwise alignments were carried out using *b2seq* of the BLAST suite (version 2.2.16). Regions were considered alignable if their match length was greater than 50 nts and if the comparison had an e-value of less than 0.01. Each sequence within an alignable region was then parsed for its triplet conservation using custom Perl scripts in conjunction with the Bioperl library [126]. A sliding window of three nucleotides incremented in one nucleotide steps was used.

The Gene Ontology for *A. thaliana* was downloaded from ftp://ftp.arabidopsis.org/home/tair/Ontologies/ on 4 September 2009. Each orthologous group was assigned a GO categorization base on the *A. thaliana* ortholog. uAUG enrichment was assessed as the average number of uAUGs per sequence in the group divide by the average sequence length in the group.

Assessing uAUG enrichment

Using the angiosperm orthologous groups described above and their associated 5' UTRs, groups were given a triplet enrichment score ('AUG' and control 'GUA') based on the following algorithm (see Figure 3.7A): 1) Each 5' UTR was searched for the number of triplets. 2) The median value was taken with regard to all sequences representing each individual species. 3) Median values were rounded to the nearest integer. 4) If the resultant value was one or greater, the species received a score of 1; otherwise, 0. 5) Scores were tabulated for each orthologous group. Only orthologous groups with 11 species represented were used, resulting in 2,094 groups.

Hypothetical distributions underlying the AUG distribution were modeled such that the 'neutral' category was defined by the GUA distribution (Figure 3.7B). The 'enriched' and 'depleted' categories were modeled as binomial distributions. The relative contribution of each category to the actual distribution was a free parameter ranging from 0 to 1. The *p* parameter in the 'enriched' and 'depleted' models was also a free parameter, ranging from 0 to 1. In total this resulted in 4 free parameters for the NDE-model. The parameters were fit by minimizing the squared difference between the prediction and the actual frequency of each enrichment-score. Minimization was performed using an evolutionary strategy implemented in Perl (http://search.cpan.org/~pjb/Math-Evol-1.12/Evol.pm). The two parameters for the ND-model (*p* for 'depleted' category and relative contribution) were fit in a similar manner.

GO term enrichment

Arabidopsis accessions from groups with an enrichment score of 0 were tested for GO term enrichment using the 'Term Enrichment' webservice (http://amigo.geneontology.org/). The 'TAIR' background model was used with default parameters. We additionally used *Arabidopsis* accessions from groups with an enrichment score of 5 through 8 as a background model, with default parameters as well.

Group-specific curation

Groups with precedent for conserved AUGs in the 5' UTR (see *Results*) were aligned as full transcipts using ClustalW with default parameters. ORFs were then mapped back to the alignment using custom Perl scripts in conjuction with the Bioperl library [126]. Local alignments of the 5' UTR (Figure 3.12) were generated using MEME (Version 4.4.0) [127] and a background model of dinucleotide frequencies based on all *Arabidopsis* 5' UTRs.

Chapter 4: The mechanics of translation initiation

Section 1 of this chapter has been previously published in:

Roy B, Vaughn JN, Kim B-H, 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.

Small portions of the data and analysis in Section 3 have been previously published in:

Kim B-H, Cai X, Vaughn JN, von Arnim AG (2007) On the functions of the h subunit of eukaryotic initiation factor 3 in late stages of translation initiation. Genome Biol 8: R60.

Though I had an advisory role, the computational pipeline used to analyze secondary structure in the 5' UTR was implemented as software by Qidong Jia at the University of Tennessee

Abstract

The multiplicity of events that affect a single uORF initiation-reinitiation cycle can confound intuitive predictions of the system being tested. This complexity grows substantially with the addition of multiple uORFs. Ideally, we could formalize these events into a model, which would make quantitative predictions based on known elemental processes. To address this need, we developed a 'sum-histories' computational model of the scanning process that incorporates the experimentally observed effects of uORF length, intercistonic length, and start codon sequence context (Section 1). Data from dual luciferase assays involving an allelic series of 21 mutant 5'UTR constructs with various uORF structures of the Arabidopsis AtbZip11 5' UTR as a model system were used to estimate the quantitative parameters of the model. According to these estimates, ~75% of encounters between the ribosomal preinitiation complex and a start codon result in elongation, and this occurs regardless of sequence context. In addition, estimates were made in order to identify salient defects associated with a mutation in the subunit h of eukaryotic initiation factor 3 (eIF3h). Based on these result, we concluded that eIF3h buffers against the loss of reintiation competence that has been observed to occur during uORF elongation. Though the luciferase reporter system described in Section 1 was useful for characterizing molecular defects for a single gene, ideally we could apply existing models to all endogenous transcripts of the entire Arabidopsis transcriptome. The translation state of all expressed mRNAs in Arabidopsis can be measured using polysome fractionation followed by microarray hybridization. In Section 2, we examine 5' UTR features in light of emerging data based on this technique, finding that secondary structure around the mRNA 5' cap is a major determinant of translation state. In Section 3, we used high-resolution data from yeast in order to test computational methods that will maximize the biological insight gained from these assays while minimizing the large cost associated with such multi-fraction microarray experiments. Though motivated by technical improvement, the result in Section 3 further suggest a decoupling of translationally active and quiescent pools of mRNA.

Section 1: The mechanics of initiation/reinitiation can predict the extent of uORF-mediated repression and the molecular defects associated with a mutation in eIF3h

Introduction

Quantitative models of translation have extensive precedent in the literature, but, to our knowledge, only two of these models deal explicitly with the events preceding translation initiation (though some of these models do allow a single generic initiation rate to vary between mRNAs [128,129]. Both Skjøndal-Bar & Morris [130] and Dimelow and Wilkinson [131] incorporated known initiation factors into a large system of differential equations in an attempt to explore initiation affects manifested by variations in the concentration and phosphorylation state of these factors. The later study concludes that, based on currently available data, useful estimates of the rates of most of these reactions cannot be determined. Neither of these models deals with uORFs explicitly, in spite of a clear correlation between initiation efficiency and uORF content [22,132]. Hence, in order to assess the degree to which these events can explain our system, we took a more abstract approach by modeling the effect of uORFs based on the known repercussions of various uORF-associated parameters, such as uAUG context, uORF length, and length of the spacer sequence between uORFs.

The von Arnim lab has undertaken numerous in vivo dual luciferase assays to test the effects of varying uORF structures on translation initiation efficiency. The variations are derived from the *Arabidopsis AtbZip11* 5'UTR, which harbors a cluster of phylogenetically conserved uORFs (Figure 4.1A). In brief, the translation efficiency was determined by taking the ratio between the activities of two luminescent reporter proteins. Firefly luciferase was expressed from a uORF-containing 5'UTR (condition) and Renilla luciferase was expressed from an unstructured, uORF-less 5'UTR (control). Each reporter construct was driven by the same promoter to control for transcript abundance. These reporter constructs were then co-transformed into seedlings and given time to express. These seedlings are then ground and their lysate is assayed for luciferase ratios.

To facilitate quantitatively rigorous conclusions for these numerous experiments, we constructed a computational model of translation initiation on the uORF-containing AtbZip11 leader (Figure 4.1B). The five variables of the model correspond to canonical events of the scanning model of translation and represent the following mechanistic events (Figure 4.2). Variables p_{cs} and p_{cw} represent the probability that the 40S ribosome recognizes an AUG in strong or weak sequence context, respectively. Variable k_1 describes the rate per nucleotide (nt) at which the ribosome loses its reinitiation competence, permanently, during uORF translation. The remaining ribosomes, said to have suffered only conditional loss of competence due to loss of Met-tRNA-eIF2-GTP (ternary complex), will terminate translation and resume scanning. Variable k_2 describes the rate per nt scanned at which the ribosome regains its full reinitiation competence, in part by acquiring a ternary complex. Finally, certain uORFs trigger permanent loss of reinitiation competence in a fashion dependent on the peptide sequence, independent of their length. This is true for uORF2b of AtbZip11. Variable p_{2b} represents the probability of escape from the attenuation caused by translating the peptide of uORF2 of AtbZip11, where $p_{2b}=0$ indicates dissociation of the ribosome from the mRNA every time, i.e. no escape.

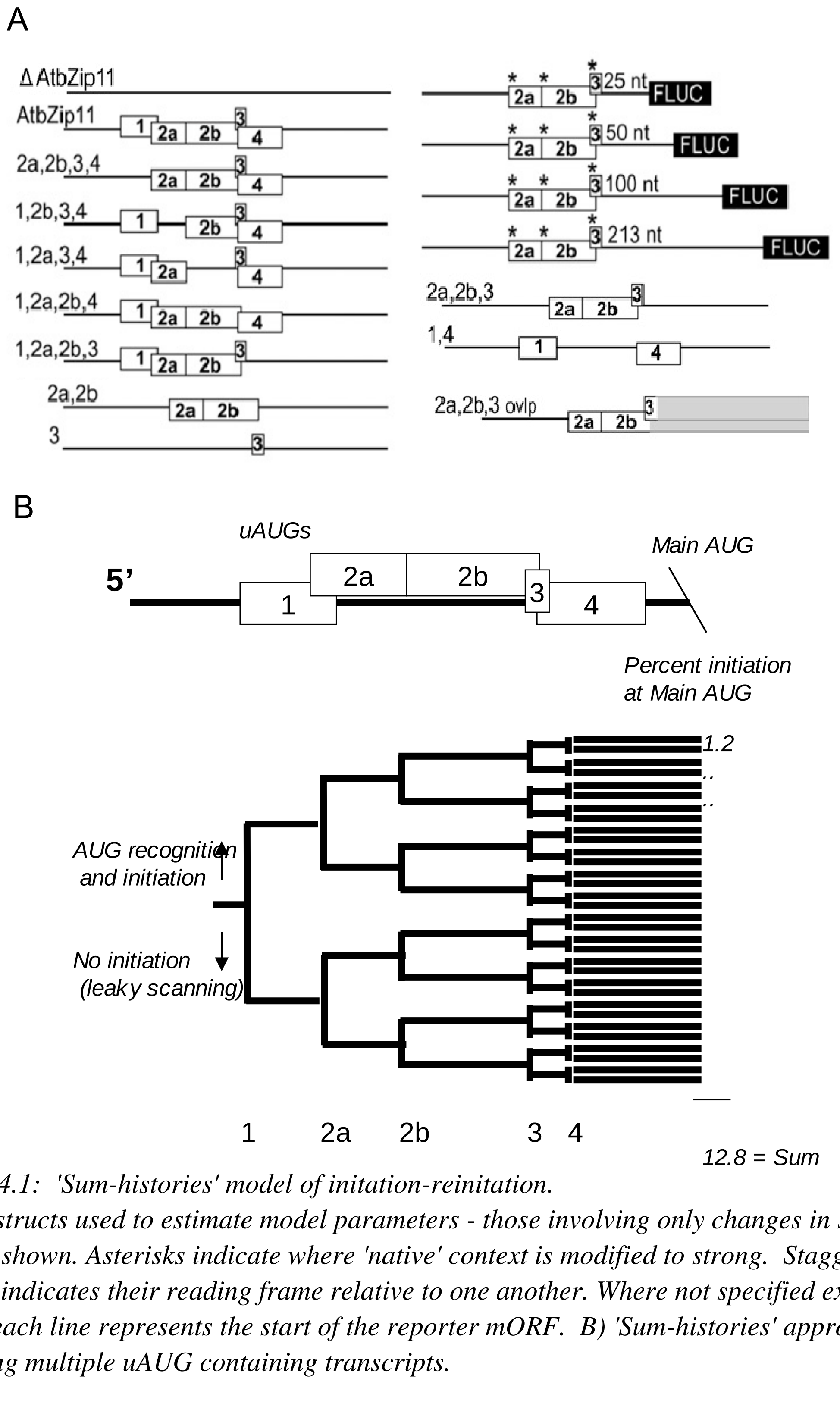


Figure 4.1: 'Sum-histories' model of initation-reinitation.
A) Constructs used to estimate model parameters - those involving only changes in start context are not shown. Asterisks indicate where 'native' context is modified to strong. Staggering of uORFs indicates their reading frame relative to one another. Where not specified explicitly, the end of each line represents the start of the reporter mORF. B) 'Sum-histories' approach to modeling multiple uAUG containing transcripts.

Results

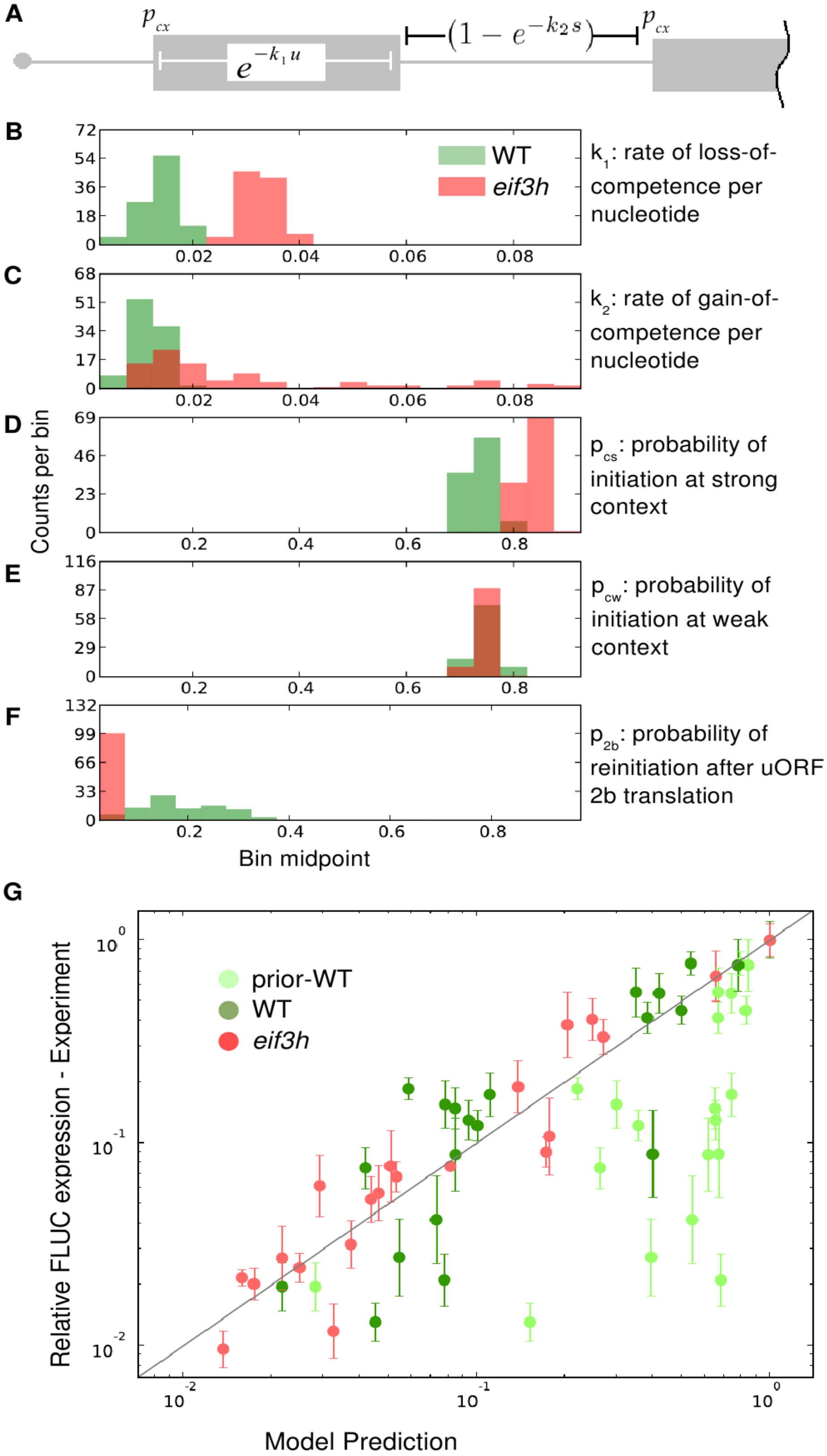
Parameters culled from the literature (Table 4.1) yielded a poor fit between model and wild-type experimental data (Figure 4.2G). We therefore adopted an evolutionary algorithm to generate estimates for each parameter for both wild-type and *eif3h* mutant plants, using as a fitness criterion the least sum-squares fit between model and experimental data (see Methods). To assess uncertainty with regard to experimental variation, we generated 95% confidence intervals (CIs) for these parameters by bootstrapping from the experimental data (Figure 4.2B-F). Significantly, the maximum likelihood estimate (MLE) for k_1, the rate at which reinitiation competence is lost during uORF translation, was lower in wild-type plants than in eif3h mutant plants (Table 4.1). Thus, in the wild type, 50% loss of competence was estimated at 58 nt of uORF translated, whereas in the mutant 50% loss of competence was more rapid, at 22 nt. In contrast, the parameter estimates for AUG recognition (p_{cx}) were affected little by eIF3h. Weak context was the same between the two genotypes. Strong context differed by a small, yet statistically significant, margin (Table 4.1). The escape from attenuation at uORF2 (MLE(p_{2b}) = 0.16 for wild type) suggests that translation of uORF2b allows only one out of six wild-type ribosomes to retain reinitiation competence. In contrast, in eif3h p_{2b} was effectively nil (Figure 4.2F), indicating complete loss. The low MLEs for p_{2b} suggest that uORF2 strips ribosomes of their reinitiation competence more effectively than suggested by its length. It stands to reason that the mechanisms driving k_1 and p_{2b} are intimately related. Loss of competence (k_1) and uORF2 dissociation penalty (p_{2b}) were negatively correlated in WT conditions, i.e. modeling trials in which loss of competence was mild tended to assign a very strong dissociation penalty to uORF2 and vice versa (not shown). The k_2 parameter for regain of initiation competence during scanning was smaller in wild type than in *eif3h*, i.e. slower recovery in the wild type. However, k_2 had a weak effect on fit, and hence was poorly constrained, in the *eif3h* background (Figure 4.2C; Table 4.1). In summary, the new parameters generally improved the fit between model and experimental data (Figure 4.2G). The modeling work supports two conclusions; (i) translation initiation on the AtbZip11 leader can be explained to a large degree by translation of inhibitory uORFs, reinitiation, and leaky scanning; and (ii) the most evident molecular function of eIF3h is the retention of reinitiation competence during uORF translation.

Table 4.1: Estimates of parameters in model of translation initiation

	Wild type			***eif3h* mutant**	
Parameter	**Estimate from literature**	**MLE**	**95% CI**	**MLE**	**95% CI**
Rate of loss-of-competence (k_1)	0.008 [133,134]	0.012	0.006 - 0.017	0.031	0.026 - 0.036
Rate of gain-of-competence (k_2)	0.015 [135]	0.008	0.004 - 0.014	0.009	0.008 - 0.092
Probability of initiation, strong context (p_{cs})	0.950 [11]	0.71	0.67 - 0.76	0.82	0.78 - 0.84
Probability of initiation, weak context (p_{cw})	0.250 [11]	0.72	0.69 - 0.76	0.72	0.69 - 0.74
uORF 2b penalty (p_{2b})	set to 1 *	0.16	0.026 - 0.31	0	0 - 0

MLE, Maximum likelihood estimate
CI, confidence interval
* Value could not be estimated from literature data.

Figure 4.2 (next page): Computational model of the translational defect in eif3h mutant plants.
(A) Model parameters. (Gray boxes) ORFs. Spanning bars indicate the range over which a given term applies. (B–F) Distributions of parameter estimates. x-axis length reflects the manually set boundaries of possible parameter estimates. y-axis counts indicate the number of times out of 100 trials that a parameter fell into one of 20 x-axis bins. (B) Loss rate of reinitiation competence (k1) as a function of uORF length (u [nt]). (C) Regaining of reinitiation competence (k2) as a function of intercistronic spacer length (s [nt]). (D) Probability of AUG recognition in a strong context (pcs). (E) AUG recognition in a weak context (pcw). (F) Escape from attenuation upon translation of uORF2b peptide (p2b). (G) Scatter plot illustrating the match between model output using Maximum Likelihood Estimates (x-axis) and experimental data (y-axis, bars are one standard error) for all 21 AtbZip11 leader constructs in wild-type (dark green) and eif3h plants (red). (Light green symbols) Predicted expression values using model parameters culled from the literature (see Table 4.1).


A
p_{cx}
$e^{-k_1 u}$
$(1 - e^{-k_2 s})$
p_{cx}
B
72
54
36
18
0
0.02
0.04
0.06
0.08
🟩 WT
🟥 *eif3h*
k_1: rate of loss-of-competence per nucleotide
C
68
51
34
17
0
0.02
0.04
0.06
0.08
k_2: rate of gain-of-competence per nucleotide
D
Counts per bin
69
46
23
0
0.2
0.4
0.6
0.8
p_{cs}: probability of initiation at strong context
E
116
87
58
29
0
0.2
0.4
0.6
0.8
p_{cw}: probability of initiation at weak context
F
132
99
66
33
0
0.2
0.4
0.6
0.8
p_{2b}: probability of reinitiation after uORF 2b translation
Bin midpoint
G
● prior-WT
● WT
● *eif3h*
Relative FLUC expression - Experiment
10^{0}
10^{-1}
10^{-2}
10^{-2}
10^{-1}
10^{0}
Model Prediction

The 2b portion of uORF2 is an inhibitory attenuator peptide that is activated when AtbZip11 translation is repressed by sucrose. The computer model incorporated an equivalent uORF2 penalty. Although we worked at a comparatively low sucrose concentration of 1%, changing the peptide sequences of uORF2 and 3 via compensatory frameshift mutations resulted in translational derepression, the extent of which was more pronounced in *eif3h* than in wild type. Despite alteration of the uORFs' peptide sequences, translation remained dependent on eIF3h albeit at a diminished level. Several versions of the frameshifted uORF with slightly different coding sequences gave similar results [136]. In conclusion, the role of eIF3h in reinitiation is not restricted to the uORF2 peptide. Instead, as suggested by the model, eIF3h helps to retain reinitiation competence in peptide sequence-dependent and sequence-independent ways.

Discussion

The computational model is founded on the notion that four types of variables drive initiation efficiency at the main start codon: the context of uORF start codons , the length of the previously translated uORF , any attenuation caused by the nature of the uORF peptide , and the spacer length between a uORF stop codon and the next AUG start codon . Granted that eIF3h stimulates reinitiation, is it not possible that eIF3h simply increases the affinity between the 40S subunit and eIF3, such that a post-termination 40S subunit can effectively recruit a fresh eIF3 complex from the soluble cytosolic pool? Speaking against this is that eIF3h's effect is conditional on attributes of the uORF, and in two distinct ways. First, eIF3h suppresses the permanent loss of reinitiation competence during uORF translation. Specifically, it reduces the rate parameter, k_1, by about two-fold (Figure 4.2B). Second, eIF3h reduces the additional loss of reinitiation competence caused by uORF2 (p_{2b}, Figure 4.2F). The effect of eIF3h on parameters k_1 and p_{2b} may well be due to one and the same molecular activity of eIF3h. We propose eIF3h facilitates post-initiation retention of eIF3 on the ribosome, and ribosomes that have retained eIF3 will be more likely to resume scanning and reinitiate than ribosomes that have not.

Discrepancies between model and experiment were dispersed over the entire data set and were generally within two standard errors of the experimental mean (Figure 4.2G). However, there were exceptions, which are evidence of mechanistic events that have yet to be modeled. The set of variables was deliberately kept to the minimum that is well supported by prior knowledge. For example, uAUG2a is very inhibitory, especially in the eif3h mutant, even though it is in a weak context and masked by uORF1. This might point to ribosome-ribosome interactions. Let us consider that the uORF2 attenuator peptide slows the progression of elongating or terminating 80S ribosomes, consistent with the uORF2 penalty in both wild type and mutant (Figure 4.2F). We now postulate that ribosome occupancy by uORF2 affects the trajectory of upstream ribosomes that are poised to reinitiate after uORF1. The block on uORF2 would block 40S ribosomes that are scanning downstream from uORF1 and this might cause them to dissociate from the mRNA, possibly in an eIF3h-dependent way. Another plausible mechanism is that stacking of initiation-competent 40S ribosomes may foster AUG recognition at uAUG2a or 2b, which would exacerbate exponentially the eIF3h-dependent inhibition of expression. AUG recognition by 40S ribosomes can also be enhanced when the ribosome is blocked by RNA secondary structure [137]. Stacking on top of uAUG2a might arise from a block of eif3h mutant ribosomes upon termination of uORF1, a block of 60S subunit joining on uAUG2b, a block in elongation over a triplet of rare arginine codons present in uORF2b, or a combination of these. These possibilities remain to be tested. At pcw= 0.78, our parameter estimate for initiation at AUG in a weak context was fairly high, but not unprecedented, compared to published values . The postulated ribosome stacking effect may be the reason underlying the high AUG recognition and may explain why leaders harboring uORF1 are generally eIF3h dependent, but in a fashion dependent on uAUG2a.

Methods

We developed a suite of Perl objects to handle the data management, automate model calculations, and estimate model parameters (evolModel available at http://web.utk.edu/~jvaughn7/code/). Each 5'UTR sequence was parsed for uORFs using a custom Perl module (uORF.pm described above). Initiation efficiency was modeled as the sum probability of a strongly binary tree representing all possible initiation events that a ribosome could experience prior to encountering the start codon of the main ORF. Hence the model is referred to as the 'sum-history' model (Figure 4.1B). The probability of initiation at a given AUG was calculated as the context of that AUG (weak or strong) multiplied by the effects of the previously translated uORF's length and distance from that previously translated uORF's stop:

$$P(u,s)=p_{cx}e^{-k_1u}(1-e^{-k_2s})$$

where,

P(u,s) = probability of initiation; p_{cx} = probability of initiation based on context (strong or weak); k_1 = rate of loss-of-competence per nucleotide; k_2 = rate of gain-of-competence per nucleotide; *u* = length of last translated uORF; *s* = length from last translated uORF stop to the start of the current ORF

See Results and for further description of parameters.

AUG contexts were considered strong (p_{cs}, [GA]nnAUGG, [GA]nnAUGn or nnnAUGG), and weak (p_{cw}, nnnAUGn), with n being any other nucleotide and brackets indicating alternatives . The uORF2b penalty (p_{2b})was applied when uORF2b sequence was translated, which included the translation of uORF2a. The experimental data are from 21 AtbZip11 reporter constructs. The five parameters in the model were optimized for best fit with the experimental data using an evolutionary strategy implemented in Perl with Math::Evol (http://www.cpan.org/authors/id/P/PJ/PJB/Math-Evol-1.10.tar.gz) and run for 80 cpu seconds maximum with default parameters and with a relative and absolute convergence criteria of 10^{-11} and 10^{-16}, respectively; if fit did not improve by the given criteria within the last 25 generations then optimization was terminated. In order to correct for differential efficiency of reporters, fit was evaluated based on the following equation:

$$\sum_{j=1}^{m}\sum_{i=1}^{n}[\ln(\frac{x_{i,j}}{x_{norm}})-\ln(\frac{y_j}{y_{norm}})]^2$$

where,

n = the number of experiments per condition; m = the number of conditions; x_{norm} = experiment median of the uORF-less leader condition; y_{norm} = model prediction of the uORF-less leader; x = the experimental result; y = model prediction

Each of the 100 optimization trials was started with parameter values selected from a uniform distribution delimited by the biologically relevant minima and maxima described below. The sampled range and initial value of the adjustable step-size, in parentheses, for all parameters are as follows: weak (p_{cw}) and strong (p_{cs}) contexts as well as p_{2b}, 0 - 1 (0.204); k_1 and k_2, 0 - 0.10 (0.025). To obtain confidence intervals for our estimates, the reference dataset used to evaluate fit was bootstrapped from the original dataset for each trial of parameter estimation (Hillier, 2005; Hunt 1998). Increasing the trial number further had only a nominal effect on confidence intervals. To assure parameter convergence, 100 trials were run against the set of 21 conditions without bootstrapping. All but one of the resulting parameter estimates varied to within <0.001% of the parameter mean, suggesting that a global minimum (within the bounded region) was consistently found. The one exception was rate of gain-of-competence (k_2) in eif3h, which found local minima 12 out of 100 times and, excluding these values, varied to within 1% of the parameter mean. Meta-analysis suggests that this is not a result of rough topology around the MLE but of k_2's lack of impact on model fit to the eIF3h data once the parameter exceeds ~0.009.

Section 2: Translation state and UTR features

Introduction

In Section 1, we have shown that repressive effects of uORFs effectively explain most of the variation in translation rates among the luciferase constructs assayed. Ideally, we could challenge our models of initiation with observations concerning the translation state of endogenous transcripts. In pursuit of this aim, our lab is attempting to measure this value for all mRNAs in aerial plant organs using polysome fractionation and microarray technology.

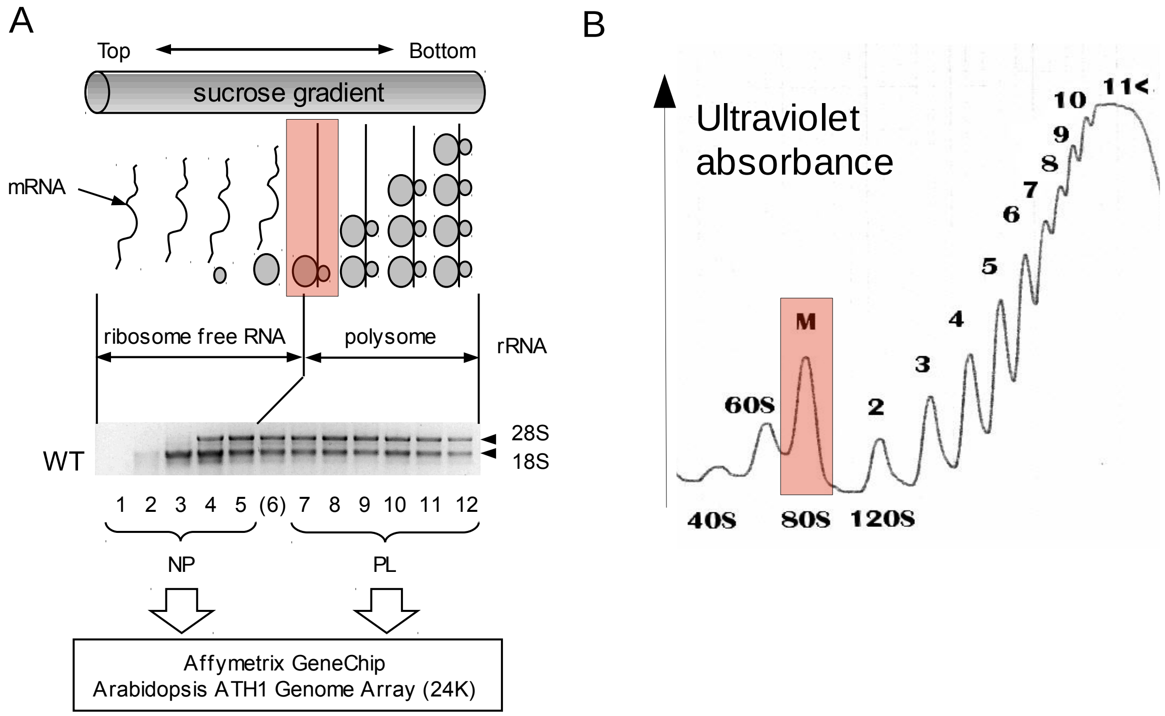


Figure 4.3: Polysome-microarray schematic.
A) Schematic of experimental pipeline for polysome microarray. B) Illustrative ultraviolet absorbance profile of total RNA after ribosomes have been fractionated by centrifugation through a sucrose gradient. 'M' indicates the monosomal fraction. Transparent red boxes in (A) and (B) indicate the same fraction of the gradient. Figure modified from [132].

Results and Discussion

Secondary structure in the 5' UTR influences translation state more than any generic mRNA feature

5' UTR, CDS, and 3' UTR sequences were downloaded from TAIR (Version 9). Sequences features were extracted using custom programs. Secondary structural features in the 5' UTR were predicted using UNAFold software [138]. Prior work has looked at the relationship between sequence features and translation state [139]. In that analysis, each feature was looked at in isolation, yet many transcript features are correlated. For example, uORF number and 5' UTR length are predictors of one another, but increases in sequence length also correlate with the probability of containing secondary structure. Thus, we began this analysis by plotting each attribute against every other attribute as well as against the TL of 3,774 Arabidopsis transcripts present in 4 biological replicates (Figure 4.4).

As expected there is a clear relationship between uORF number and the length of the 5'UTR, as well as between length of the 5' UTR and folding energy of its secondary structure ('Energy'). The free energy of the most stable secondary structure is distinctly related to the translation state, as seen previously [139]. This supports the notion that secondary structures in the 5' UTR have an inhibitory effect on translation efficiency. The length of the coding sequence and translation state are not correlated. This is somewhat contrary to the *a priori* assumption that a longer CDSs will generally have more ribosomes and, thus, should find itself in the NP state less frequently. As described in Chapter 1, some of this discrepancy is related to the reduction in ribosome density as CDSs get longer [13].

Whereas random expectation predicts only 12% of CDSs will begin with an AUG in strong context, we found that 67% of genes have this feature. This finding indicates that strong initiation is important, but, interestingly, AUG start codon context was not related to TL.

Similarly, though uORFs are known to repress translation initiation in all eukaryotic systems, uORF number appears to have only a small effect on TL. This may be due to the fact that uORFs attract ribosomes and therefore push mRNAs into the polysomal fraction, without actually contributing to the formation of gene product from the main ORF. Additionally, it is possible that measurements of uORF effect are dampened because TL indicates an mRNA's ribosome-occupancy, which may be only partially related to the initiation rate. Another interesting relationship involves TL and 3' UTR length. TL is optimal for mRNAs with 3' UTRs in the 200-400 nt range. Initiation context, though not related to TL, does appear to have many non-linear relationships with other sequence features, most notably the sharp drop associated with 3' UTR lengths less than 300 nts long. Together these data indicate that an optimally expressing mRNA in *Arabidopsis* will have little 5' UTR secondary structure, strong AUG context, and a 3' UTR length of 300-400 nts.

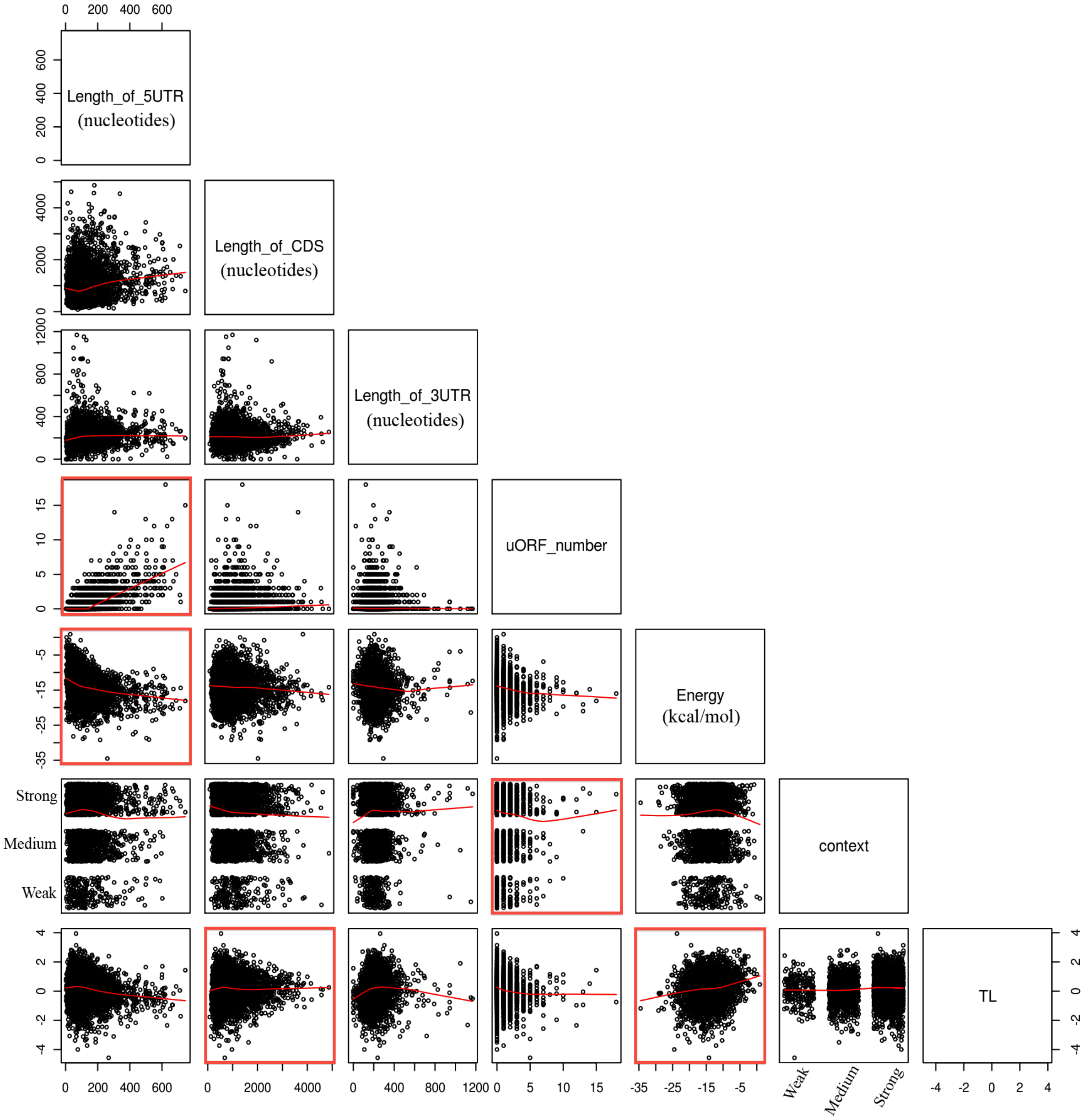


Figure 4.4: mRNA sequence features and translation state.
Scatterplots of mRNA sequence features shown in all possible pairwise combinations. Red lines are the LOWESS regression, which creates a locally-weighted polynomial fit to the data. The "Energy" attribute indicates the lowest energy (kcal/mol) from among all 60 nt windows within the 5' UTR.

mRNA secondary structure around the cap reduces translation state by 32%

In spite of many weak, nonlinear relationships between TL and other sequence features, secondary structure within the 5'UTR appears to have the greatest effect on whether or not an mRNA is associated with ribosomes. In a monumental experiment in *E. coli*, 154 versions of green flourescent protein were created such that they differed only in their synonymous site nucleotides [140]. The mRNA to protein ratios of these different constructs could vary by 250-fold. Strikingly, little of this variation is explained by codon adaption index, while 50% is explained by the RNA secondary structures these mutations produce around the 5' terminus of the mRNA. Though these experiments were carried out in a bacterial system, the results are born out in computational studies of yeast, where structure around the cap is selected against [141].

Table 4.2: Structural features at the cap are associated with a 32% reduction in translation state.

Structure (< -15 kcal/mol) within 60nts of cap				Structure (< -15 kcal/mol) anywhere in 5' UTR			
TL(+)	**TL(-)**	**total pairs**	**p-value**	**TL(+)**	**TL(-)**	**total pairs**	**p-value**
-0.259	0.291	44	0.0094	0.273	0.326	570	0.35

* TL (+) and TL (-) give translation state with and without secondary structures, respectively. TL is the log-ratio of mRNA abundance in polysomal and non-polysomal fractions. TL=log(PL/NP)

Based on results in the previous subsection, we addressed whether or not the observed relationship between translation state and secondary structure has a cap bias. In order to separate the role of secondary structure on translation state from that of other co-variables, such as uORFs and 5'UTR length, we collected pairs of sequences, one with secondary structure ($\Delta G \leq$ -15kcal/mol) and one without, in such a way that co-variables such as uORF number etc. were effectively constant within each pair (see *Methods*).

Table 4.2 shows the results from 44 pairs with secondary structure near the cap of the mRNA. The log-difference in translation state of 0.550 indicates that structure around the transcript cap reduces translation state by 32%. No significant difference was seen when transcripts with structures of equal or greater strength anywhere within the 5' UTR are used as the condition group. As predicted by experiments *in vitro [142]*, the scanning complex may be able melt most structures within the 5' UTR but may not be able to bind if the transcript cap is in a double-stranded state.

Methods

Translation state data were averaged from four independent replicates of polysome microarray data collected from 10-day-old wild-type Arabidopsis seedlings ([132] and unpublished data). This analysis was based on 3,774 genes that yielded expression data in each of the four experiments. Sequence features were parsed using custom Perl scripts and 5' UTR sequence data downloaded via TAIR (version 10). Each 5'UTR sequence was parsed for uORFs using a custom Perl module (uORF.pm). AUG contexts were considered strong (p_{cs}, [GA]nnAUGG, [GA]nnAUGn or nnnAUGG), and weak (p_{cw}, nnnAUGn), with n being any other nucleotide and brackets indicating alternatives . Secondary structure in the 5' UTR was assessed by extracting the 5’ UTR and the first 100 nt of the coding sequence (CDS). Then, we used UNAFold [138] with default parameters to predict the free energy of the secondary structure across a sliding window size of 60 nts, step size 20 nts, from the 5' cap to the beginning of the CDS.

To analyze the effect of cap structure, we selected sequences with secondary structure in the first 60 nt ($\Delta G \leq -15$kcal/mol) and paired each of them with a control sequence lacking secondary structure ($\Delta G > -15$kcal/mol). Sequence pairs were selected such that they had: 1) the same start codon context, 2) no uORFs, and 3) nearly identical length of 5’ UTR (± 20 nt), CDS (± 40 nt), and 3’ UTR (± 30 nt). This filtering resulted in 44 sequence pairs. A comparable approach was used for mRNAs possessing structure anywhere within the 5' UTR, which resulted in 570 pairs; this number is much higher than the 44 pairs because there are many more transcripts with 5' UTR secondary structure when the criteria is not limited by cap-proximity.

Section 3: The decoupling of ribosome- occupancy and ribosome-density

Introduction

In Section 2, we used translation state (TL=log(PL/NP)) as an indirect measure of ribosome occupancy, the fraction of mRNA molecules occupied by at least one ribosome. However, this method does not distinguish well between highly translated and moderately translated mRNAs. Alternatively, each of the 12-14 fractions from a polysome gradient - each corresponding to a specific number of ribosomes - can be applied to a microarray, thus obtaining the relative concentrations of all expressed mRNAs within that sample. The most common number of ribosomes with which an mRNA species is associated can be divided by that mRNA's main ORF length to get the ribosome-density of a transcript [13]. While ribosome-density does not reveal the kinetics of translation reactions, if one assumes fairly constant elongation and termination rates, ribosome-density is directly proportional to the initiation rate of mRNAs in the translational pool [143]. It is evident that high-resolution fractionation is more accurate in measuring translational efficiency than low-resolution fractionation and ribosome occupancy, but the cost is prohibitive. Using high-resolution data from yeast, we assessed the degree to which ribosome-occupancy predicts ribosome-density and how best to exploit polysome fractionation data in order to feasibly arrive at a more accurate measures of ribosome-density for the *Arabidopsis* transcriptome.

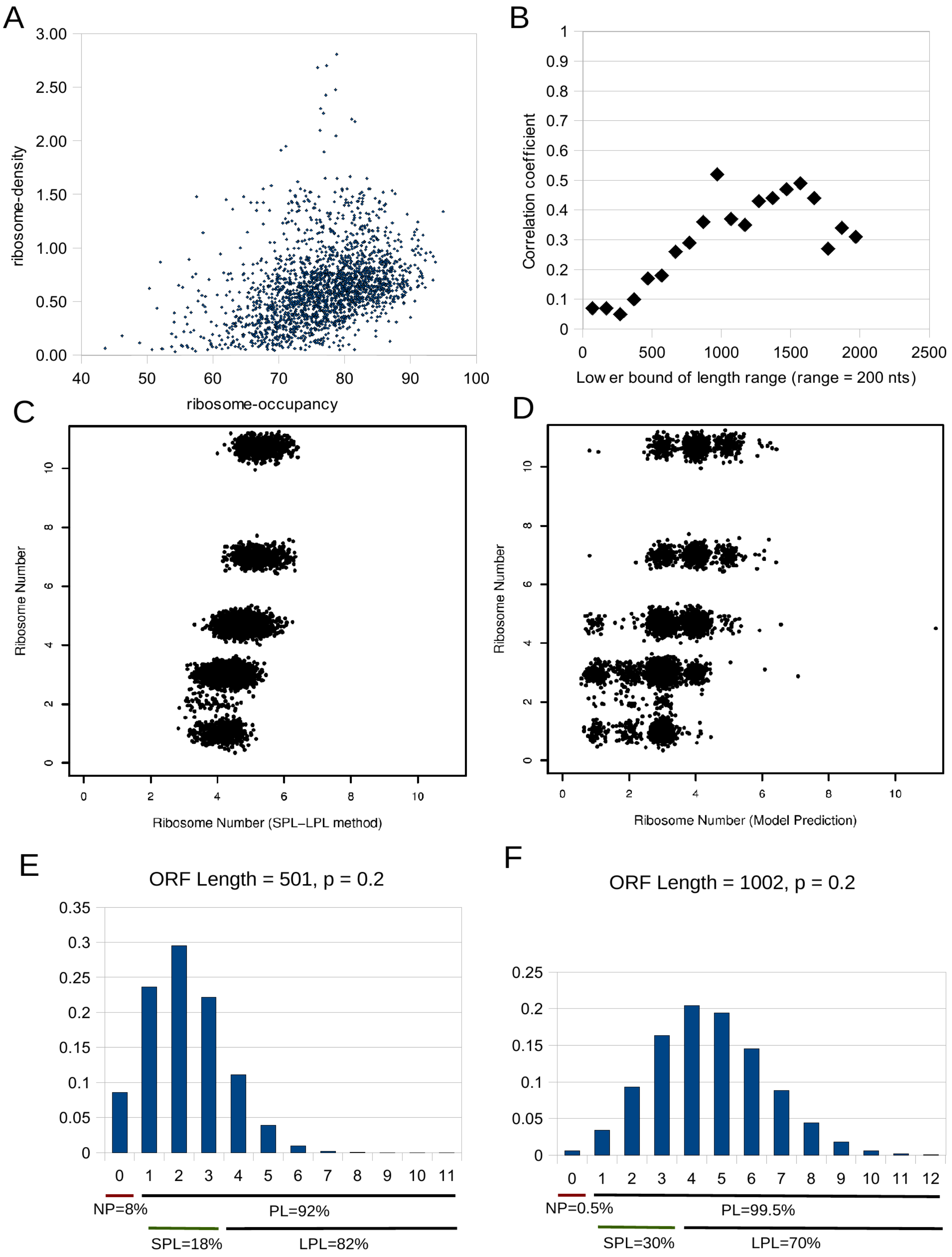
Results and Discussion

Ribosome-occupancy is a poor predictor of ribosome-density

As it stands, no studies have been carried out in higher eukaryotes using high-resolution polysome fractionation. Such resolution has been achieved in yeast, and we used that study in what follows [13]. Briefly, yeast cells were harvested at log-phase growth conditions and lysed. Lysate was centrifuged through a sucrose gradient in order to differentiate mRNAs with varying numbers of associated ribosomes. Each fraction was then applied to a two-color microarray. Each fractionated mRNA (red dye) was compared with a common reference (green dye) of 15mg total mRNA. Relative concentrations were then normalized to a spiked-in control of known concentration, a step that allows comparisons across fractions. Ribosome-*occupancy* is the fraction of a gene's mRNA associated with at least one ribosomes relative to the total amount of mRNA for that gene. Ribosome-*density* is calculated as the ribosome number associated with the peak fraction among the polysomal fractions divided by (mORF length / 100); thus, the units of ribosome-density are number of ribosomes per 100 nts.

Ribosome-occupancy is a poor predictor of density (Figure 4.5A), suggesting that, even in simple eukaryotes there exist two pools of mRNAs that should be considered independently. In fact, the molecular underpinnings of variation in ribosome occupancy are somewhat unclear. Many transcripts may be in a non-translating state, either because they are highly transcribed and have not been exposed to the cytoplasm ("kinetic effect") or because they are actively partitioned in the nucleus or cytoplasm and made inaccessible to ribosomes ("partitioning effect") [19].

Figure 4.5 (next page): Assessing ribosome-density estimation methods using yeast data A) Scatterplot of ribosome-density versus ribosome-occupancy. B) Correlation coefficient between density (ribosomes per 100 nts) and occupancy (percent of mRNAs associated with 1 or more ribosomes) plotted against specific ORF length ranges. C) Scatterplot of the number of ribosomes most frequently associated with a transcript versus the same value as predicted by the SPL-LPL method. D) As in (C) except that the binomial model is used instead of the SPL-LPL method. In both (C) and (D), because the actual data and the binomial model prediction are integers, these were randomly shifted around the integer value so that all the data can be visualized. E) Illustration of the binomial model and its predictions for an ORF length of 501 and a p parameter set to 0.2. F) Similar illustration using ORF length of 1002.



A
3.00
2.50
2.00
1.50
1.00
0.50
0.00
ribosome-density
40
50
60
70
80
90
100
ribosome-occupancy
B
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Correlation coefficient
0
500
1000
1500
2000
2500
Low er bound of length range (range = 200 nts)
C
Ribosome Number
0
2
4
6
8
10
Ribosome Number (SPL–LPL method)
D
Ribosome Number
0
2
4
6
8
10
Ribosome Number (Model Prediction)
E
ORF Length = 501, p = 0.2
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0 1 2 3 4 5 6 7 8 9 10 11
NP=8%
PL=92%
SPL=18%
LPL=82%
F
ORF Length = 1002, p = 0.2
0.25
0.2
0.15
0.1
0.05
0
0 1 2 3 4 5 6 7 8 9 10 11 12
NP=0.5%
PL=99.5%
SPL=30%
LPL=70%

Correlation between ribosome occupancy and ribosome density depends on ORF-length

Because very short transcripts can only harbor one or two ribosomes, it is assumed that, if ribosome-occupancy and ribosome-density are related, short transcripts will show the highest correlation between these two attributes. For example, if the mORF is only 100 nts long, then, because of spatial constraints, it can harbor ~2 ribosomes at most. Thus, any reduction in translation efficiency will move many mRNAs into the non-polysomal fraction as well as reducing the ribosome-density. We tested this expectation by looking at transcripts categorized by their lengths. Each category has a range of 200 nts and overlaps the neighboring categories by 100 nts.

Strikingly, our results were effectively opposite our expectation (Figure 4.5B). For shorter transcripts, there is significant decoupling between ribosome-occupancy and ribosome-density. The distinct rise in correlation as the ORF length range increases, suggested that shorter (and, to a lesser degree, longer) transcripts are more susceptible to this decoupling. One explanation is that shorter transcripts produce highly expressed proteins such as ribosomal and histone subunits, and these may be under unique forms of translation control [42,144]. Additionally, the differential decoupling of shorter transcripts may reflect kinetic effects associated with higher transcription rates: many mRNAs have not yet been exposed to ribosomes.

The use of small and large polysomal fractions reduces the true variation in ribosome density

Because ribosome-occupancy is a poor predictor of ribosome-density, we are exploring cost-effective ways to acquire a more accurate measure of initiation efficiency. In prior studies, researchers generated a small-polysome and a large-polysome fraction (SPL and LPL, respectively) and calculated the ribosome density based on an estimate of the average or median number of ribosomes found within each fraction [143]. In other words, if an mRNA is found in the SPL (or LPL) fraction, one must ask, in the absence of other information, what is the best guess as to the number of ribosomes to which this mRNA is attached. Such a guess can come from a profile of the UV spectrum (Figure 4.3B), where the height of a particular fraction represents the contribution of the given ribosome number to the entire polysome population. We explored the efficacy of the SPL-LPL approach using the high-resolution yeast dataset. Moreover, since we knew the contribution of each ribosome fraction to the pooled SPL and LPL fractions, we could generate our best guesses based on these elemental values; thus, we evaluated this approach using the best possible guess.

The SPL-LPL method over-predicted the ribosome number of mRNAs associated with <2 ribosomes (Figure 4.5C). Alternatively, the method under-predicted the ribosome number when an mRNA was actually associated with >6 ribosomes. Such bias was anticipated: because the best guess of ribosome number for the SPL fraction was 2.2 (see *Methods*), the ribosome-density of transcripts that harbor only 1 ribosome gets inflated. Likewise, because the best guess for the LPL fraction was 6.9, higher values are deflated. Since ribosome number (and resultant ribosome-density) is the value we would be attempting to estimate, there is no clear way to correct for this bias using the SPL-LPL approach. It should also be noted that the bias is not linear, suggesting that factors other than statistical averaging are confounding this method.

A simple mathematical model of ribosome-density improves resolution

It is clear from the previous subsection that we lost substantial information by applying a best guess as to the ribosome number across all transcripts. If certain assumptions are made about the distribution of the data then a more accurate estimate of ribosome number might be achieved that alleviates this difficulty. The initiation event that starts protein polymerization is a binary event with a certain probability of occurring or not occurring. If a binary event has a fixed probability (such as a coin-flip), then the total number of events expected to occur in a particular number of trials has a binomial distribution. One can imagine the ORF of an mRNA as having a certain number of slots into which a ribosome can fit. Whether or not a slot is filled is a result of whether or not a ribosome has initiated. At steady state conditions, the number of possible slots in an ORF can be considered the number of trials. The ribosome density profile can then be modeled as a binomial distribution which is determined by the probability of initiation (Figure 4.5E-F).

We assumed a maximum effective ribosome size of 40nts. In order to generate the number of slots for the model, we divided each ORF by this value. Given a random starting point between 0 and 1, the *p* parameter was fit for each gene, such that the squared difference between the SPL fraction as predicted by the model and as seen in the data was minimized.

The binomial model improves on the estimation of mRNAs associated with 1 and 2 ribosomes, but these are still over-predicted in general (Figure 4.5D). Additionally, the binomial model does not resolve the problem of underestimating mRNAs with a high ribosome number. A closer inspection of the data used to predict ribosome number indicated that the SPL fraction for mRNAs with a high ribosome number is simply too high if in fact the binomial model is a somewhat accurate representation of initiation process. For example, if the ribosome number is 10, the binomial model would require a very small frequency of a gene's mRNAs to be in the SPL fraction. This is rarely the case. It is possible that a significant proportion of the SPL and LPL signal is simply a result of underlying noise in the data. Such signal would drive the binomial model to over-prediction of low ribosome numbers and, more dramatically, under-prediction of higher ribosome number. We tested this possibility by subtracting the lowest signal across all ribosome fractions from each ribosome fraction, and rerunning the analysis. Though there was a minor improvement (not shown), fit was still comparable to Figure 4.5D. In summary, this discrepancy may speak to a violation in our assumptions concerning the steady-state conditions of mRNAs in the translating pool. We described above how the kinetic effects of transcription can affect polysome occupancy. They may also affect the perceived rate of initiation because many mRNAs have not had sufficient time to accumulate ribosomes and so fall into the SPL fraction. In yeast, it has been estimated that the median time to steady-state ribosome loading is ~2 min, whereas the median half-life of an mRNA is ~1.5 hrs [145]. This finding speaks against a kinetic explanation of SPL enrichment as most mRNAs will be in steady state relative to their ribosome number. An alternative explanation is that the "pioneer round" of translation that occur upon exit of the nuclear pore has much slower kinetics than cytosolic initiation and elongation [146].

Outlook on improved resolution

Pooled fraction polysome microarray experiments are commonly used to assess the translation state of all mRNAs in the transcriptome. Typically, these are used in terms of condition/control experiments [22,23,132], but they are also used for absolute quanitification of translation rates [143]. Our results have shown that, depending on the conclusions being drawn, variation in NP/PL ratios may be indicative of other factors besides initiation rate, emphazing the need to consider ribosome-occupancy and ribosome-density as two distinct biological states. Additionally, SPL-LPL methods are likely to give a false impression of the true variation in initiation rates across the transcriptome.

Is it possible to predict the number of ribosomes with which an mRNA is associated? We have shown that two methods, one used previously [143], and one developed by us based on underlying biological assumptions, were both unsuccessful. Future elaborations of models that account for transcription rate may improve fit. Additionally, more sophisticated noise reduction may make the binomial model more appropriate. It should be noted that the microarrays used for the yeast data are more likely to be susceptible to cross-hybridization error than those in current use in most polysome assays, which account for mismatches .

It should be noted that the binomial model presented here has the useful feature that it predicts the component of the translationally-active pool of mRNAs that are found in the nonpolsomal, NP, state. Thus, it would be able to account for this component of the profile in attempting to diagnose molecule defects related to ribosome-occupancy.

Methods

Yeast data was downloaded from http://genome-www.stanford.edu/yeast_translation/data.shtml. Summary data was directly used to assess the Pearson correlation coefficient of ribosome-occupancy and ribosome-density. For length analysis, the Pearson correlation coefficient were assessed for all genes within a particular length range (Figure 4.5B). A sliding window of 200nts was used, starting with the shortest mORF (70 nts) and moving in 100nt increments.

For the SPL-LPL method (Figure 4.5C), fractions containing 1-3 ribosomes were pooled (SPL). Fractions containing >3 ribosomes (LPL) were also pooled. The relative proportion of each was multiplied by best guess for the number of associated ribosomes in a pooled fraction. The best guess was calculated from the data by counting all replicates from all genes that fell within a particular peak fraction. These values were pooled based on where we decided to divide the PL fraction (>3 ribosomes), and the mean was taken for each pool separately. This is expressed in the following equation; note, $s + l$ does not equal 1 because there is a certain frequency of mRNAs in the NP fraction:

$$R[s,l]=\frac{sr_s+lr_l}{s+l}$$

where,

R = number of ribosomes associated with a transcript, s = frequency of a transcript in SPL fraction, l = frequency of transcript in LPL fraction, r_s = best guess of ribosome number for SPL fraction, and r_l = best guess of ribosome number for LPL fraction.

For the binomial method (Figure 4.5D-F), an effective ribosome size of 40 nts was used.

Each ORF was divided by this value to give the number of potential slots or trials. The frequency of successful initiation events (*p*) was then fit to the data using an evolutionary strategy and, alternatively, the downhill simplex algorithm, both of which were implemented in Perl and are available at http://search.cpan.org/~pjb/Math-Evol-1.12/Evol.pm and http://search.cpan.org/~tom/Math-Amoeba-0.05/lib/Math/Amoeba.pm, respectively. For both of these algorithms, default parameters were used with limits on *p* between 0 and 1. *p* was found for each gene by minimizing the following fitness function:

$$f[p]=\sum_{n=r}^{i=1}(s_i-m[p])^2$$

where,

$$m[p]=\frac{\boldsymbol{B}[3,p,\frac{l}{40}]-\boldsymbol{B}[0,p,\frac{l}{40}]}{1-\boldsymbol{B}[0,p,\frac{l}{40}]}$$

and,

r = number of replicates, s_i = frequency of a transcript in the SPL fraction for an individual replicate, m = the predicted frequency a transcript in the SPL fraction as a function of the probability of initiation at its main ORF, x = the ribosome number, p = the probability of initiation, l = ORF length, and $\boldsymbol{B}$ = the cumulative binomial probability as a function of number-of-positions-filled, probability of initiation, and number-of-possible-positions.

In brief, the numerator of the *m*[*p*] function represents the probability of an mRNA having 1, 2, or 3 ribosomes attached. This value is then divided by the probability of an mRNA being in the polysomal fraction (1 - probability_of_an_mRNA_being_in_the_non-polysomal_fraction).

For consistency, we chose to use the same fractions to combine for the SPL (and, hence, LPL) pools as was used in the SPL-LPL method assessment.

Addendum: A nucleotide-specific kinetic model of translation initiation

Introduction

When given a system of many constituents and the interactions of those constituents, human intuition often fails to make the correct prediction as to how that system will behave over time. Yet, scientists can recognize when the correct behavior has been predicted. Thus, a critical test of systems biology is whether or not it can take the same input and do better than intuition at predicting emergent properties or phenotypes. Translation, particularly with regard to a large class of mRNAs, is just such a system. The translation of these mRNAs is known across many conditions and in a variety of genetic backgrounds. Many of the constituents and their interactions are known. Yet, whether the data fit what is known or not is still unclear.

Results in Chapter 4, Section 1 imply nucleotide-specific molecular events commonly occur during translation initiation on the AtbZip11 5'UTR. Our 'sum-histories' model did not have the granularity to estimate such events. For example, the weak start-codon context parameter was estimated to be much higher than known values; ribosome stacking resulting from multiple initiation/termination events in the 5'UTR could be responsible for the enhanced initiation efficiency [137]. Moreover, other regulatory events involving mRNA secondary structure and/or RNA-protein interactions are intimately related to the presence of scanning and elongating ribosomes, both in the 5'UTR and in the coding region [137,140,141]. In brief, as the ribosome reads the message, it may encounter steric hindrance to its forward progression from other ribosomes or from proteins bound at specific locations. This mixture of elongation and stalling results in complex, nonlinear relationships. Adding further complexity, some mRNAs are under combinatorial control by multiple interacting proteins [47,77,78]. While these may not have a dramatic effect on bZip11 translation, we would like for our models of initiation to be as general as possible. Though we can model these events indirectly by the inclusion of constants that reduced efficiency in sequence specific manner - as with the uORF2b penalty in the 'sum-histories' model - ideally we could be more biochemically exact.

As discussed in Introduction to Section 1, pre-existing models do account for more specific molecular events related to initiation [130,131]. Yet these models focus on the sequential binding and activation of factors that interact with a translated mRNA. They are not robust to the removal or addition of sequence elements that might affect translation. Moreover, when processivity effects and ribosome-ribosome interactions are modeled, they are considered to scale across the entire mRNA region, whereas in reality, these are highly localized events. As described below, our approach is sequence-centric in that the user can specify particular sequence patterns at which molecular events occur; the computational framework then accounts for any spatial constraints related to the molecules involved. This approach allows a biologist to quickly test their intuition concerning the potential effects of a sequence element or a mutation.

Results and Discussion

An exact stochastic simulation of initation and reinitiation

Our modeling approach extended the work of Niemitalo, *et al.* in which each nucleotide position of an mRNA encoding a human protein disulfide isomerase was monitored for endoribonuclase exposure [147]. Our model took as starting conditions a single 5'UTR (mRNA molecule) and a saturating population of the other elemental components: large and small ribosomal subunits, eIF3, and eIF2-tRNA. Simulations end with start codon recognition at the mainORF. The 5' UTR was subdivided into its constituent nucleotides. Because there is only one representative species of each nucleotide position, all resultant complexes are either present or absent (1 or 0). This, in turn, was how ribosome stalling by other ribosomes was modeled: forward movement can not occur if the upcoming nucleotide is unavailable. Important processes not yet addressed include RNA secondary structure, rare codon stalling effects, peptide stalling effects, and internal ribosomal entry sites. The robust framework of this model would allow these mechanisms to be incorporated in the future without fundamentally altering the current configuration. In fact, because the model treats the initiation/elongation process as a set of chemical reactions perpetuating from a single nucleotide position to one of two neighboring positions (forward or backward), any of these processes could be accounted for with minor effort.

Instead of manually generating all chemical reaction equations, they were generated by a Perl script using a spreadsheet template (Table 4.3). This method of organization was necessary because of the highly repetitive nature of processes that occur during initiation, such as forward scanning, elongation, etc., coupled with the point processes that can occur concurrently. After the reactions were generated, the system was simulated using the Gibson-Bruck Next Reaction algorithm, which is an optimized version of the Gillespie First Reaction algorithm [148]. Briefly, the propensity of a reaction is calculated as the population size of each reactant multiplied by the given reaction constant. This propensity value describes a function of probable times. A random number is generated and used in the time function; each reaction then has a time. The reaction are sorted based on this time value. The smallest value is selected to occur (hence "First Reaction Method"). The Gibson-Bruck Next Reaction Method uses a graph data structure to significantly decrease the amount of updates required for each cycle.

Simulation can recapitulate many features of uORF-mediated suppression

Initially, to test the utility of this simulation framework, we used data from transfected mammalian cell cultures that directly characterized how re-initation depends on uORF length [133] and the length of space between a uORF stop and the next downstream start [135]. These uORF length effects (Figure 4.6A) and spacer effects (Figure 4.6B) were reproduced by the simulation. In the case of changes in spacer length, the first-order kinetics are apparent. Alternatively, uORF length effects appear to have a fairly precise threshold, suggesting that a sequential step may be required in order to loose re-initiation competence during uORF translation. Running the model on a series of 5'UTRs of Arabidopsis AtbZip11 that differ in their uORF pattern [136] (see Figure 4.1A) yielded a fit that, though not replicating the breadth of values, approximates their values relative to one another (Figure 4.6C). It should be noted that the parameters for this test were not rigorously optimized; doing so might further improve the ability of the model to predict the full range of initiation efficiency seen in the AtbZip11 data (Figure 4.6C). In any case, this test showed that the 'sum-histories' model from Section 1, which used an extended set of this data, and these simulations are complementary.

Table 4.3: Reaction template for exact stochastic simulation of translation initiation.

For 'reactant' and 'product', complexes are indicated by underscores between constituents. 'forward' and 'reverse' are rates used in the propensity function of the simulation. 'zone' describes where on the mRNA a reaction is allowed to occur; reactions can occur anywhere on the mRNA that matches the given regular expression. 'diffusion behavior' defines if and how a given chemical species moves on the mRNA. 'bind size' defines the number of nts covered by the chemical species.

molecular process	**reactant**	**product**	**forward**	**reverse**	**zone**	**diffusion behavior**	**bind size**
preinitaion complex formation	eIF2+tRNA	eIF2_tRNA	100	0	SINK	0	0
preinitaion complex formation	eIF3+eIF2_tRNA	eIF3_eIF2_tRNA	100	0	SINK	0	0
preinitaion complex formation	eIF3_eIF2_tRNA+small	eIF3_eIF2_tRNA_small	100	0	SINK	0	0
mRNA cap binding	eIF3_eIF2_tRNA_small	eIF3_eIF2_tRNA_small	100	0	S[ATGCVXYZ]	BIND	30
scanning	eIF3_eIF2_tRNA_small	eIF3_eIF2_tRNA_small	30	10	.	move,1	30
start site identification (weak context)	eIF3_eIF2_tRNA_small	eIF3_tRNA_small+eIF2	10	0	.{$bindPosition}[^VYAG].{2}ATG[^G]	0	30
start site identification (moderate context)	eIF3_eIF2_tRNA_small	eIF3_tRNA_small+eIF2	20	0	(.{$bindPosition}[VYAG].{2}ATG[^G]\|.{$bindPosition}[^VYAG].{2}ATGG)	0	30
start site identification (strong context)	eIF3_eIF2_tRNA_small	eIF3_tRNA_small+eIF2	30	0	.{$bindPosition}[VYAG].{2}ATGG	0	30
translation initiation	eIF3_tRNA_small+large	eIF3_tRNA_small_large	2	0	.{$bindPosition}ATG	0	30
translation initiation	eIF3_tRNA_small_large	eIF3_small_large+tRNA	1000	0	.{$bindPosition}ATG	0	30
elongation	eIF3_small_large	eIF3_small_large	10	0	.{$bindPosition}[ACTGN]{3}	move,3	30
subunit loss during elongation	eIF3_small_large	small_large+eIF3	3	0	.{$bindPosition}[ACTGN]{3}	0	30
elongation	small_large	small_large	10	0	.{$bindPosition}[ACTGN]{3}	move,3	30
termination (no resumption of scanning)	small_large	TOSINK	100	0	.{$bindPosition}(TAG|TGA|TAA)	0	30
termination	eIF3_small_large	eIF3_small	100	0	.{$bindPosition}(TAG|TGA|TAA)	0	30
resumption of scanning	eIF3_small	eIF3_small	30	10	.	move,1	30
reacquisition of the ternary complex	eIF3_small+eIF2_tRNA	eIF3_eIF2_tRNA_small	3	0	.	0	30
main ORF translated (protein made)	eIF3_small_large	Protein+TOSINK	10000	0	B	0	30
main ORF translated (protein made)	small_large	Protein+TOSINK	10000	0	B	0	30
main ORF skipped	eIF3_eIF2_tRNA_small	TOSINK	10000	0	B	0	30
main ORF skipped	eIF3_small	TOSINK	10000	0	B	0	30

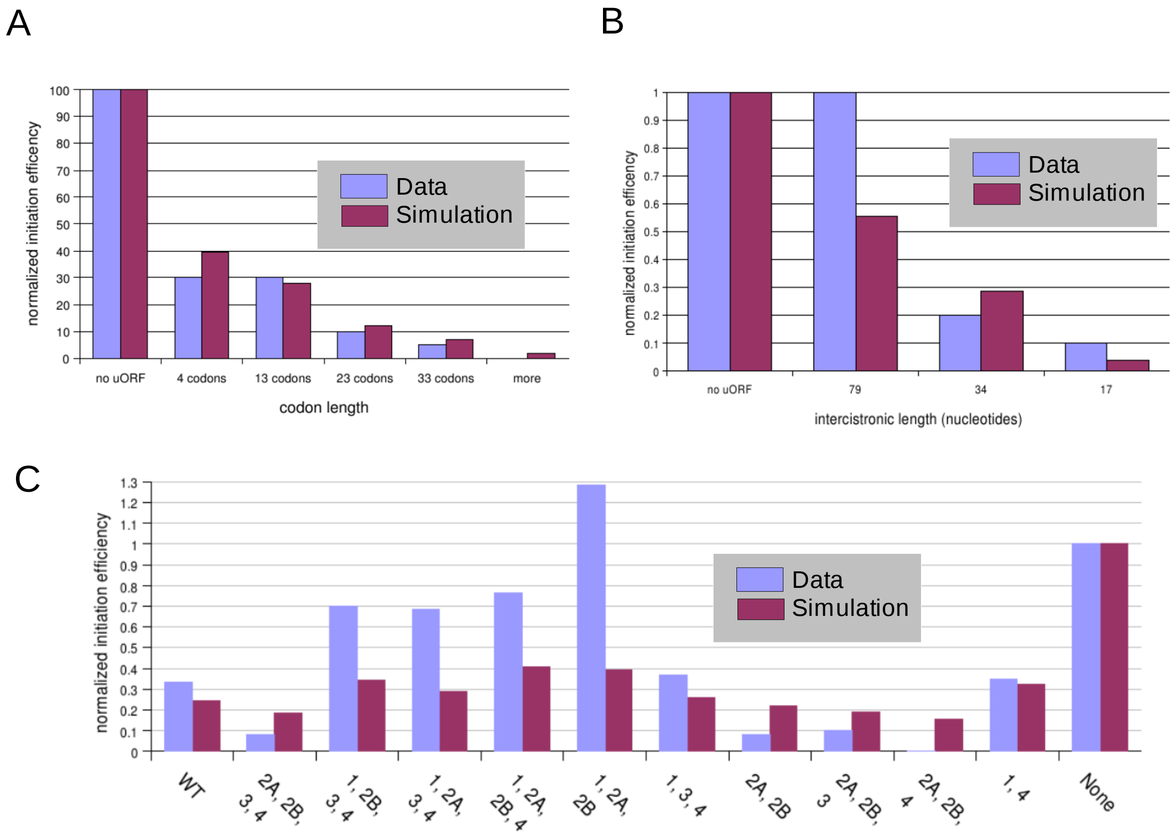


Figure 4.6: Comparison of actual data with results from simulated translation initiation. (A) and (B) published experimental data that measured the effect of (A) uORF length [133] and (B) length of the spacer between uORF and main ORF [135] on translation of the mainORF downstream. (C) Experimental data from a series of 5' UTRs of the Arabidopsis bZip11 gene that vary in their uORF pattern [136] (see Figure 4.1B). Each of the 12 constructs is labeled with the numbers of the uORFs that are present. WT AtbZip11 has 5 uAUGs that form 4 uORFs, 1 2a/2b, 3, and 4. None, no uORFs.

Model can be used to make predictions concerning ribosomal footprints

As discussed in the Introduction, technical advances in the RNA sequencing coupled with older ribosome protection assays have allowed for a genome-wide characterization of the translation process by ribosome footprinting [16]. Specifically, the likelihood of finding a ribosome on a specific nucleotide has been measured across the entire transcriptome of a yeast cell. These data have nucleotide-specific resolution: the kinetics associated with the first position in a codon - which are driven by the time needed to acquire the correct aminoacylated tRNA - can be clearly differentiated from those of the second and third positions, which are driven by ribosome translocation after peptide bond formation. Such resolution can speak to many biochemical events, but currently it is difficult to formulate such complex kinetic models. Our simulation framework allows some of that formulation to be automated.

Figure 4.7 illustrates results from simulation of initiation on a simple 5' UTR using the reaction template in Table 4.3, scoring whether a nucleotide is protected by a ribosome or exposed. As expected, there is a brief delay until the positions farther from the cap begin to interact with the scanning complex. Position 1 is exposed much more frequently than other sites. This is due to the ability of the simulation to account for spatial aspects of the scanning complex. In order for cap binding to occur, at least 18 nucleotides must be exposed ('bind size' and 'bind position' in Table 4.3). As modeled, initiation occurs at roughly the same speed as the forward biased scanning rate; thus there is little ribosome stacking associated with this step. Because the rate of cap-binding is not limiting, stochastic differences in scanning rate do result in ribosome-ribosome interactions as well as erratic dynamics in site exposure. Additionally, these lead to a 5' to 3' bias in exposure, excluding cap-proximal positions.

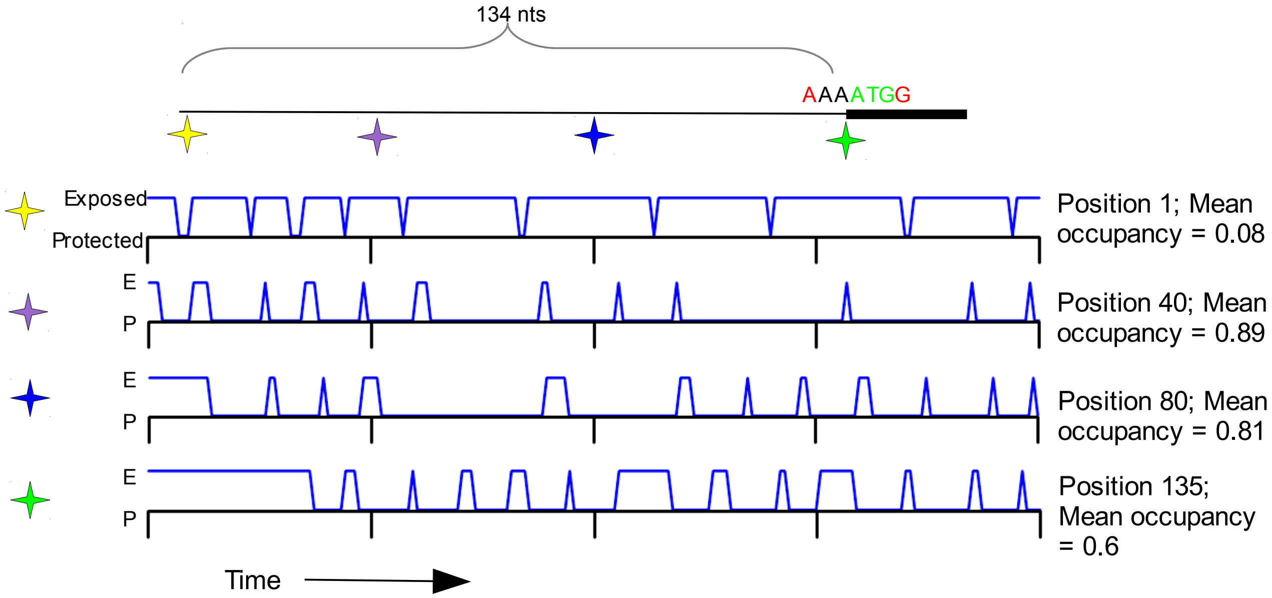


Figure 4.7: Using kinetic simulation framework to make nucleotide-specific predictions concerning ribosome occupancy on a generic 5' UTR sequence.
Schematic of the sequence used to simulated the initiation events. Stars indicate positions for which the protected-exposed plots are shown.

Methods

Reactions and their rates were entered via a custom template (Table 4.3). Because these rates are in most cases unknown, parameters were manually selected to reflect their presumed values relative event rates that are fairly well known, such as the elongation rate of ~30 nts/sec [21]. The input sequences - a 5' UTR containing the main start codon plus its +4 position for context assessment - were parsed to produce a list of the position of all uAUGs, their respective stop codon positions, and the context of the respective start codons. uAUGs that overlap the main AUG have the full sequence length as their stop codon position. This list was used by another module to build a list of chemical reactions that coincide with the positional constraints. The total system of reactions was then simulated using the Gibson-Bruck next reaction algorithm, implemented under the Dizzy simulation library written in Java [149] and ported to Perl via

Inline::Java. Initiation rate of a 5' UTR was reported as the stable slope of the initiation events versus time. Stability was considered achieved when the Pearson correlation coefficient of initiation versus time was >0.75 (not including the time before the first initiation event) and statistically significant ($<10^{-2}$).

Chapter 5: Perspectives and Future Directions

Post-transcriptional regulation is emerging as a major factor in predicting gene expression levels. Translation initiation is a focal point of post-transcription control. The canonical model of translation initiation predicts that amino-acid polymerization will begin at the first AUG codon encountered by the scanning pre-initiation complex. Surprisingly, ~30% of plant mRNAs contain an initial AUG that is not associated with the major ORF. The resultant upstream open reading frame (uORF) should dramatically inhibit protein expression. Using numerous translation assays and a computational model of initiation, we showed that eukaryotic initiation factor 3 contributes to uORF tolerance during the elongation phase (Chapter 4, Section 1). Additionally, our computational model indicates that known elemental processes of initiation and re-initiation are sufficient to predict the inhibitory effects of complex uORF structures.

The presence of factors, such as eIF3, that diminish the effect of uORFs, led us to assess if uORFs are in fact neutral features in the transcriptome (Chapter 3). Based on patterns of triplet bias, we found that a large proportion of gene families cannot tolerate uORFs. These families appear to be related to fundamental cellular processes such as translation and nucleotide synthesis. Interestingly, a comparison of orthologous 5' UTRs revealed that AUGs are conserved at a higher frequency than any other triplet in the 5' UTR, indicating that uORFs have been exploited by Nature to regulate a subset of genes.

To address which genes might be under uAUG/uORF-mediated regulation, we identified 5' UTR elements that are conserved between Arabidopsis and five other families of dicot plants (Chapter 2). uORFs with peptide-independent function appear to be conserved at least as often as those that encode a peptide with synonymous/non-synonymous substitution bias. We additionally created a computational pipeline to categorize all conserved UTR elements emerging from the study. Other conserved elements in the 5' UTR are common, particularly purine-rich sequences. For contrast, we used the same pipeline to categorized conserved motifs in the 3' UTR. This region generally harbors more complex motifs included likely PUF-binding elements and sequences that are likely to be involved in the localization of Expansin mRNAs. These data have implications for the RNA regulon concept in plants.

These findings, their implications, and their experimental and theoretical precedent in the literature are summarized in Table 5.1.

Table 5.1: Major findings of this dissertation.

Findings	Implications	Precedent
Elemental processes of initiation-reinitiation can explain most of the suppressive effects of uORFs.	Even in complex uORF arrangements the canonical scanning model is sufficient explanation for reduced expression levels.	
Molecular defects of eIF3h are related to the uORF elongation phase.	eIF3h supports eIF3-ribosomal interaction during early phase of elongation.	[5,44]^{p}
uAUGs are depleted from the 5' UTR.	Nature selects against uAUG in a large class of genes.	[86,116,118]^{vf}
uAUGs, when present, are conserved across short evolutionary distances	In some case, nature selects against removal of uAUGs once established.	[86,116]^{vf}
5% of genes require a uAUG across the large evolutionary distances between monocots and dicots.	uAUGs have fundamental importance to flowering-plant biology.	[117]^{f}
uAUG-depleted genes function in translation and nucleotide synthesis, as well as auxin synthesis and stress response.	These proteins cannot tolerate even mild suppressive effects of an uAUGs.	[24]^{v}
The 5' UTR contains additional coding potential in the form of Conserved Peptide (CP)uORFs.	Because it is not a part of the CDS, this coding potential regulates expression, likely through cis-acting mechanisms.	[25-27,119]^{pvifb}
Additional conserved coding sequences start with non-AUG start sites and encode N-terminal extensions of the main ORF.	These N-terminal extensions lead to functionally distinct protein isoforms.	[76,119,150] ^{pvb}
Many conserved 5' sequence elements contain AUG, but the downstream uORF is not conserved at the peptide level.	Peptide-independent repression via uAUGs and uORFs is also functionally constrained in evolution.	[24,86,117]^{vf}
Polypurine repeats are conserved in the 5' UTR.	These repeats are functioning at the post-transcriptional level.	
Putative PUF-protein binding elements are conserved in the 3' UTRs of many targeted mRNAs.	Plants use Puf elements and presumably PUF proteins to regulate gene expression of functionally diverse mRNAs.	[49] ^{p}
Multiple Expansin genes contain a conserved 3' UTR element.	The conserved element is responsible for the known subcellular localization of some expansin transcripts.	[81]^{p}
Ribosome-occupancy and ribosome-density are highly decoupled in yeast, particularly for short transcripts.	mRNAs exist is two main states, translationally active and quiescent.	[151]^{f}
RNA secondary structure around the cap reduces translation state by 32%.	Cap-structure channels mRNA into a translationally quiescent state.	[140,142,151,152]^{vfb}
The binomial model is a poor predictor of ribosome-density.	Kinetic effects of transcription and/or large variation in ribosome number confound simple model of ribosome-density.	

^{p}From plants, ^{v}vertebrates, ^{i}invertebrates, ^{f}fungi or ^{b}bacteria

These results suggest a more refined set of biological questions. I discuss these questions below and present future analyses and experiments, which may begin to address them.

Exploring bioinformatic leads

mRNA specificity for a majority of the ~200 RNA-binding proteins in *Arabidopsis* is unknown [153]. In terms of those that are known, *Arabidopsis* has 26 PUF-domain-containing proteins [154]. While a subclass of PUF domains bind non-canonical motifs, most bind TGTA-containing sequence elements in the 3' UTR [35]. Interestingly, we only find 16 conserved putative PUF binding sites (Table 2.7). Comparably, a fungal genome comparisons within the genus *Aspergillus* revealed 48 such conserved sites [155]. While we have certainly missed some true positives, the relationship between PUF-protein and client mRNA may be near one-to-one in plants. This is suggested by yeast-3-hybrid assays in plants [41,49,156] but not by *Drosophila* pull-down experiments [48], which showed that one PUF-protein may bind mRNAs from ~1,000 different genes. It still remains to be seen what proportion of these 1,000 interactions are evolutionarily constrained, but our comparative sequence analysis would predict a fairly small fraction.

We generated an extensive list of conserved mRNA sequence elements, which we hypothesized to be acting at the post-transcriptional level. Though some of these elements have pre-existing support for such a hypothesis, most are uncharacterized. Some Expansin mRNAs show subcellular localization. These mRNAs also have clearly defined mRNA elements in the 3' UTR (Figure 2.6). Do these fit the *XYZ* model of translation repression and localization whereby an RNA-binding protein bound to the 3' UTR suppresses translation initiation (see Figure 1.2 and Chapter 1 text)? If so, it would, to our knowledge, be the first demonstration that this model is applicable to plants. Confirming and extending this model in plants might entail 1) Identifying Protein *X* via pulldown with the Expansin target [157], 2) Identify transcriptome wide targets using Protein *X* as bait and performing RNA-seq or microarray on the resultant elution [35], 3) Assay sub-cellular localization of identified transcripts [81], 4) Assess the translation state of all genes with regard to the effect of Protein *X* knockout or knockdown [132].

Coupling functional and comparative transcriptomics

In Chapters 2 and 3, I described RNA sequence elements that are critical to plants, an inference based on their resistance to mutations across large evolutionary distances. In Chapter 4, I discussed genome-wide assays of translation state. These two approaches naturally represent two orthogonal dimensions by which to elucidate post-transcriptional pathways in plants, one reinforcing the other [158]. Co-regulated genes can be searched for overrepresented motifs and these motifs compared across lineages. Emerging polysome microarray data from multiple conditions and genetic backgrounds lend themselves to element identification in a manner that has helped understand transcriptional regulatory modules [159]. Currently, the search for motifs within plant transcripts that are co-regulated at the translational level has been inconclusive [22,23], but more sophisticated approaches that integrate comparative sequence analysis could dramatically sharpen resolution (see [160] and Chapter 2, Addendum). Our Anchored-MEME approach is appealing because it incorporates phylogenetic data into an already established and well-maintained tool for identifying motifs in co-regulated and/or co-bound genes [127].

Still another exciting avenue of research would couple comparative sequence analysis with comparative functional genomics [161]. Though polysome microarrays have yet to be performed in rice, this would be an ideal touchstone with regard to similiar work in *Arabidopsis*, allowing us to begin to map divergence in post-transcriptional modulation against divergence between monocots and dicots. As indicated by Chapters 2 and 3, there are clearly cases where elements have been retained across these branches and elements that are specific to each branch. Are such distinctions in conservation reflected in functional assays?

Lineage-specific element retention and loss

Using eleven plant transcriptomes, we have estimated that 5% of genes across major branches of angiosperms are under regulation of uAUGs (Figure 3.8B). Additionally, nearly half of genes show significant uAUG depletion, suggesting that lack of uAUGs is important for function. It is of particular interest how these proportions would change with regard to changes in phylogenetic scope. For example, as suggested by our analysis of angiosperms, do only 5% of the 35% of uAUG-containing genes in *Arabidopsis* require an uAUG. Alternatively, if only the Brassicaceae lineage was used to assess uAUG enrichment, would the estimate of constraint be higher? If so, it would mean that certain uORFs are conserved among the Brassicaceae, but not beyond. In turn, this would suggest that uAUGs are important in defining phenotypic differences between major clades. We are fairly confident from this analysis and previous studies that peptide-dependent uORFs are deeply conserved ([26] and Table 2.5) suggesting that, if there are substantial differences with regard to phylogenetic scope, they will be related to peptide-independent effects.

The "1000 Plants" initiative aims to sequence 1000 transcriptomes from a broad sampling of the plant phylogeny (http://www.onekp.com/). Although it is unclear when these sequences will be made available (Neal Stewart, personal communication), because they will contain UTR regions, they will add a vast amount of potential information to the analyses presented here. The software developed for the research above is robust to the expansion in data and also lends itself to parallelization. As it stands the rate limiting step for this pipeline involves clustering sequences into orthologous groups after the pairwise BLAST has been performed. (This assumes that BLAST searches were performed in parallel on a super-computer, as was done in our angiosperm analysis in Chapter 3.) A recent version of OrthoMCL (version 2.0) claims to have optimized this step to accommodate hundreds of proteomes, but, having used version 1.4, I cannot confirm this improvement.

We were not able to identify with confidence peptide-independent uAUG/uORF conservation using pan-angiosperm comparisons, although we showed that they likely exist in 5% of genes (Figure 3.8). The sequence coverage associated with the onekp project would allow us to resolve gene-specific examples. Such examples would indicate viable hypotheses of uAUG/uORF function based on their patterns of conservation.

Exploiting initiation models to understand patterns of uORF conservation

The previous passage describes examining uORF patterns of conservation with regard to functional hypotheses. Such an analysis would be fairly *ad hoc* in the absence of a rigorous method for predicting the molecular repercussions of conserved uORFs on protein production. Assuming that parameters associated with initiation-reinitation apply across flowering plants, the model described in Chapter 3, Section 1 would be highly suited to this task. Many of the functional explanations for uORF activity are dependent on tightly defined spatial constraints; thus, an obvious approach might be to score uORFs relative to a common point-of-reference, such as the start of the mORF. Unfortunately, this approach fails to appreciate the fact that the first uORF seen may mask the effects of downstream uORFs. Alternatively, a short uORF that is close to the mORF can repress expression as effectively as a longer uORF that is farther from the mORF. Only an approach that accounts for the mechanics of initiation-reinitation would be able to determine that the long and short uORFs have the same effect on translation. These and similiar insights might be critical to understanding why particular uORF patterns (Figure 3.11 and 3.12) are enriched among the genes that harbor them.

Translationally active versus quiescent transcripts

Ribosome-occupancy is an indicator of the mRNAs in a cell that are translationally active. Alternatively, under assumptions of constant elongation rates, ribosome-density (number of ribosomes per unit length of mRNA) is an indirect estimate of the number of proteins produced per mRNA per unit time. Although stated as such, these values are clearly distinct, they are also experimentally related by the fact that some translationally active mRNAs will be found in the non-polysomal fraction of a polysome gradient. We have shown that sequence features that affect occupancy may be different than those that affect density. For example, secondary structure around the cap negatively affects translation state, which, as defined in Chapter 4, is a relative measure of ribosome-occupancy (Figure 4.4). Yet, uORFs, which we know can drastically reduce density in yeast [16], have a much more subdued effect on occupancy. Models of translation, particularly initiation, will help to deconvolve occupancy and density effects. Additionally, these models could help to reduce the cost of such assays that are based on expensive microarray analysis of density gradient fractions. If a model predicts the frequency of mRNAs in lighter fractions versus heavier fractions, then only three density gradient fractions could be used to arrive at ribosome occupancy and density (instead of 14 fractions used previously). As a first pass, we have used the simplest possible empirical model of initiation - represented as the *p* parameter in a binomial distribution (Figure 4.5). Though some gains were made in improving resolution, more sophisticated models that address kinetic aspects of gene expression may be required. These are currently under active development (Michael Gilchrist, personal communication).

References

[1] 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.
[2] 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.
[3] Moore MJ (2005) From birth to death: the complex lives of eukaryotic mRNAs. Science 309: 1514-1518.
[4] M.B. Mathews, N. S. & Hershey, J. (2007). Origins and Principles of Translational Control. In: Michael B. Mathews, Nahum Sonenberg, J. W. H. (Ed.), Translational Control in Biology and Medicine, Cold Spring Harbor Laboratory Press.
[5] Pöyry TAA, Kaminski A, Connell EJ, Fraser CS, Jackson RJ (2007) The mechanism of an exceptional case of reinitiation after translation of a long ORF reveals why such events do not generally occur in mammalian mRNA translation. Genes Dev 21: 3149-3162.
[6] Larson DR, Singer RH, Zenklusen D (2009) A single molecule view of gene expression. Trends Cell Biol 19: 630-637.
[7] 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.
[8] Hentze, M. W.; Gebaur, F. & Preiss, T. (2007). cis-Regulatory Sequences and trans-Acting Factors in Translation Control. In: Mathews, M. B.; Sonenberg, N. & Hershey, J. W. (Ed.), Translational Control in Biology and Medicine, Cold Spring Harbor Laboratory Press.
[9] Brooks SA (2010) Functional interactions between mRNA turnover and surveillance and the ubiquitin proteasome system Wiley Interdisciplinary Reviews - RNA 1: 240-252.
[10] Spirin AS (2009) How does a scanning ribosomal particle move along the 5'-untranslated region of eukaryotic mRNA? Brownian Ratchet model. Biochemistry 48: 10688-10692.
[11] 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.
[12] Kozak M (2005) Regulation of translation via mRNA structure in prokaryotes and eukaryotes. Gene 361: 13-37.
[13] Arava Y, Wang Y, Storey JD, Liu CL, Brown PO, Herschlag D (2003) Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A 100: 3889-3894.
[14] Brandt F, Carlson L-A, Hartl FU, Baumeister W, Grünewald K (2010) The three-dimensional organization of polyribosomes in intact human cells. Mol Cell 39: 560-569.
[15] Eldad N, Yosefzon Y, Arava Y (2008) Identification and characterization of extensive intra-molecular associations between 3'-UTRs and their ORFs. Nucleic Acids Res 36: 6728-6738.
[16] 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.
[17] Arava Y, Boas FE, Brown PO, Herschlag D (2005) Dissecting eukaryotic translation and its control by ribosome density mapping. Nucleic Acids Res 33: 2421-2432.
[18] 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.
[19] Gonzalez, C. I.; Wilusz, C. J. & Wilusz, J. (2007). The Interface between mRNA Turnover

and Translation Control. In: Mathews, M. B.; Sonenberg, N. & Hershey, J. W. (Ed.), Translational Control in Biology and Medicine, Cold Spring Hrbor Laboratory Press.
[20] Larson DR, Zenklusen D, Wu B, Chao JA, Singer RH (2011) Real-time observation of transcription initiation and elongation on an endogenous yeast gene. Science 332: 475-478.
[21] Siwiak M, Zielenkiewicz P (2010) A comprehensive, quantitative, and genome-wide model of translation. PLoS Comput Biol 6: e1000865.
[22] Kawaguchi R, Girke T, Bray EA, Bailey-Serres J (2004) Differential mRNA translation contributes to gene regulation under non-stress and dehydration stress conditions in Arabidopsis thaliana. Plant J 38: 823-839.
[23] Branco-Price C, Kawaguchi R, Ferreira RB, Bailey-Serres J (2005) Genome-wide analysis of transcript abundance and translation in Arabidopsis seedlings subjected to oxygen deprivation. Ann Bot 96: 647-660.
[24] 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 U S A 106: 7507-7512.
[25] 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.
[26] Hayden CA, Bosco G (2008) Comparative genomic analysis of novel conserved peptide upstream open reading frames in Drosophila melanogaster and other dipteran species. BMC Genomics 9: 61.
[27] 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.
[28] 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.
[29] 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.
[30] Buchon N, Vaury C (2006) RNAi: a defensive RNA-silencing against viruses and transposable elements. Heredity 96: 195-202.
[31] Cerutti H, Casas-Mollano JA (2006) On the origin and functions of RNA-mediated silencing: from protists to man. Curr Genet 50: 81-99.
[32] 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.
[33] Jones-Rhoades MW, Bartel DP, Bartel B (2006) MicroRNAs and their regulatory roles in plants. Annu Rev Plant Biol 57: 19-53.
[34] Serganov A, Patel DJ (2008) Towards deciphering the principles underlying an mRNA recognition code. Curr Opin Struct Biol 18: 120-129.
[35] 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.
[36] Bayer M, Nawy T, Giglione C, Galli M, Meinnel T, Lukowitz W (2009) Paternal control of embryonic patterning in Arabidopsis thaliana. Science 323: 1485-1488.
[37] Condeelis J, Singer RH (2005) How and why does beta-actin mRNA target? Biol Cell 97: 97-110.
[38] Hüttelmaier S, Zenklusen D, Lederer M, Dictenberg J, Lorenz M, Meng X, Bassell GJ, Condeelis J, Singer RH (2005) Spatial regulation of beta-actin translation by Src-dependent phosphorylation of ZBP1. Nature 438: 512-515.
[39] Crofts AJ, Washida H, Okita TW, Ogawa M, Kumamaru T, Satoh H (2004) Targeting of proteins to endoplasmic reticulum-derived compartments in plants. The importance of RNA localization. Plant Physiol 136: 3414-3419.
[40] Washida H, Sugino A, Kaneko S, Crofts N, Sakulsingharoj C, Kim D, Choi S-B, Hamada S, Ogawa M, Wang C, Esen A, Higgins TJV, Okita TW (2009) Identification of cis-localization elements of the maize 10-kDa delta-zein and their use in targeting RNAs to specific cortical endoplasmic reticulum subdomains. Plant J 60: 146-155.
[41] Okita TW, Choi SB (2002) mRNA localization in plants: targeting to the cell's cortical region and beyond. Curr Opin Plant Biol 5: 553-559.
[42] Bailey-Serres (1999) Selective translation of cytoplasmic mRNAs in plants. Trends Plant Sci 4: 142-148.
[43] de Sousa Abreu R, Penalva LO, Marcotte EM, Vogel C (2009) Global signatures of protein and mRNA expression levels. Mol Biosyst 5: 1512-1526.
[44] 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.
[45] Oyama M, Itagaki C, Hata H, Suzuki Y, Izumi T, Natsume T, Isobe T, Sugano S (2004) Analysis of small human proteins reveals the translation of upstream open reading frames of mRNAs. Genome Res 14: 2048-2052.
[46] 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.
[47] Keene JD (2007) RNA regulons: coordination of post-transcriptional events. Nat Rev Genet 8: 533-543.
[48] 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 U S A 103: 4487-4492.
[49] Francischini CW, Quaggio RB (2009) Molecular characterization of Arabidopsis thaliana PUF proteins--binding specificity and target candidates. FEBS J 276: 5456-5470.
[50] Pichersky E (2005) Is the concept of regulation overused in molecular and cellular biology? Plant Cell 17: 3217-3218.
[51] Chen H, Blanchette M (2007) Detecting non-coding selective pressure in coding regions. BMC Evol Biol 7 Suppl 1: S9.
[52] Freeling M, Subramaniam S (2009) Conserved noncoding sequences (CNSs) in higher plants. Curr Opin Plant Biol 12: 126-132.

[53] Doniger SW, Fay JC (2007) Frequent gain and loss of functional transcription factor binding sites. PLoS Comput Biol 3: e99.
[54] Blanchette M, Tompa M (2002) Discovery of regulatory elements by a computational method for phylogenetic footprinting. Genome Res 12: 739-748.
[55] 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.
[56] 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.
[57] Tran MK, Schultz CJ, Baumann U (2008) Conserved upstream open reading frames in higher plants. BMC Genomics 9: 361.
[58] Brown RH, Gross SS, Brent MR (2005) Begin at the beginning: predicting genes with 5' UTRs. Genome Res 15: 742-747.
[59] 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.
[60] Mignone F, Gissi C, Liuni S, Pesole G (2002) Untranslated regions of mRNAs. Genome Biol 3: REVIEWS0004.
[61] Lynch M, Scofield DG, Hong X (2005) The evolution of transcription-initiation sites. Mol Biol Evol 22: 1137-1146.
[62] Lockton S, Gaut BS (2005) Plant conserved non-coding sequences and paralogue evolution. Trends Genet 21: 60-65.
[63] Lawson MJ, Zhang L (2006) Distinct patterns of SSR distribution in the Arabidopsis thaliana and rice genomes. Genome Biol 7: R14.
[64] 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.
[65] Santi L, Wang Y, Stile MR, Berendzen K, Wanke D, Roig C, Pozzi C, Müller K, Müller J, Rohde W, Salamini F (2003) The GA octodinucleotide repeat binding factor BBR participates in the transcriptional regulation of the homeobox gene Bkn3. Plant J 34: 813-826.
[66] 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.
[67] 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.
[68] Fahlgren N, Howell MD, Kasschau KD, Chapman EJ, Sullivan CM, Cumbie JS, Givan SA, Law TF, Grant SR, Dangl JL, Carrington JC (2007) High-throughput sequencing of Arabidopsis microRNAs: evidence for frequent birth and death of MIRNA genes. PLoS One 2: e219.
[69] Howell MD, Fahlgren N, Chapman EJ, Cumbie JS, Sullivan CM, Givan SA, Kasschau KD, Carrington JC (2007) Genome-wide analysis of the RNA-DEPENDENT RNA POLYMERASE6/DICER-LIKE4 pathway in Arabidopsis reveals dependency on miRNA- and tasiRNA-directed targeting. Plant Cell 19: 926-942.
[70] Kasschau KD, Fahlgren N, Chapman EJ, Sullivan CM, Cumbie JS, Givan SA, Carrington JC (2007) Genome-wide profiling and analysis of Arabidopsis siRNAs. PLoS Biol 5: e57.

[71] 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.
[72] Dezulian T, Remmert M, Palatnik JF, Weigel D, Huson DH (2006) Identification of plant microRNA homologs. Bioinformatics 22: 359-360.
[73] Zhang B, Pan X, Cannon CH, Cobb GP, Anderson TA (2006) Conservation and divergence of plant microRNA genes. Plant J 46: 243-259.
[74] Brodersen P, Voinnet O (2009) Revisiting the principles of microRNA target recognition and mode of action. Nat Rev Mol Cell Biol 10: 141-148.
[75] 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 U S A 101: 11511-11516.
[76] 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.
[77] Farley BM, Pagano JM, Ryder SP (2008) RNA target specificity of the embryonic cell fate determinant POS-1. RNA 14: 2685-2697.
[78] 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 U S A 106: 20252-20257.
[79] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25-29.
[80] 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.
[81] Im KH, Cosgrove DJ, Jones AM (2000) Subcellular localization of expansin mRNA in xylem cells. Plant Physiol 123: 463-470.
[82] 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.
[83] 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.
[84] 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.
[85] Shabalina SA, Ogurtsov AY, Rogozin IB, Koonin EV, Lipman DJ (2004) Comparative analysis of orthologous eukaryotic mRNAs: potential hidden functional signals. Nucleic Acids Res 32: 1774-1782.
[86] Neafsey DE, Galagan JE (2007) Dual modes of natural selection on upstream open reading frames. Mol Biol Evol 24: 1744-1751.
[87] 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.
[88] Roy SW, Penny D, Neafsey DE (2007) Evolutionary conservation of UTR intron boundaries in Cryptococcus. Mol Biol Evol 24: 1140-1148.
[89] 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.
[90] 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.
[91] 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.
[92] Palmieri L, Picault N, Arrigoni R, Besin E, Palmieri F, Hodges M (2008) Molecular identification of three Arabidopsis thaliana mitochondrial dicarboxylate carrier isoforms: organ distribution, bacterial expression, reconstitution into liposomes and functional characterization. Biochem J 410: 621-629.
[93] Allas U, Tenson T (2010) A method for selecting cis-acting regulatory sequences that respond to small molecule effectors. BMC Mol Biol 11: 56.
[94] Takahashi K, Maruyama M, Tokuzawa Y, Murakami M, Oda Y, Yoshikane N, Makabe KW, Ichisaka T, Yamanaka S (2005) Evolutionarily conserved non-AUG translation initiation in NAT1/p97/DAP5 (EIF4G2). Genomics 85: 360-371.
[95] 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.
[96] Emanuelsson O, Brunak S, von Heijne G, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc 2: 953-971.
[97] Lin R-C, Park H-J, Wang H-Y (2008) Role of Arabidopsis RAP2.4 in regulating light- and ethylene-mediated developmental processes and drought stress tolerance. Mol Plant 1: 42-57.
[98] Hyvönen MT, Uimari A, Keinänen TA, Heikkinen S, Pellinen R, Wahlfors T, Korhonen A, Närvänen A, Wahlfors J, Alhonen L, Jänne J (2006) Polyamine-regulated unproductive splicing and translation of spermidine/spermine N1-acetyltransferase. RNA 12: 1569-1582.
[99] 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.
[100] Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13: 2178-2189.
[101] 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.
[102] 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.
[103] Pavesi G, Mauri G, Pesole G (2004) In silico representation and discovery of transcription factor binding sites. Brief Bioinform 5: 217-236.

[104] Blanchette M, Tompa M (2003) FootPrinter: A program designed for phylogenetic footprinting. Nucleic Acids Res 31: 3840-3842.
[105] Siddharthan R, Siggia ED, van Nimwegen E (2005) PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. PLoS Comput Biol 1: e67.
[106] Wang T, Stormo GD (2003) Combining phylogenetic data with co-regulated genes to identify regulatory motifs. Bioinformatics 19: 2369-2380.
[107] 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.
[108] Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13: 555-556.
[109] 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.
[110] 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.
[111] Dai X, Zhuang Z, Zhao PX (2010) Computational analysis of miRNA targets in plants: current status and challenges. Brief Bioinform . In press.
[112] Koonin, E. V. and Galperin, M. Y.,2003. Sequence - Evolution - Function. Kluwer Academic Publishers, .
[113] Tagle DA, Koop BF, Goodman M, Slightom JL, Hess DL, Jones RT (1988) Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. J Mol Biol 203: 439-455.
[114] Prakash A, Tompa M (2005) Discovery of regulatory elements in vertebrates through comparative genomics. Nat Biotechnol 23: 1249-1256.
[115] Ludwig MZ, Palsson A, Alekseeva E, Bergman CM, Nathan J, Kreitman M (2005) Functional evolution of a cis-regulatory module. PLoS Biol 3: e93.
[116] 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.
[117] Cvijovic M, Dalevi D, Bilsland E, Kemp G, Sunnerhagen P (2007) Identification of putative regulatory upstream ORFs in the yeast genome using heuristics and evolutionary conservation. BMC Bioinformatics 8: 295.
[118] Iacono M, Mignone F, Pesole G (2005) uAUG and uORFs in human and rodent 5'untranslated mRNAs. Gene 349: 97-105.
[119] Cruz-Vera LR, Sachs MS, Squires CL, Yanofsky C (2011) Nascent polypeptide sequences that influence ribosome function. Curr Opin Microbiol 14: 160-166.
[120] Nyikó T, Sonkoly B, Mérai Z, Benkovics AH, Silhavy D (2009) Plant upstream ORFs can trigger nonsense-mediated mRNA decay in a size-dependent manner. Plant Mol Biol 71: 367-378.
[121] Jackson, R. J.; Kaminski, A. & P¿yry, T. (2007). Coupled Termination-initiation Events in mRNA Translation. In: Mathews, M. B.; Sonenberg, N. & Hershey, J. W. (Ed.), Translational Control in Biology and Medicine, Cold Spring Harbor Laboratory Press.

[122] Kochetov AV, Ahmad S, Ivanisenko V, Volkova OA, Kolchanov NA, Sarai A (2008) uORFs, reinitiation and alternative translation start sites in human mRNAs. FEBS Lett 582: 1293-1297.
[123] Fraser HB, Hirsh AE, Giaever G, Kumm J, Eisen MB (2004) Noise minimization in eukaryotic gene expression. PLoS Biol 2: e137.
[124] Chaw S-M, Chang C-C, Chen H-L, Li W-H (2004) Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes. J Mol Evol 58: 424-441.
[125] Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature 434: 338-345.
[126] Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JGR, Korf I, Lapp H, Lehväslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E (2002) The Bioperl toolkit: Perl modules for the life sciences. Genome Res 12: 1611-1618.
[127] Bailey TL, Baker ME, Elkan CP (1997) An artificial intelligence approach to motif discovery in protein sequences: application to steriod dehydrogenases. J Steroid Biochem Mol Biol 62: 29-44.
[128] Basu A, Chowdhury D (2007) Traffic of interacting ribosomes: effects of single-machine mechanochemistry on protein synthesis. Phys Rev E Stat Nonlin Soft Matter Phys 75: 021902.
[129] Zouridis H, Hatzimanikatis V (2007) A model for protein translation: polysome self-organization leads to maximum protein synthesis rates. Biophys J 92: 717-730.
[130] Skjï¿½ndal-Bar N, Morris DR (2007) Dynamic model of the process of protein synthesis in eukaryotic cells. Bull Math Biol 69: 361-393.
[131] Dimelow RJ, Wilkinson SJ (2009) Control of translation initiation: a model-based analysis from limited experimental data. J R Soc Interface 6: 51-61.
[132] Kim B-H, Cai X, Vaughn JN, von Arnim AG (2007) On the functions of the h subunit of eukaryotic initiation factor 3 in late stages of translation initiation. Genome Biol 8: R60.
[133] Kozak M (2001) Constraints on reinitiation of translation in mammals. Nucleic Acids Res 29: 5226-5232.
[134] Rajkowitsch L, Vilela C, Berthelot K, Ramirez CV, McCarthy JEG (2004) Reinitiation and recycling are distinct processes occurring downstream of translation termination in yeast. J Mol Biol 335: 71-85.
[135] Kozak M (1987) Effects of intercistronic length on the efficiency of reinitiation by eucaryotic ribosomes. Mol Cell Biol 7: 3438-3445.
[136] Roy B, Vaughn JN, Kim B-H, Zhou F, Gilchrist MA, Arnim AGV (2010) The h subunit of eIF3 promotes reinitiation competence during translation of mRNAs harboring upstream open reading frames. RNA 16: 748-761.
[137] Kozak M (1990) Downstream secondary structure facilitates recognition of initiator codons by eukaryotic ribosomes. Proc Natl Acad Sci U S A 87: 8301-8305.
[138] Markham NR, Zuker M (2008) UNAFold: software for nucleic acid folding and hybridization. Methods Mol Biol 453: 3-31.
[139] Kawaguchi R, Bailey-Serres J (2005) mRNA sequence features that contribute to translational regulation in Arabidopsis. Nucleic Acids Res 33: 955-965.

[140] Kudla G, Murray AW, Tollervey D, Plotkin JB (2009) Coding-sequence determinants of gene expression in Escherichia coli. Science 324: 255-258.
[141] Tuller T, Waldman YY, Kupiec M, Ruppin E (2010) Translation efficiency is determined by both codon bias and folding energy. Proc Natl Acad Sci U S A 107: 3645-3650.
[142] Kozak M (1989) Circumstances and mechanisms of inhibition of translation by secondary structure in eucaryotic mRNAs. Mol Cell Biol 9: 5134-5142.
[143] Piques M, Schulze WX, Höhne M, Usadel B, Gibon Y, Rohwer J, Stitt M (2009) Ribosome and transcript copy numbers, polysome occupancy and enzyme dynamics in Arabidopsis. Mol Syst Biol 5: 314.
[144] Hamilton TL, Stoneley M, Spriggs KA, Bushell M (2006) TOPs and their regulation. Biochem Soc Trans 34: 12-16.
[145] Wang Y, Liu CL, Storey JD, Tibshirani RJ, Herschlag D, Brown PO (2002) Precision and functional specificity in mRNA decay. Proc Natl Acad Sci U S A 99: 5860-5865.
[146] Maquat LE (2004) Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. Nat Rev Mol Cell Biol 5: 89-99.
[147] Niemitalo O, Neubauer A, Liebal U, Myllyharju J, Juffer AH, Neubauer P (2005) Modelling of translation of human protein disulfide isomerase in Escherichia coli-A case study of gene optimisation. J Biotechnol 120: 11-24.
[148] Gibson MA, Bruck J (2000) Efficient Exact Stachastic Simulation of Chemical Systems with Many Species and Many Channels Journal of Physical Chemistry 104: 1876-1889.
[149] Ramsey S, Orrell D, Bolouri H (2005) Dizzy: stochastic simulation of large-scale genetic regulatory networks. J Bioinform Comput Biol 3: 415-436.
[150] 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.
[151] Sagliocco FA, Laso MRV, Zhu D, Tuite MF, McCarthy JE, Brown AJ (1993) The influence of 5'-secondary structures upon ribosome binding to mRNA during translation in yeast. J Biol Chem 268: 26522-26530.
[152] Laso MRV, Zhu D, Sagliocco F, Brown AJ, Tuite MF, McCarthy JE (1993) Inhibition of translational initiation in the yeast Saccharomyces cerevisiae as a function of the stability and position of hairpin structures in the mRNA leader. J Biol Chem 268: 6453-6462.
[153] Fedoroff NV (2002) RNA-binding proteins in plants: the tip of an iceberg? Curr Opin Plant Biol 5: 452-459.
[154] Tam PPC, Barrette-Ng IH, Simon DM, Tam MWC, Ang AL, Muench DG (2010) The Puf family of RNA-binding proteins in plants: phylogeny, structural modeling, activity and subcellular localization. BMC Plant Biol 10: 44.
[155] Galagan JE, Calvo SE, Cuomo C, Ma L-J, Wortman JR, Batzoglou S, Lee S-I, Bastürkmen M, Spevak CC, Clutterbuck J, Kapitonov V, Jurka J, Scazzocchio C, Farman M, Butler J, Purcell S, Harris S, Braus GH, Draht O, Busch S, D'Enfert C, Bouchier C, Goldman GH, Bell-Pedersen D, Griffiths-Jones S, Doonan JH, Yu J, Vienken K, Pain A, Freitag M, Selker EU, Archer DB, Peñalva MA, Oakley BR, Momany M, Tanaka T, Kumagai T, Asai K, Machida M, Nierman WC, Denning DW, Caddick M, Hynes M, Paoletti M, Fischer R, Miller B, Dyer P, Sachs MS, Osmani SA, Birren BW (2005) Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae. Nature 438: 1105-1115.

[156] Yang Z, Watson JC (1993) Molecular cloning and characterization of rho, a ras-related small GTP-binding protein from the garden pea. Proc Natl Acad Sci U S A 90: 8732-8736.
[157] de Silanes IL, d'Alcontres MS, Blasco MA (2010) TERRA transcripts are bound by a complex array of RNA-binding proteins. Nat Commun 1: 33.
[158] Sinha S, Blanchette M, Tompa M (2004) PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. BMC Bioinformatics 5: 170.
[159] Vingron M, Brazma A, Coulson R, van Helden J, Manke T, Palin K, Sand O, Ukkonen E (2009) Integrating sequence, evolution and functional genomics in regulatory genomics. Genome Biol 10: 202.
[160] Wang X, Haberer G, Mayer KFX (2009) Discovery of cis-elements between sorghum and rice using co-expression and evolutionary conservation. BMC Genomics 10: 284.
[161] Wohlbach DJ, Thompson DA, Gasch AP, Regev A (2009) From elements to modules: regulatory evolution in Ascomycota fungi. Curr Opin Genet Dev 19: 571-578.References

Vita

Justin Vaughn was born in Chattanooga, Tennessee in 1980. He graduated from Samford University in 2003 with a Bachelor of Sciences degree in Biology. Between receiving his undergraduate degree and returning to graduate school at the University of Tennessee, he worked for the National Park Service in the Great Smoky Mountains National Park as a field technician. He joined the Department of Biochemistry, Cellular, and Molecular Biology in 2005 and joined Albrecht von Arnim's lab in 2006. In 2007, he took a sabbatical to the University of Milan to receive training in bioinformatics under Flavio Mignone. In 2011, he received his Ph. D. with an emphasis on plant genomics and moved to the University of Georgia, Athens to pursue postdoctoral training under Jeff Bennetzen.