De novo computational identification of stress-related sequence motifs and microRNA target sites in plant untranslated regions (UTRs)

Prabhakaran Munusamy

Department of Plant Science

Faculty of Agricultural and Environmental Sciences

McGill University

Montreal, Quebec, Canada

April 2015

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science

Abstract

Gene regulation at the transcriptional and translational level leads to diversity in phenotypes and functional organisms. Regulatory DNA or RNA sequence motifs adjacent to the gene coding sequence act as binding sites for proteins that in turn enable or disable expression of the gene. Whereas the known DNA and RNA binding proteins range in the thousands, only a few motifs are known and have been experimentally verified. In this study, we have predicted putative regulatory motifs in groups of untranslated regions (UTR) from genes regulated at the translational level in *Arabidopsis thaliana* under normal and stressed conditions. The test group of sequences was divided into random sub-groups and subjected to three *de novo* motif finding algorithms (Seeder (Fauteux *et al*., 2008), Weeder (Pavesi *et al*., 2004) and MEME (Bailey *et al*., 2009)). In addition to identifying sequence motifs, using an *in silico* tool we have predicted microRNA target sites in the 3′ UTRs of the translationally regulated genes, as well as identified upstream open reading frames (uORFs) located in the 5′ UTRs. Our bioinformatics strategy and the knowledge generated contributes to understanding gene regulation during stress, and can be applied to disease and stress resistant plant development.

Résumé

La régulation de la transcription des gènes et la régulation de leur traduction contribuent à la diversité des phénotypes et des organismes. Les motifs de séquence ADN ou ARN dédiés à ces régulations et avoisinant les séquences codantes des gènes, sont des sites de liaison pour des protéines capables de déclencher ou d'empêcher l'expression de ces gènes. Bien que l'on recense des milliers de protéines de liaison à l'ADN ou à l'ARN, seuls quelques motifs de liaison sont connus et ont été testés du point de vue fonctionnel. Dans la présente étude, nous avons prédéterminé les potentiels motifs de régulation des régions non traduites (UnTranslated Regions, UTR) d'un groupe de gènes dont le niveau de traduction est régulé différemment entre condition normale ou de stress, chez *Arabidopsis thaliana*. L'ensemble de ces séquences a été réparti en sous-groupes et soumis à trois algorithmes de détection de motifs *de novo* (Seeder (Fauteux et al., 2008), Weeder (Pavesi et al., 2004) et MEME (Bailey et al., 2009)). En plus d'identifier les motifs de régulation, nous avons pu déterminer, grâce à une méthode *in silico*, les potentielles séquences cibles des microARNs au sein des régions 3′ UTRs des gènes en question. Nous avons également déterminé les fenêtres de lecture en amont (upstream open reading frames (uORFs)) situées dans les régions 5′ UTRs de ces gènes. Notre stratégie bioinformatique ainsi que les nouvelles connaissances apportées, contribuent à mieux comprendre les mécanismes de régulations génétiques dans les conditions de stress. Elles seront applicables dans les domaines des maladies et de la résistance au stress des plantes.

Acknowledgements

I would like to take this opportunity to thank my supervisor Dr. Martina Stromvik for her guidance and support throughout this entire work, especially in providing valuable suggestions and feedback when needed and in writing this thesis.

I would like to thank my advisory committee member Dr. Kevin Wade for his advice and being supportive in pursuing this project work.

I would also like to acknowledge the contribution of my collaborators Dr. Peter Moffett and Louis-Valentin Meteignier from University of Sherbrooke, and I thank them for providing the data to carry out this project.

I gratefully thank my lab members Yevgen, Andrew, Marcia, Eva and Hector for their constructive comments and friendly advice whenever needed.

Last but not least, I am very grateful for the enormous support and encouragements provided by my friends and family members.

Table of Contents

List of tables

List of figures

List of abbreviations

ABI	Abcisic acid insensitive
AdoMetDC	*S*-adenosylmethionine decarboxylase
AFP	ABI five binding protein
AP2	Apetala 2
APUM	Arabidopsis pumilio
AtBRNs	Arabidopsis Bruno-like proteins
Avr	Avirulence
bzip	basic leucine zipper
CC	Coiled-coil
CCA1	circadian clock associated 1
CCS	copper chaperone for superoxide dismutase
CMV	Cucumber mosaic virus
CSD	copper/zinc superoxide dismutase
DCL1	Dicer-like 1 enzyme
eIFs	eukaryotic initiation factors
flg	flagellins
FPKM	Fragments per kilo bases of exons for per million mapped reads
HR	Hypersensitive response
IRES	Internal ribosome entry site
LRR	Leucine rich repeat
MEME	Multiple Expectation Maximization for Motif Elicitation
miRNA	microRNA
mRNA	messenger RNA
NB	Nucleotide binding
PAMP	Pathogen associated molecular patterns
Pri-miRNA	primary miRNA
psRNAtarget	plant small RNA target analysis server
PWM	Position weight matrix
R	Resistance

RIN	RPM1-interacting protein
RISC	RNA induced silencing complex
RNA	Ribonucleic acid
RPM	Resistance to *Pseudomonas syringae pv maculicola*
RPS	Resistance to *Pseudomonas syringae*
SBP-box	Squamosa promoter binding protein–box
SOC1	Suppressor of constans
SPL	Squamosa promoter binding protein–like
Tomtom	Tool for motif to motif comparison
uORFs	upstream open reading frames
UTRs	Untranslated regions

Contribution of authors

The project was designed and supervised by Dr. Martina Stromvik. Wet-lab work, sample collection and differential expression analysis was done by our collaborators Dr. Peter Moffett and Louis-Valentin Meteignier. Python scripts were written by Prabhakaran Munusamy with support from Yevgen Zolotarov. Prabhakaran Munusamy performed data analysis and results were interpreted with the guidance of Dr. Martina Stromvik.

Chapter 1. Introduction

1.1 General introduction

Precise regulation of gene expression is important for plants to survive varying environmental conditions. Plant cells have to synthesize proteins in response to biotic and abiotic stresses. Studies have revealed that the level of mRNA transcript produced does not always correlate to the protein synthesized, which could be attributed to variable mRNA translation efficiency. Several features of mRNA influence the translation activity, most importantly the regulatory elements harbored in the 5' and 3' untranslated regions. Untranslated regions are located upstream and downstream of the main protein coding regions (Figure 1.1). These regions have features, such as a 5' methyl cap and a 3' poly-A tail, which have been found to play a significant role in regulating gene expression at the translational level (Munroe and Jacobson, 1990; Gallie, 1991; Green, 1993). Other influencing features include the length of 5' and 3' UTRs, secondary structure, presence of start codon, and upstream open reading frames (uORFs) (Tanguay and Gallie, 1996; Mignone et al., 2002). A number of studies have been carried out to understand the translational mechanism occurring in plants in response to stress. It has been observed that ribosome loading of mRNAs is affected globally under abiotic stresses such as salt, drought, hypoxia, light and darkness (e.g. Mazzcotelli et al., 2008; Dickey et al., 1998). Some mRNAs also seem to escape from the translational block and are thus regulated during stress. For example, under heat stress in plants, mRNAs encoding heat shock proteins was found to be highly regulated whereas some mRNAs that are bound to heat shock granule complexes were translationally repressed (Nover et al., 1989). A further study (Matsuura et al., 2008 & 2013) on this identified a sequence element located in the 5' UTR responsible for active mRNA translation under heat shock.

Plant miRNAs are known to trigger mRNA cleavage or translational repression by binding to target sites found in the 5' and 3' UTR, and protein coding regions (Broderson et al., 2008; Carrington and Ambros, 2003). Studies have identified various roles of miRNAs in plant growth, and development, and its response to stress (Jones-Rhoades et al., 2006; Mallory and Vaucheret, 2006). For example, an Arabidopsis SPL3 gene encoding an SBP-box transcription factor is translationally regulated by miR156/miR157 which inhibits SPL3 gene expression by binding to its target complementary site in the 3' UTR leading to early flowering phenotype (Gandikota et al., 2007). In addition to miRNAs, UTRs have regulatory motifs or sequence patterns to which regulatory or RNA binding proteins bind and mediate translational control (Figure 1.2). For example, Bruno-like proteins encoded by the *AtBRN1* and *AtBRN2* genes in *Arabidopsis thaliana* led to delayed flowering time. In a mutant study, it was discovered that a Bruno-like protein binds to a sequence element or motif in the 3' UTR of *SOC1* mRNA and represses its expression, thereby causing delayed flowering time (Kim et al., 2013). However, understanding the control and regulation of translational process is still at its early stage. Studies have discovered the existence of several parallel mechanisms in plants, which are utilized in response to stress, but a complete picture of their regulation is yet to be understood. There is a vast knowledge gathered on the transcriptional level of genes but only limited at the translational level. The identification of the key elements of stress-regulated genes and their characteristics would help in elucidating the translational mechanism and could lead to producing a stress resistant or tolerant plants. Most of the evidence on translational control of genes are provided by mutant and genetic screening studies. Not many computational studies have been carried to collect information about the gene features with respect to the translational level especially the regulatory motifs located in the UTRs. As well there are not many databases available to annotate the predicted motifs compared to the databases available for transcriptional regulatory

elements. Thus, it requires laborious wet lab work to identify and validate genes regulated by stress.

In this study we set out to use a unique custom bioinformatics approach and different computational tools to analyze genes that are translationally regulated under stress in Arabidopsis to identify various regulatory elements potentially responsible for mediating translational regulation. The gene features identified in this study could help differentiate between the stress regulated genes versus non-stress regulated genes.

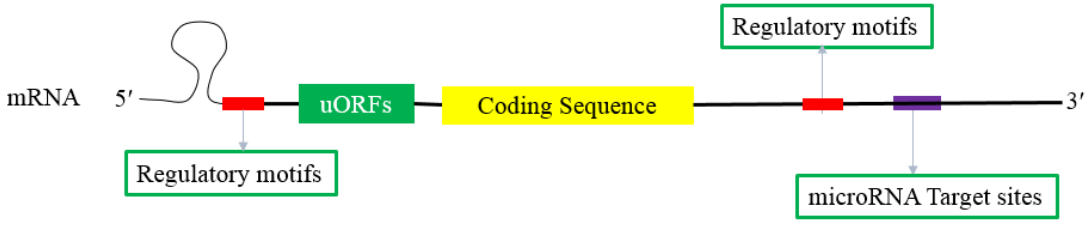


Figure 1.1: Diagram of an mRNA with possible regulatory elements located in the 5′ and 3′ UTR which could be involved in translational process.

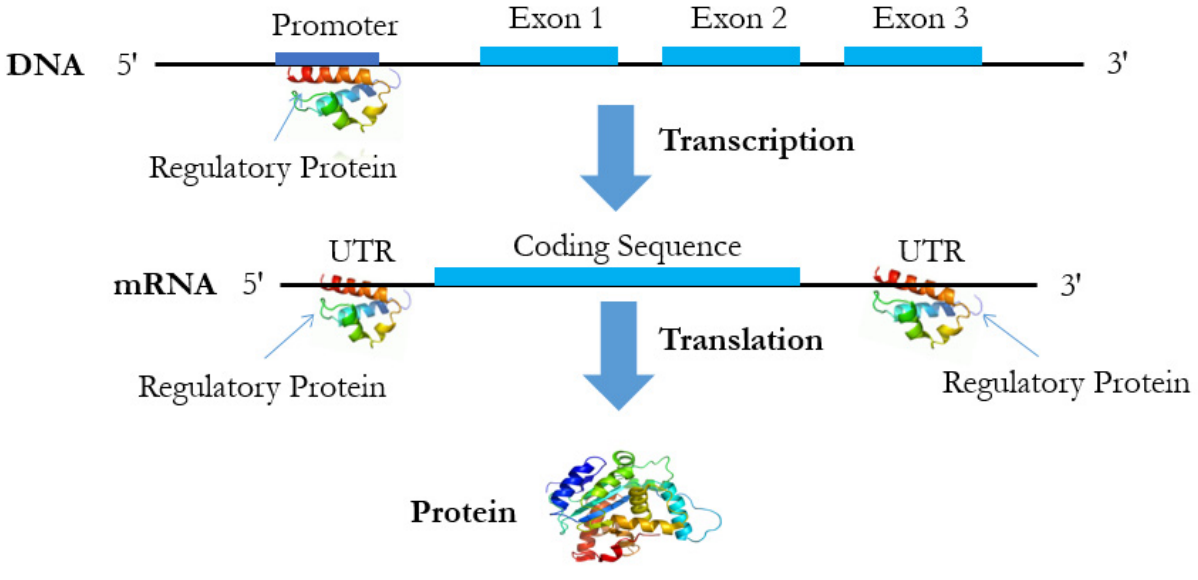


Figure 1.2: Regulatory proteins can bind to the promoter (in DNA), 5′ UTR or 3′ UTR in the mRNA and regulate gene expression.

1.2 Hypotheses

1. Regulatory elements such as uORFs, regulatory motifs and miRNA binding sites are present in the untranslated regions of potentially translationally regulated genes in plants under stress.
2. Conserved regulatory motifs in the UTRs can be detected by dividing the sets of sequences into random subgroups and subjecting them to *de novo* motif analysis.

1.3 Objectives

1. To implement the approach of dividing the sequences of interest into random sub-groups for efficient *de novo* motif discovery.
2. To analyze *Arabidopsis thaliana* genes with significant potential translational efficiency during stress and
    i. To identify regulatory motifs in their 5' and 3' UTRs using the *de novo* motif discovery tools Seeder, Weeder and MEME,
    ii. To predict upstream open reading frames in their 5' UTRs using UTRscan
    iii. To identify miRNA target binding site in their 3' UTRs using psRNATarget
3. To predict the role of *de novo* motifs identified by matching against motifs already known, and through literature search.

Chapter 2. Literature review

2.1 Overview

In an organism, the genetic information required to survive and maintain homeostasis is stored as DNA. Cells use this information through transcription and translation to produce proteins and enzymes for various biological functions. It is important to understand the way genes function and how they are regulated to produce protein. The process in which the DNA is converted into messenger RNA (mRNA) is known as transcription. The mRNA is translated to produce proteins. When a gene is expressed, it is not necessary for the level of mRNA to be equivalent to protein formation. There are many factors that affect the expression of a gene at the transcription as well as at the translational level. In the case of plants, it is very important for cells to synthesize proteins in response to fluctuating environmental conditions. Many abiotic and biotic factors such as salt, drought, soil quality, water, and light intensity, infection by pathogens like bacteria and virus, as well as damage by herbivores cause stress in plants. Plants respond to these stresses by modulating gene expression to protect themselves and survive, correct translation is crucial for protecting them.

2.2 Translational Regulation

Translational regulation is very important for many physiological processes in cells during its development, differentiation, and proliferation, in controlling metabolic pathways, and protecting it from various external modulating environmental conditions. In a cell, translational control is one among the vital mechanism in transcribing the genetic information that affects the overall rate of protein synthesis from the mRNA. Many factors including stress such as oxygen deprivation, salt stress, and nutrient conditions that affect gene expression are referred to as global control. Translation regulation controlled by proteins such as initiation and elongation factors that interact with the *cis*-acting elements in the 5′ and 3′ untranslated region (UTR) of the mRNA are called as mRNA-specific control. Depending on the translational activity of the mRNA, the fate of the physiological process of the cell changes to maintain homeostasis.

Three major steps involved in the mRNA translation process includes initiation, elongation, and termination. Initiation occurs in two ways i.e., via cap-dependent and cap-independent initiation. In the former method, the eukaryotic initiation factors (eIFs) recognize the methylated 5′ cap structure or the 3′ poly-A tail, and in-turn activates the mRNA for the small 40S ribosomal containing the methionyl-tRNA to bind to it, and scans the 5′ leader sequence for the AUG codon. After recognition of the start AUG codon on the mRNA, 60S ribosomal subunit joins the 40S, forming a large 80S ribosomal initiation complex. Next, during the elongation process, the aminoacyl-tRNA binds to the initiation complex, and the peptide synthesis takes place. The elongation process is terminated when the termination codon is recognized. Many protein factors are involved during the process of translation. Unavailability of those factors results in inhibiting translation and essentially halts protein synthesis. In addition, in the cap-dependent process, apart from the proteins, presence of certain RNA secondary structures in the 5′ UTR would also restrict the ribosomal scanning mechanism that leads to poor translation or inhibition of the

mRNA. The 5′ UTR contains internal ribosome entry sites (IRES) that are mostly in the form of a hairpin-loop structure, which recruit the ribosomal units for scanning the mRNA and the downstream initiation, elongation and termination for translation of the mRNA happens. Under starvation or stress conditions, certain mRNAs after the transcription process actively recruits the ribosomal subunits along with the protein factors, they further interact with the regulatory elements harbored in the 5′ and 3′ UTR of the mRNA, and produce the peptides in the cell to adapt itself to the modifying environmental conditions. Thus, identifying and understanding the function of regulatory elements located in the UTR would be one way to fine tune the process of translation thereby the synthesis of protein required to combat stress. Apart from regulatory elements, small RNAs such as microRNAs and upstream open reading frames (uORFs) have been shown to affect the translational activity of the mRNA (Fabian et al., 2010; Carrington and Ambros, 2003; Morris and Geballe, 2000; Von Arnim et al., 2014).

2.3 miRNAs

2.3.1 miRNAs and their biogenesis

MicroRNAs (miRNAs) have proved to be important regulators of gene expression especially involved in translational and mRNA decay processes. MiRNAs are small non-coding RNA molecules, usually 20 - 24 nucleotides in length (Bartel, 2004). They were initially discovered in *C*. *elegans*, in which lin-4 and let-7 miRNA were identified to be involved in regulating the timing of larval development in *C*. *elegans* (Lee et al., 1993; Reinhart et al., 2000). Later using techniques like cloning, forward genetic screening experiments, as well as bioinformatic approaches with certain specific criteria followed by experimental tests more miRNAs were identified in plants and animals, and they seem to be conserved among different species. Studies

have found that miRNAs play a major role in plant organ development and morphogenesis, and that they are involved in various stress responses (Fujii et al., 2005; Sunkar et al., 2006). MiRNAs affect gene regulation through the mechanism of base-pairing with the target mRNAs for either cleavage or translation inhibition activity.

Plant miRNAs mostly come from the genomic regions rather than from the introns of protein-coding genes, which is the case of animal miRNAs mostly (Reinhart et al., 2002). Plant miRNAs possess their own transcriptional units and are very rarely located closer to each other in the genome sometimes leading to transcription of multiple miRNA genes from a single transcript. The synthesis of a mature miRNA is a multi-step process requiring multiple enzymes. Initially, using Polymerease II enzymes (Lee et al., 2004) the miRNA gene is transcribed to produce the primary miRNA (pri-miRNA) transcript containing several hundred nucleotides. Then, the pri-miRNA transcript is processed by the Dicer-like 1 (DCL 1) enzyme, an endonuclease with cleavage activity, to form a stem-loop intermediate known as pre-miRNA. A DCL 1 knock-out mutant study in Arabidopsis has shown reduced level of miRNA being produced in the system, suggesting its vital importance in the pathway of miRNA synthesis (Schauer et al., 2002). The pre-miRNA transcript, before being transported to the cytoplasm from nucleus, the dicer-like enzyme again cuts the flanking stem-loop sequence of the pre-miRNA producing a miRNA-miRNA* duplex. Then, using the nucleocytoplasmic transporter HASTY, the miRNA duplex is exported to the cytoplasm (Bollman et al., 2003). Further, in the cytoplasm, the helicase enzyme unwinds the double stranded miRNA duplex to yield a mature single stranded miRNA of usually 20 -24 nucleotides long. Finally, the mature miRNA gets loaded into the RNA-induced silencing complex that possesses ARGOUNAUTE protein among other molecules and carries out its function of target gene expression by binding to the target mRNA molecules (Bartel, 2004).

A number of studies has been carried out to elucidate the role and mechanism of miRNAs. One of the best examples to describe the miRNAs function would be the study that discovered the role of *lin*-4, a small RNA negatively regulating the *lin*-14 target gene involved in controlling the timing of larval development in *C*. *elegans* (Wightman et al., 1993). The study found that the 3'UTR of *lin*-14 mRNA contains multiple target sites that are complementary to the *lin*-4 miRNA molecule. Binding of the *lin-4* molecule to the target site in *lin-14* leads to down regulation of *lin*-14. In another study, it was observed that the presence of *lin*-4 along with *lin-14* mRNA affects its protein level during the first and second larval stages in *C .elegans* but not its mRNA level (Olsen and Ambros, 1999). This study confirmed that *lin-4* affects gene regulation at the level of translation specifically after translation initiation process.

2.3.2 Studies on plant gene regulation by miRNAs

MiRNAs have been identified to regulate various developmental processes, and control gene regulation during stress in plants. Studies have demonstrated the targeted activity of miR172 on APETALA 2 and AP2-like genes responsible for flowering time and floral identity in Arabidopsis plants (Chen, 2004; Aukerman and Sakai, 2003). Under normal conditions, the miR172 target genes were found to act as floral repressors. Experiments on mutants overexpressing the miR172 were found to bind the AP2 and AP2-like genes possessing target complimentary sites for the miRNA thereby suppressing its activity, leading to early flowering and effects on floral identity. Results also suggest that miRNAs not only regulate developmental processes but they are also induced by external environmental conditions. The activity of miR169 that targets the gene NFYA5 that encodes a ubiquitous transcription factor in Arabidopsis has been studied during drought and salt stress (Li et al., 2008). In the study

transgenic Arabidopsis plants overexpressing NFYA5 were more resistant to drought, whereas miR169 overexpressing transgenic lines were sensitive to drought. It was demonstrated that miR169 overexpression suppresses the activity of NFYA5 transcript accumulation by binding to the miRNA target region. However, under drought condition the expression of miR169 is down-regulated thereby relieving NFYA5 from repression by miRNA169. Studies have also observed plants response to oxidative stress as well as copper deprivation mediated by miR398 that targets the copper/zinc superoxide dismutase genes (CSD1 and CSD2) required for the removal of superoxide radicals, and CCS1 that encodes copper chaperone for superoxide dismutase required for delivering copper to CSD1 and CSD2 (Sunkar and Zhu, 2004.) Further studies have shown that reactive oxygen species being removed by the action CSD1 and CSD2 and reduced levels of miR398 during oxidative stress (Sunkar et al., 2006; Yamasaki et al., 2007; Beauclair et al., 2010). In contrast, under copper starvation, miR398 up-regulation was observed, which binds CSD1, CSD2 and CCS1 and triggers cleavage of the mRNA, in order to save copper for essential proteins. As well, plants resistance to pathogens regulated by miRNAs have also been demonstrated. For example, the miR393 mediated repression of F-box auxin receptors (TIR1, AFB1, and AFB2) down-regulates auxin signalling (Navarro et al., 2006.)

2.4 Upstream Open reading frames (uORFs)

2.4.1 uORF background

Post-transcriptional gene regulation or translational regulation can be controlled by various regulatory signals or elements present in the untranslated regions of mRNAs. Upstream open reading frames (uORFs) are control elements located upstream of the main protein coding sequence of a gene in the 5′ untranslated regions (5′ UTR). uORFs, like the main open reading frames, have an in-frame start codon followed by stop codon which might sometime extend into

the downstream ORF that could be either the second uORF in the UTR or main ORF. uORFs affect gene regulation either by triggering mRNA decay process or repress translation by different mechanisms depending on the uORF recognition by ribosome and the possible outcomes would be i) disassociation of the ribosome after translating the uORF, ii) stalling the elongation or termination process after uORF translation initiation creating a translational block or inducing mRNA decay process, and iii) translation of the uORF after which the ribosome remains associated with the mRNA for downstream translation re-initiation process (Barbosa et al., 2013). A study conducted in yeast (Yun *et al*., 2012) observed that introducing a mutation in the 5′ upstream region produced uORFs that extends to the main ORF, resulting in down regulation of the gene producing reduced levels of mRNA and protein in the system and hence inhibiting gene expression at the transcriptional and translational level.

2.4.2 Different types of uORFs

Different types of uORFs affect gene expression through different mechanisms. There are two types of functional uORFs; sequence-dependent and sequence-independent uORFs. Sequence-dependent uORFs encode a peptide, which binds to the translational machinery. This either prevents the ribosome from further scanning the coding region and thereby inhibiting the translational process, or it disrupts the peptidyl transferase enzyme component part of the ribosome from forming a peptide bond, which in turn triggers mRNA decay via early termination. Sequence-independent uORFs on the other hand, do not encode small peptides, and instead control translation through features such as uORF start codon recognition based on the sequence context, its distance from start of the 5′ UTR, uORF length, GC content around the stop

codon, and the downstream inter-cistronic length and structure (Morris et al., 2000; Vilela et al., 2003; Tran et al., 2008).

Sequence-independent uORFs might control translation by triggering the non-sense mediated mRNA decay process, which occurs when the ribosome encounters a pre-termination codon after the translation of the uORF (Nyiko et al., 2009). Studies have shown that rich GC content around the stop codon triggers the release of ribosomal subunits from the mRNA required for the reinitiation of the downstream ORF possibly exposing the mRNA for decay in a mechanism known as post-termination mediated decay process (Hinnebusch, 1997).

2.4.3 Examples of uORFs role in plants

An Arabidopsis gene *bzip11* encoding a transcription factor, has multiple uORFs in the 5' untranslated region upstream of the main coding sequence. High levels of sucrose were found to induce translation of the second uORF in the 5' UTR, producing a sucrose control peptide, which stalls the ribosome from efficient translation of the downstream ORF, eventually repressing the translational activity of the *bzip11* gene (Rahmani et al., 2009). Another example, the plant *S*-adenosylmethionine decarboxylase (AdoMetDC) enzyme required for ploymine synthesis containing a tiny and small uORFs (size of 10 and 46 codons respectively) was studied in Arabidopsis (Hill and Morris, 1993). Under low level of polyamines, the tiny uORF is translated and gets terminated immediately, and further reinitiates the downstream ORF synthesizing the AdoMetDC enzyme, thereby rising the levels of polyamines. Under increased levels of polyamines the small uORF AUG codon is recognized by the ribosome and upon its translation yields a peptide that prevents or stalls the ribosome from further downstream scanning, which leads to translational repression of the downstream ORF (Hanfrey et al., 2005).

2.5 Regulatory motifs involved in translational control

Cis-regulatory elements are often described as sequence motifs. They are a short stretch of nucleotides, usually of fixed length (e.g. six or eight nucleotides) with a recurring sequence pattern (Priest et al., 2009). Motifs that are over represented sequence patterns in a set of biological sequence are also referred to as sequence signals. Not much has been discovered with respect to the sequence motifs that are thought to regulate translational process in plants. Studies have observed that variation in light condition for *Arabidopsis thaliana* plays a vital role in translational control of the mRNAs (Liu et al., 2012; Gamm et al., 2014). In the study by Liu et al., 2012, to understand the effect of light on gene regulation they compared the transcriptomic and polysome-bound mRNA levels. As a result, they observed genes that encode ribosomal proteins were highly regulated at the level of translation and, possesses longer half-life and short mRNA length. Also in their study, they found a *cis*-element TAGGGTTT located in the 5′ UTR, and genes containing this regulatory motif had significant translational efficiency, however the mechanism is not clear yet. Likewise in another study in Arabidopsis, the effect of sucrose and light on mRNA translation was experimented (Gamm et al., 2014). It was found that in the presence of light a selective set of mRNAs were translationally regulated compared to dark conditions, and most of those mRNAs codes ribosomal proteins among others. It was also observed in the study that motifs like [GA][GA]AGA[GA] and [TCA]CG[GCA]CG[GA][CA]G were enriched in the 5′ UTR of translationally regulated genes (Gamm et al., 2014) which could be involved in its translational control.

A study by Huh et al., 2013 has demonstrated the role of a *cis*-regulatory element UGUA (A/C/U) AUA recognized by puf proteins in preventing the viral infection on Arabidopsis plants using mutant study. Transgenic plants carrying the APUM5, an Arabidopsis puf gene showing reduced susceptibility to *Cucumber mosaic virus (CMV)* was observed. And it was indentified in

the study that the APUM5 protein binds to the pumilio-binding region in the 3′ UTR of the *CMV*. This in-turn suppresses the translational activity of viral RNA thereby by its replication in the host (Huh et al., 2013).

A RNA-binding protein known as Bruno initially reported in Drosophila is shown to act as repressor of translation of the *oskar* mRNA by binding to Bruno-response element located in its 3′ UTR (Kim-Ha *et al*., 1995). Several homologs of Bruno genes were found in other species including two genes in Arabidopsis namely AtBRN1 and AtBRN2. Arabidopsis plants carrying AtBRNs double mutants and transgenic plants overexpressing AtBRNs were used in a study to role of Bruno-like protein and its mechanism (Kim et al., 2013). It was observed in the study, that plants carrying mutants showed early flowering phenotype whereas the plants expressing Bruno-like proteins exhibited delayed flowering time. Further, using yeast-three hybrid assay, it was identified that Bruno protein targets the *SOC1* gene responsible for regulating flowering time, and the protein repressed its activity by binding to a motif, Bruno-response element (UAUGUAU) in the 3′ UTR of the SOC1 mRNA (Kim et al., 2013). Therefore, it is very essential to analyze the untranslated regions of the mRNA for regulatory motifs that could play a significant role in the translational activity of the gene.

2.6 Defense system in plants

Plants encounter varied environmental conditions, and stresses such as biotic and abiotic, to protect themselves it has developed different defense mechanisms. There are two mechanisms exhibited by plants depending on the type of attack or infection namely pre-formed mechanism and infection-induced mechanism. Plants also have tolerance towards certain pathogen attack. The pre-formed mechanism is also known as passive defense which comprises the physical and

chemical barrier such as the trichomes, cell walls, enzymes to breakdown bacterial toxins, and antimicrobial proteins. Infection-induced mechanism is activated by the recognition of pathogenic compounds by the receptors present on the surface and inside the cell. The receptors based on the molecules it encounters triggers different type of immune response such as pathogen-associated molecular pattern triggered immunity and effector triggered immunity discussed in detail below(Chisholm et al., 2006).

2.6.1 Pathogen-associated molecular pattern triggered immunity

Pathogenic microbes upon infecting plants encounters invokes immune response during its access to plants interior. Initially, the pathogen either penetrates the leaf or root surface or by wounding plant cells at the surface. Upon successful breach of the initial barrier it encounters the rigid cell wall and chemical compounds produced by the cells. Penetration of this barrier exposes the plant plasma membrane possessing receptors to the pathogen and the cell surface receptor posseses the ability to distinguish between the self and non-self molecules to invoke immune response Pathogens produce molecules known as pathogen-associated molecular patterns (PAMP). Cell surface receptors also known as pattern recognition receptors recognize the lipopolysaccharides and flagellins produced by the bacteria and activate the innate defense mechanism (Nurnberger et al., 2004). Flagellins, a protein produce by the bacterial flagellum is required for the virulence and mobility of the pathogen in the host. Felix et al., 1999 has identified in Arabidopsis plants, a 22 amino acid peptide known as flg22, has a conserved amino terminus, to be very essential for the trigger of host-pathogen defense activation. In an another study, it was identified upon perception of the flg22 peptide by the host receptors activates the MAP kinase signalling pathway and transcription of WRKY transcription factors, eventually conferring higher resistance to the pathogen (Asai et al., 2002). It was also identified from

studies on Arabidopsis plants carrying fls2 mutants that lacked FLS2, a flagellin receptor in Arabidopsis, showed more susceptibility to the pathogen. Whereas the fls2 mutants treated with the PAMP producing bacterial extracts exhibited resistance and inhibited the growth of the bacteria. It was concluded in the study, other than flagellin receptors, there are other PAMP molecule receptors available in the host to resist the pathogen (Zipfel et al., 2004). For example, a bacterial protein, elongation factor Tu has been demonstrated in a study to be recognized by the cell-surface receptors in Arabidopsis plants and triggers defense response similar to flagellin (Kunze et al., 2004). It is important to note that some pathogens has evolved to possess the ability to evade these resistance mechanisms and cause infection in plants.

2.6.2 Effector triggered immunity

Studies have discovered that pathogens produce effector proteins in order to manipulate the host plants and produce pathogenicity. To overcome the effects of the effector or virulence protein produced by the pathogens, plants have developed the ability to synthesize resistance (R) proteins to counteract them, which comes under the second layer of defense in plants. Pathogens uses a type III secretion system to deliver the effector proteins into the host plants, and they encode genes known as avirulence (avr) genes. The gene interaction between the host and pathogen was initially demonstrated by H.H. Flor in 1946. Since then, many R gene interaction in various host plants has been explored (Nimchuck et al., 2003). Most of the R genes encode proteins that contains nucleotide binding (NB) site and the leucine-rich repeat (LRR) domains. The NB site share motifs recognized by the bacterial elicitor proteins and activates defense mechanism. Based on the N-terminus of the R protein, it is further classified into coiled-coil (CC) NB-LRR and Toll-interlukin-1 receptor NB-LRR (Chisholm et al., 2006). More than 200 genes that codes for R protein has been discovered in Arabidopsis genome. Recognition of the

Avr gene products by the R gene in plants produces a heightened defense response leading localized host cell death known as hypersensitive (HR) response. The HR response occurs in order to prevent the spread of the pathogen in the host and its growth. Lot of work has been done to understand the mechanism of Avr protein recognition by the R proteins and how they regulate the defense response in plants.

Studies conducted in Arabidopsis against *Pseudomonas syringae* infection discovered a disease resistance gene which encodes a NB-LRR protein i.e., Resistance to *Pseudomonas syringae pv maculicola 1 (RPM1)* conferring resistance to two avirulent genes AvrB and AvrRpm1 of *P.syringae* (Chen et al., 2010). It is found the RPM1-mediated resistance requires RIN4 (RPM1-interacting protein) for its production in plants. Whereas plants expressing RPM1 and rpm1 mutants showed constitutive RIN4 production independent of RPM1. Thus, upon infection by *P.syringae* it was observed, the AvrB and AvrRPM1 phosphorylates the RIN4 protein which in turn activates RPM1 leading to hypersensitive response (Mackey et al., 2002). In another study, RIN4 was identified to activate the Resistance to *P. syringae* 2 (RPS2) gene which provides resistance against *P.syringae* expressing AvrRpt2 protein (Axtell & Staskawicz, 2003). Thus, NB-LRR proteins interact directly or indirectly with the bacterial type III effectors thereby preventing pathogenicity to plants.

2.7 Computational prediction of regulatory elements in the UTRs

2.7.1 uORF detection using UTRscan

It is known that uORFs regulate gene expression, and play vital role under certain environmental conditions, thus it is very important to identify them in order to understand their mechanism. As wet-lab techniques are expensive to use, many computational algorithms and tools has been

developed to predict uORFs. One such program is UTRscan, available under UTRsite specifically developed to analyze the 5′ and 3′ UTR sequences. The UTRscan program searches the input sequences for matches to predefined patterns or motifs from the UTRsite database, which is built based on experimental evidence and literature reports concerning ORFs. In the case of uORFs, the program scans for an in-frame start and stop codon. The database contains 473,330 5′ UTR and 527,323 3′ UTR entries of eukaryotic mRNAs, obtained from 483,605 genes across 79 species (Grillo *et al*., 2010).

2.7.2 Detection of miRNA target sites

Computational identification of miRNA target site involves searching the target sequence for similarity or complementary binding region in the target sequence. Recently lot of miRNA prediction tools has been developed based on different significant features of the miRNA such as sequence similarity, binding free energy, flanking sequences around binding site, and base pair mismatches. In this study we chose to use a plant small RNA target analysis web-server available at http://plantgrn.noble.org/psRNATarget/ (Dai and Zhao, 2011). The pipeline contains three different programs to predict miRNA targets. One such program is miRU (Zhang 2005), which predicts a miRNA target, and it calculates the complementarity between the target and miRNA. This tool when used alone lacks the ability of high-throughput analysis. The second program employed by the server, RNAup from Vienna package (Muckstein *et al*., 2006), is used to calculate the energy required by the miRNA to open-up the secondary of the target mRNA and access its binding site. This feature is unique to this tool, which has not been taken into account by other miRNA prediction tools (Patscan (Dsouza *et al*., 1997), miRNAassisst (Xie *et* al., 2007) as an important factor, but which has a great impact on target recognition. Ssearch, a third

program in the pipeline calculates the number of miRNA target sites (multiplicity) per sequence. Genes with multiple target sites has been reported to trigger transacting-siRNA biogenesis (Brodersen *et al*., 2008; Axtell *et al*., 2006). It has also been studied that a mismatch in the centre region prevents the cleavage action of RISC complex, whereas the complete binding leads to cleavage of the target (Brodersen *et al*., 2008). The server takes into account these features and based on it produces output predicting the possible function of the miRNA (cleavage or inhibition). Hence, for these reasons psRNATarget server clearly seems to be a better tool compared to other miRNA prediction tools available and which is why it is used in our analysis.

2.7.3 *De novo* motif discovery tools

Many computational tools are available to identify overrepresented pattern sequences in a gene. Among them, we plan to employ three tools, Seeder (Fauteux et al., 2008), Weeder (Pavesi et al., 2004), and MEME (Bailey et al., 2009), for the *de novo* motif analysis. Seeder is a tool that uses a discriminative motif identification algorithm that identifies over represented motifs in a group of sequences as compared with a background set of sequences (Fauteux et al., 2008). Example, for the analysis of a group of seed-specific promoters from three different plant families, the Seeder algorithm was applied, and as a result the two most significant motifs discovered in this group of promoters correspond to cis-regulatory elements that have been experimentally characterized in some promoters of seed-storage protein genes (Fauteux et al., 2008). This tool has been proved to be very efficient in performance equal to or better than other well-performing widely used motif discovery tools in benchmarking analysis.

The Multiple Expectation Maximization for Motif Elicitation (MEME) is another tool used to identify novel motifs in a set of related sequences (Bailey et al., 2009). The motif

algorithm in MEME searches for a short stretch of sequence. It takes input a set of DNA or protein sequences and aligns them to find patterns of repeated, un-gapped sequence among them. In order to use MEME for the discovery of regulatory motif, a higher order background sequence model has to be used. The order of the Markov model is the number of preceding positions considered when calculating the character frequencies for the current position. For instance, if we want to find a motif with length six, the order of markov model to choose should not be higher than two.

Finally, the Weeder tool uses exhaustive enumeration search approach to predict novel motifs (Pavesi et al., 2004). The Weeder algorithm searches for a match to a pattern with allowed number of substitutions in every sequence (Pavesi et al., 2001). The program finally outputs best top two motifs that occur among the sequence. In a computational analysis study by Tompa et al., 2005, a single data set was analyzed using different motif prediction algorithm, it has been identified that MEME and Weeder are among the top best performing motif finding tools. It has also been suggested in the study, to use combination of various motif discovery tools instead of one to identify significant over-represented motifs (Tompa et al., 2005).

Chapter 3. Materials and Methods

3.1 Source of transcriptome and translatome data

Transcriptome and translatome data was obtained from Dr. Peter Moffett at University of Sherbrooke and was generated as follows. Transgenic *Arabidopsis thaliana* plants were generated by the Moffett lab to contain a dexamethasone inducible promoter::Avrpm1 protein construct. When the construct is induced (by dexamethasone), the Avrpm1 protein is expressed and recognized by the Rpm1 protein, which leads to stimulation of rapid defense response in plants. The Avrpm1 transgenic plants were crossed with plants transgenic for Rpl18-FLAG that expresses an epitope tagged version of the RPL18 ribosomal protein, which was used to purify mRNAs bound to the ribosomes. Plants were subjected to two hours of treatment in both the presence and absence (control) of dexamethasone, respectively, and mRNAs were purified from total RNA and from ribosome bound mRNA, respectively. The four different mRNA samples (treatment, total transcriptome; treatment, ribosome-bound mRNA; control, total transcriptome; control, ribosome-bound mRNA) were sequenced using the Illumina sequencing platform. Based on the cuffdiff-derived FPKM values (Trapnell et al., 2012), the differential levels of the genes were calculated at the transcriptional and translational level in both the treatment and control samples. The translational efficiency is measured as the ratio of number of reads from total RNA (total transcriptome) to the number of reads from translatome i.e., mRNAs bound to ribosomes.

A nomenclature of YES-NO (YN) was used to represent genes with significant translational activity in control plants and plants expressing defense response (treated with dexamethasone). An assumption in this nomenclature is that under normal gene regulation, the transcriptome level of a gene would be equal to the level of that gene in the translatome (ribosome-bound mRNAs). A gene is labelled ‘YES’ (Y), when that gene has a significant translational efficiency

(transcriptome over translatome ratio). This translational efficiency can be either down or up-regulation at the translational level in comparison to the transcriptional level. A label 'NO' (N) is given to the gene if it is normally regulated based on its translational efficiency (transcriptome over translatome ratio). To investigate if any gene is under translational control in treatment or control plants, based on the logarithmic ratio of translational efficiency, each gene is denoted with YES-NO (YN), NO-YES (NY) or YES-YES (YY) designation.

As per the nomenclature and based on the log ratio of translational activity, the three main groups or lists of genes of interest are:

YES-NO (YN) - genes that are regulated at the translational level in plants expressing defense response (under treatment) but are normally regulated in control plants.

NO-YES (NY) - genes that are normally (transcriptionally only) regulated in treated plants but are regulated at the translational level in control plants.

YES-YES (YY) – genes that are significantly regulated at the translational level in both control and treated plants.

(A fourth list, NO-NO (NN) – genes that are not translationally regulated in either treated or control plants, is not of interest in this study).

Furthermore, the genes in each group were also identified as up- or down-regulated based on the comparison of their log ratio of transcript abundance at the translational level in control to the defense plants versus the log ratio of transcript abundance at the transcriptional level in both plants. In total there are thus six lists of genes: YNup, YNdown, NYup, NYdown, YYup, and YYdown.

3.2 Sequence retrieval

For each of the six gene lists in the dataset, using the gene identifiers (e.g., AT3G54220), the 5′ and 3′ UTR fasta sequences of *Arabidopsis thaliana* were retrieved using the BioMart tool from Phytozome (http://www.phytozome.net/) (Goodstein et al., 2012). Genes with no sequence content and those with very short sequences (<10 nucleotide) were removed using a custom written Python script (Appendix III).

3.3 Length and nucleotide composition analysis

The length and nucleotide composition of the genes in our dataset were calculated using in house python script. Significance of difference in the nucleotide content in the 5' UTR versus the 3' UTR regions as well as their length was calculated using a two tail Z-test (Sprinthall, 2011). The Z-score obtained from the statistical test was used to determine their significance.

3.4 uORF prediction

The 5′ UTR sequences of 463 genes differentially regulated at the translational level in *A. thaliana* under stress condition were used for upstream open reading frame (uORF) analysis. The sequences in FASTA format were subjected to open reading frame (ORF) prediction using the UTRscan tool (http://itbtools.ba.itb.cnr.it/utrscan) (Pesole et al., 2002) resulting in an output file with the ORF sequence and its position in the 5′ UTR of the genes. UTRscan searches the input sequences for matches to predefined patterns or motifs from the UTRsite database, which is built based on experimental evidence and literature reports concerning ORFs. The database contains 473,330 5′ UTR and 527,323 3′ UTR entries of eukaryotic mRNAs, obtained from 483,605 genes across 79 species (Grillo *et al*., 2010).

3.5 microRNA prediction using plant small RNA target analysis server

The miRNA target sites in the 3′ UTRs of translationally regulated genes were predicted using a plant small RNA target analysis (psRNATarget) server (Dai and Zhao, 2011; http://plantgrn.noble.org/psRNATarget/).

The 3′ UTR sequences of genes with potential translational activity were uploaded to the psRNATarget server, and compared with the 337 published miRNAs of *Arabidopsis thaliana* from miRNABase (Griffiths-Jones et al., 2006) already available at the server. The analysis was run with default parameters using a score of 3.0 for maximum expectation (mismatch value accepted), a length of 20bp for complementarity scoring, an un-pairing energy threshold value of 25.0. A range from 9 to 11 nt was set to find any mismatch in the central region, which would predict the activity of the miRNA.

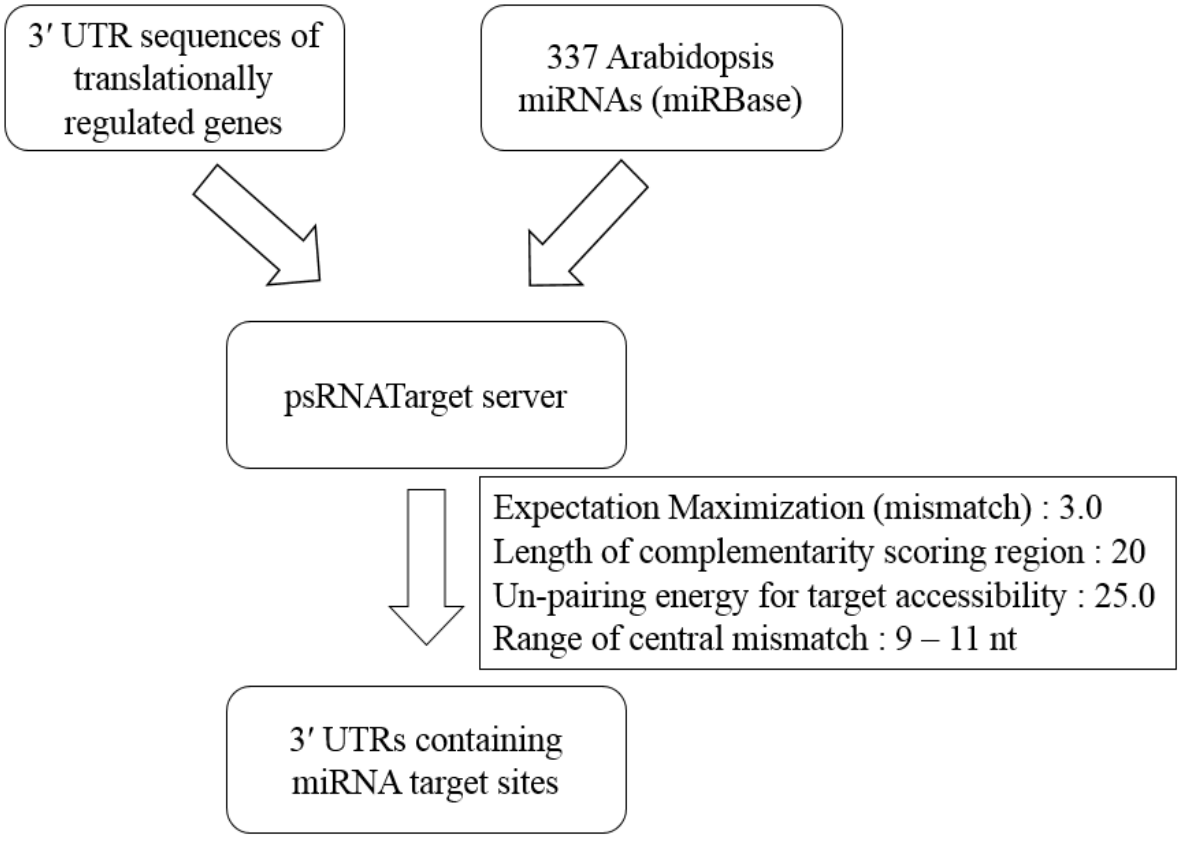


Figure 3.1: Flowchart showing steps involved in the miRNA prediction process

3.6 *De novo* motif discovery in the 3′ and 5′ UTRs

The 3′ UTR sequences (20,346) from the whole *Arabidopsis thaliana* genome were downloaded from Phytozome (Goodstein et al., 2012). The Seeder::Background module was used to generate the background distribution of seed length 6 and 8 (6- and 8-mers) for computational prediction of motifs in the dataset using Seeder (Fauteux et al., 2008). A sixth-order Markov Model was created for MEME (Bailey et al., 2009) using the 20,346 3′ UTR sequences of *Arabidopsis thaliana*. The motif discovery analysis was performed using MEME with a system of 8GB RAM and 4 core processor. The Seeder software was used for detecting significant motifs of length 6 (6-mers) with the same system. Background computation for analysis of 8-mers was performed on a Rocks Linux desktop cluster system with four compute nodes each with RAM memory of 16, 8, 8, and 8 GB respectively. Random subgroups of the 3′ UTR sequences were created from each of the six gene lists. A Python script was written to produce 500 subgroups for each dataset. Each subgroup contained ten 3′ UTR sequences of the genes that are randomly chosen from each of the six gene lists.

The 20,346 3' UTR sequences of *Arabidopsis thaliana*, were used to create frequency files of 6- and 8-mers for Weeder (Pavesi et al., 2004), which was used to predict *de novo* motifs in the entire set of each of the six lists of 3′ UTRs (i.e. no subgroups were created).

Likewise, 19,128 5′ UTR sequences of *Arabidopsis thaliana* were downloaded from Phytozome. These were used to create the Seeder and MEME background distribution as well as the Weeder frequency files for the 6- and 8-mer analysis.

3.7 Comparison of position weight matrices and creation of sequence logos

To eliminate redundant motifs identified in the motif discovery analysis, a pairwise comparison tool, Tomtom (Gupta et al., 2007), available under the MEME suite was used. The motifs obtained were matched against each other to find their similarity based on a threshold E-value cut-off of 0.05. The motifs that were found to be similar were clustered using a python script to produce an average position weight matrix of all the similar motifs. Again, using the Tomtom tool, the clustered average matrix was queried against the motifs with known function from literature to find matches and annotate the motifs. Finally, the Weblogo software (Crooks et al., 2004) was used to create sequence logos, which display information content in bits at each nucleotide position of 6- or 8-mers.

Chapter 4. Results

4.1 Identification of stress-regulated genes at the translational level

In order to study gene regulation at the translational level during stress, differential expression analysis was carried out on RNA-Seq data (Illumina) sequencing reads of total and ribosome-bound mRNAs from control and dexamethasone treated *Arabidopsis thaliana* plants.

The differentially expressed genes were classified into six different groups depending on them being likely transcriptionally or translationally regulated in control and/or treated plants (representing stress), and whether this regulation was up (more transcripts) or down (less transcripts). The YNup and YNdown groups contain genes that are translationally up-regulated and down-regulated, respectively in treated plants (i.e. under stress) but are normally regulated (transcriptionally) in control plants The NYup and NYdown groups contains genes with significant translational up-regulation and down-regulation, respectively, in control plants but with normal regulation (transcriptional) in treated plants. The YYup and YYdown groups contain genes that are up-regulated and down-regulated, respectively at the translational level in both control and treated plants. Based on the cuff-diff derived FPKM values, in total 514 genes (Appendix I) are noticeably regulated at the translational level during stress, spread between 12-265 genes depending on the group (Table 4.1). Among them, the NYdown group had the highest number of genes (265 / 514) around 51.5% whereas the YYup had the least number with only 12 genes in its group (Table 4.1).

Since the untranslated regions (UTRs) of genes are of great importance to the ability of a transcript to be translated, the UTRs of the six different groups of genes were investigated for potential conserved sequences, such as regulatory elements, miRNA target motifs and upstream open reading frames (uORFs). Genes with 3' and 5'UTR sequences longer than 10 nucleotides

were selected using a custom written Python script (Appendix III). The resulting number of genes with 5′ and 3' UTR sequences for each group is shown in Table 4.2.

4.2 The length of 5′ and 3′ UTRs may impact translation of the transcript

The untranslated regions of the 514 likely translationally regulated genes, were investigated for length, as this may have an impact on translation of the transcript (Jackson and Standart, 1990; Fabian et al., 2010; Von Arnim et al., 2014). Out of these, 455 genes have a 5′ UTR sequence with an overall length ranging from about 10 to 745 nucleotides. The average length of the 5′ UTR is 138 nucleotides and the average length of whole 5′ UTR is 140 nucleotides.

Likewise, 470 genes have 3′ UTR sequences ranging from 12 to 1048 nucleotides, with an average length of 237 nucleotides. The average length of all 3′ UTR is 227 nucleotides.

4.3 Base composition of the 3′ and 5′ untranslated regions

The nucleotide content of the UTRs of genes exhibiting significant translational activity was calculated, since the nucleotide composition can influence the specific function of the gene (Kawaguchi and Bailey-Serres, 2005; Kochetov et al., 2002). As a result, considerable differences in nucleotide content were found between the 5′ and 3′ UTR. The 5′ UTRs have an average 62.78 % A+T content and a 37.22 % G+C content, while the 3′ UTRs have an average of 68.52% A+T content and 31.48 % G+C content. Notably, the G+C content is higher in the 5′ UTR in comparison to its 3′ UTR. Based on statistical two-tail z-test, the difference was found to be highly significant. The nucleotide content may play a significant role in the mRNA secondary structure formation, stability and its function (Kawaguchi and Bailey-Serres, 2005; Kochetov et al., 2002).

Table 4.1 Number of translationally regulated genes in Arabidopsis plants treated with (Defense) and without (Control) dexamethasone

Group	NYup	NYdown	YNup	YNdown	YYup	YYdown
Number of genes with differential translational efficiency	90	265	65	58	12	24

Up, up-regulated genes; down, down-regulated genes
NY represents genes translationally regulated in control but normally regulated in treated plants
YN represents genes translationally regulated in treated plants but normal in control plants
YY represents genes translationally regulated in both control and treated plants

Table 4.2 Number of differentially regulated genes possessing 5′ and 3' UTR sequences

Group	NYup	NYdown	YNup	YNdown	YYup	YYdown
Number of genes with 3′ UTR sequence	85	241	59	54	10	21
Number of genes with 5′ UTR sequence	79	237	58	53	6	22

Up, up-regulated genes; down, down-regulated genes
NY represents genes translationally regulated in control but normally regulated in treated plants
YN represents genes translationally regulated in treated plants but normal in control plants
YY represents genes translationally regulated in both control and treated plants

4.4 Prediction of uORFs in the 5′ UTR

4.4.1 Presence of uORFs in the 5′ UTR of stress-regulated genes

Upstream open reading frames (uORFs) are thought to influence translation of downstream protein coding regions of a gene. Therefore, the 5′ UTRs of the genes in the set of translationally regulated genes were searched for open reading frames (ORFs) using the UTRscan program (Pesole et al., 2002). The results show that approximately 19 % (89 / 455) of the genes contain one or more uORFs. In total, 106 uORFs were identified from 89 genes (Table 4.3) exhibiting significant translational regulation during stress. Of those uORFs, 65 (or 73%) were located 20 base pairs away from the start of the 5′ UTR sequence (i.e., from the 5′ end of the 5′ UTR), which could possibly be regulating the translational process of the genes containing them. The full list of uORFs are detailed in Appendix II. Among the genes predicted to have uORFs in the 5′ UTR, some of them were identified to contain multiple number of uORFs (Table 4.4).

4.4.2 Start codon sequence context of the predicted uORFs

Previous studies have shown that the nucleotides around the start (AUG) codon of the uORFs play a major role in determining the translation efficiency of the uORF and its effect on the main coding region in the mRNA (Kozak, 1987 & 1999). In particular, the A(A/G)CC**A$_{+1}$UG**GC called kozak signal, seems to be of importance. Therefore the sequences surrounding the start codons of the predicted uORFs were analyzed. As seen in Figure 4.1, eight uORFs with strong sequence context around the start AUG codon with nucleotides 'A / G' at -3 and 'G' at +4 positions (A of AUG codon is designated +1) were identified. In addition, approximately 18 % of the identified uORFs were found to have a sub-optimal sequence context around its AUG codon, which could possibly be recognized by the ribosome under some specific conditions and initiate translation process.

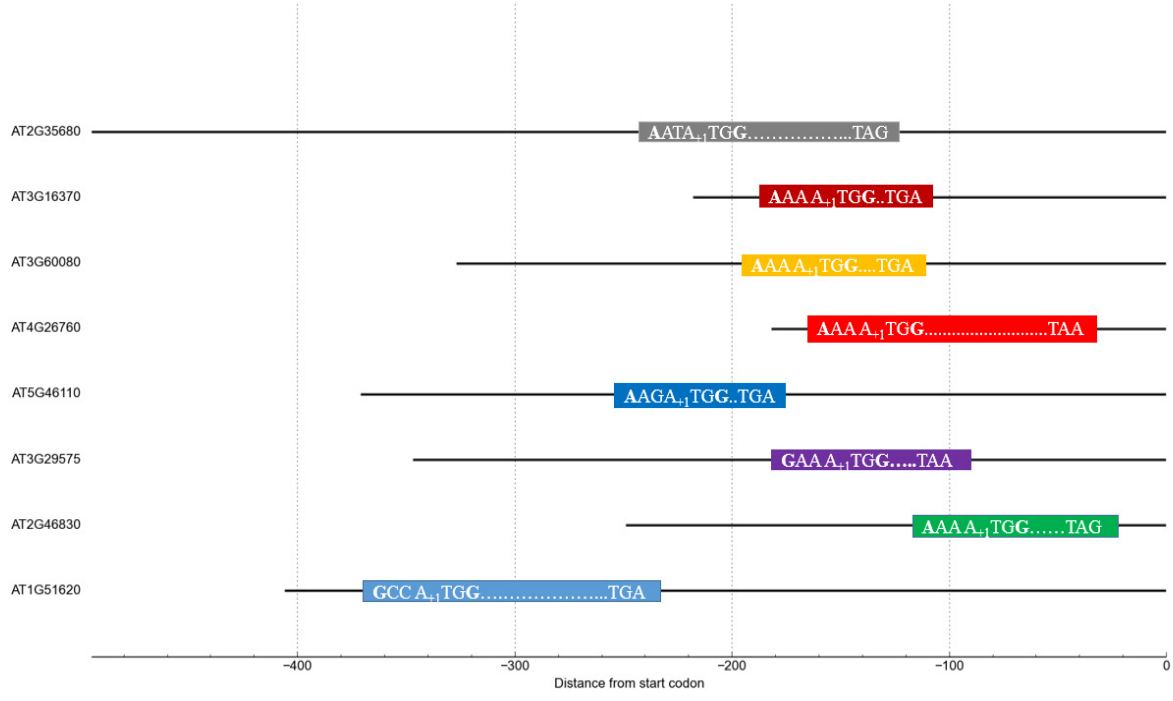


Figure 4.1: Position of uORFs possessing strong kozak sequence context.

Nucleotides [A/G] at position -3 and G at +4, where A of uORF ATG codon is designated +1 is the kozak signal. 0 indicates the start of the main protein coding sequence.

Table 4.3 Number of uORFs found in the 5′ UTRs of translationally regulated genes during defense

Dataset	Number of 5′ UTR sequences	Number of 5′ UTRs containing uORFs	Number of predicted uORFS
YNdown	53	15	18
YNup	59	8	9
NYdown	241	50	62
NYup	80	13	14
YYdown	23	2	2
YYup	7	1	1

Table 4.4 Examples of genes containing one or more uORFs in their 5′ UTR

Gene	Gene description	Number of uORFs	uORF location in the 5′ UTR
AT2G46830	Circadian clock associated 1, a transcription factor	1	[147 - 232]
AT1G54260	Winged-helix DNA-binding transcription factor family protein	2	[23 - 97], [156 - 245]
AT4G02280	Sucrose synthase enzyme 3	2	[3 - 107], [145 - 222]
AT3G29575	ABI five binding protein 3	2	[86 - 169], [178 - 267]
AT1G51620	protein kinase family protein	2	[35 - 169], [237 - 365]
AT1G74088	Unknown protein	3	[45 - 116], [132 - 197], [213 - 326]

4.5 microRNA target sites located in the 3′ UTR

Several studies have demonstrated the activity of miRNA in plant gene expression during plant stress and development process. In this study, miRNA targets were analyzed in the 3′ UTRs of genes predicted to be translationally regulated under stress, using the psRNAtarget web server (Dai and Zhao, 2011). In total, 26 miRNA target sites were predicted in the down-regulated groups (YNdown, NYdown, and YYdown) whereas only 5 target sites were identified in the upregulated NYup group (Table 4.5). The genes in the YNup and YYup groups do not have any miRNA target site in their 3′ UTRs. Specifically, the number of miRNA target sites identified in each group were 22, 2, 2 and 5 in the NYdown, YNdown, YYdown, and NYup, respectively. Based on this analysis, approximately 8% (26 out of 316) of the translationally down-regulated genes have miRNA target sites, as do around 3% (5 out of 154) of the translationally up-regulated genes. The presence of more miRNA binding sites in the 3′ UTRs of down-regulated genes is consistent with the role of miRNA, i.e. they either suppress or inhibit gene expression.

Most of the identified target sites match perfectly with complementary regions of miRNAs. The least mismatch is ≤ 3 base pairs between them. Importantly, the mismatches are present at either end of the miRNA complementary region, and very rarely at the central region between 9 – 11 nucleotides, which determines the activity of the miRNA in either cleaving or inhibiting the mRNA expression.

Genes AT4G12080, AT1G53160, and AT2G03750 had target sites for the miR156 / miR157 in their 3′ UTRs. The AT1G53160 encodes the transcription factor SPL (SQUAMOSA promoter binding protein-like), which is involved in the regulation of flowering specifically during the floral induction in plants. This gene has been identified in our analysis, and from previous studies to be regulated by the miR156 (Cardon et al., 1997, Ruiz-Ferrer & Voinnet, 2009).

The AT3G25660 gene encodes an amidase family protein, and it is among those predicted to be translationally down-regulated under the impact of stress response in our Arabidopsis transgenic plants. It has a miRNA binding region in its 3′ UTR between 134 – 153 bp for miR5021 (Fig 4.2). Analyzing the 5′ and 3′ end around the binding site of the miRNA (miR5021) revealed a complete and a partial mismatch at either end of the complement, and a full complementarity in the central region. Interestingly another gene AT3G54220, a basic-leucine zipper domain containing protein, functioning similar to a DNA binding protein involved in root radial organization and leaf development was predicted to have a target site for the miR5021 in the region 1 – 20 of the 3′ UTR. Analyzing its complementary binding region, a nucleotide mismatch was identified in the middle region of the mRNA-miRNA duplex. (Figure 4.2).

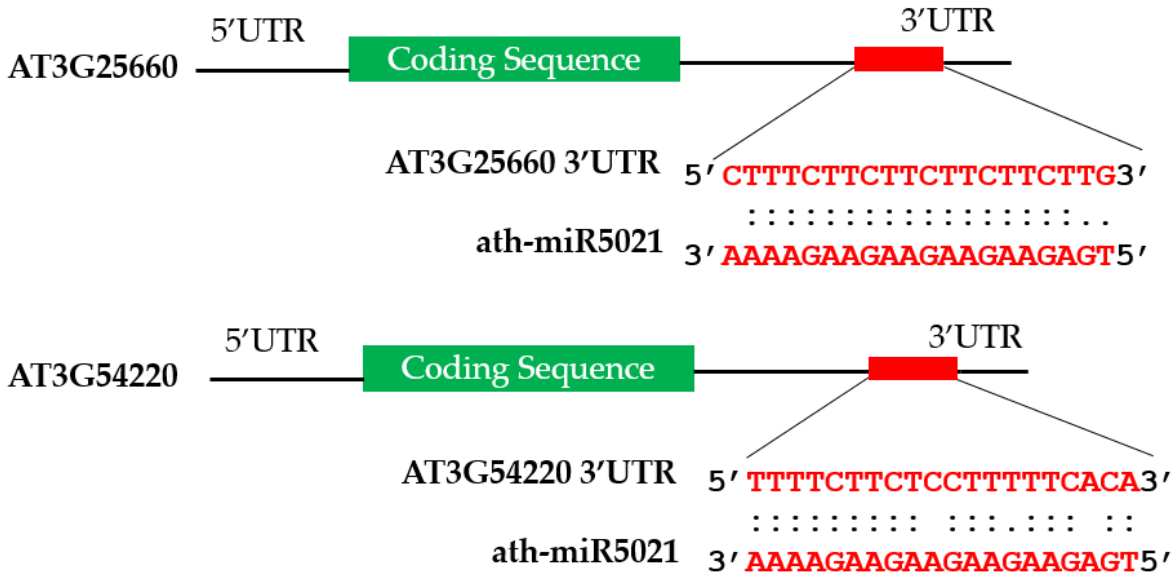


Figure 4.2: Using psRNA target web server, genes of YNdown group AT3G25660 and AT3G54220 were predicted to have miRNA binding site for the microRNA ath-miR5021 in their 3′ UTR. Gene AT3G25660 has three nucleotide mismatches present on either end of their binding region which could lead to miRNA cleavage activity; whereas the gene AT3G54220 has a nucleotide mismatch in the central region of the binding site possibly leading to mRNA translation inhibition.

Table 4. 5: Translationally regulated genes containing miRNA target sites in their 3' UTR

miRNA accession[1]	Target accession[2]	UPE[3]	Target start[4]	Target end[4]	miRNA aligned fragment[5]	Target aligned fragment[6]	Inhibition[7]
ath-miR157a	AT4G12080	7.942	11	30	TTGACAGAAGATAGAGAGCA	AATTGTCTTTTATGTCTCTT	Cleavage
ath-miR157d	AT2G03750	9.025	10	29	TGACAGAAGATAGAGAGCAC	GTGTTTTTTTTCTTTTGTCA	Translation
ath-miR161.2	AT3G43740	11.944	120	139	TCAATGCATTGAAAGTGACT	AGTAACTTGCAATGTATTGG	Cleavage
ath-miR172a	AT3G14770	12.993	53	73	AGAATCTTGATGATGCTGCAT	ATGTGGTGACATCAAGATTCT	Cleavage
ath-miR156h	AT2G03750	9.025	10	29	TGACAGAAGAAAGAGAGCAC	GTGTTTTTTTTCTTTTGTCA	Cleavage
ath-miR414	AT1G60870	10.465	5	25	TCATCTTCATCATCATCGTCA	TGATGATGATGATGAATATGA	Cleavage
ath-miR414	AT1G15690	22.818	22	42	TCATCTTCATCATCATCGTCA	TGACGATGATGAAGAAGAAGA	Translation
ath-miR415	AT1G15690	8.295	189	208	AACAGAGCAGAAACAGAACA	TGTTTTCTTTTTGTTCTGTT	Cleavage
ath-miR837-3p	AT3G28180	12.194	50	70	AAACGAACAAAAAACTGATGG	TCACTATTTTTTTGTTTGTTT	Cleavage
ath-miR837-3p	AT3G60080	9.603	21	40	AAACGAACAAAAAACTGATG	TATGAGTTTTTTTTTTGTTT	Cleavage
ath-miR1886.2	AT4G02280	10.606	223	243	TGAGATGAAATCTTTGATTGG	TCAATGGAACATTTCATTTCA	Cleavage
ath-miR5021	AT1G20696	12.441	110	129	TGAGAAGAAGAAGAAGAAAA	TTTTCTTTTTCTTTTTCTTA	Cleavage
ath-miR5021	AT4G30993	15.186	747	766	TGAGAAGAAGAAGAAGAAAA	TCTTCTTCTTCTTCTTCTTA	Cleavage
ath-miR5021	AT4G38600	6.647	277	296	TGAGAAGAAGAAGAAGAAAA	TTTTTTTCTTCTTCTTCTTC	Cleavage

ath-miR5021	AT5G27490	18.273	28	47	TGAGAAGAAGAAGAAGAAAA	TTTTTTTTTTCTTCTTTTCC	Cleavage
ath-miR5021	AT5G28050	14.765	19	38	TGAGAAGAAGAAGAAGAAAA	CATTCTTCTTCTTCTTCTCC	Cleavage
ath-miR5021	AT3G23150	7.714	131	150	TGAGAAGAAGAAGAAGAAAA	TTTTTTTGTTCTTTTTCTCA	Cleavage
ath-miR5021	AT5G46110	12.128	178	197	TGAGAAGAAGAAGAAGAAAA	TTTTATTCTGCTTCTTCTCT	Translation
ath-miR5021	AT1G33240	4.847	180	199	TGAGAAGAAGAAGAAGAAAA	TTTTGTTCTTTTCTTTCTCA	Cleavage
ath-miR156j	AT2G03750	9.025	10	29	TGACAGAAGAGAGAGAGCAC	GTGTTTTTTTTCTTTTGTCA	Cleavage
ath-miR5654-3p	AT1G04680	11.760	18	37	TGGAAGATGCTTTGGGATTT	AAATTCAAAACCATCTTCCA	Translation
ath-miR5658	AT4G38740	10.473	24	43	ATGATGATGATGATGATGAA	TTTAGCATCATCGTCGTCAT	Cleavage
ath-miR156a	AT1G53160	6.849	10	29	TGACAGAAGAGAGTGAGCAC	CTGCTCTCTCTCTTCTGTCA	Cleavage
ath-miR855	AT1G70320	11.692	127	147	AGCAAAAGCTAAGGAAAAGGA	TCTATTTCCTTAGATTTTGTT	Cleavage
ath-miR4243	AT1G67090	16.617	244	263	TTGAAATTGTAGATTTCGTA	TATGAAATTTACAAGTTTAA	Cleavage
ath-miR156j	AT1G53160	6.849	10	29	TGACAGAAGAGAGAGAGCAC	CTGCTCTCTCTCTTCTGTCA	Cleavage
ath-miR5658	AT5G41700	13.744	127	147	ATGATGATGATGATGATGAAA	TTTTATTATCATCATCTTTAT	Cleavage
ath-miR5021	AT3G25660	10.698	134	153	TGAGAAGAAGAAGAAGAAAA	CTTTCTTCTTCTTCTTCTTG	Cleavage

ath-miR5021	AT3G54220	6.604	1	20	TGAGAAGAAGAAGAAGAAAA	TTTTCTTCTCCTTTTTCACA	Translation
ath-miR5658	AT5G14740	7.535	45	64	ATGATGATGATGATGATGAA	ATCATCATCGTCGTCATCAT	Cleavage
ath-miR5998a	AT4G33010	7.352	158	177	ACAGTTTGTGTTTTGTTTTG	GAAAACAAAACACAAACTTT	Cleavage

[1]Accesion number of the miRNA
[2]Accesion number of the target gene
[3]Un-pairing energy required by the miRNA to unfold the secondary structure and access the target binding site
[4]Indicates start and end position on the target 3' UTR
[5]Sequence of the miRNA that binds to the target
[6]Sequence of the 3' UTR complementary to the miRNA
[7]Function predcited based on the target-miRNA duplex match

4.6 Genes regulated at the translational level contains regulatory motifs

In order to understand the mechanisms of gene regulation at the translational level, putative regulatory motifs were predicted in the 5′ and 3′ UTRs of the mRNA. The approach was to divide the set of sequences into random sub-groups and analyze those with the three *de novo* motif discovery tools Seeder (Fauteux *et al*., 2008), Weeder (Pavesi *et al*., 2004) and MEME (Bailey *et al*., 2009) using the untranslated sequences of the whole Arabidopsis genome as background distribution model. Based on the statistical threshold cut-off value such as Q-value ≤ 0.05 for Seeder, E-value ≤ 0.05 for MEME and top significant motifs produced by the Weeder adviser, a number of significant motifs were found in the 5′ and 3′ UTRs, detailed below.

4.6.1 Motifs identified in the 3′ and 5′ UTR

The random sub-groups containing the 3′ and 5′ UTR sequences were analyzed for the presence of motifs of length 6 and 8.

Several motifs were detected with high significant cut-off value in the 3'UTRs. Seeder identified over-represented 6- and 8-mer sequence motifs among the subgroups of genes in the YNup, YNdown, NYup, NYdown, and YYdown groups, but not in the YYup group. MEME identified 6-mers only in subgroups of the NYdown and NYup groups, 8-mers in the NYdown, NYup and YNdown groups, and no motifs in the YNup, YYup and YYdown groups. Weeder detected 8-mers in all six groups, and 6-mers in three of the six groups (YNup, YYdown, and YYup). The exact number of significant motifs detected in each group is shown in Table 4.6 and 4.7.

The number of 6- and 8-mer motifs discovered in the 5′ UTRs also varied between the three different tools used (Table 4.6 and 4.7). The 8-mer motifs were detected in almost all the

groups by the three motif prediction tools except that motifs were not identified by MEME in the YYup group. In the prediction of 6-mers, significant motifs were found in five out of six groups by Seeder and MEME, and none were identified in the YYup group. Finally, Weeder was able to produce top significant 6-mer motif for the groups YNup, YYdown and YYup, but no over-represented motifs were predicted in the NYdown, NYup, and YNdown groups.

Many identified motifs are similar to each other within the same group. We carried out a pairwise comparison analysis using the position weight matrices of the motifs to determine their similarity. Using the Tomtom tool (Gupta et al., 2007) for pairwise comparison, based on a threshold cut-off value of 0.5, the motifs that are similar were clustered to produce an average position weight matrix containing the information content of the nucleotides from each and every matrix combined into one. The clustered average matrix was then used to find genes having those motifs in the individual dataset as well as in the whole Arabidopsis genome.

Finally, in order to predict the function of the discovered motifs, the literature was searched to find similar motifs as there is no database analogous to TRANSFAC (Wingender et al., 1996) or JASPAR (Sandelin et al., 2004) available for regulatory motifs identified in the UTR. As a result two motifs relevant for 3′ UTRs and one for 5′ UTRs were identified that are experimentally tested and found to play a significant role in gene regulation at the translational level. The motifs in the 3′ UTR are [TGTA (A/C/U) ATA], a Pumilio protein binding motif (Huh et al., 2013) and [T(G/A)T(A/G)T(G/A)T], Bruno-like protein binding motif (Kim et al., 2013); and the motif in the 5′ UTR is [TAGGGTTT] (Liu et al., 2012). Position weight matrices were made of these three motifs and used to compare against the motifs discovered in the study. As a result, 23 motifs identified from the 3′ UTRs were found to match the binding motifs of Pumilio and Bruno-like protein (Supplemental Table 1-6). Examples of genes containing these two motifs are mentioned in Table 4.8 and 4.9. Likewise, 7 motifs from the 5′ UTR matched to the

single motif [TAGGGTTT] identified from literature (Supplemental Table 7-12). For instance, motif S29 identified by Seeder in the 5′ UTR matched to the reverse complement of the motif [TAGGGTTT]. Motif S29 predicted from the NYup group was found to be present in 12 out of 79 genes in the specific group. Among the genes of interest containing the S29 motif, gene AT4G27310, encodes a B-box type zinc finger family protein possessing DNA binding transcription factor activity is involved in transcription regulation (Riechmann et al., 2000); whereas genes AT2G07706 and AT1G20970 encodes an uncharacterized protein. Gene AT5G59880, encodes actin depolymerizing factor 3 (ADF3) known to play a role in biological processes such as depolymerisation of actin filament, gluconeogenesis as well as in response to specific stress in plants (Feng et al., 2006; Sweetlove et al., 2002). Another gene, AT1G10940 encodes a protein kinase similar to the calcium/calmodulin-dependent protein kinase subfamily and the SNF1 kinase subfamily (SnRK2) is involved in plant response to stress (Kulik et al., 2012).

Table 4.6 Number of significant motifs identified in the 3′ UTRs

Motif length	*De novo* motif discovery tool	NYdown	NYup	YNdown	YNup	YYdown	YYup
6-mer	**Seeder**	3	16	3	1	3	None
	MEME	1	1	None	None	None	None
	Weeder	None	None	None	1	1	1
8-mer	**Seeder**	56	62	29	7	22	None
	MEME	6	3	2	None	None	None
	Weeder	1	1	1	1	1	1

Table 4.7 Number of significant motifs identified in the 5′ UTRs

Motif length	*De novo* motif discovery tool	NYdown	NYup	YNdown	YNup	YYdown	YYup
6-mer	**Seeder**	36	91	79	36	86	None
	MEME	87	28	81	140	27	None
	Weeder	None	None	None	1	1	1
8-mer	**Seeder**	57	39	50	42	49	None
	MEME	6	8	8	7	4	None
	Weeder	1	1	1	1	1	1

Table 4.8 Genes with motifs [U(G/A)U(A/G)U(G/A)U] recognized by Bruno-like protein in their 3′ UTR

Gene ID	Gene description	Position	Sequence that matches to Bruno-like protein motif
AT2G06520	Encodes a protein similar to spinach photosystem II subunit PsbX	-94	C **TGTGAT**
AT4G38460	Geranylgeranyl reductase involved in isoprenoid biosynthetic process	-169	C **TGTGAT**
AT5G57350	Arabidopsis H(+)-ATPase	-230	C **TGTGAT**
AT2G06520	Encodes a protein similar to spinach photosystem II subunit PsbX	-93	**TGTGTAT** T
AT5G53030	Unknown protein	-70	**TGTGTAT** A
AT2G29670	Tetracopeptide repeat-like superfamily protein	-256	T **TATGTAT**
AT4G38460	Geranylgeranyl reductase involved in isoprenoid biosynthetic process	-19	C **TATGTAT**
AT4G38460	Geranylgeranyl reductase involved in isoprenoid biosynthetic process	-15	G **TATGTAT**
AT4G25570	Alpha/beta-Hydrolases superfamily protein	-143	T **TATGTAT**
AT4G14500	Polyketide cyclase/dehydrase and lipid transport superfamily protein	-54	A **TATGTAT**
AT1G56280	Unknown protein	-289	C **TATGTAT**

AT3G54220	Similar to DNA binding protein containing basic-leucine zipper region	-95	T **TATGTAT**
AT4G30400	RING/U-box superfamily protein	111	**ATACATA** T
AT3G46450	Cytosolic factor family protein / Phosphoglyceride transfer family protein	59, 95	**ATACATA** G
AT3G54220	Similar to DNA binding protein containing basic-leucine zipper region	138	**ATACATA** G
AT2G34070	Member of TRICHOME BIREFRINGENCE-LIKE gene family	11	A **ATACACA**
AT2G29670	Tetracopeptide repeat-like superfamily protein	66	T **ATACACA**

*Bold and underlined represents the sequence that matches to motif

Table 4.9 Genes with motifs [UGUA (A/C/U) AUA] recognized by Pumilio protein in their 3′ UTR

Gene ID	Gene description	Position	Sequence that matches to Pumilio protein motif
AT4G25570	Encodes cytochrome b561	-85	TG **TGTAAA**
AT3G01400	ARM repeat superfamily protein	-165	GG **TGTATA**
AT5G53030	Unknown protein	-70	TG **TGTATAT**
AT1G78630	Embryo defective 1473 involved in embryo development	206	TG **ACATGT**
AT3G49670	Encodes a CLAVATA1-related kinase-like protein	204	A **TATGTAC**
AT4G30400	RING/U-box superfamily protein	111	AT **ACATAT**
AT4G00490	Encodes a chloroplast beta-amylase enzyme	-220	TT **ACATAT**
AT4G17080	Histone H3 K4-specific methyltransferase SET7/9 family protein	-132	GT **ACATAT**

*Bold and underlined represents the sequence that matches to motif

4.6.2 A conserved sequence motif identified using MEME found in the down-regulated set of genes present in the miRNA binding site

Using a *de novo* motif discovery tool, MEME, a motif (Figure 4.3) was discovered in 26 genes among the set of 54 down-regulated genes in the 3′ UTR identified based on FPKM values in plants under stress. Interestingly, this motif is in the miRNA target site of the two genes that were predicted to bind to ath-miR5021. In addition, this motif is present in the 5′ end region of the miRNA, which is highly responsible for miRNA recognition of the target mRNA, called a seed region (2-11 of 20 nucleotides) (Figure 4.4). It is possible that other genes containing this motif might also be regulated by miRNAs except that it was not predicted in our analysis for reasons such as the complementarity, and the un-pairing energy optimal threshold value (Kertesz et al., 2007; Dai and Zhao, 2011) used to predict miRNA sites might vary in the *in vivo* conditions of plants.

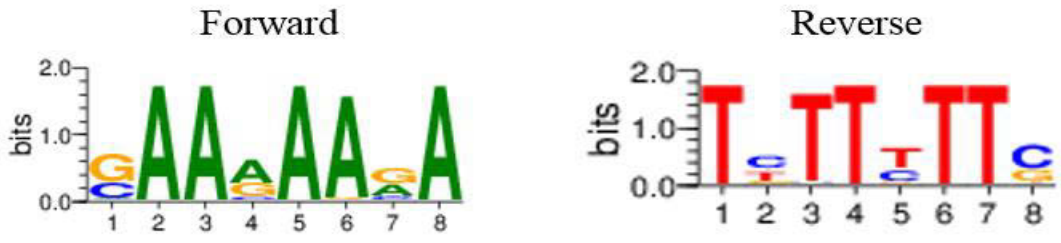


Figure 4.3: A *de novo* sequence motif identified using MEME motif discovery tool found to be conserved in the 3′ UTR of 26 out of 54 genes that belong to YNdown group which are computationally predicted to be down-regulated in plants during stress.

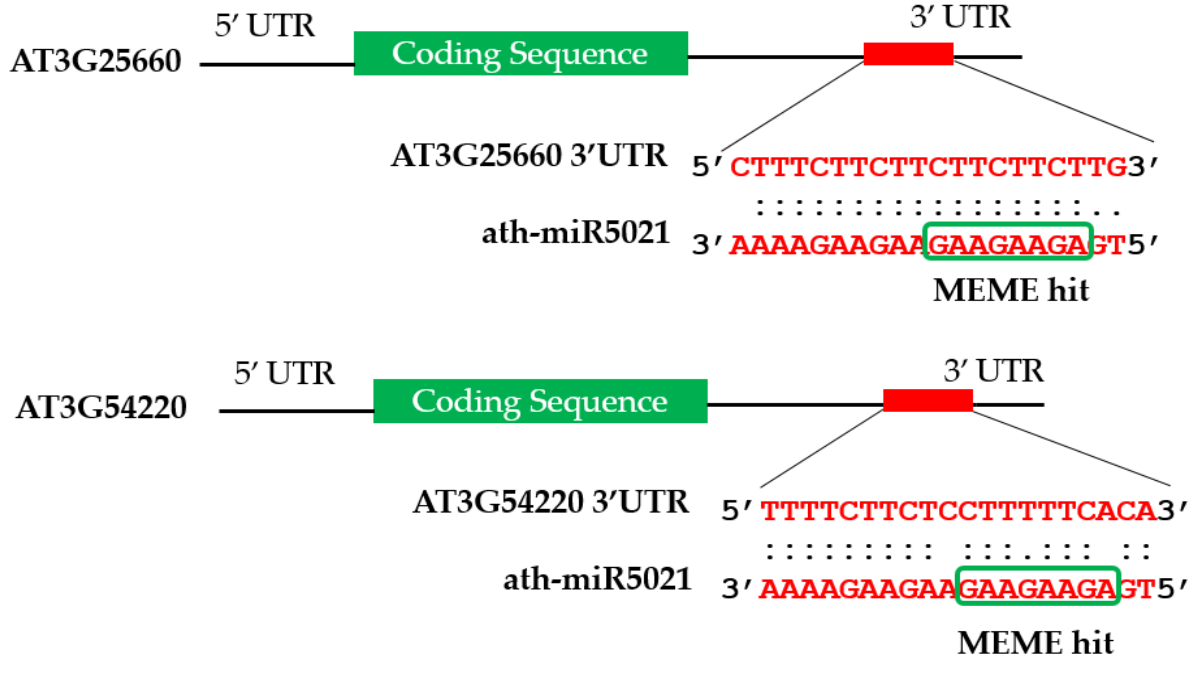


Figure 4.4: Two genes in the YNdown group predicted to be translationally down-regulated was identified in the analysis to contain a binding site for the microRNA (ath-miR5021). In addition a sequence motif discovered by MEME was found to be conserved within the miRNA binding region in their 3′ UTR. The motif is present between 2 – 11 nucleotides from 5′ end of the miRNA, which is known as the seed region, important for miRNA activity.

Chapter 5. Discussion

5.1 UTR length and its nucleotide composition plays a deterministic role in the mRNA function

To understand the role of untranslated regions in the process of gene translation, the length of the UTR and its nucleotide base content was analyzed. Studies have demonstrated that longer 3′ UTR is essential for the stimulation of translation under certain stress conditions (Tanguay et al., 1996; Kozak, 1988; Pesole et al., 2001). From the analysis, the average length of the 3'UTR of the translationally regulated genes under stress were higher in comparison to the non-translationally regulated genes. According to statistical z-test, the difference in UTR length between them was found to be significant. Studies indicate high GC content as characteristics for formation of stable secondary structure, which could hinder the movement of ribosome during the scanning process of gene translation (Mignone et al., 2002; Pesole et al., 2001). Likewise, based on the statistical test, the G+C content of the 5′ UTR as well as the 3′ UTR was considerably higher in translationally regulated genes when compared to non-translationally regulated genes.

5.2 uORFs located in the 5′ UTRs may be involved in translational regulation

In order to understand the role of uORFs during gene translation under stress, several uORFs were detected in the 5′ UTR of translationally regulated genes expressed under stress. Studies have reported that sequences around the AUG start codon of the uORF play an important role in promoting translation initiation in plants (Joshi et al., 1997; Kawaguchi and Bailey-Serres, 2005). There is a high probability for the uORFs with strong kozak signal to be translated upon ribosome recognition whereas the uORFs with weak or sub-optimal nucleotide context might get

translated under certain specific conditions in plants, however the exact mechanism of uORF start codon recognition by the ribosome is not yet clear (Nyiko et al., 2009). Evidence from several studies show that the uORFs position in the 5′ UTR, among other factors, has great impact on its functional role (Van den Heuvel et al., 1989). Out of the 106 predicted uORFs, 65 were located 20 nucleotides away from the start of the 5' UTR. Studies have shown this to be the optimal distance for uORF translation initiation to occur (Van den Heuvel et al., 1989). It is also to be noted that these 65 uORFs are positioned relatively at a distance of approximately 20 nucleotides from the downstream main ORF (Vilela et al., 1999; Child et al., 1999).

Based on the factors that determine uORF functionality, some of the predicted uORFs could act in a sequence-dependent manner. One of the translationally regulated genes with a detected uORF, AT2G46830 encodes *circadian clock associated 1* (CCA1), a transcription factor protein involved in regulating circadian system of Arabidopsis, and is required for sensing changing environmental conditions such as light and temperature (Kangisser et al., 2013). A functional study (Zhou et al., 2015) in Arabidopsis against the downy mildew pathogen, *Hyaloperonospora arabidopsis (Hpa),* looked at the resistance genes and found significant enrichment of the evening element, which is regulated by CCA1. A mutant study was carried to confirm the role of CCA1, it was discovered that *cca1* mutant had no resistance against *Hpa,* while the mutant line with overexpressed CCA1 showed enhanced resistance (Zhou et al., 2015). The study hypothesised that regulation of defense genes by CCA1 acts like a signal for plants to anticipate the infection and prevent damage (Zhou et al., 2015). AT2G46830 contains a uORF between positions 147 – 232 in the 249 nt long 5' UTR. The uORF of this gene has a strong kozak sequence required by the ribosome for translation initiation. In addition, this uORF is located a few nucleotides away from the main ORF, which could result in translation of the uORF after which translation reinitiation process might occur to translate the main ORF located downstream

of it. From the differential gene expression analysis (Chapter 3), this gene is predicted to be down-regulated at the translational level during stress. Therefore it is highly likely that the uORF upon translation would either produce a peptide and in-turn stall the ribosome from further scanning thereby inhibiting the main ORF translation, or the ribosome after uORF translation initiation would encounter a pre-termination codon triggering a non-sense mediated decay process. Studies show that reduced levels or loss of CCA1 function in Arabidopsis affects various developmental processes regulated by light, change in flowering time and disrupts the function of circadian clock and its related gene expression (Wang and Tobin, 1998; Kangisser et al., 2013; Green and Tobin, 1999).

Two other genes of interest are AT1G74088, encoding a protein of unknown function, and AT1G54260, coding for winged-helix DNA-binding transcription factor family protein involved in nucleosome assembly (Wierzbicki and Jerzmanowski, 2005). These two genes are among the YNdown genes, and they have three and two uORFs, respectively. It is possible that the expression of these two genes might be inhibited by the presence of more than one uORFs through stalling the ribosome from scanning the mRNA transcript. In addition, the time required to transverse the 5' UTR of these genes by the scanning ribosome would be very long. Thus, it is possible that translation of the main open reading frame of those two genes would be inhibited in the process under stress condition. Likewise, for instance a gene AT4G02280 that belong to NYdown group had two uORFs in their 5′ UTR. AT4G02280 encodes a sucrose synthase enzyme 3 involved in processes like starch metabolism, sucrose biosynthesis and metabolism; and studies have demonstrated the requirement of sucrose synthase activity during hypoxia and water deprivation condition in plants (Bieniawska et al., 2007). It has been discovered that the circadian clock, which senses environmental conditions and triggers immune response, regulates many sugar-metabolizing enzymes (Blasing et al., 2005). Modulation of sucrose under stress

affects the sugar specific signalling pathways including the flowering time of the plant (Seo et al., 2011). Additional genes of interest from the up-regulated groups, NYup and YNup, such as AT3G29575 and AT1G51620 contain two uORFs. AT3G29575 encodes an ABI five binding protein 3 (AFP3) and studies have discovered that ABI five related proteins play a major role in abscisic acid signalling and in various stress response to modulate the seedling development and growth where it is highly expressed (Li et al., 2008; Garcia et al., 2008). Gene AT1G51629, a protein kinase family protein, has been discovered in studies to be highly regulated during resistance against bacterial *P*. *syringae* infection on plants (Mohr and Cahil, 2007). As these two genes AT3G29575 and AT1G51620 are found to play significant role during stress, the possible mechanism that could occur with the presence of uORFs in their 5′ UTR is that under stress the ribosome might by-pass without recognizing the uORF start codon through leaky scanning mechanism and may directly translate the main ORFs.

5.3 The 3′ UTRs of translationally regulated genes contain microRNA targets

Several miRNA prediction tools are available for animal data (Watanabe et al., 2007; Mendes et al., 2009). For plant miRNA prediction, however, only a few are available. Using psRNAtarget server (Dai and Zhao., 2011), several miRNA targets were identified in the 3′ UTRs of genes regulated at the translational level during stress. Genes specifically down-regulated during stress contained pre-dominantly higher numbers of microRNA target sites compared to the up-regulated genes.

It is evident from the mechanism of miRNAs action on the targets, that these miRNAs upon binding to the target site in the 3′ UTR, would recruit the argonaute proteins and they in-turn either cleave the mRNA using RNA-induced silencing complex mechanism or

translationally repress the activity of the mRNA by sequestering to it (Bartel, 2004), based on the complementarity around the target regions between the miRNA and the mRNA target.

As mentioned in the results (Chapter 4), gene AT3G25660 of YNdown group was predicted to bind to ath-miR5021 with complete complementarity in the central region of the binding site Studies have discovered (Kertesz et al., 2007, Yamasaki et al., 2013) perfect complementarity in the middle of the miRNA-target duplex would allow access for the argounate proteins and RISC complex to bind and cleave the mRNA. Whereas in the case of gene AT3G54220, a basic-leucine zipper domain containing protein, revealed a mismatching nucleotide base pair in the middle region of the miRNA binding site for the same microRNA (ath-miR5021), which could lead to a bulge formation. Thereby, the miRNA would only be able to repress the translational activity as the RISC complex would be blocked from interacting with the target. Thus, these two genes AT3G25660 and AT3G54220 with a miRNA binding site at the 3′ UTR has a very good chance of being translationally repressed by miRNAs in the event of stress induction in plants.

5.4 Highly conserved sequence motifs in the UTR

Motif search in the untranslated regions of translationally regulated genes yielded motifs with consensus sequences that were rich in [GA] and [CT] nucleotide repeats both in the 5′ and 3′ UTRs. Studies suggest that single / di-nucleotide repeats plays a role at the transcriptional as well as at the translational level (Santi et al. 2003; Meister et al. 2004; Kooiker et al. 2005; Yamamoto et al. 2009). Important factors such as the location of the repeat and the nucleotide content of the UTR has been observed to affect its functionality and not much has been studied on sequence repeats in plants (Lawson and Zhang, 2006). For example, a CAG repeat located in the 5′ UTR

of human calmodulin-1 (*hCALM1*) gene when disrupted has been observed to reduce its expression level significantly (Toutenhoofd et al., 1998). BASIC PENTACYSTEINE1 (BPC1), a regulatory protein is known to bind the *Arabidopsis thaliana gene SEEDSTICK (STK)* and regulate the ovule identity in Arabidopsis (Kooiker et al., 2005). In a study conducted on a *bpc1* mutant an increased *STK* expression was observed and it was revealed in the study that BPC1 induces conformational change to *STK* gene upon binding to the GA repeats in wild-type plants (Kooiker et al., 2005). It is possible that the predicted sequence motifs containing di-nucleotide repeats may be involved in translational regulation but experimental validation is needed for confirmation.

5.5 Motifs similar to Pumilio and Bruno-like protein motifs found in our analysis

Based on the *de novo* motif discovery analysis of untranslated regions (UTRs) of genes regulated at the translational level during (simulated) defense, using our approach of dividing the set of sequences into random subgroups, a number of significant regulatory motifs have been identified in the 3′ UTRs using the Seeder software (Fauteux *et al*., 2008). Among the several significant motifs discovered, we found few motifs that match a sequence / *cis*-regulatory element that was experimentally characterized and validated in two different studies (Kim *et al*., 2013 and Huh *et al*., 2013). In a study conducted in Arabidopsis, a Bruno-like protein was found to bind a motif [U(G/A)U(A/G)U(G/A)U] located in the 3′ UTR of *SOC1* (SUPPRESSOR *OF OVEREXPRESSION OF CONSTANS1*) mRNA, which encodes a MADS box transcription factor modulates the flowering time in plants (Kim *et al*., 2013). In another study, during viral infection on Arabidopsis, a *cis*-regulatory element [UGUA (A/C/U) AUA] in the 3′ UTR of viral RNA

interacts with the Pumilio protein, which suppresses the viral infection in plants (Huh *et al*., 2013). Interestingly, both the motifs identified in the literature are highly similar to each other but are recognized by two different proteins under different conditions. Out of the 54 genes in our dataset of translationally down-regulated (YNdown) genes, 18 genes had motifs in their 3′ UTR similar to the ones recognized by the Bruno-like protein (mentioned in Table 5.1). As mentioned previously, the Bruno-like protein is involved in translational down-regulation of genes responsible for the regulation of flowering time by binding to the *cis*-regulatory element in their 3′ UTR. It is possible the genes with motif binding site for the Bruno-like protein is down-regulated under stress and it might affect the flowering time in plants. With this evidence, and from our bioinformatics analysis findings, there is a high probability that the regulatory motifs located in the 3′ UTR in our set of genes might be responsible for gene down-regulation at the translational level. Likewise, in the same dataset, we discovered 8 genes (mentioned in Table 5.2) containing the motif recognized by the Pumilio protein in Arabidopsis, which could be involved in translational control of the genes. Overall, the down-regulation of genes during defense might be associated with the Pumilio and Bruno-like protein motif in their 3′ UTRs.

Chapter 6. Summary and Conclusion

The role of untranslated regions in post-transcriptional or translational control of plant genes is not thoroughly known. Experimentally, it is difficult and tedious to find the specific regulatory elements involved gene translational regulation. Computational methods can help form hypotheses, and make experimental validation more targeted and precise. In this study, we used various computational methods to predict statistically significant regulatory elements in UTRs of genes that are likely to be translationally regulated.

In the first objective of our study, for the prediction of *de novo* motifs in the untranslated regions (5′ and 3′ UTR), we developed a novel bioinformatics approach to elucidate the conserved sequence motifs in our genes of interest. The objective was carried out on genes that are differentially regulated at the translational level in plants under stress. As a result, several over-represented motifs were identified in the 5′ and 3′ UTRs. Interestingly, a higher number of conserved and distinct motifs were discovered using our approach. Some of the motifs predicted matched to the experimentally validated motifs, Pumilio and Bruno-like protein binding motifs, which could be potentially involved in translational regulation of genes. As well as some of the genes found to contain conserved motifs play a significant role in plant immune response.

For the second objective, the computational tool psRNAtarget was used to predict the miRNA target sites in the 3′ UTRs. The Arabidopsis miRNAs publicly available in the miRBase database were used to search for miRNA complementary sites in the 3′ UTRs of our genes. As a result, several miRNA target sites were predicted and most of them were found in the translationally down-regulated genes of our dataset. It is very highly likely that these genes with miRNA target sites could be down-regulated as the miRNA upon binding to the target either cleaves the mRNA transcript or represses it translational activity.

An interesting finding in this study is that one of the discovered miRNA target sites was also identified in our *de novo* motif discovery analysis as a statistically significant overrepresented motif. Studies suggest that the seed region is of high importance for the miRNA binding and correct function, and based on our finding from YNdown group, it seems to be conserved. Two genes that were predicted to be down-regulated at the translational level, and that contain the ath-miR5021 miRNA target site in their 3′ UTR and a conserved motif in the miRNA binding region are currently being tested for the miRNA activity in a transient transformation experiment (P. Moffett lab).

In addition to miRNA and regulatory motif identification, the 5′ UTRs were analyzed for uORFs, which have demonstrated to play a significant role in translation initiation. We have identified several uORFs, some with a strong kozak sequence context necessary for ribosome recognition, and they are located at an optimal distance that could facilitate its translation. Some of the genes with uORFs in their 5′ UTR also had miRNA target sites in the 3′ UTR. It is possible depending on the specific condition, either one of these two regulatory elements could be involved in gene translation.

In conclusion, our bioinformatics approach for motif identification has detected several significant motifs, few of them were found to match the experimentally verified motifs. The regulatory elements such as miRNAs, uORFs as well as the regulatory motifs predicted in this study needs to be further validated using experimental techniques in the lab to determine their exact role(s) in translation of genes under stress.

List of references

Asai, T., Tena, G., Plotnikova, J., Willmann, M. R., Chiu, W. L., Gomez-Gomez, L., & Sheen, J. (2002). MAP kinase signalling cascade in Arabidopsis innate immunity. *Nature*, *415*(6875), 977-983.

Aukerman, M. J., & Sakai, H. (2003). Regulation of flowering time and floral organ identity by a microRNA and its APETALA2-like target genes. The Plant Cell Online, 15(11), 2730-2741.

Axtell, M. J., & Staskawicz, B. J. (2003). Initiation of RPS2-specified disease resistance in Arabidopsis is coupled to the AvrRpt2-directed elimination of RIN4. *Cell*, *112*(3), 369-377.

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., ... & Noble, W. S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic acids research, gkp335.

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. cell, 116(2), 281-297.

Barbosa, C., Peixeiro, I., & Romão, L. (2013). Gene expression regulation by upstream open reading frames and human disease. PLoS genetics, 9(8), e1003529.

Beauclair, L., Yu, A., & Bouché, N. (2010). microRNA-directed cleavage and translational repression of the copper chaperone for superoxide dismutase mRNA in Arabidopsis. The Plant Journal, 62(3), 454-462.

Bieniawska, Z., Paul Barratt, D. H., Garlick, A. P., Thole, V., Kruger, N. J., Martin, C., ... & Smith, A. M. (2007). Analysis of the sucrose synthase gene family in Arabidopsis. *The Plant Journal*, *49*(5), 810-828.

Bisgrove, S. R., Simonich, M. T., Smith, N. M., Sattler, A., & Innes, R. W. (1994). A disease resistance gene in Arabidopsis with specificity for two different pathogen avirulence genes. *The Plant Cell Online*, *6*(7), 927-933.

Bläsing, O. E., Gibon, Y., Günther, M., Höhne, M., Morcuende, R., Osuna, D., ... & Stitt, M. (2005). Sugars and circadian regulation make major contributions to the global regulation of diurnal gene expression in Arabidopsis. *The Plant Cell*, *17*(12), 3257-3281.

Bollman, K. M., Aukerman, M. J., Park, M. Y., Hunter, C., Berardini, T. Z., & Poethig, R. S. (2003). HASTY, the Arabidopsis ortholog of exportin 5/MSN5, regulates phase change and morphogenesis. Development, 130(8), 1493-1504.

Brodersen, P., Sakvarelidze-Achard, L., Bruun-Rasmussen, M., Dunoyer, P., Yamamoto, Y. Y., Sieburth, L., & Voinnet, O. (2008). Widespread translational inhibition by plant miRNAs and siRNAs. *Science*, *320*(5880), 1185-1190.

Carrington, J. C., & Ambros, V. (2003). Role of microRNAs in plant and animal development. *Science*, *301*(5631), 336-338.

Chen, X. (2004). A microRNA as a translational repressor of APETALA2 in Arabidopsis flower development. Science, 303(5666), 2022-2025.

Chen, N. W., Sévignac, M., Thareau, V., Magdelenat, G., David, P., Ashfield, T., ... & Geffroy, V. (2010). Specific resistances against Pseudomonas syringae effectors AvrB and AvrRpm1 have evolved differently in common bean (Phaseolus vulgaris), soybean (Glycine max), and Arabidopsis thaliana. New Phytologist, 187(4), 941-956.

Child, S. J., Miller, M. K., & Geballe, A. P. (1999). Translational control by an upstream open reading frame in the HER-2/neu transcript. Journal of Biological Chemistry, 274(34), 24335-24341.

Chisholm, S. T., Coaker, G., Day, B., & Staskawicz, B. J. (2006). Host-microbe interactions: shaping the evolution of the plant immune response. *Cell*,*124*(4), 803-814.

Collier, S. M., & Moffett, P. (2009). NB-LRRs work a "bait and switch" on pathogens. *Trends in plant science*, *14*(10), 521-529.

Crooks, G. E., Hon, G., Chandonia, J. M., & Brenner, S. E. (2004). WebLogo: a sequence logo generator. Genome research, 14(6), 1188-1190.

Dai, X., & Zhao, P. X. (2011). psRNATarget: a plant small RNA target analysis server. Nucleic acids research, 39(suppl 2), W155-W159.

Dickey, L. F., Petracek, M. E., Nguyen, T. T., Hansen, E. R., & Thompson, W. F. (1998). Light regulation of Fed-1 mRNA requires an element in the 5′ untranslated region and correlates with differential polyribosome association.*The Plant Cell Online*, *10*(3), 475-484.

Fabian, M. R., Sonenberg, N., & Filipowicz, W. (2010). Regulation of mRNA translation and stability by microRNAs. *Annual review of biochemistry*, *79*, 351-379.

Dsouza, M., Larsen, N., & Overbeek, R. (1997). Searching for patterns in genomic data. Trends in Genetics, 13(12), 497-498.

Fauteux, F., Blanchette, M., & Strömvik, M. V. (2008). Seeder: discriminative seeding DNA motif discovery. Bioinformatics, 24(20), 2303-2307.

Felix, G., Duran, J. D., Volko, S., & Boller, T. (1999). Plants have a sensitive perception system for the most conserved domain of bacterial flagellin. *The Plant Journal*, *18*(3), 265-276.

Feng, Y., Liu, Q., & Xue, Q. (2006). Comparative study of rice and Arabidopsis actin-depolymerizing factors gene families. Journal of plant physiology, 163(1), 69-79.

Fujii, H., Chiou, T. J., Lin, S. I., Aung, K., & Zhu, J. K. (2005). A miRNA involved in phosphate-starvation response in Arabidopsis. Current Biology, 15(22), 2038-2043.

Fütterer, J., & Hohn, T. (1996). Translation in plants—rules and exceptions. In *Post-Transcriptional Control of Gene Expression in Plants* (pp. 159-189). Springer Netherlands.

Gallie, D. R. (1991). The cap and poly (A) tail function synergistically to regulate mRNA translational efficiency. *Genes & Development*, *5*(11), 2108-2116.

Gamm, M., Peviani, A., Honsel, A., Snel, B., Smeekens, S., & Hanson, J. (2014). Increased sucrose levels mediate selective mRNA translation in Arabidopsis. *BMC plant biology*, *14*(1), 306.

Gandikota, M., Birkenbihl, R. P., Höhmann, S., Cardon, G. H., Saedler, H., & Huijser, P. (2007). The miRNA156/157 recognition element in the 3′ UTR of the Arabidopsis SBP box gene SPL3 prevents early flowering by translational inhibition in seedlings. *The Plant Journal*, *49*(4), 683-693.

Garcia, M. E., Lynch, T., Peeters, J., Snowden, C., & Finkelstein, R. (2008). A small plant-specific protein family of ABI five binding proteins (AFPs) regulates stress response in germinating Arabidopsis seeds and seedlings. *Plant molecular biology*, *67*(6), 643-658.

Green, P. J. (1993). Control of mRNA stability in higher plants. *Plant physiology*, *102*(4), 1065.

Green, R. M., & Tobin, E. M. (1999). Loss of the circadian clock-associated protein 1 in Arabidopsis results in altered clock-regulated gene expression.*Proceedings of the National Academy of Sciences*, *96*(7), 4176-4179.

Grillo, G., Turi, A., Licciulli, F., Mignone, F., Liuni, S., Banfi, S., ... & Pesole, G. (2010). UTRdb and UTRsite (RELEASE 2010): a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. Nucleic acids research, 38(suppl 1), D75-D80.

Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L., & Noble, W. S. (2007). Quantifying similarity between motifs. Genome biology, 8(2), R24.

Hanfrey, C., Elliott, K. A., Franceschetti, M., Mayer, M. J., Illingworth, C., & Michael, A. J. (2005). A dual upstream open reading frame-based autoregulatory circuit controlling polyamine-responsive translation. Journal of Biological Chemistry, 280(47), 39229-39237.

Hill, J. R., & Morris, D. R. (1993). Cell-specific translational regulation of S-adenosylmethionine decarboxylase mRNA. Dependence on translation and coding capacity of the cis-acting upstream open reading frame. *Journal of Biological Chemistry*, *268*(1), 726-731.

Hinnebusch, A. G. (1997). Translational regulation of yeast GCN4 A window on factors that control initiator-tRNA binding to the ribosome. Journal of Biological Chemistry, 272(35), 21661-21664.

Huh, S. U., Kim, M. J., & Paek, K. H. (2013). Arabidopsis Pumilio protein APUM5 suppresses Cucumber mosaic virus infection via direct binding of viral RNAs. Proceedings of the National Academy of Sciences, 110(2), 779-784.

Jackson, R. J., & Standart, N. (1990). Do the poly (A) tail and 3′ untranslated region control mRNA translation?. Cell, 62(1), 15-24.

Jones, D. A., & Takemoto, D. (2004). Plant innate immunity–direct and indirect recognition of general and specific pathogen-associated molecules. *Current opinion in immunology*, *16*(1), 48-62.

Jones-Rhoades, M. W., Bartel, D. P., & Bartel, B. (2006). MicroRNAs and their regulatory roles in plants. *Annu*. *Rev*. *Plant Biol*., *57*, 19-53.

Joshi, C. P., Zhou, H., Huang, X., & Chiang, V. L. (1997). Context sequences of translation initiation codon in plants. Plant molecular biology, 35(6), 993-1001.

Kangisser, S., Yakir, E., & Green, R. M. (2013). Proteasomal regulation of CIRCADIAN CLOCK ASSOCIATED 1 (CCA1) stability is part of the complex control of CCA1. *Plant signaling & behavior*, *8*(3).

Kawaguchi, R., & Bailey-Serres, J. (2005). mRNA sequence features that contribute to translational regulation in Arabidopsis. Nucleic Acids Research,33(3), 955-965.

Kim, H. S., Abbasi, N., & Choi, S. B. (2013). Bruno-like proteins modulate flowering time via 3′ UTR-dependent decay of SOC1 mRNA. *New Phytologist*,*198*(3), 747-756.

Kim-Ha, J., Kerr, K., & Macdonald, P. M. (1995). Translational regulation of oskar mRNA by bruno, an ovarian RNA-binding protein, is essential. *Cell*, *81*(3), 403-412.

Kochetov, A. V., Sirnik, O. A., Rogosin, I. B., Glazko, G. V., Komarova, M. L., & Shumny, V. K. (2002). Contextual features of higher plant mRNA 5′-untranslated regions. Molecular Biology, 36(4), 510-516.

Kooiker, M., Airoldi, C. A., Losa, A., Manzotti, P. S., Finzi, L., Kater, M. M., & Colombo, L. (2005). BASIC PENTACYSTEINE1, a GA binding protein that induces conformational changes in the regulatory region of the homeotic Arabidopsis gene SEEDSTICK. The Plant Cell Online, 17(3), 722-729.

Kozak, M. A. R. I. L. Y. N. (1988). Leader length and secondary structure modulate mRNA function under conditions of stress. *Molecular and cellular biology*, *8*(7), 2737-2744.

Kozak, M., Evans, M., Gardner, P. D., Flores, I., Mariano, T. M., Pestka, S., ... & Schkgger, H. (1991). Structural features in eukaryotic mRNAs that mod.*Biol Chem*, *266*, 19867-19870.

Kulik, A., Anielska-Mazur, A., Bucholc, M., Koen, E., Szymańska, K., Żmieńko, A., ... & Dobrowolska, G. (2012). SNF1-related protein kinases type 2 are involved in plant responses to cadmium stress. Plant physiology, 160(2), 868-883.

Kunze, G., Zipfel, C., Robatzek, S., Niehaus, K., Boller, T., & Felix, G. (2004). The N terminus of bacterial elongation factor Tu elicits innate immunity in Arabidopsis plants. *The Plant Cell Online*, *16*(12), 3496-3507.

Lawson, M. J., & Zhang, L. (2006). Distinct patterns of SSR distribution in the Arabidopsis thaliana and rice genomes. Genome biology, 7(2), R14.

Lee, R. C., Feinbaum, R. L., & Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. cell, 75(5), 843-854.

Lee, Y., Kim, M., Han, J., Yeom, K. H., Lee, S., Baek, S. H., & Kim, V. N. (2004). MicroRNA genes are transcribed by RNA polymerase II. The EMBO journal, 23(20), 4051-4060.

Li, Y., Zhu, Y., Liu, Y., Shu, Y., Meng, F., Lu, Y., & Guo, D. (2008). Genome-wide identification of osmotic stress response gene in Arabidopsis thaliana. *Genomics*, *92*(6), 488-493.

Li, W. X., Oono, Y., Zhu, J., He, X. J., Wu, J. M., Iida, K., ... & Zhu, J. K. (2008). The Arabidopsis NFYA5 transcription factor is regulated transcriptionally and posttranscriptionally to promote drought resistance. The Plant Cell Online, 20(8), 2238-2251.

Liu, M. J., Wu, S. H., Chen, H. M., & Wu, S. H. (2012). Widespread translational control contributes to the regulation of Arabidopsis photomorphogenesis. *Molecular systems biology*, *8*(1).

Mackey, D., Holt, B. F., Wiig, A., & Dangl, J. L. (2002). RIN4 interacts with Pseudomonas syringae type III effector molecules and is required for RPM1-mediated resistance in Arabidopsis. *Cell*, *108*(6), 743-754.

Mallory, A. C., & Vaucheret, H. (2006). Functions of microRNAs and related small RNAs in

plants. *Nature genetics*, *38*, S31-S36.

Matsuura, H., Shinmyo, A., & Kato, K. (2008). Preferential translation mediated by Hsp81-3 5′-UTR during heat shock involves ribosome entry at the 5′-end rather than an internal site in Arabidopsis suspension cells. Journal of bioscience and bioengineering, 105(1), 39-47.

Matsuura, H., Takenami, S., Kubo, Y., Ueda, K., Ueda, A., Yamaguchi, M., ... & Kato, K. (2013). A computational and experimental approach reveals that the 5′-proximal region of the 5′-UTR has a Cis-regulatory signature responsible for heat stress-regulated mRNA translation in Arabidopsis. *Plant and Cell Physiology*, *54*(4), 474-483.

Mazzucotelli, E., Mastrangelo, A. M., Crosatti, C., Guerra, D., Stanca, A. M., & Cattivelli, L. (2008). Abiotic stress response in plants: When post-transcriptional and post-translational regulations control transcription. *Plant Science*, *174*(4), 420-431.

Meister, R. J., Williams, L. A., Monfared, M. M., Gallagher, T. L., Kraft, E. A., Nelson, C. G., & Gasser, C. S. (2004). Definition and interactions of a positive regulatory element of the Arabidopsis INNER NO OUTER promoter. The Plant Journal, 37(3), 426-438.

Mendes, N. D., Freitas, A. T., & Sagot, M. F. (2009). Current tools for the identification of miRNA genes and their targets. Nucleic acids research, 37(8), 2419-2433.

Meyers, B. C., Kozik, A., Griego, A., Kuang, H., & Michelmore, R. W. (2003). Genome-wide analysis of NBS-LRR–encoding genes in Arabidopsis. *The Plant Cell Online*, *15*(4), 809-834.

Mignone, F., Gissi, C., Liuni, S., & Pesole, G. (2002). Untranslated regions of mRNAs. *Genome Biol*, *3*(3), 1-10.

Moffett, P., Farnham, G., Peart, J., & Baulcombe, D. C. (2002). Interaction between domains of a plant NBS–LRR protein in disease resistance-related cell death. *The EMBO journal*, *21*(17), 4511-4519.

Mohr, P. G., & Cahill, D. M. (2007). Suppression by ABA of salicylic acid and lignin accumulation and the expression of multiple genes, in Arabidopsis infected with Pseudomonas syringae pv. tomato. *Functional & integrative genomics*, *7*(3), 181-191.

Morgante, M., Hanafey, M., & Powell, W. (2002). Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nature genetics, 30(2), 194-200.

Morris, D. R., & Geballe, A. P. (2000). Upstream open reading frames as regulators of mRNA translation. *Molecular and Cellular Biology*, *20*(23), 8635-8642.

Mückstein, U., Tafer, H., Hackermüller, J., Bernhart, S. H., Stadler, P. F., & Hofacker, I. L. (2006). Thermodynamics of RNA–RNA binding. Bioinformatics, 22(10), 1177-1182.

Munroe, D., & Jacobson, A. (1990). mRNA poly (A) tail, a 3'enhancer of translational initiation. *Molecular and Cellular Biology*, *10*(7), 3441-3455.

Navarro, L., Dunoyer, P., Jay, F., Arnold, B., Dharmasiri, N., Estelle, M., ... & Jones, J. D. (2006). A plant miRNA contributes to antibacterial resistance by repressing auxin signaling. Science, 312(5772), 436-439.

Nimchuk, Z., Eulgem, T., Holt Iii, B. F., & Dangl, J. L. (2003). Recognition and response in the plant immune system. *Annual review of genetics*, *37*(1), 579-609.

Nover, L. U. T. Z., Scharf, K. D., & Neumann, D. I. E. T. E. R. (1989). Cytoplasmic heat shock granules are formed from precursor particles and are associated with a specific set of mRNAs. *Molecular and Cellular Biology*, *9*(3), 1298-1308.

Nürnberger, T., Brunner, F., Kemmerling, B., & Piater, L. (2004). Innate immunity in plants and animals: striking similarities and obvious differences.*Immunological reviews*, *198*(1), 249-266.

Nyikó, T., Sonkoly, B., Mérai, Z., Benkovics, A. H., & Silhavy, D. (2009). Plant upstream ORFs can trigger nonsense-mediated mRNA decay in a size-dependent manner. Plant molecular biology, 71(4-5), 367-378.

Olsen, P. H., & Ambros, V. (1999). The lin-4 regulatory RNA controls developmental timing in Caenorhabditis elegans by blocking LIN-14 protein synthesis after the initiation of translation. Developmental biology, 216(2), 671-680.

Pavesi, G., Mauri, G., & Pesole, G. (2001). An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics*, *17*(suppl 1), S207-S214.

Pavesi, G., Mereghetti, P., Mauri, G., & Pesole, G. (2004). Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic acids research*, *32*(suppl 2), W199-W203.

Pesole, G., Mignone, F., Gissi, C., Grillo, G., Licciulli, F., & Liuni, S. (2001). Structural and functional features of eukaryotic mRNA untranslated regions. *Gene*, *276*(1), 73-81.

Priest, H. D., Filichkin, S. A., & Mockler, T. C. (2009). Cis-regulatory elements in plant cell signaling. Current opinion in plant biology, 12(5), 643-649.

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 Physiology, 150(3), 1356-1367.

Riechmann, J. L., Heard, J., Martin, G., Reuber, L., Jiang, C. Z., Keddie, J., ... & Yu, G. L. (2000). Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. Science, 290(5499), 2105-2110.

Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W., & Lenhard, B. (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic acids research, 32(suppl 1), D91-D94.

Santi, L., Wang, Y., Stile, M. R., Berendzen, K., Wanke, D., Roig, C., ... & Salamini, F. (2003). The GA octodinucleotide repeat binding factor BBR participates in the transcriptional regulation of the homeobox gene Bkn3. The Plant Journal, 34(6), 813-826.

Schauer, S. E., Jacobsen, S. E., Meinke, D. W., & Ray, A. (2002). DICER-LIKE1: blind men and elephants in Arabidopsis development. Trends in plant science, 7(11), 487-491.

Seo, P. J., Ryu, J., Kang, S. K., & Park, C. M. (2011). Modulation of sugar metabolism by an INDETERMINATE DOMAIN transcription factor contributes to photoperiodic flowering in Arabidopsis. *The Plant Journal*, *65*(3), 418-429.

Sprinthall, R. C. (2011). *Basic Statistical Analysis* (9th ed.). Pearson Education

Sunkar, R., Kapoor, A., & Zhu, J. K. (2006). Posttranscriptional induction of two Cu/Zn superoxide dismutase genes in Arabidopsis is mediated by downregulation of miR398 and important for oxidative stress tolerance. The Plant Cell Online, 18(8), 2051-2065.

Sunkar, R., & Zhu, J. K. (2004). Novel and stress-regulated microRNAs and other small RNAs from Arabidopsis. The Plant Cell Online, 16(8), 2001-2019.

Sweetlove, L. J., Heazlewood, J. L., Herald, V., Holtzapffel, R., Day, D. A., Leaver, C. J., & Millar, A. H. (2002). The impact of oxidative stress on Arabidopsis mitochondria. The Plant Journal, 32(6), 891-904.

Tanguay, R. L., & Gallie, D. R. (1996). The effect of the length of the 3′-untranslated region on expression in plants. *FEBS letters*, *394*(3), 285-288.

Tompa, M., Li, N., Bailey, T. L., Church, G. M., De Moor, B., Eskin, E., ... & Zhu, Z. (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nature biotechnology*, *23*(1), 137-144.

Toutenhoofd, S. L., Garcia, F., Zacharias, D. A., Wilson, R. A., & Strehler, E. E. (1998). Minimum CAG repeat in the human calmodulin-1 gene 5′ untranslated region is required for full expression. Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression, 1398(3), 315-320.

Tran, M. K., Schultz, C. J., & Baumann, U. (2008). Conserved upstream open reading frames in higher plants. BMC genomics, 9(1), 361.

van den Heuvel, J. J., Bergkamp, R. J., Planta, R. J., & Raué, H. A. (1989). Effect of deletions in the 5′-noncoding region on the translational efficiency of phosphoglycerate kinase mRNA in yeast. Gene, 79(1), 83-95.

Vilela, C., & McCarthy, J. E. (2003). Regulation of fungal gene expression via short open reading frames in the mRNA 5′ untranslated region. Molecular microbiology, 49(4), 859-867.

Vilela, C., Ramirez, C. V., Linz, B., Rodrigues-Pousada, C., & McCarthy, J. E. (1999). Post-termination ribosome interactions with the 5′ UTR modulate yeast mRNA stability. The EMBO journal, 18(11), 3139-3152.

Von Arnim, A. G., Jia, Q., & Vaughn, J. N. (2014). Regulation of plant translation by upstream open reading frames. *Plant Science*, *214*, 1-12.

Wang, Z. Y., & Tobin, E. M. (1998). Constitutive expression of the CIRCADIAN CLOCK ASSOCIATED 1 (CCA1) gene disrupts circadian rhythms and suppresses its own expression. *Cell*, *93*(7), 1207-1217.

Watanabe, Y., Tomita, M., & Kanai, A. (2007). Computational methods for microRNA target prediction. Methods in enzymology, 427, 65-86.

Wierzbicki, A. T., & Jerzmanowski, A. (2005). Suppression of histone H1 genes in Arabidopsis results in heritable developmental defects and stochastic changes in DNA methylation. *Genetics*, *169*(2), 997-1008.

Wightman, B., Ha, I., & Ruvkun, G. (1993). Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. Cell, 75(5), 855-862.

Wingender, E., Dietze, P., Karas, H., & Knüppel, R. (1996). TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic acids research, 24(1), 238-241.

Xie, F. L., Huang, S. Q., Guo, K., Xiang, A. L., Zhu, Y. Y., Nie, L., & Yang, Z. M. (2007). Computational identification of novel microRNAs and targets in Brassica napus. Febs Letters, 581(7), 1464-1474.

Yamasaki, H., Abdel-Ghany, S. E., Cohu, C. M., Kobayashi, Y., Shikanai, T., & Pilon, M. (2007). Regulation of copper homeostasis by micro-RNA in Arabidopsis. Journal of Biological Chemistry, 282(22), 16369-16378.

Yamasaki, T., Voshall, A., Kim, E. J., Moriyama, E., Cerutti, H., & Ohama, T. (2013). Complementarity to an miRNA seed region is sufficient to induce moderate repression of a target transcript in the unicellular green alga Chlamydomonas reinhardtii. The Plant Journal, 76(6), 1045-1056.

Yun, Y., Adesanya, T. A., & Mitra, R. D. (2012). A systematic study of gene expression variation at single-nucleotide resolution reveals widespread regulatory roles for uAUGs. Genome research, 22(6), 1089–1097.

Zhang, Y. (2005). miRU: an automated plant miRNA target prediction server. Nucleic acids research, 33(suppl 2), W701-W704.

Zhou, M., Wang, W., Karapetyan, S., Mwimba, M., Marqués, J., Buchler, N. E., & Dong, X. (2015). Redox rhythm reinforces the circadian clock to gate immune response. *Nature*, *523*(7561), 472-476.

Zipfel, C., & Felix, G. (2005). Plants and animals: a different taste for microbes?. *Current opinion in plant biology*, *8*(4), 353-360.

Zipfel, C., Robatzek, S., Navarro, L., Oakeley, E. J., Jones, J. D., Felix, G., & Boller, T. (2004). Bacterial disease resistance in Arabidopsis through flagellin perception. *Nature*, *428*(6984), 764-767.

Appendix I. List of translationally regulated genes belonging to different groups based on differential expression analysis.

Group	Gene IDs
NYup	AT2G05710, AT5G59880, AT3G29575, AT5G60440, AT3G48000, AT1G79440, AT3G62290, AT1G04800, AT1G06530, AT1G10940, AT1G20920, AT1G20970, AT1G23130, AT1G25275, AT1G31540, AT1G36730, AT1G54410, AT1G64370, AT1G74250, AT2G04520, AT2G07706, AT2G17240, AT2G22370, AT2G23120, AT2G26440, AT2G31410, AT2G39400, AT2G39570, AT3G04300, AT3G05220, AT3G08530, AT3G09160, AT3G17800, AT3G28940, AT3G57930, AT3G60540, AT4G08850, AT4G27290, AT4G27310, AT5G10625, AT5G10840, AT5G15940, AT5G17190, AT5G24230, AT5G38070, AT5G42220, AT5G61390, AT5G63370, AT5G63390, AT5G67410, AT3G51800,AT3G19190, AT5G28540, AT3G56240, AT4G11920, AT5G58940, AT1G08450, AT4G26430, AT5G49740, AT1G17220, AT4G36810, AT1G05200, AT5G53460, AT5G35630, AT2G33210, AT3G02875, AT2G02130, AT4G37300, AT2G35630, AT4G37260, AT1G54330, AT5G47640, AT4G24020, AT3G52525, AT5G54160, AT5G43980, AT1G55670, AT1G67090, AT1G21450, AT4G29160, AT5G67250, AT1G53160, AT1G80070,AT5G49190, AT2G18260, AT3G16640, AT2G43520, AT5G41700, AT1G70320, AT5G64810
NYdown	AT1G36160, AT4G12080, AT5G19140, AT4G35450, AT5G46110, AT5G19220, AT1G01320, AT1G02305, AT1G03140, AT1G04680, AT1G07080, AT1G09310, AT1G11240, AT1G11360, AT1G15230, AT1G16000, AT1G18620, AT1G23390, AT1G25570, AT1G26761, AT1G26800, AT1G26900, AT1G34110, AT1G47970, AT1G48300, AT1G50950, AT1G56423, AT1G61740, AT1G64510, AT1G66100, AT1G66840, AT1G67250, AT1G68730, AT1G69980, AT1G75280, AT1G77020, AT1G77170, AT1G78230, AT1G80130, AT1G80180, AT1G80780, AT1G80940, AT2G03505, AT2G03750, AT2G04340, AT2G15020, AT2G16595, AT2G17350, AT2G23200, AT2G24090, AT2G26390, AT2G28600, AT2G35640, AT2G35680, AT2G37750, AT2G41870, AT2G42700, AT2G45750, AT2G46780, AT3G02400, AT3G03150, AT3G05900, AT3G07230, AT3G09980, AT3G10020, AT3G11930, AT3G13590, AT3G14067, AT3G14190, AT3G14420, AT3G14770, AT3G16350, AT3G16370, AT3G16650, AT3G19440, AT3G22440, AT3G23450, AT3G24490, AT3G26450, AT3G27100, AT3G27460, AT3G29240, AT3G30390, AT3G43740, AT3G47070, AT3G48210, AT3G52500, AT3G56510, AT3G57400, AT3G57490, AT3G57910, AT3G58840, AT3G60080, AT3G61260, AT4G00026, AT4G00895, AT4G09950, AT4G10300, AT4G11320, AT4G13840, AT4G15030, AT4G19420, AT4G21450, AT4G21920, AT4G23670, AT4G26630, AT4G27450, AT4G28080, AT4G28240, AT4G30993, AT4G36420, AT4G37470, AT4G40030, AT5G01200, AT5G01520, AT5G02160, AT5G02410, AT5G03660, AT5G06530, AT5G09450, AT5G10350, AT5G14510, AT5G15750, AT5G19590, AT5G20700, AT5G20820, AT5G21430, AT5G22580, AT5G23760, AT5G25010, AT5G25280, AT5G25630,

	AT5G26610, AT5G27490, AT5G28050, AT5G38420, AT5G38980, AT5G40500, AT5G42680, AT5G43180, AT5G43830, AT5G46870, AT5G47550, AT5G49700, AT5G50390, AT5G51570, AT5G52740, AT5G52990, AT5G53880, AT5G58260, AT5G65207, AT5G65220, AT5G65380, AT5G66490, AT4G16520, AT4G10120, AT1G15690, AT4G36540, AT3G01500, AT1G42990, AT3G01500, AT2G21890, AT4G37970, AT3G56240, AT4G34050, AT1G02800, AT4G25080, AT1G48260, AT4G15560, AT5G43860, AT5G50920, AT2G47400, AT2G46410, AT4G04720, AT5G40380, AT3G56940, AT1G08830, AT3G28180, AT4G17090, AT3G54900, AT5G23530, AT5G23060, AT3G26125, AT5G42800, AT1G08930, AT1G20900, AT3G23150, AT4G33630, AT4G30950, AT5G53170, AT1G17220, AT5G53170, AT1G42970, AT2G18380, AT4G36240, AT1G03970, AT5G48030, AT4G02520, AT2G47730, AT1G33240, AT2G22430, AT1G20696, AT5G23420, AT1G04100, AT4G02670, AT1G64790, AT4G38600, AT1G01790, AT3G13850, AT5G01530, AT1G61520, AT5G01530, AT5G44610, AT4G26760, AT2G33820, AT1G11760, AT4G00950, AT1G60870, AT1G33520, AT4G04840, AT2G33480, AT4G37260, AT1G21640, AT5G47640, AT2G19620, AT3G25882, AT5G55810, AT1G79280, AT4G20260, AT2G26020, AT1G04980, AT5G64070, AT4G11840, AT5G38150, AT2G42120, AT5G45070, AT5G61770, AT2G42600, AT1G31330, AT1G52230, AT5G66570, AT4G20360, AT1G13260, AT1G01360, AT2G15080, AT5G25610, AT4G38740, AT1G52240, AT1G60620, AT5G49010, AT2G46340, AT5G24150, AT2G32765, AT4G02280, AT3G55980, AT5G42980, AT2G47770, AT5G46740, AT2G06530, AT5G64530, AT5G64570, AT5G41000, AT1G58340, AT1G56590, AT1G10970
YNup	AT3G28970, AT1G09580, AT1G51620, AT1G52800, AT1G55152, AT1G61170, AT1G66530, AT1G67170, AT1G67800, AT1G72810, AT1G76240, AT2G15300, AT2G17970, AT2G24650, AT2G30105, AT2G36170, AT2G44370, AT2G44580, AT3G04970, AT3G09035, AT3G09070, AT3G11720, AT3G23540, AT3G44220, AT3G57000, AT4G04950, AT4G09490, AT4G16695, AT4G17430, AT4G18400, AT4G25320, AT5G01380, AT5G01790, AT5G08430, AT5G16720, AT5G17610, AT5G25210, AT5G45590, AT5G46160, AT5G46680, AT5G56940, AT5G67090, AT1G80420, AT2G26000, AT4G15130, AT1G25425, AT5G39210, AT2G45080, AT4G18750, AT2G36640, AT4G05110, AT2G37640, AT1G10240, AT1G79460, AT4G02290, AT3G51910, AT1G48050, AT4G26890, AT5G44160, AT5G15860, AT5G02010, AT4G34410, AT3G28150, AT2G26580, AT3G21175
YNdown	AT4G25570, AT1G62700, AT1G01630, AT1G07280, AT1G25440, AT1G54260, AT1G56570, AT1G60000, AT1G68400, AT1G68520, AT1G70180, AT1G74088, AT2G29670, AT2G32560, AT2G44230, AT3G01400, AT3G01660, AT3G23900, AT3G25660, AT3G46450, AT3G62050, AT4G04880, AT4G14500, AT4G17080, AT4G17616, AT4G30400, AT5G45670, AT5G53030, AT5G56850, AT2G25900, AT3G49670,AT4G00490, AT2G43360, AT5G18930, AT2G46830, AT1G29395, AT4G11460, AT3G26300, AT1G56280, AT1G72630, AT2G24762, AT4G38460, AT5G57350, AT1G17560, AT1G15820, AT1G22640, AT2G32100, AT1G04520,

	AT4G03280, AT1G32440, AT2G06520, AT4G02080, AT3G54220, AT1G75540, AT4G13460, AT2G34070, AT2G46030, AT1G78630
YYup	AT1G48090, AT2G07687, AT2G07698, AT2G07707, AT2G07727, AT2G07732, AT2G07738, AT4G21020, AT3G02260, AT1G75830, AT5G26000, AT5G25980
YYdown	AT4G29140, AT1G06080, AT1G74840, AT2G18300, AT2G31141, AT3G54500, AT5G07140, AT5G14740, AT5G10930, AT1G09340, AT2G21330, AT4G38970, AT1G14070, AT3G26650, AT4G33010, AT2G26080, AT5G04140, AT1G07900, AT2G34430, AT2G34420, AT3G47470, AT5G54270, AT4G10340, AT2G39800

Up, up-regulated genes; down, down-regulated genes
NY represents genes translationally regulated in control but normally regulated in treated plants
YN represents genes translationally regulated in treated plants but normal in control plants
YY represents genes translationally regulated in both control and treated plants

Appendix II. Translationally regulated genes containing uORFs in their 5′ UTR

Gene ID	5′ UTR Length	uORF Position	uORF Sequence
AT3G19440.1	83	uORF [2,70]	ATGACTATCGGTTCCATCTTTCCCAAGTTCCTCTGCT CCAAAACCTCTTAAACCCTAATTCGAATT TAG
AT4G10300.1	102	uORF [3,74]	ATGCTATACGATTATATTGGACAGAAAATGAAAAAT TTAAAATGGAGTTTTCTATTGCATTTTGTAAGC TGA
AT4G36540.1	116	uORF [2,82]	ATGGATTACTTTTCTTCCACACATATATAACTTAACT CTCTCTTTTTCTCTTTTGCTTTAACCCCCTCAAAGAAA AGA TAA
AT1G60620.1	119	uORF [6,74]	ATGAAATTGAAGAACGAAATTGAAGTGAAAAGTGA AATCGTGCAATTTTCATCTTCCTCTTCTCCT TAA
AT1G21640.1	129	uORF [13,105]	ATGAAGAAGAGGACGAATCCACGAGGTCCACCAAAT TCCTGCAAACTTCCAAAACACCATCTCAAATATCCAC ATTCTCCGCCATTACCC TGA
AT5G19140.1	129	uORF [17,88]	ATGTCCGATATATCGTTACCAGAGAAAGAGAGCTCA GAGATTTAGTTAGCTTAATTTTTCTTTTCGTTT TGA
AT3G11930.1	134	uORF [1,111]	ATGCATATCTTATTGCACGTGGCCGGTGCCCTTTGAG ACGTATCACCCTCTACATGCAATATAAAAAGCATTG CAAATGAACCAAAGTTTGATCACATCCATCACTTA TAA
AT3G03150.1	147	uORF [12,86]	ATGCATATATATAAGCACACATTAAACGGAGTTAAA CTCATCATACATTACTGACTAAAAGAAGCAAAATCT TGA
AT4G23670.1	148	uORF [11,100]	ATGATTTAAAGAAACGGTAGAAACCTTGTTTTTAACC GTCACGGTCCCTATATAAACCTATGAGCTAGCGACG ACAAGTATTCAGTA TAA
AT4G13840.1	160	uORF [41,127]	ATGTCCAAGCAAAGGTTCATATAATTTTATGTTATTA AATTTGCTTTTTTTTTGTTAAAGTGTTACATACATCGA TCGACTTGT TAG

AT1G56423.1	163	uORF [17,136]	ATGGCTTTTGGAGTCTTTGAATATTCCTTCCCTCTGA CTTATGGTTTCCAGAAACCTTCCTCACACTCAAATGT TTTGTTCTTCTTAATCACAATTTCAGAATTGGTCTTTT ACGGT TAA
AT2G42700.1	163	uORF [1,72]	ATGTTCGTCGTCGGTGAGATTTCCCGGCGAAGCGAA AGCAGTGTTTCTCCGGTGGTTCAGGCATCAGTC TGA
AT5G42980.1	165	uORF [4,90]	ATGTTTTATTAATTATTATTATTCCATTTGCATTA AAATATAATAAAAAGATGTAAGATTCTCTCCGCCTC CTTCTCTCTA TAA
AT1G01360.1	167	uORF [4,126]	ATGTCGGTGGGTTCATTAAAAGAAACCAAAAAACAT AAACAAGTAATTTTGTTTTGGCATAACGAAGCATCTT CTTCTTCTTCCTTGTATTTATTATCCATTTCCAGAGAT TCTCCCTTG TGA
AT1G80130.1	168	uORF [41,106]	ATGCGGCTTACCATTATATAAGACTCTGGTAGACTAC TCTCATTATATACATTATAAAGATAC TGA
AT1G42970.1	171	uORF [8,73]	ATGTATCTTACCTACATAATAGCCACATATTACTTTG CTTCACTTTGAACCATTGGTTGGTTT TGA
AT4G26760.1	182	uORF [4,159]	ATGGGTCCCACTTTTCCCATTTCCCAACCTCCATAAC CAAAGTTGGAGATTTCTCTGTTTCCTATGCTCTCTCTC TCTCTTACTCTCACAGTGACTATTCGTCGCTTCAGAT CTAAAAGAGAGGAAGATAAACCATTTGGATTCAATA TCGAT TAA
AT1G11360.1	189	uORF [15,98]	ATGGTTTTGGTTATATATTTTTACAGCCGTTAGATCTT CAAACCGATCAATCTAACGGTTCTTAACAATATCCA ATTCTCA TAA
AT1G10970.1	206	uORF [13,87]	ATGCGTCAGTCTCTCATCAGTGTTCAACTGCCACGGA GCGAACCGATTCCTAATTGCAACGTCCCGAGTCCA TAG
AT1G10970.1	206	uORF [89,172]	ATGTCGACACTCTTTCACTCTTTCTCCAAGTTGCCTCC TTTGAGTCCTTTCTCATATTTTATAGACTCACTTTCTG TTTCT TGA
AT2G22430.1	218	uORF [7,150]	ATGGAAAACGCTCATCAACCTCAACAATCTCTCTCCT CTCTCTCTCTGTATATAGAAGAATCTCCATTGTCTTT AATTTCTCTCCATTTCTCTTTCTCTCTTCCTTCAAAGT TTTCTCTTTTCTTGATGGGTTTAAGAGAG TAA

AT3G16370.1	218	uORF [42,116]	ATGGTTGGAGAAATATTGAATAAGAGAGAGAGAGA GAGAGAGAGAGATAGAGTTTCAGAGTTGACACAACA T TGA
AT1G52240.1	222	uORF [35,136]	ATGCTTATACGTACATTTCTTCTTGAAAAGACTTTGA ATCCAAGAACCCGTGAGGAAGTAATCCCAGTGAAGA GAAGACGACAAAGAAAAAGGAGCCTA TAA
AT1G33240.1	226	uORF [133,210]	ATGGTTTTATGAGATTTATATCAAAAAACATTGAGG AGCTAGAGAGAAAGAGAGAGTGTGTGTGTAGAAAA AGAT TGA
AT2G17350.1	226	uORF [32,118]	ATGTCTCTGTCCCTTTCTTCCTCAGTTCTCACGCAAGC TTCAACGAGAAAAGATCGGAATCTCGCCGTCGGTAA AGCGCCGCCG TAA
AT3G47070.1	226	uORF [23,193]	ATGTATCCAAGTGATTCAAAGAAAAATGTGTGAAGG GAAAAAAGAAAGAGGATAAGACACTTGTGGAGAGC TTGTGCCAAAAAAATCCAACGGCTCAGATTCAACCA ACCTCGTTATAATCCCTTCGTAACAAATGCCACACGA AGACACACATACACACACTACCGA TAA
AT5G49010.1	233	uORF [144,221]	ATGTTTTTATTTCCCGCTCGAGCAAATTGATCTGATC CGAGTCTTTCAAGCGGGGAGAGAAAGAGAGAGATTC GT TAG
AT2G16595.1	236	uORF [6,137]	ATGTTATAACACCTATGGGCAAGGTGTTTCTAAATTG GTAAGTCGGAAAACGACGTCGTTTACAAAATCCGAT CTGCACACACACTCACATTCATATGTCGTTCTTCGTC AACGTCGGTGAAGAAAGCA TAA
AT4G02280.1	239	uORF [3,107]	ATGTACTCTCCTCTAACATAAACACGTCACTTGTAGC GAAAACAGTATCAAGAAAAAGAGAAGATCAAACAC GTCTTCTTTTCTCTCTCTCTTTGTCGCC TAA
AT4G02280.1	239	uORF [145,222]	ATGATTTTTCCTTTTAGTAGCAATCGTTGGTGATTCG AAAAACCAAACTTTTCTCGGACTAGGATTCTAGGGTT T TAG
AT1G18620.1	243	uORF [24,227]	ATGTGGCTGTCTTTTTGTTTTGTTCTCTGAGCAGAAC CAGGAGGAAACATTTTGCAGAGGATTTACTGGAAGC ACAATAAAGAAGAGTTAAGACACTGTCACTACAGAG AGAAGAGTTCAAATTCTCTCCTTTTTGTTACCAGTAG CAAGTTCCGGCGCCGGAGATTCGCCGGAGAACCCTC CTGGTTTCAGGTGAACTTC TAG
AT1G80780.1	250	uORF [119,217]	ATGAAACCAGATCGGGTCTTCCTCTAGATCCGTTCTA GGTTTTCTATTTCCACCGTCTTATCATTATCATCATCA TCAATTCATCATCATCATCAA TAA

AT3G24490.1	252	uORF [7,93]	ATGTTCACCGGTGGTGGTAGCTTTTCTTCTACTCCGA CCCGGCTGAAAATGCAAATCCCTAACCCTAGACACG GCCATGGTTTA TAA
AT5G52990.1	260	uORF [18,83]	ATGATCTCCATTTCTGTTATTTTCCCTAGAAATTTTCG CAAACCCTAAAAAGAGAGAGAAACA TAA
AT5G52990.1	260	uORF [128,217]	ATGTATTTCTAAAAAGAGAGAGAAATCCTCTGATCT CCTCTCTGATTCGATCGAGATCGCATCATTTCCCTAA TGATTTCAAAAACC TAA
AT1G34110.1	277	uORF [19,96]	ATGTAAAGACCTCACTCTTAGCTTCACATGGTTATCT ACAGTTATCGCTCTTGATGTTCTTGTTCTGCAATGGT T TAG
AT3G28180.1	281	uORF [20,106]	ATGATCCTCGTCTCTCTTTCTTTCTCTCTCTCAATCAC AAGTTTCAGTTACACAGCTGAACCCAAAGTATCTCA CATTCTGATC TAA
AT1G17220.1	300	uORF [148,243]	ATGCTTTCGGATTTGATTCGTCGATAACCCCAATTTC TTTACAAAGTTGAAGTTTTTATGTCATTTCTGGTTCG AAATTCGTTGCAATTGGAA TGA
AT4G19420.1	300	uORF [197,283]	ATGTGCTGAAACTGAGTCATCGTCCGAGTTAGCAAA GTGAAACTTCGAAGTTTCTCTTACACTAATCTCAGGT TTGGGATTTTG TAA
AT2G46340.1	310	uORF [88,219]	ATGTAATTATCAATAGTATATGACAAATTTATAGCCT CACCTTTCTTCTTCTTCTTCTTCTTCTGTGTGTTTT TGAATTTGCTTTCTTCAAGAGGTTTCTTCTCCACACC GCTTCGGTGTATGATC TGA
AT4G15560.1	311	uORF [37,144]	ATGAAGTTGGCTTTCTTGTCGTTTTACTTCATCACCCC ATTTTTTTAAAGTCTCCATCTTTATACTTCTTCAACTC TCCACCACCACCATTGTCACCACCACATT TAA
AT3G60080.1	327	uORF [127,192]	ATGGTAGGCCCTTTATAAACCAACTTTAATCCATTTT AACTTTTATATATTAAAAAAGGGCCT TGA
AT3G01500.1	330	uORF [102,326]	ATGTCGACCGCTCCTCTCTCCGGCTTCTTTCTCACTTC ACTTTCTCCTTCTCAATCTTCTCTCCAGAAACTCTCTC TTCGTACTTCTTCCACCGTCGCTTGCCTCCCACCCGC CTCTTCTTCTTCCTCATCTTCCTCCTCCTCGTCTTCCC GTTCCGTTCCAACGCTTATCCGTAACGAGCCAGTTTT TGCCGCTCCTGCTCCTATCATTGCCCCTTATTGG TGA

AT3G52500.1	333	uORF [42,140]	ATGTATTGGCTGGCTCTGTATCATCATGTCTACGTAG ACAGAGATGTATCATGTCGACTAAACAAAAATTACA TTTCCACGTATAAGATTTACTTA TAG
AT5G40380.1	349	uORF [16,210]	ATGAATAGTAGTTTTGTAATTAATTAATCAAGAATTA AAAAAAGTTCTATCAAAATCTAAATATATAGTCTTTT GTTTGTAATATGTTCTGCTTTTTACAAATATGGACTTT TGGTATCAATCATATTCACTTGTTCAGTATTATTCGA ATTTCTTGAATCATATAAAAACGAATTATGTAATAAA ACGATA TAA
AT5G40380.1	349	uORF [232,300]	ATGTCAATTGTCATATAACTATATAGTATATATATAT AAACCTTTCTGCAAATACAATCCTATATT TAA
AT5G01520.1	365	uORF [196,267]	ATGGAAATGGGAGAGAGAATCAATTCGAGTTGTTGT TAGGGGGTTTGCGATATTTTAGGCTTCTGTCGG TGA
AT5G46110.1	371	uORF [49,123]	ATGATGATGGTGATATGGAACTTCGATTGGCTAATAT TCACTGTGTCTCTAAAAACCATCCACTTATCAAGA TAA
AT5G46110.1	371	uORF [125,202]	ATGGACCCTACACTCATCCAATCTAAACCAGTATCTC AAGATTCTTATCTAATTACATCATTCTCTACCGTTAG A TGA
AT3G23150.1	396	uORF [39,107]	ATGGCGGTTTTCCGGCACTAATCATCTCCGGCATATA TAAATAAACGTACTTCACGTTTTTTTATA TAA
AT3G23150.1	396	uORF [239,316]	ATGAGAGGAAGATCGGAATGTCGAAGAGAATTAGA AGATTCTCGTACATCACTTCGTTGGAATTTCACAGGT CGA TGA
AT2G41870.1	417	uORF [48,206]	ATGTGAAAAATACAATCGATCGCATTATCTTTATCCC TAGCTAATCATTCATGTACAAGCATGTCTCCGAAGGT TAAAAGCAGTCGCTATTTACCGGACCAACGTAGTTTT CTCGAAGTGGTGGTCCGTTGTCATATTTTAAATTTAT CACCTTCT TGA
AT2G41870.1	417	uORF [234,344]	ATGTAGTGTATATTTTTTCCTCTAACCTAATTAAAAT CAAAACAAAATCCTTTGACCCAATTAGCTTCGCGAT ATATCAGAAGAGATCAAACTACTTTGATCAGACCA TGA
AT1G20900.1	468	uORF [29,133]	ATGAATCTCAAGCTTCTCTCTCCTTTTTTTCCCATAGC ACATCAGAATCGCTAAATACGACTCCTATGCAAAGA AGAAGCTACTTCTTTCTCTTGCCCTAAT TAA

AT1G20900.1	468	uORF [165,314]	ATGAGAGAGATCATTTAACATAAGTCACCTTTTTTAT ATCTTTTGCTTCGTCTTTAATTTAGTTCTGTTCTTGGT CTGTTTCTATATTTTGTCGGCTTGCGTAACCGATCAC ACCTTAATGCTTTAGCTATTGTTTCCTCAAAATCA TGA
AT2G35680.1	495	uORF [251,370]	ATGGTGGTGATCTCCCTGAAGATTTTGAATTCAAAAT TTCTTCGGTTGCTTCCTCTTGACTTTATCGCGATTTCG ATTACGTGTAGACTTTCTTCTTGGCTTTGTCTTCGAAT CAAT TAG
AT3G14770.1	502	uORF [1,156]	ATGATTTTCAAAACCTTTCGAGTTTTTTCAATCTTTTT AATCAATGGGATCAAATTTCGAATATCATTAAGACTT TTAAGAGAATTTTTAGACCATTATGCATATGTTTACC TTAATTTTTTCATGATCTATATTATTTTAAATGGAATT TAT TAA
AT3G14770.1	502	uORF [162,305]	ATGTATAACGTTACGACACTTTTTCGTAATTAATTTG TTTGTAAGATAGTTAAACAGTAGACAAAGCCAAATA TGGACCGCATATATATAGTCGTACACATTCAGCCTTT TGCTCCTTACCTTTTTAAGAAAGTCTTCTTT TAA
AT2G47730.1	745	uORF [6,89]	ATGAATTGGTTTGATAACTTTCGATTTTACTTGGAAA TTAACTATTACAAGAAAAATCTAATTTCAAATAGTG GAATGATT TGA
AT2G47730.1	745	uORF [111,326]	ATGATAACAATTACAAAATAAAAATGAGACGGCCGG CTGTAGAATAATATGACGACGTGAAGCAAAAGCAGG AGTGCGACAGAAACCTCTCTAACCAAATTGCTTTTAG TTATATTCTTCTTCGTCTTGACGCTACTCTTTACCTTT TTTTATCTGATTTCAGATTTCTCCAGAATCTTCAGATT CCCATGGCCACTGTTCATGTTGTCGGTC TGA
AT2G47730.1	745	uORF [375,500]	ATGCAATTCATGCTGCTTTATCACCGTCATCATCTCT ATCATTACTCAGAGGAGTGTTACCAAAGTATTTCAGC TCAGAGTTGTCAGTGGCTCAGTTCCGGTTGGTCGAAG GATCTTAATACA TAG
AT2G47730.1	745	uORF [513,611]	ATGAGAAGAACACAGATTCTTGGGATGGATGGGAGC TTCTACAAAGGCCAGTTTTGATCCAGTACACGGAGG AGAGATGTCTCAGCTCGAGTATCA TAA
AT4G36810.1	139	uORF [23,94]	ATGGCCAATCCAATTCTCCATTCATAATTTTGACTCC ACTTTCCAAAAAAAACAAAAAAAAATAAATAC TAA
AT5G10625.1	147	uORF [13,126]	ATGGCTATAAATACAGAGCCCGCACACAAACTCTCT CATAATTCATAAGTAAACACCAAAGCAATCGCTCGA GACCCCTTACAAATATCGATCTCTCTCTTTCTATATA CA TAA

AT5G59880.1	151	uORF [1,132]	ATGTTTACTTATTCGGACTAGAGAGCTTCCGCATAAA GCTGAGGAAAAAAAAGAGAGAAGACGCACACCGAA GAAAACACACAAGACTCCATTATTCTCCTGCTTCTTC GTCTCTTCAATTATTTTCTC TGA
AT4G37300.1	164	uORF [19,108]	ATGAAATTTGTAATGAGAAAATTAAAGGACTAAAAA GGCCCACTAAAAGCCCATATACATTTTCTTTTAAACA CAACTAAAAACCAG TGA
AT3G48000.1	177	uORF [12,125]	ATGTGTTAGAGATATCCGGAGATCTCAACCGTTGGA TTCTTTCTCCTTCAAATCAAATTATAAATCCCATCAA ACAGAGAGAGAGAGAGAAGAGGAGGAGAATTCGAA GAA TAA
AT2G43520.1	186	uORF [34,102]	ATGATAATTTCTAGACCCAACGATAAAGATAAAATA GTAGCATCGCACCACAACAAGTGTTGAGCA TGA
AT1G05200.1	227	uORF [16,132]	ATGTAAAGCACCAGTTTCTAGAATCAAAGGGCCTCA TATAAATACATTATTTTATACACTATAAATAAATATA CAGACACAAAAATCTGTATACACCAATCTCTCTACA GAATC TAA
AT4G26430.1	237	uORF [54,155]	ATGCTATAATCTTTTCCATTATTTTTCTAACAAAAGA AAGTTATATATGGGCTAAACTTAGAAGGAAAGCCCA AGAGGCCAAGACTATATAAAGATTCA TGA
AT3G17800.1	278	uORF [35,100]	ATGGTCTCAAAATATTTAAACGAAAAATATAATTTTC TGAGTCCTTTCATTTATTATTATAAA TAA
AT1G17220.1	300	uORF [148,243]	ATGCTTTCGGATTTGATTCGTCGATAACCCCAATTTC TTTACAAAGTTGAAGTTTTTATGTCATTTCTGGTTCG AAATTCGTTGCAATTGGAA TGA
AT5G35630.1	324	uORF [15,194]	ATGAGTATTGAAGTTGAGATAGAGGAGGTACAAGGA GACCTTATCTGCAGAAGACAAAAAGCCATTTTTAGC AAAACTAAAGAAAGAAAAAAGATTGAAACACAAAT ATGCGCCACTCGTAGTCCACCCCTATCTCTTTGGCAA AAGCCACTTCACTCTTTTTCCCTTTTTATATAT TAA
AT3G29575.1	347	uORF [86,169]	ATGTTGAAATTTTGTGGGATTTTTTTTTATTTCTTTAT TAAACTTTTTTTTATTGAATTTATAAAAAGGGAAGGT CGTCAT TAA
AT3G29575.1	347	uORF [178,267]	ATGGAATCTTCCAAAATTTGATATTTTGCTGTTTTCTT GGGATTTGAATTGCTCTTTATCATCAAGAATCTGTTA AAATTTCTAATC TAA

AT1G36730.1	470	uORF [264,437]	ATGTCTGAACAGGCTTCTCTCTTTCAATCATCTATCC GGCTTTTAGGAGAGTACTGTCCTATATCTAGGAAAA AGTCCACTCATAAAGTTGGTGGTTTGCGTTTTGAGGA GGACATCACGAGCTGTATTGGATTTCAGTTTATGCAA CTGAGTGATTATCACTGCCTCTTC TGA
AT1G29395.1	82	uORF [6,74]	ATGCAGTGATATAGTCGTTTAACAGAAAGATCACTC CGTCAATCTTTCCACCACTGTGACTAAATC TAA
AT4G00490.1	102	uORF [3,74]	ATGATTCTTGGATCCCATGAATAAGCCTTCACGTGGC TTTTGATTTCACATTCCACGAAAATCCAATTT TAA
AT4G03280.1	168	uORF [8,79]	ATGTGGAAGCAGAGGCTGGCGCGTGCCTTATCTGCT TCTCTCTCTCCAGTGGTCTCAAATCTCTCTGTG TGA
AT1G25440.1	176	uORF [17,109]	ATGCGCTTCATGCGGGTCATCCTCTTAATCTCAAACT CTCTAGGACTACACTAAATCTAACTTTTTGCAGAGAG CAAAAGATTCAATAAT TGA
AT1G01630.1	193	uORF [5,94]	ATGAGTAGAATTTTTGTAAGGAAGAGAGTAGTTGGT GAGCGTCGCCATTGAAGGGGAAAGCAGCCCAACTTT GCATTGAATGACACT TAA
AT3G23900.1	193	uORF [82,162]	ATGGTAATGGAATCTTAGGCTCTGCTTCTTGTTGATT GTAAATTCTCGAATTAGGGATTTAGCTCTCATTTGAG AGAT TGA
AT4G14500.1	215	uORF [98,205]	ATGTCGCTTAAATTTTTAGCGATCTTCGAATCCCATT TTCCTTCGGGAATCTGACATTTTTTCTCGGGAAATTT TTTTATCCAAATCGGAGTAAAGTATCGAATC TAG
AT3G46450.1	241	uORF [43,138]	ATGAACTTCGGAGAGAAGACGAAATCTCGCTCGCTT TCGCCGCCGTGAAATTTGATTCTCCGATCACTACAAC GATTCCGATCTGTCCTCCGT TGA
AT5G53030.1	245	uORF [57,188]	ATGTCTTCTCTAACCTCTCCTCCCACTATACATCGTCC TTAACTTTTCCTTTTTCTCGTTCAAATGAATCCAATTC CTATAAAGCTTTCTTTAATTTTTCTCTCCTATTTAACC CAAACCCATTTCATA TAA
AT2G46830.1	249	uORF [142,237]	ATGGAATCTTTATCGAATCCAAGCTGATTTTGTTTCT TTCATTGAATCATCTCTCTAAAGTGGAATTTTGTAAA GAGAAGATCTGAAGTTGTG TAG

AT1G04520.1	274	uORF [12,98]	ATGGTCATCAAAAGCCAAAGCCAAAGCCAAAAAAA ACAAAGAACCCAGAAACCTAGAAGACAGAAACAAA AATCAAAGCTTTAC TGA
AT1G54260.1	276	uORF [23,97]	ATGGACGGTGGAAGAAGGAAGCCCTTCTCGCCGGAC TAGCCAAGCACGGTTCCGGAAAATGGACAAGTATTT TAG
AT1G54260.1	276	uORF [156,245]	ATGGCGTAATATAAAGAAATCAAGGACACAATCACT CGAATCTACCTCCATTTTCACTTCACCTTCAGAGACT TCTCGTTCTGTTGT TAG
AT3G54220.1	392	uORF [17,247]	ATGAATAGAGATAGAAAGAGTCATTAAATGTACGAA GCGACATTCACAATAATTCGAAAGGTGGAAGACGAC TTAGATACGGCCAGGCTTCACTGTCCTCCTCGTCCTC CTCAATTACCCCTAACCCCTTTTTCCGGGATTCATCT CCAACCCACATCCTTCCAAATTCTCACCCCCTCACTG AGTTTTTGCTTTTTCTCCTCATCGGAGATCGTGAAGA CGATCAAG TAA
AT2G25900.1	394	uORF [6,173]	ATGGATAAAGAAAAAATATTTTTATTTTGTTTTATTA TTCTCTTTTAAAGTTAAAGACTTTAGGAAAAAAACA AAAGAAAAAAAGTAAATGGAGAGAAAAAAGCTTTT AAAACTTTAAAAATTAATAAAGTTTTAAAAAAAAAA AAAAAAAAAAAAAGACGAACA TGA
AT1G74088.1	404	uORF [45,116]	ATGTCACTCTCTAGTCTCTCCTCCATCTCTCGATTAAT CCCAGCAAGATCCTACTTGAAGGAGACGCAA TAG
AT1G74088.1	404	uORF [132,197]	ATGACCGTACAAATTTCACTGAGAAGGCAACAAGCG CTTCCGCGTACTCCTGACAATGTTAGA TGA
AT1G74088.1	404	uORF [213,326]	ATGTTTCTATGACCAAAGAACACAGTAATCTTCGTTT AGCTTCTGCTCAATAACACAAAAATCAGAAATTGCC GCTTCTGCTCTCAAAAGCTGCAATCAAAAGATGACA TA TAA
AT3G11720.1	192	uORF [12,89]	ATGTATATTCCCGAGAAAATGAGGGAAAGTATCTTC TAATCGGGAAAAACATCTCCACGTTCCTTGCCCGTCT CG TAG
AT1G80420.1	224	uORF [21,146]	ATGATTTCGATCAAAGAGAAGCCACGGGGGGAGCGC TAGCACTTAAAGATTTGCACCGTTTTCTCTCTGGCTT TGTGTTGTGCTAGATGCGATGCAATAGCAATATTTTC CTGGAGTTTCAAG TAA

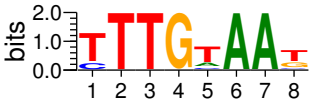
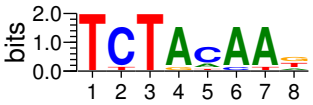
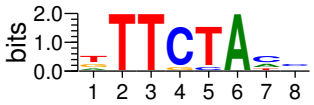
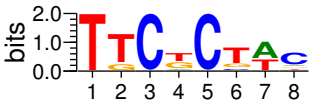
AT4G26890.1	233	uORF [35,124]	ATGTCGTATCCGACTTCGAACCAAAGCCCCATGCCA AAGTCACATACACACACTTCTACATACCCTAAACAC ACACACTCGAATACA TAA
AT3G09035.1	247	uORF [43,138]	ATGTCTCTGCTTTCTCCAGAAGATGATATTTACTTTTT GTGTCAATTAGCTGTATAACGTCTTCTTCACATAAAA CAAGAACACTCCACTTAC TAA
AT5G08430.1	274	uORF [171,254]	ATGTATTCGCCTTTCATTGCAGTAGTAGGGTTTATAG TTTCTTAAGCTAGAGGTGTGAGTTTTTTTCTTCTCCTT AGGTTT TGA
AT5G02010.1	340	uORF [61,285]	ATGTTGAAGATGCAGAGAGATCGAGGAAGACAGAG CAAGAGAACCTCTTCACTCACAGACCCACTCACTGC ATTTTCTCTTCTTTTCTCAGAAAATACTTTTTTCCCCC GAGAAAATGTTCCACAAGAATCAGATCGAACAATAA GGGTTCTCATAAATGTAATGCGATTCAAGCACGAGA ATATCCACAGAGAGAGATACAGAAGGCTCCTTTTGA GTATT TGA
AT1G51620.1	406	uORF [35,169]	ATGGTGTGTATTGGTGTTTCTAGGAAACAGATAAAC CAGTTTGCATTCTAGGCGATGACAGAAGCGACAGGC CTGAGGCGTTTCTTTTCTAAGTTCATTTGCTGTGTGTC CAATGCACAAGATCCATTTTTA TGA
AT1G51620.1	406	uORF [237,365]	ATGCAGACGATTCGACTATTGAATATGAACCATATG TGTACATTCCAGCGAATAGTTACTCGTATGCCGAGGT TACGAAAATTACAAACAAGTTTAATAGAGTTCATGG CAAAGGAGGGTTTGGTG TAG
AT4G25320.1	411	uORF [141,227]	ATGTATGTTTTAGGTCGAATTTTCTGAAATTAAGATT CATTCCTCCATGGAAGAAGCTCTGTTTTTATTCTCTTT AGCTTAGCT TAG
AT5G10930.1	208	uORF [53,148]	ATGGAATTGCTCTACTTCTTCACCTATGAGAGTAATA TATCCTCATGTACTCAAAATACTAATCAAACTTCATG CTTCAACACCACCTTTTCT TAA
AT5G14740.1	285	uORF [27,125]	ATGTTTTTGTTTAATCAACAAGAGGCGGAGATACGG GAGAAATTGCATGTGTAATCATAAAATGTAGATGTT AGCTTCGTCGTTTTTACTATAGTT TAG
AT2G07727.1	285	uORF [41,151]	ATGAATCTAAGAAATTTAGGTCTCTGCCCGCTTGAAA GATTCTTCTTTCCTTTTCGGTGAAAGAGGGCAAAAGT GTGTAGGAGAAAGAATTCTAAAAACGTCGACGCT TAA

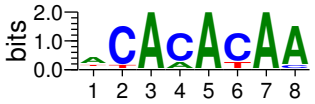
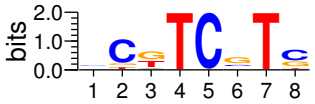
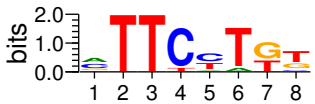
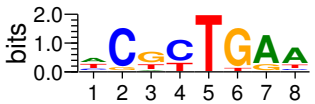
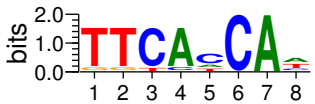
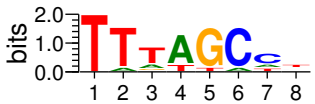
*Genes highlighted in bold represents that they have strong kozak signal

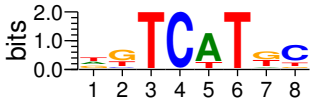
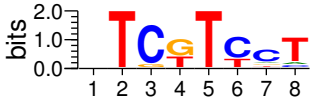
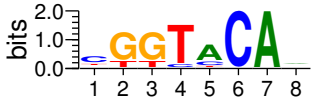
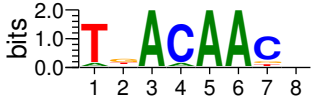
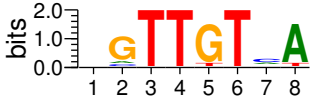
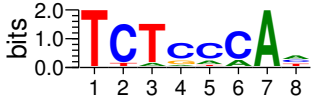
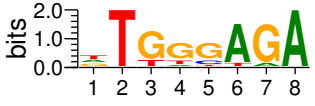
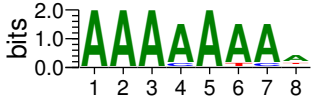
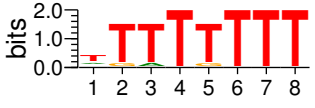
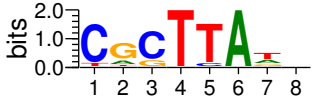
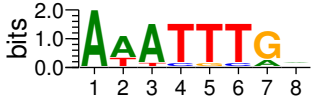
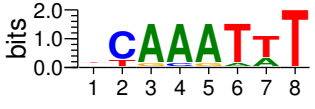
Appendix III. Custom Python script to extract sequence with length greater than10 nucleotides

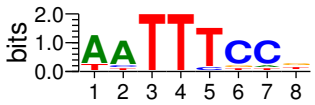
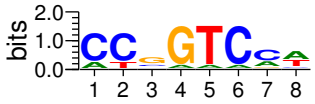
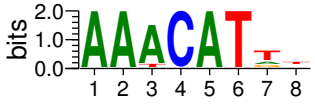
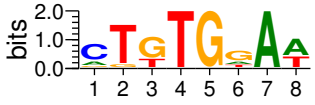
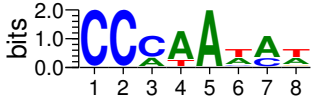
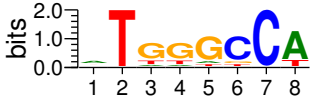
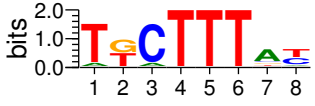
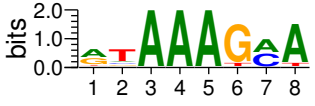
```
from Bio import SeqIO

from sys import argv

for sequence in SeqIO.parse(open(argv[1], "r"), "fasta"):

    if ".1" in sequence.id and "Sequence" not in sequence.seq
and "UTR" not in sequence.seq:

        print (">" + str(sequence.id))

        print sequence.seq


input_seq_iterator = SeqIO.parse(open("def.fas", "rU"), "fasta")

long_seq_iterator = (record for record in input_seq_iterator \

                     if len(record.seq) > 10)


output_handle = open("filename", "w")

SeqIO.write(long_seq_iterator, output_handle, "fasta")

output_handle.close()
```

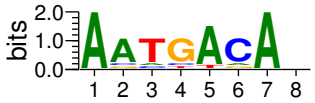
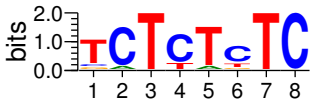
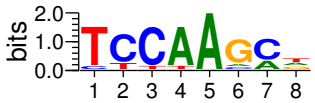
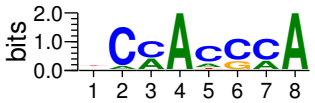
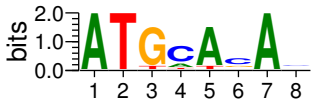
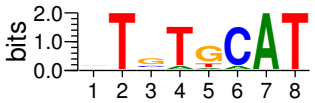
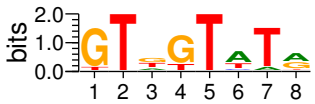
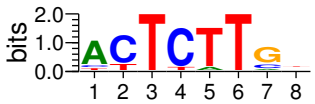
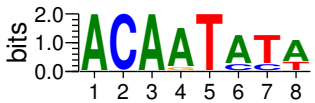
Supplemental Table 1: *De novo* motifs of length 8 found in the 3' UTRs of genes that belong to NYdown group

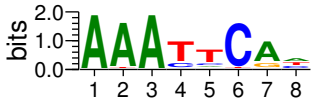
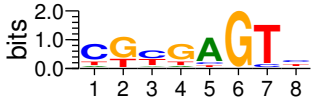
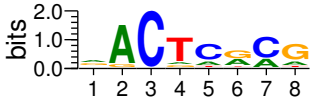
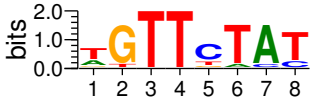
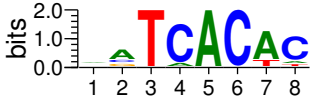
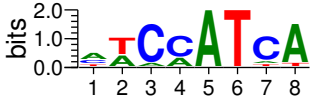
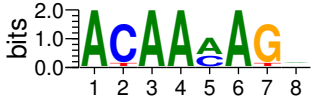
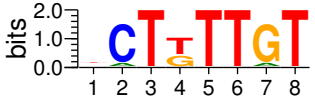
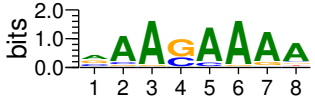
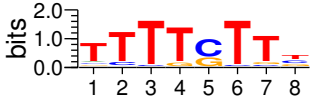
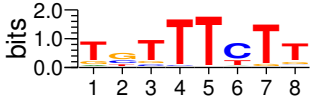
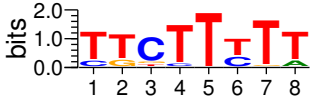
***De novo* motif**						
Motif ID[1]	**Software**[2]	**Forward**[3]	**Reverse**[4]	**Occurence in the group**[5]	**Occurence in the whole genome**[6]	**Possible function**[7]
S10	Seeder			51 / 241	657 / 20346	-
S11	Seeder			126 / 241	1625 / 20346	-
S12	Seeder			125 / 241	2509 / 20346	-
S13	Seeder			47 / 241	112 / 20346	Matches Bruno motif, regulation of translation
S14	Seeder			48 / 241	498 / 20346	-
S15	Seeder			59 / 241	234 / 20346	-
S16	Seeder			128 / 241	2368 / 20346	-
S17	Seeder			28 / 241	460 / 20346	-
S18	Seeder			35 / 241	405 / 20346	-

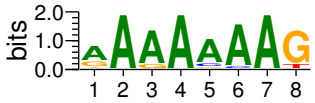
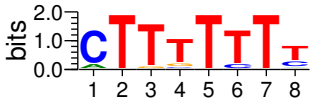
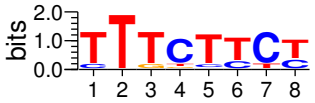
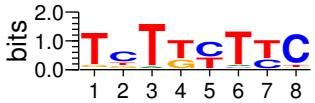
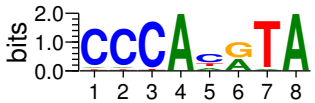
S19	Seeder			104 / 241	1476 / 20346	Simialr to Pumilio and Bruno motif, translation repression
S1	Seeder			79 / 241	798 / 20346	-
S20	Seeder			39 / 241	175 / 20346	-
S21	Seeder			40 / 241	278 / 20346	-
S22	Seeder			82 / 241	761 / 20346	-
S23	Seeder			42 / 241	231 / 20346	-
S24	Seeder			115 / 241	351 / 20346	-
S25	Seeder			46 / 241	160 / 20346	-
S26	Seeder			64 / 241	805 / 20346	-
S27	Seeder			49 / 241	424 / 20346	-

S28	Seeder			117 / 241	493 / 20346	-
S29	Seeder			52 / 241	209 / 20346	-
S2	Seeder			56 / 241	98 / 20346	-
S30	Seeder			47 / 241	255 / 20346	-
S31	Seeder			19 / 241	151 / 20346	-
S32	Seeder			76 / 241	390 / 20346	-
S33	Seeder			54 / 241	428 / 20346	-
S34	Seeder			463 / 241	12492 / 20346	-
S35	Seeder			45 / 241	409 / 20346	-
S36	Seeder			173 / 241	2757 / 20346	-

S37	Seeder			38 / 241	373 / 20346	-
S38	Seeder			18 / 241	80 / 20346	-
S39	Seeder			134 / 241	5148 / 20346	-
S3	Seeder			42 / 241	198 / 20346	Matches Bruno motif, regulation of translation
S40	Seeder			155 / 241	5142 / 20346	-
S41	Seeder			36 / 241	115 / 20346	
S42	Seeder			117 / 241	1184 / 20346	Matches Bruno motif, regulation of translation
S43	Seeder			138 / 241	2754 / 20346	-
S44	Seeder			34 / 241	168 / 20346	-
S45	Seeder			70 / 241	989 / 20346	-

S46	Seeder			47 / 241	556 / 20346	Matches Pumilio motif, translation repression
S47	Seeder			76 / 241	720 / 20346	-
S48	Seeder			56 / 241	1498 / 20346	-
S49	Seeder			33 / 241	303 / 20346	-
S4	Seeder			37 / 241	179 / 20346	-
S50	Seeder			48 / 241	369 / 20346	Matches Bruno motif, regulation of translation
S51	Seeder			30 / 241	150 / 20346	-
S52	Seeder			66 / 241	474 / 20346	-
S53	Seeder			45 / 241	497 / 20346	-
S54	Seeder			58 / 241	1481 / 20346	-

S55	Seeder			96 / 241	676 / 20346	-
S56	Seeder			331 / 241	2423 / 20346	-
S5	Seeder			5 / 241	8 / 20346	-
S6	Seeder			101 / 241	718 / 20346	-
S7	Seeder			51 / 241	1469 / 20346	-
S8	Seeder			65 / 241	923 / 20346	-
S9	Seeder			74 / 241	2645 / 20346	-
1	MEME			182 / 241	13219 / 20346	-
2	MEME			53 / 241	3931 / 20346	-
3	MEME			128 / 241	7311 / 20346	-

4	MEME			50 / 241	3862 / 20346	-
5	MEME			58 / 241	3237 / 20346	-
6	MEME			54 / 241	3061 / 20346	Similar to Bruno and Pumilio motif, translation repression
W1	Weeder			16 / 241	630 / 20346	-

[1]Name of the motif

[2]*de novo* discovery software that was used to locate the motif

[3]Forward sequence of the motif; Motif logo representing the occurence of a specific nucleotide at a respective position. The x-axis represents the position of a nucleotide and y-axis represents the amount of information in bits

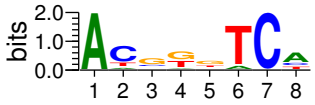
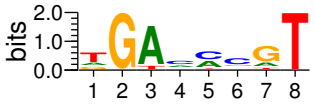
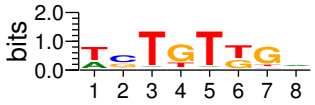
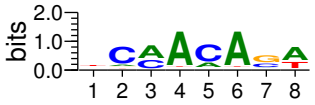
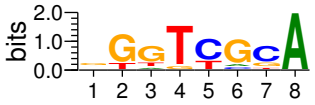
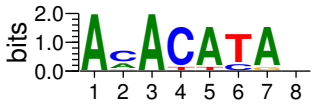
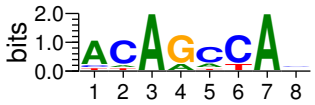
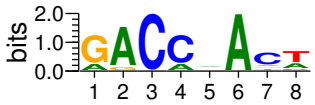
[4]Reverse sequence of the motif; Motif logo representing the occurence of a specific nucleotide at a respective position.The x-axis represents the position of a nucleotide and y-axis represents the amount of information in bits

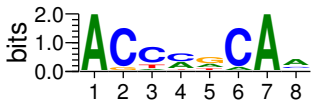
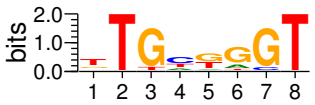
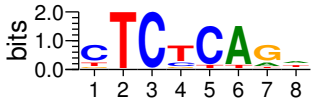
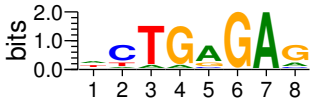
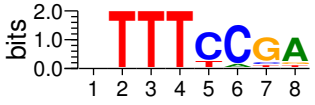
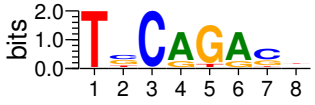
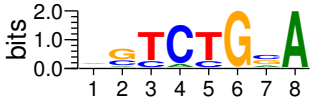
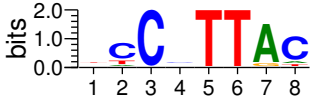
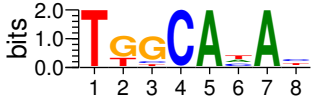
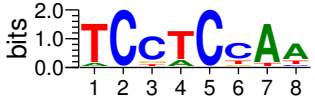
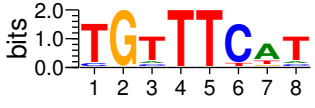
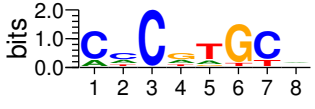
[5]Number of 3' UTRs containing the motif in the specific group. Note: Motif might occur more than once per sequence

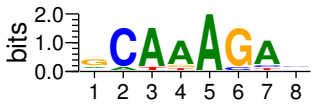
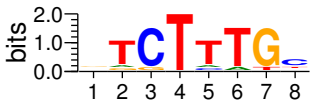
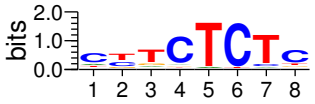
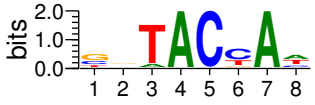
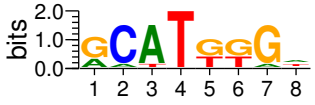
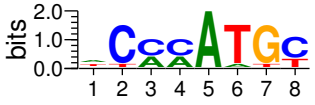
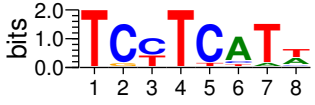
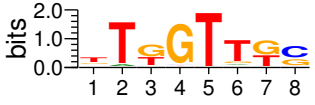
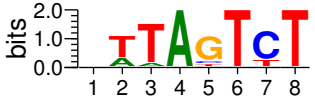
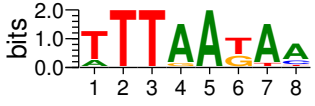
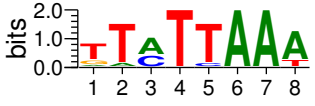
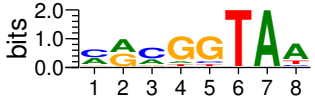
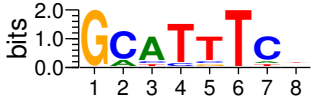
[6]Number of 3' UTRs containing the motif in the whole Arabidopsis genome. Note: Motif might occur more than once per sequence

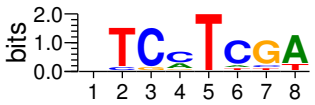
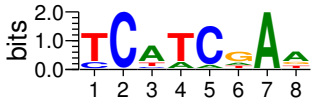
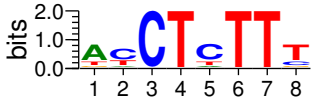
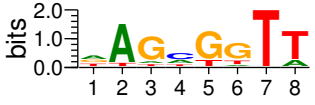
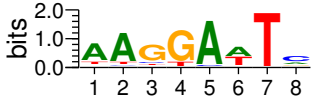
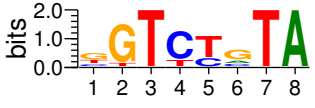
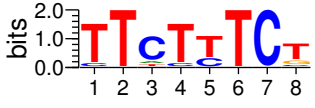
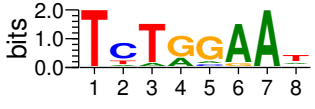
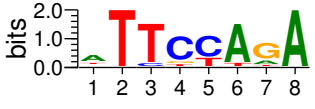
[7]Function of the motif obtained from literature

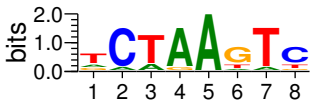
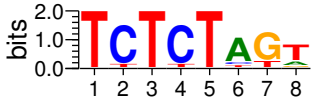
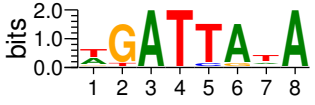
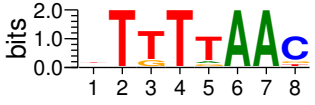
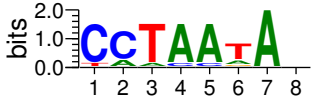
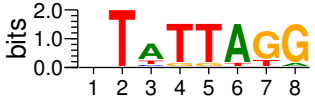
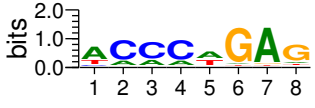
Supplemental Table 2: *De novo* motifs of length 8 found in the 3' UTRs of genes that belong to NYup group

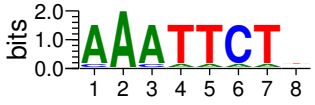
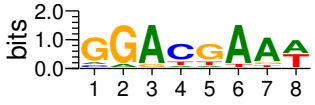
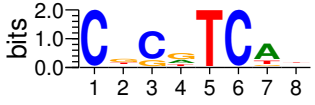
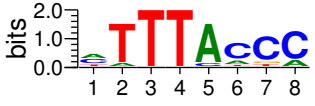
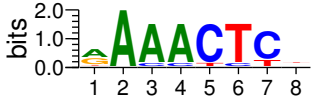
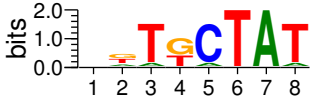
***De novo* motif**						
Motif ID[1]	**Software**[2]	**Forward**[3]	**Reverse**[4]	**Occurence in the group**[5]	**Occurence in the whole genome**[6]	**Possible function**[7]
S10	Seeder			11 / 85	1816 / 20346	-
S11	Seeder			77 / 85	10766 / 20346	-
S12	Seeder			7 / 85	701 / 20346	-
S13	Seeder			21 / 85	2962 / 20346	Matches Pumilio motif, translation repression
S14	Seeder			59 / 85	9676 / 20346	Matches Bruno motif, regulation of translation
S15	Seeder			14 / 85	2150 / 20346	Matches Bruno motif, regulation of translation
S16	Seeder			58 / 85	9875 / 20346	Matches Bruno motif, regulation of translation
S17	Seeder			32 / 85	4459 / 20346	-
S18	Seeder			18 / 85	2729 / 20346	-

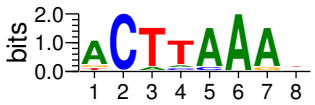
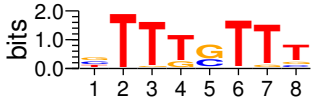
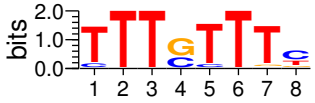
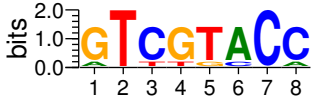
S19	Seeder			17 / 85	2900 / 20346	Matches Bruno motif, regulation of translation
S1	Seeder			19 / 85	4415 / 20346	-
S20	Seeder			16 / 85	2014 / 20346	-
S21	Seeder			18 / 85	3209 / 20346	-
S22	Seeder			17 / 85	2514 / 20346	-
S23	Seeder			18 / 85	2326 / 20346	-
S24	Seeder			14 / 85	2832 / 20346	-
S25	Seeder			12 / 85	2108 / 20346	-
S26	Seeder			62 / 85	6369 / 20346	-
S27	Seeder			15 / 85	1718 / 20346	-

S28	Seeder			45 / 85	5772 / 20346	-
S29	Seeder			70 / 85	12919 / 20346	-
S2	Seeder			32 / 85	5376 / 20346	-
S30	Seeder			17 / 85	2822 / 20346	-
S31	Seeder			27 / 85	4888 / 20346	-
S32	Seeder			113 / 85	18252 / 20346	-
S33	Seeder			16 / 85	4072 / 20346	-
S34	Seeder			40 / 85	6751 / 20346	-
S35	Seeder			10 / 85	1417 / 20346	-
S36	Seeder			21 / 85	3351 / 20346	-

S37	Seeder			15 / 85	1878 / 20346	-
S38	Seeder			34 / 85	4778 / 20346	-
S39	Seeder			25 / 85	2618 / 20346	-
S3	Seeder			28 / 85	5349 / 20346	-
S40	Seeder			11 / 85	1497 / 20346	-
S41	Seeder			17 / 85	2546 / 20346	-
S42	Seeder			38 / 85	8148 / 20346	-
S43	Seeder			17 / 85	1822 / 20346	Matches Pumilio motif, translation repression
S44	Seeder			49 / 85	8049 / 20346	-
S45	Seeder			17 / 85	3290 / 20346	-

S46	Seeder			10 / 85	2175 / 20346	-
S47	Seeder			28 / 85	4378 / 20346	-
S48	Seeder			34 / 85	7331 / 20346	-
S49	Seeder			16 / 85	2066 / 20346	-
S4	Seeder			33 / 85	6096 / 20346	-
S50	Seeder			37 / 85	7078 / 20346	-
S51	Seeder			22 / 85	4249 / 20346	-
S52	Seeder			67 / 85	9153 / 20346	Matches Bruno motif, regulation of translation
S53	Seeder			20 / 85	4175 / 20346	Matches Bruno motif, regulation of translation
S54	Seeder			105 / 85	17627 / 20346	-

S55	Seeder			19 / 85	4958 / 20346	-
S56	Seeder			56 / 85	8935 / 20346	-
S58	Seeder			45 / 85	10448 / 20346	-
S59	Seeder			21 / 85	2755 / 20346	-
S5	Seeder			26 / 85	3372 / 20346	-
S61	Seeder			17 / 85	2747 / 20346	-
S62	Seeder			16 / 85	2100 / 20346	-
S6	Seeder			52 / 85	9649 / 20346	-
S7	Seeder			30 / 85	2809 / 20346	-
S8	Seeder			17 / 85	3318 / 20346	-

S9	Seeder			62 / 85	11917 / 20346	-
M1	MEME			43 / 85	4252 / 20346	Matches Bruno motif, regulation of translation
M2	MEME			68 / 85	7193 / 20346	-
M3	MEME			26 / 85	3169 / 20346	-
W1	Weeder			8 / 85	438 / 20346	-

[1]Name of the motif

[2]*de novo* discovery software that was used to locate the motif

[3]Forward sequence of the motif; Motif logo representing the occurence of a specific nucleotide at a respective position. The x-axis represents the position of a nucleotide and y-axis represents the amount of information in bits

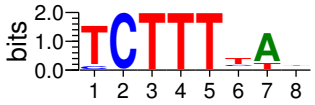
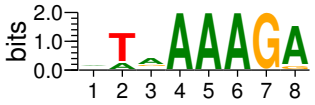
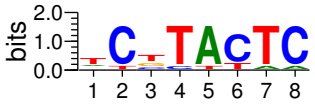
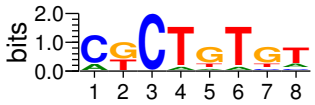
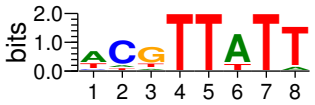
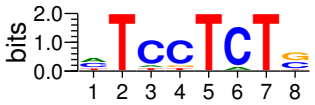
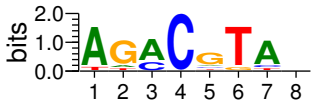
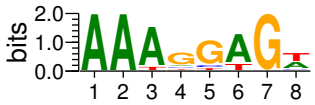
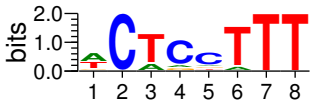
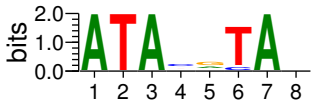
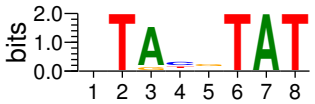
[4]Reverse sequence of the motif; Motif logo representing the occurence of a specific nucleotide at a respective position.The x-axis represents the position of a nucleotide and y-axis represents the amount of information in bits

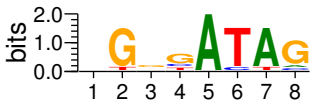
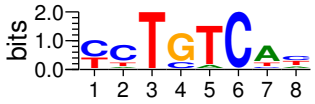
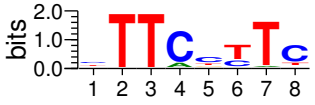
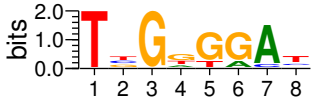
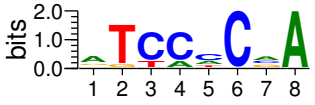
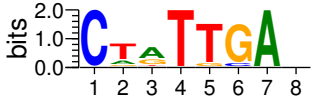
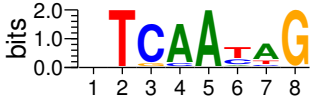
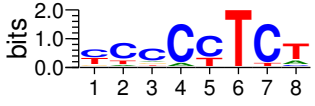
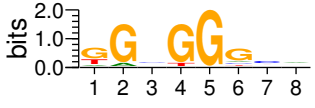
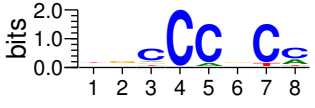
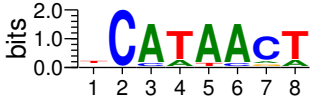
[5]Number of 3' UTRs containing the motif in the specific group. Note: Motif might occur more than once per sequence

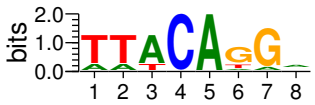
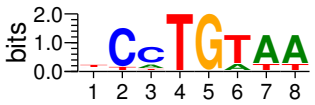
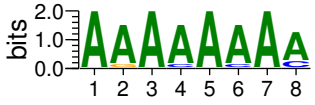
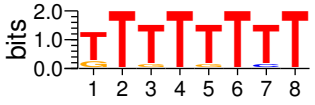
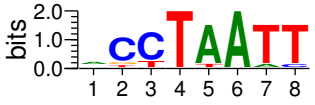
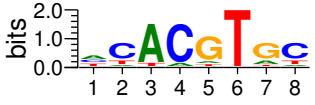
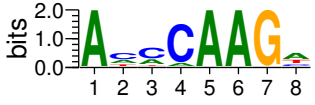
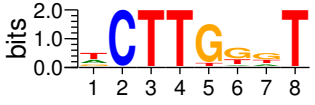
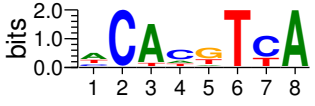
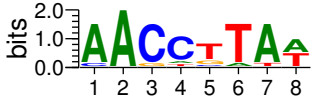
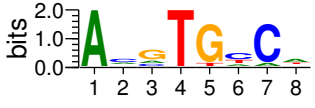
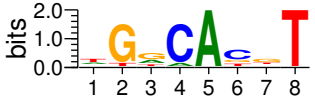
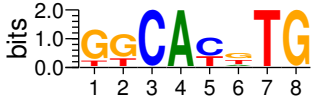
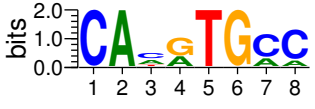
[6]Number of 3' UTRs containing the motif in the whole Arabidopsis genome. Note: Motif might occur more than once per sequence

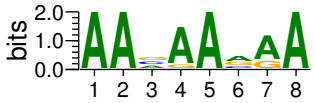
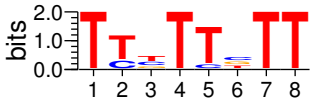
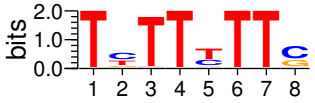
[7]Function of the motif obtained from literature

Supplemental Table 3: *De novo* motifs of length 8 found in the 3' UTRs of genes that belong to YNdown group

***De novo* motif**						
Motif ID[1]	**Software**[2]	**Forward**[3]	**Reverse**[4]	**Occurence in the group**[5]	**Occurence in the whole genome**[6]	**Possible function**[7]
S10	Seeder			67 / 54	14715 / 20346	-
S11	Seeder			17 / 54	3657 / 20346	-
S12	Seeder			8 / 54	2565 / 20346	-
S13	Seeder			4 / 54	1410 / 20346	Matches Bruno motif, regulation of translation
S14	Seeder			20 / 54	4787 / 20346	-
S15	Seeder			16 / 54	3958 / 20346	-
S16	Seeder			13 / 54	2118 / 20346	-
S17	Seeder			29 / 54	7276 / 20346	-
S18	Seeder			29 / 54	6863 / 20346	Matches Bruno motif, regulation of translation

S19	Seeder			7 / 54	1687 / 20346	-
S1	Seeder			12 / 54	2389 / 20346	-
S20	Seeder			46 / 54	13800 / 20346	-
S21	Seeder			21 / 54	3773 / 20346	-
S22	Seeder			13 / 54	3081 / 20346	-
S23	Seeder			28 / 54	5582 / 20346	-
S24	Seeder			31 / 54	7839 / 20346	Similar to Bruno and Pumilio motif, translation process
S25	Seeder			7 / 54	1458 / 20346	-
S27	Seeder			6 / 54	735 / 20346	-
S28	Seeder			22 / 54	4543 / 20346	-

S29	Seeder			19 / 54	3800 / 20346	-
S2	Seeder			14 / 54	1810 / 20346	-
S3	Seeder			110 / 54	28345 / 20346	-
S4	Seeder			19 / 54	3407 / 20346	-
S5	Seeder			7 / 54	1099 / 20346	-
S6	Seeder			21 / 54	5712 / 20346	-
S7	Seeder			13 / 54	2650 / 20346	-
S8	Seeder			19 / 54	4908 / 20346	-
S9	Seeder			15 / 54	1886 / 20346	-
W1	Weeder			10 / 54	848 / 20346	-

M1	MEME			77 / 54	21635 / 20346	-
M2	MEME			49 / 54	9820 / 20346	-

[1]Name of the motif

[2]*de novo* discovery software that was used to locate the motif

[3]Forward sequence of the motif; Motif logo representing the occurence of a specific nucleotide at a respective position. The x-axis represents the position of a nucleotide and y-axis represents the amount of information in bits

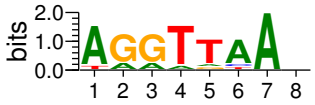
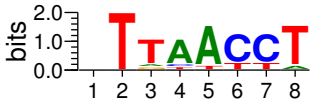
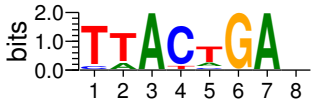
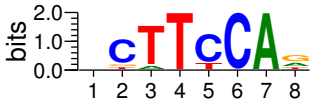
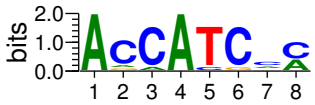
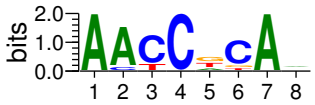
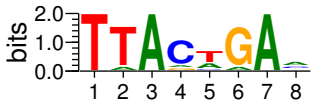
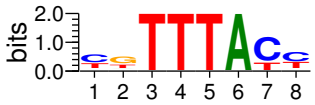
[4]Reverse sequence of the motif; Motif logo representing the occurence of a specific nucleotide at a respective position.The x-axis represents the position of a nucleotide and y-axis represents the amount of information in bits

[5]Number of 3' UTRs containing the motif in the specific group. Note: Motif might occur more than once per sequence

[6]Number of 3' UTRs containing the motif in the whole Arabidopsis genome. Note: Motif might occur more than once per sequence

[7]Function of the motif obtained from literature

Supplemental Table 4: *De novo* motifs of length 8 found in the 3' UTRs of genes that belong to YNup group

***De novo* motif**						
Motif ID[1]	**Software**[2]	**Forward**[3]	**Reverse**[4]	**Occurence in the group**[5]	**Occurence in the whole genome**[6]	**Possible function**[7]
S1	Seeder			24 / 59	4418 / 20346	-
S2	Seeder			21 / 59	4078 / 20346	-
S3	Seeder			11 / 59	2933 / 20346	-
S4	Seeder			8 / 59	1556 / 20346	-
S5	Seeder			25 / 59	5869 / 20346	-
S6	Seeder			13 / 59	3610 / 20346	-
S7	Seeder			18 / 59	3489 / 20346	-
W1	Weeder			18 / 59	2341 / 20346	-

[1]Name of the motif

[2]*de novo* discovery software that was used to locate the motif

[3]Forward sequence of the motif; Motif logo representing the occurence of a specific nucleotide at a respective position. The x-axis represents the position of a nucleotide and y-axis represents the amount of information in bits

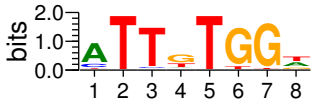
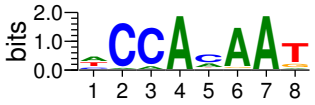
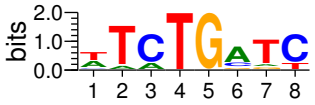
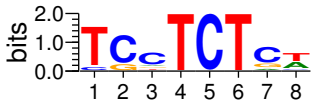
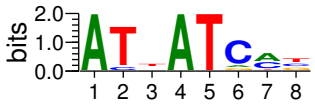
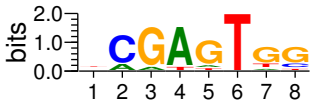
[4]Reverse sequence of the motif; Motif logo representing the occurence of a specific nucleotide at a respective position.The x-axis represents the position of a nucleotide and y-axis represents the amount of information in bits

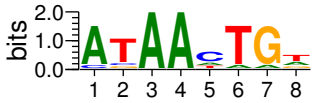
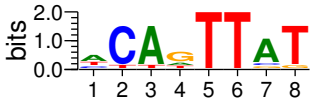
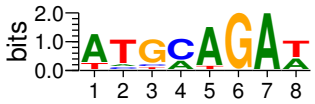
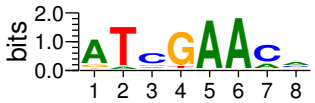
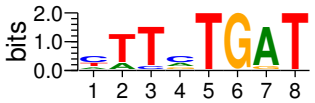
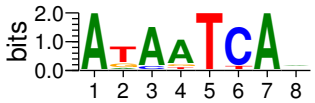
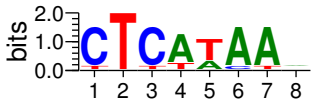
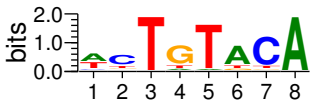
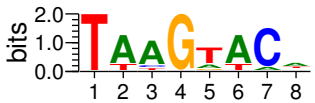
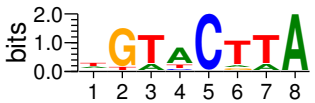
[5]Number of 3' UTRs containing the motif in the specific group. Note: Motif might occur more than once per sequence

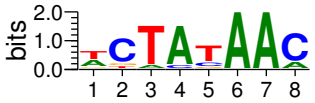
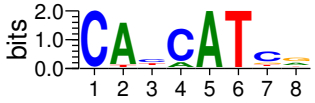
[6]Number of 3' UTRs containing the motif in the whole Arabidopsis genome. Note: Motif might occur more than once per sequence

[7]Function of the motif obtained from literature

Supplemental Table 5: *De novo* motifs of length 8 found in the 3' UTRs of genes that belong to YYdown group

***De novo* motif**						
Motif ID[1]	**Software**[2]	**Forward**[3]	**Reverse**[4]	**Occurence in the group**[5]	**Occurence in the whole genome**[6]	**Possible function**[7]
S10	Seeder			20 / 21	10853 / 20346	-
S11	Seeder			10 / 21	3927 / 20346	Matches Bruno motif, regulation of translation
S12	Seeder			15 / 21	4508 / 20346	-
S13	Seeder			14 / 21	5081 / 20346	-
S14	Seeder			9 / 21	3528 / 20346	-
S16	Seeder			7 / 21	1711 / 20346	-
S18	Seeder			20 / 21	9540 / 20346	-
S19	Seeder			6 / 21	1391 / 20346	-
S1	Seeder			20 / 21	14253 / 20346	-

S20	Seeder			20 / 21	13311 / 20346	-
S21	Seeder			14 / 21	4794 / 20346	-
S22	Seeder			12 / 21	4347 / 20346	-
S2	Seeder			11 / 21	6830 / 20346	-
S3	Seeder			15 / 21	5403 / 20346	-
S4	Seeder			19 / 21	11705 / 20346	-
S5	Seeder			18 / 21	6259 / 20346	-
S6	Seeder			20 / 21	9117 / 20346	Matches Pumilio motif, translation repression
S7	Seeder			12 / 21	3682 / 20346	-
S8	Seeder			11 / 21	4570 / 20346	Matches Pumilio motif, translation repression

S9	Seeder			12 / 21	3703 / 20346	-
W1	Weeder			13 / 21	3231 / 20346	-

[1]Name of the motif

[2]*de novo* discovery software that was used to locate the motif

[3]Forward sequence of the motif; Motif logo representing the occurence of a specific nucleotide at a respective position. The x-axis represents the position of a nucleotide and y-axis represents the amount of information in bits

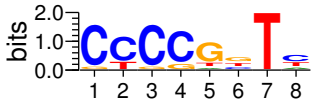
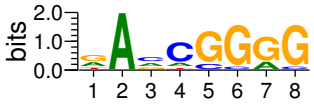
[4]Reverse sequence of the motif; Motif logo representing the occurence of a specific nucleotide at a respective position.The x-axis represents the position of a nucleotide and y-axis represents the amount of information in bits

[5]Number of 3' UTRs containing the motif in the specific group. Note: Motif might occur more than once per sequence

[6]Number of 3' UTRs containing the motif in the whole Arabidopsis genome. Note: Motif might occur more than once per sequence

[7]Function of the motif obtained from literature

Supplemental Table 6: *De novo* motifs of length 8 found in the 3' UTRs of genes that belong to YYup group

***De novo* motif**						
Motif ID[1]	**Software**[2]	**Forward**[3]	**Reverse**[4]	**Occurence in the group**[5]	**Occurence in the whole genome**[6]	**Possible function**[7]
W1	Weeder			7 / 10	566 / 20346	-

[1]Name of the motif

[2]*de novo* discovery software that was used to locate the motif

[3]Forward sequence of the motif; Motif logo representing the occurence of a specific nucleotide at a respective position. The x-axis represents the position of a nucleotide and y-axis represents the amount of information in bits

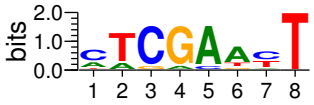
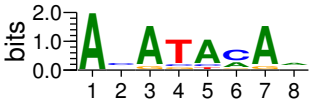
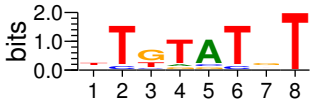
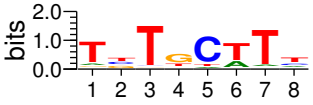
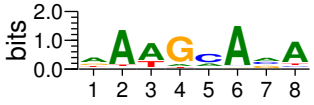
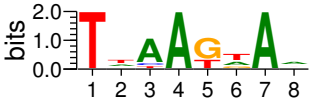
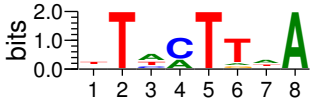
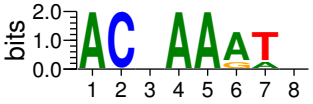
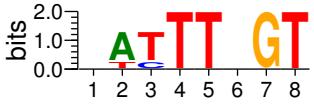
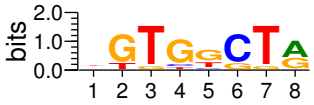
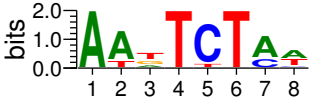
[4]Reverse sequence of the motif; Motif logo representing the occurence of a specific nucleotide at a respective position.The x-axis represents the position of a nucleotide and y-axis represents the amount of information in bits

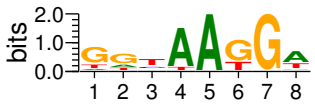
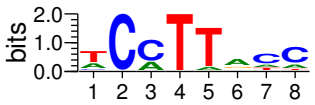
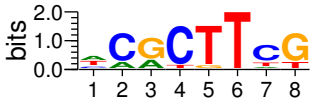
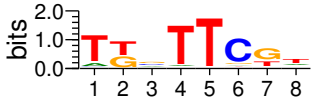
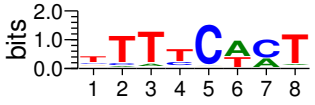
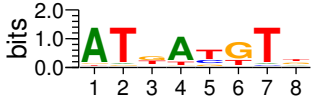
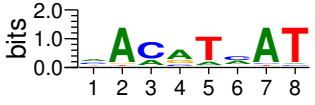
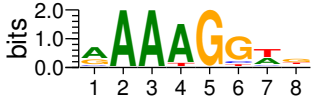
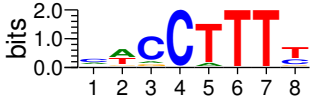
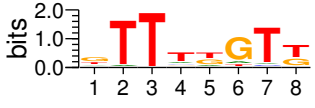
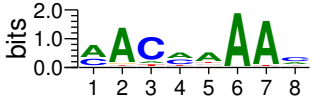
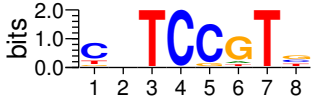
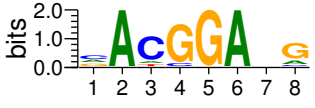
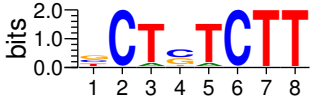
[5]Number of 3' UTRs containing the motif in the specific group. Note: Motif might occur more than once per sequence

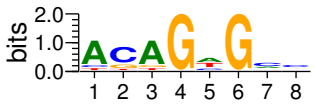
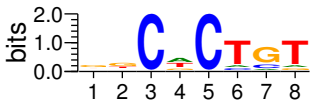
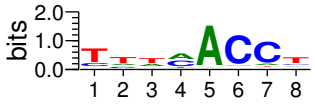
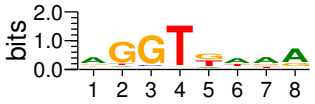
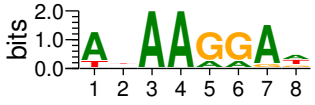
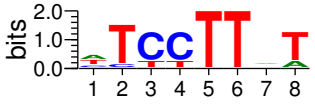
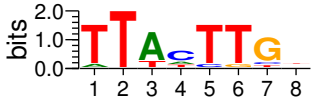
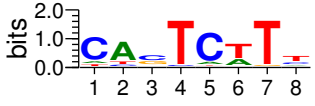
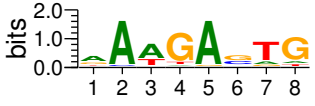
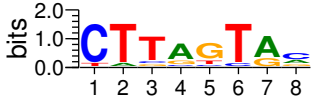
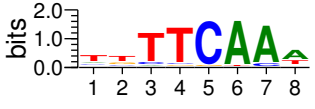
[6]Number of 3' UTRs containing the motif in the whole Arabidopsis genome. Note: Motif might occur more than once per sequence

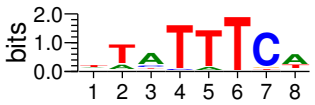
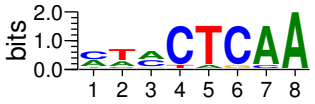
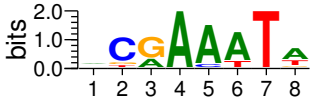
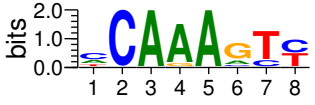
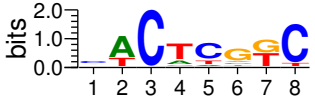
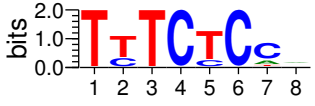
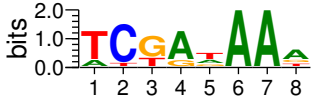
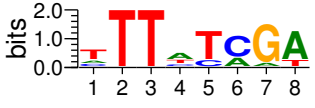
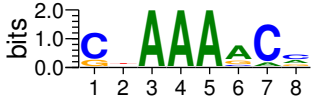
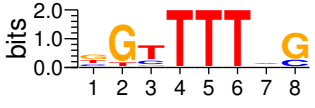
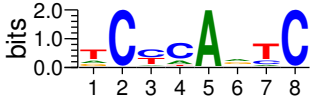
[7]Function of the motif obtained from literature

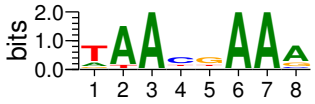
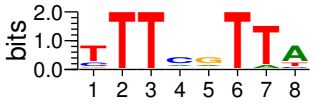
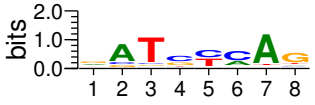
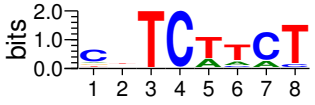
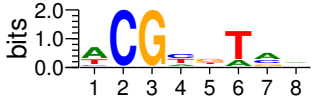
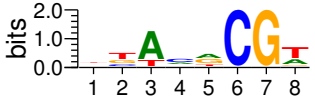
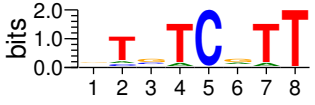
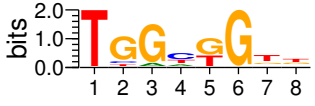
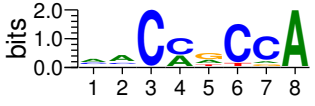
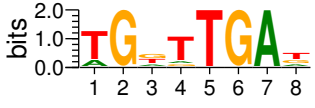
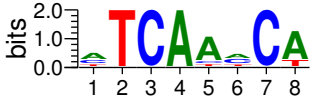
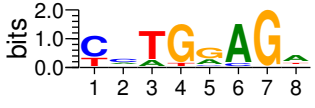
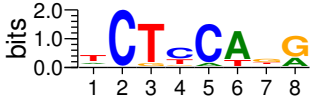
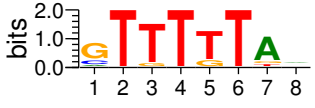
Supplemental Table 7: *De novo* motifs of length 8 found in the 5' UTRs of genes that belong to NYdown group

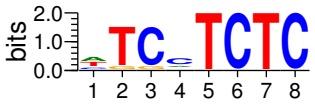
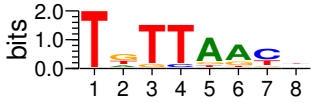
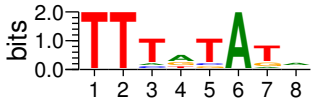
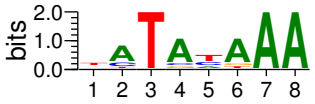
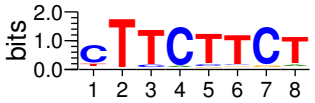
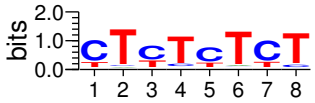
***De novo* motif**						
Motif ID[1]	**Software**[2]	**Forward**[3]	**Reverse**[4]	**Occurence in the group**[5]	**Occurence in the whole genome**[6]	**Possible function**[7]
S10	Seeder			32 / 237	2623 / 19128	-
S11	Seeder			99 / 237	5967 / 19128	-
S12	Seeder			67 / 237	5435 / 19128	-
S13	Seeder			33 / 237	2754 / 19128	-
S14	Seeder			280 / 237	21460 / 19128	-
S15	Seeder			153 / 237	12156 / 19128	-
S16	Seeder			133 / 237	10820 / 19128	-
S17	Seeder			46 / 237	2994 / 19128	-
S18	Seeder			66 / 237	6026 / 19128	-

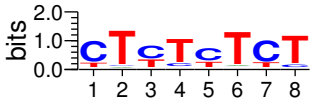
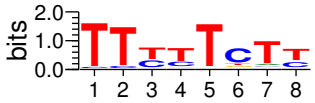
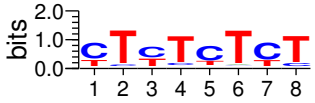
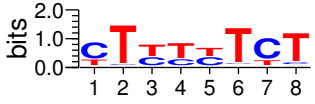
S19	Seeder			33 / 237	2335 / 19128	-
S1	Seeder			35 / 237	1662 / 19128	-
S20	Seeder			239 / 237	17199 / 19128	-
S21	Seeder			381 / 237	27130 / 19128	-
S22	Seeder			81 / 237	5173 / 19128	-
S23	Seeder			66 / 237	3612 / 19128	-
S24	Seeder			97 / 237	6026 / 19128	-
S25	Seeder			161 / 237	11943 / 19128	-
S26	Seeder			27 / 237	1953 / 19128	-
S27	Seeder			113 / 237	7517 / 19128	-

S28	Seeder			59 / 237	5027 / 19128	-
S29	Seeder			215 / 237	15106 / 19128	-
S2	Seeder			298 / 237	21549 / 19128	-
S30	Seeder			54 / 237	3560 / 19128	-
S31	Seeder			30 / 237	2251 / 19128	-
S32	Seeder			54 / 237	4298 / 19128	-
S34	Seeder			175 / 237	12507 / 19128	-
S35	Seeder			157 / 237	11320 / 19128	-
S36	Seeder			26 / 237	1617 / 19128	-
S37	Seeder			190 / 237	15941 / 19128	-

S38	Seeder			260 / 237	20037 / 19128	-
S39	Seeder			44 / 237	3528 / 19128	-
S3	Seeder			52 / 237	4285 / 19128	-
S40	Seeder			32 / 237	2132 / 19128	-
S41	Seeder			51 / 237	3615 / 19128	-
S42	Seeder			22 / 237	1947 / 19128	-
S43	Seeder			124 / 237	10461 / 19128	-
S44	Seeder			80 / 237	5476 / 19128	-
S45	Seeder			122 / 237	9103 / 19128	-
S46	Seeder			43 / 237	3088 / 19128	-

S47	Seeder			24 / 237	2116 / 19128	-
S48	Seeder			74 / 237	5313 / 19128	-
S49	Seeder			43 / 237	3914 / 19128	-
S4	Seeder			276 / 237	20878 / 19128	-
S50	Seeder			31 / 237	2874 / 19128	-
S51	Seeder			23 / 237	2799 / 19128	-
S52	Seeder			47 / 237	2214 / 19128	-
S53	Seeder			59 / 237	4529 / 19128	-
S54	Seeder			36 / 237	2912 / 19128	-
S55	Seeder			85 / 237	5784 / 19128	-

S56	Seeder			68 / 237	5558 / 19128	-
S57	Seeder			62 / 237	4736 / 19128	-
S5	Seeder			105 / 237	8238 / 19128	-
S6	Seeder			117 / 237	9691 / 19128	-
S7	Seeder			20 / 237	1982 / 19128	-
S8	Seeder			125 / 237	8116 / 19128	-
S9	Seeder			41 / 237	3654 / 19128	-
W1	Weeder			64 / 237	4258 / 19128	-
M1	MEME			339 / 237	24856 / 19128	-
M2	MEME			513 / 237	35216 / 19128	-

M3	MEME			563 / 237	39309 / 19128	-
M4	MEME			908 / 237	66925 / 19128	-
M5	MEME			537 / 237	37670 / 19128	-
M6	MEME			666 / 237	45998 / 19128	-

[1]Name of the motif

[2]*de novo* discovery software that was used to locate the motif

[3]Forward sequence of the motif; Motif logo representing the occurence of a specific nucleotide at a respective position. The x-axis represents the position of a nucleotide and y-axis represents the amount of information in bits

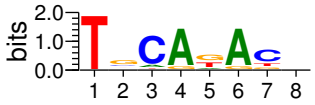
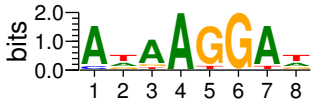
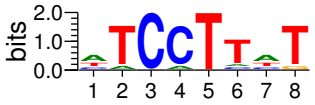
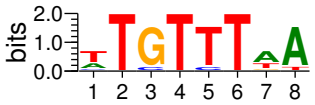
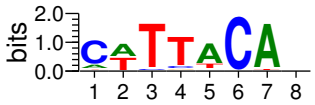
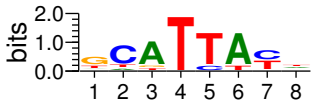
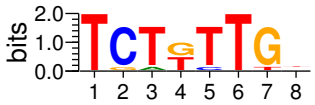
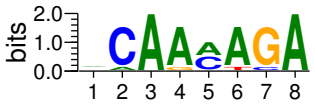
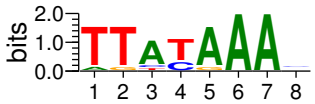
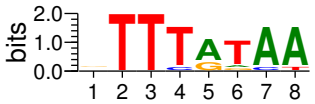
[4]Reverse sequence of the motif; Motif logo representing the occurence of a specific nucleotide at a respective position.The x-axis represents the position of a nucleotide and y-axis represents the amount of information in bits

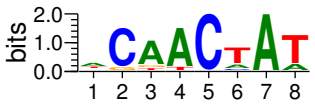
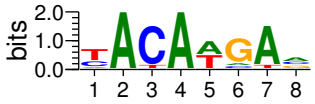
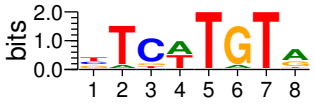
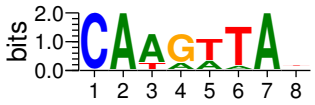
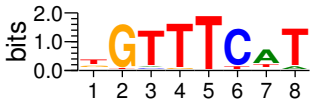
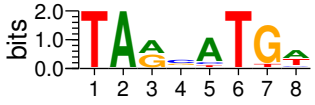
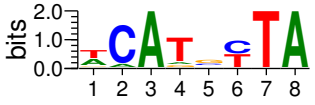
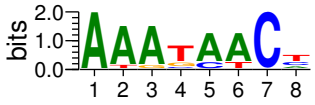
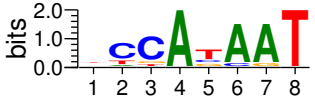
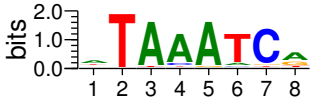
[5]Number of 5' UTRs containing the motif in the specific group. Note: Motif might occur more than once per sequence

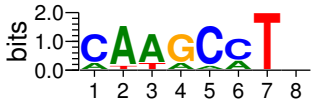
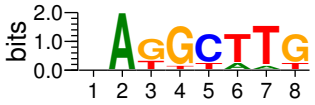
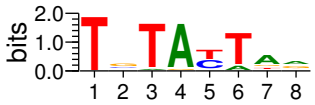
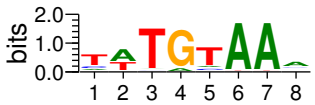
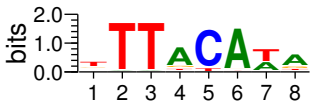
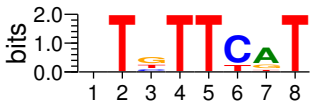
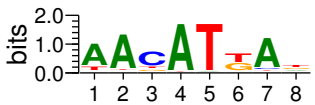
[6]Number of 5' UTRs containing the motif in the whole Arabidopsis genome. Note: Motif might occur more than once per sequence

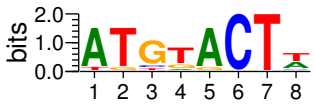
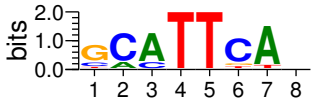
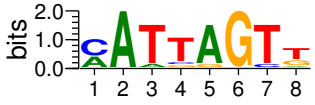
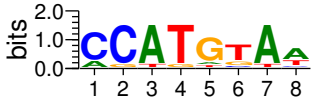
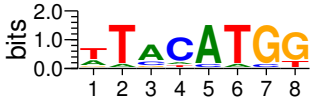
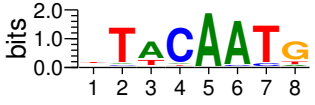
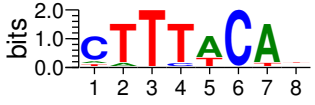
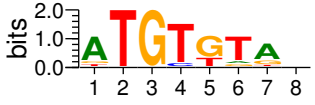
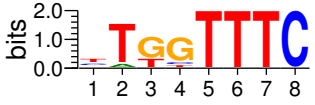
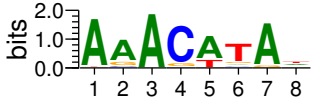
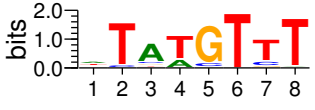
[7]Function of the motif obtained from literature

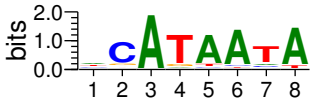
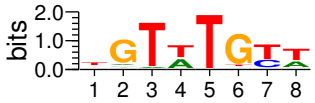
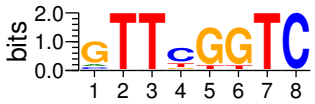
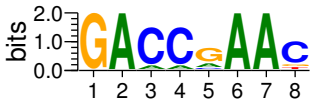
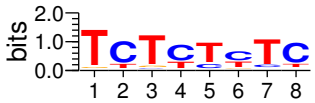
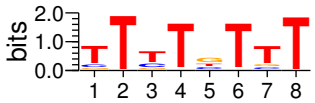
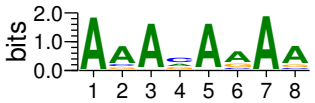
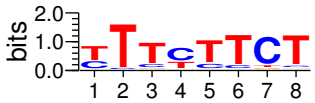
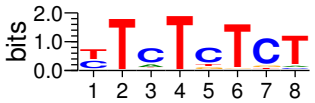
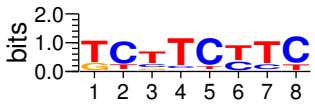
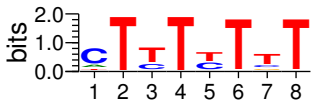
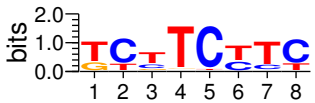
Supplemental Table 8: *De novo* motifs of length 8 found in the 5' UTRs of genes that belong to NYup group

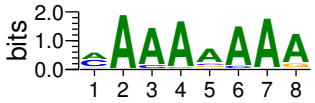
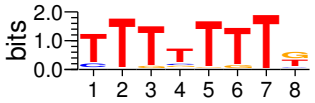
***De novo* motif**						
Motif ID[1]	**Software**[2]	**Forward**[3]	**Reverse**[4]	**Occurence in the group**[5]	**Occurence in the whole genome**[6]	**Possible function**[7]
S10	Seeder			5 / 79	1562 / 19128	-
S11	Seeder			14 / 79	3453 / 19128	-
S12	Seeder			33 / 79	7955 / 19128	-
S13	Seeder			11 / 79	4407 / 19128	-
S14	Seeder			32 / 79	6802 / 19128	-
S15	Seeder			9 / 79	3160 / 19128	-
S16	Seeder			9 / 79	2853 / 19128	-
S18	Seeder			14 / 79	3788 / 19128	-
S19	Seeder			26 / 79	4502 / 19128	-

S1	Seeder			13 / 79	1905 / 19128	-
S20	Seeder			19 / 79	4257 / 19128	-
S21	Seeder			12 / 79	2628 / 19128	-
S22	Seeder			42 / 79	9392 / 19128	-
S23	Seeder			6 / 79	664 / 19128	-
S24	Seeder			17 / 79	4529 / 19128	-
S25	Seeder			25 / 79	5208 / 19128	-
S26	Seeder			20 / 79	4528 / 19128	-
S27	Seeder			13 / 79	2639 / 19128	-
S28	Seeder			25 / 79	5072 / 19128	-

S29	Seeder			12 / 79	1251 / 19128	Matches to motif sequence TAGGGTTT, involved in translation regulation
S2	Seeder			16 / 79	3216 / 19128	-
S30	Seeder			4 / 79	662 / 19128	-
S31	Seeder			27 / 79	6254 / 19128	-
S32	Seeder			8 / 79	2779 / 19128	-
S33	Seeder			25 / 79	7134 / 19128	-
S34	Seeder			3 / 79	1283 / 19128	-
S35	Seeder			19 / 79	3574 / 19128	-
S36	Seeder			13 / 79	4087 / 19128	-

S37	Seeder			5 / 79	1179 / 19128	-
S38	Seeder			5 / 79	1732 / 19128	-
S39	Seeder			5 / 79	1515 / 19128	-
S3	Seeder			8 / 79	766 / 19128	-
S40	Seeder			9 / 79	2973 / 19128	-
S4	Seeder			16 / 79	3045 / 19128	-
S5	Seeder			6 / 79	2736 / 19128	-
S6	Seeder			18 / 79	3911 / 19128	Matches to motif sequence TAGGGTTT, involved in translation regulation
S7	Seeder			87 / 79	19231 / 19128	-

S8	Seeder			16 / 79	3623 / 19128	-
S9	Seeder			41 / 79	8515 / 19128	-
W1	Weeder			18 / 79	832 / 19128	-
M1	MEME			145 / 79	38301 / 19128	-
M2	MEME			308 / 79	60633 / 19128	-
M3	MEME			190 / 79	44507 / 19128	-
M4	MEME			182 / 79	44589 / 19128	-
M5	MEME			136 / 79	37497 / 19128	-
M6	MEME			196 / 79	41592 / 19128	-
M7	MEME			143 / 79	40533 / 19128	-

M8	MEME			233 / 79	46703 / 19128	-

[1]Name of the motif

[2]*de novo* discovery software that was used to locate the motif

[3]Forward sequence of the motif; Motif logo representing the occurence of a specific nucleotide at a respective position. The x-axis represents the position of a nucleotide and y-axis represents the amount of information in bits

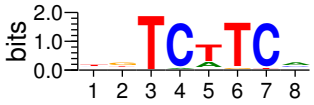
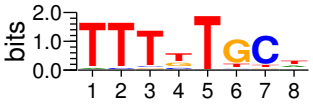
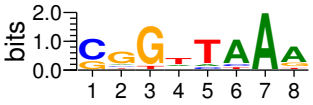
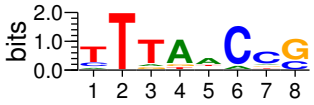
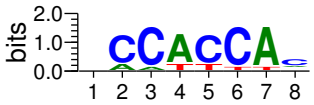
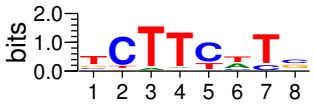
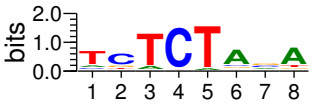
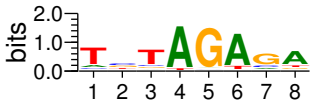
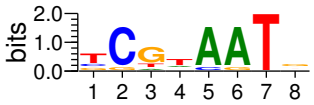
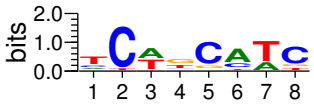
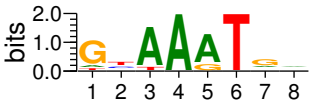
[4]Reverse sequence of the motif; Motif logo representing the occurence of a specific nucleotide at a respective position.The x-axis represents the position of a nucleotide and y-axis represents the amount of information in bits

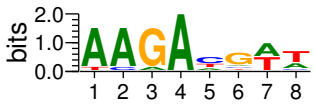
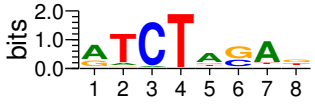
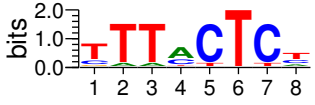
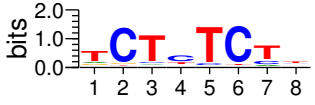
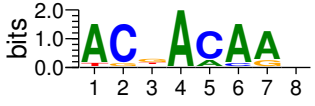
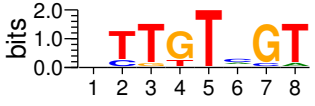
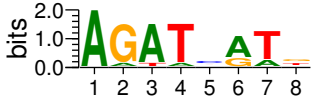
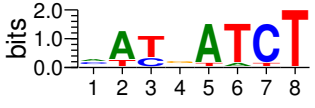
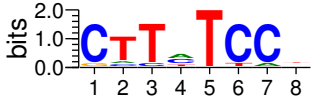
[5]Number of 5' UTRs containing the motif in the specific group. Note: Motif might occur more than once per sequence

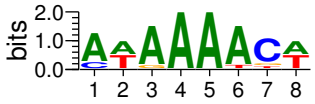
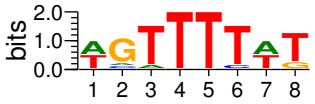
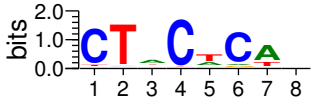
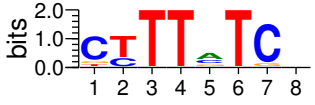
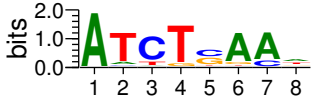
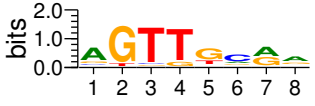
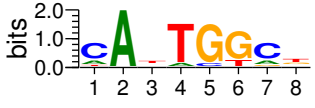
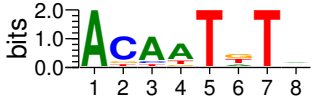
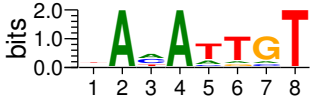
[6]Number of 5' UTRs containing the motif in the whole Arabidopsis genome. Note: Motif might occur more than once per sequence

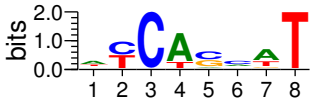
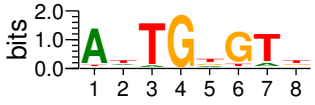
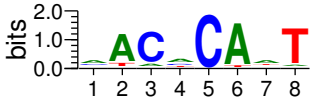
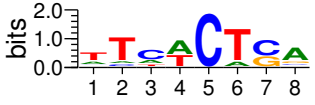
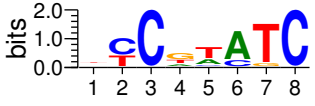
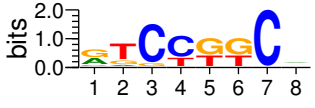
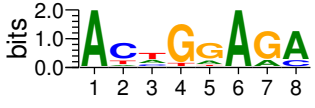
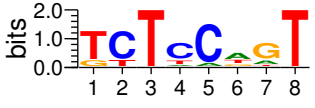
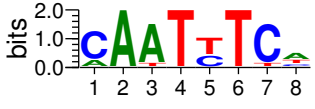
[7]Function of the motif obtained from literature

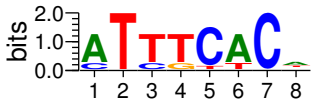
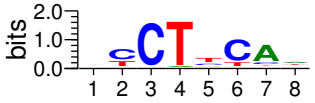
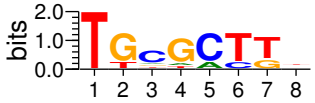
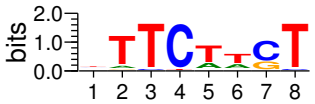
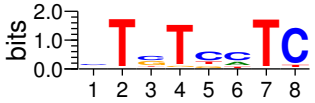
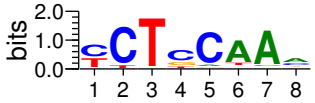
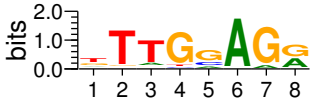
Supplemental Table 9: *De novo* motifs of length 8 found in the 5' UTRs of genes that belong to YNdown group

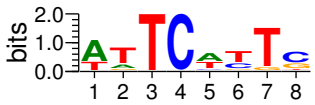
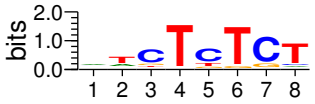
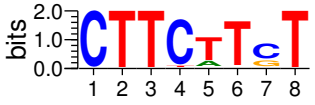
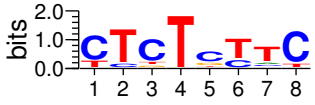
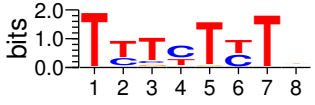
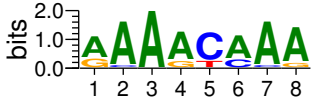
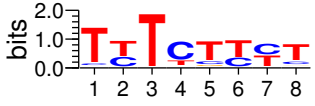
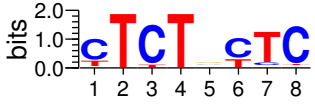
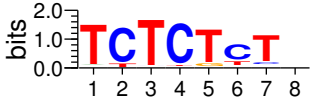
***De novo* motif**						
Motif ID[1]	**Software**[2]	**Forward**[3]	**Reverse**[4]	**Occurence in the group**[5]	**Occurence in the whole genome**[6]	**Possible function**[7]
S10	Seeder			96 / 53	31464 / 19128	-
S11	Seeder			44 / 53	12313 / 19128	-
S12	Seeder			14 / 53	2238 / 19128	-
S13	Seeder			7 / 53	1944 / 19128	-
S14	Seeder			124 / 53	42009 / 19128	-
S15	Seeder			47 / 53	11488 / 19128	-
S16	Seeder			7 / 53	2047 / 19128	-
S17	Seeder			70 / 53	21552 / 19128	-
S18	Seeder			17 / 53	3951 / 19128	-

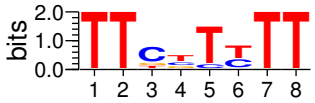
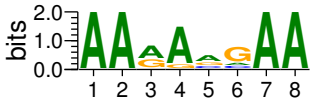
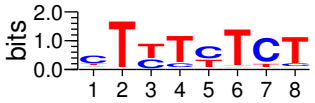
S19	Seeder			42 / 53	14140 / 19128	-
S1	Seeder			20 / 53	5581 / 19128	-
S20	Seeder			10 / 53	2094 / 19128	-
S21	Seeder			31 / 53	6396 / 19128	-
S22	Seeder			26 / 53	8139 / 19128	-
S23	Seeder			126 / 53	42638 / 19128	-
S24	Seeder			20 / 53	4180 / 19128	-
S25	Seeder			42 / 53	6270 / 19128	-
S26	Seeder			21 / 53	4364 / 19128	-
S27	Seeder			23 / 53	6138 / 19128	-

S28	Seeder			21 / 53	3654 / 19128	-
S29	Seeder			36 / 53	9895 / 19128	-
S2	Seeder			30 / 53	4470 / 19128	-
S30	Seeder			51 / 53	14542 / 19128	-
S31	Seeder			25 / 53	6856 / 19128	-
S32	Seeder			57 / 53	16723 / 19128	-
S33	Seeder			14 / 53	5614 / 19128	-
S34	Seeder			9 / 53	1918 / 19128	-
S35	Seeder			47 / 53	14497 / 19128	-
S36	Seeder			11 / 53	2317 / 19128	-

S37	Seeder			21 / 53	4575 / 19128	-
S38	Seeder			31 / 53	8057 / 19128	-
S39	Seeder			47 / 53	10071 / 19128	-
S3	Seeder			19 / 53	2973 / 19128	-
S40	Seeder			87 / 53	25809 / 19128	-
S41	Seeder			6 / 53	736 / 19128	-
S42	Seeder			17 / 53	4850 / 19128	-
S43	Seeder			9 / 53	2346 / 19128	-
S44	Seeder			19 / 53	4835 / 19128	-
S45	Seeder			24 / 53	4021 / 19128	-

S46	Seeder			12 / 53	2452 / 19128	-
S47	Seeder			14 / 53	3266 / 19128	-
S48	Seeder			12 / 53	1824 / 19128	-
S49	Seeder			25 / 53	6172 / 19128	-
S4	Seeder			84 / 53	23947 / 19128	-
S50	Seeder			11 / 53	2114 / 19128	Matches to motif sequence TAGGGTTT, involved in translation regulation
S5	Seeder			113 / 53	38165 / 19128	-
S6	Seeder			61 / 53	17479 / 19128	-
S7	Seeder			25 / 53	6896 / 19128	-

S8	Seeder			52 / 53	12452 / 19128	-
S9	Seeder			150 / 53	48878 / 19128	-
W1	Weeder			38 / 53	10355 / 19128	-
M1	MEME			69 / 53	20644 / 19128	-
M2	MEME			155 / 53	52082 / 19128	-
M3	MEME			53 / 53	12815 / 19128	Matches to motif sequence TAGGGTTT, involved in translation regulation
M4	MEME			165 / 53	55595 / 19128	-
M5	MEME			47 / 53	12280 / 19128	-
M6	MEME			152 / 53	49835 / 19128	-

M7	MEME			57 / 53	16492 / 19128	-
M8	MEME			170 / 53	55854 / 19128	-

[1]Name of the motif

[2]*de novo* discovery software that was used to locate the motif

[3]Forward sequence of the motif; Motif logo representing the occurence of a specific nucleotide at a respective position. The x-axis represents the position of a nucleotide and y-axis represents the amount of information in bits

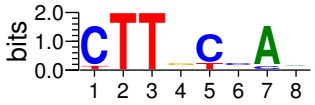
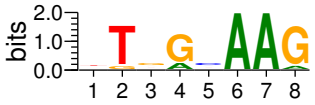
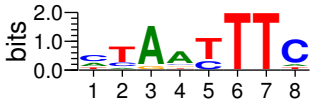
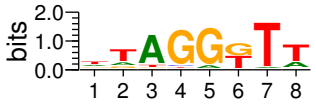
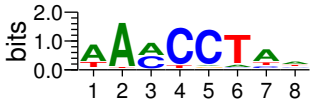
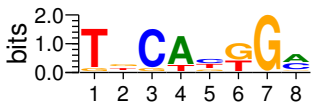
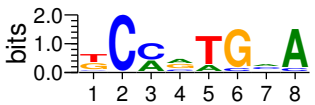
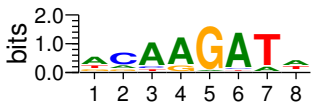
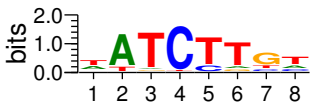
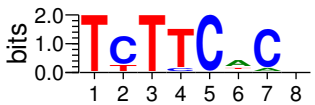
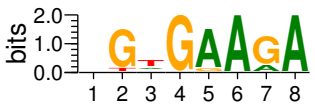
[4]Reverse sequence of the motif; Motif logo representing the occurence of a specific nucleotide at a respective position.The x-axis represents the position of a nucleotide and y-axis represents the amount of information in bits

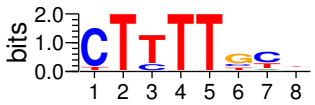
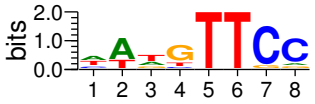
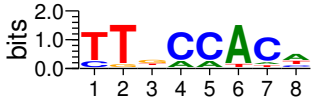
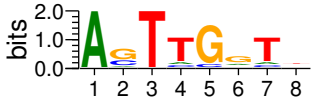
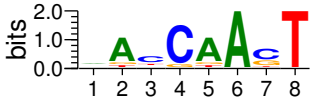
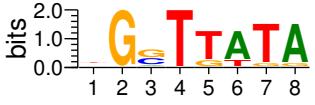
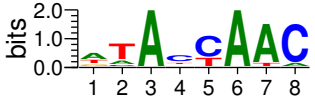
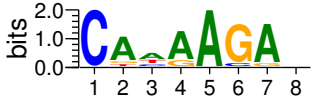
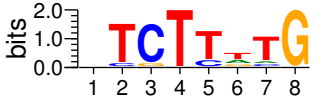
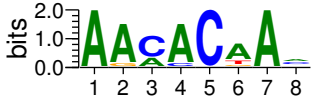
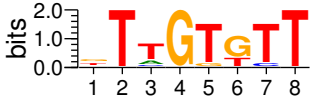
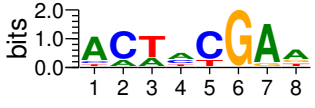
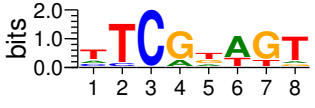
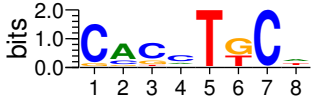
[5]Number of 5' UTRs containing the motif in the specific group. Note: Motif might occur more than once per sequence

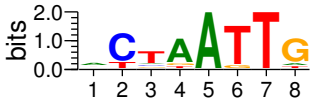
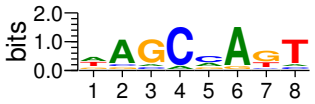
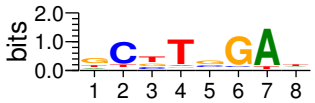
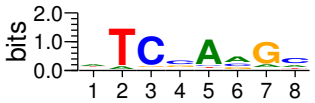
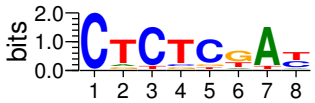
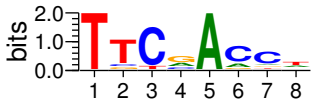
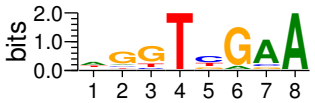
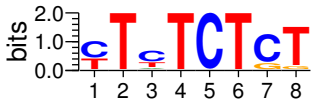
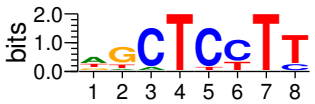
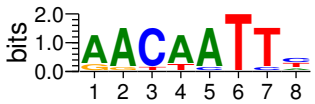
[6]Number of 5' UTRs containing the motif in the whole Arabidopsis genome. Note: Motif might occur more than once per sequence

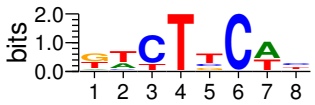
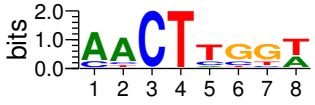
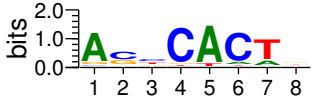
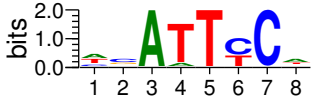
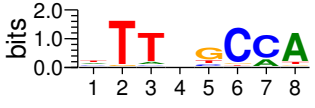
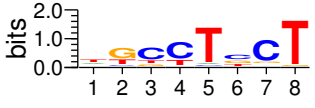
[7]Function of the motif obtained from literature

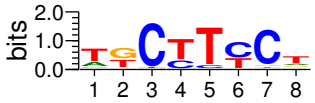
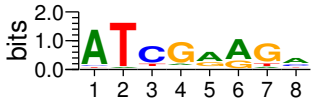
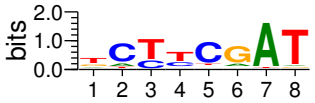
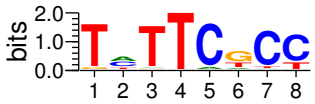
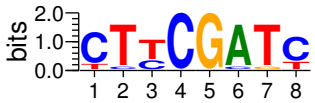
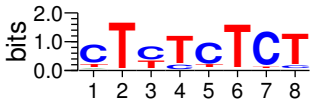
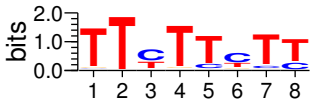
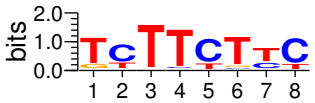
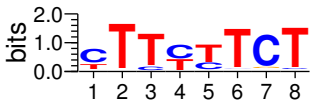
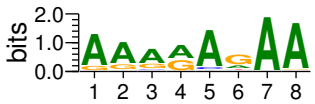
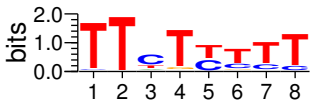
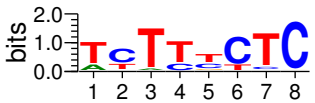
Supplemental Table 10: *De novo* motifs of length 8 found in the 5' UTRs of genes that belong to YNup group

***De novo* motif**						
Motif ID[1]	**Software**[2]	**Forward**[3]	**Reverse**[4]	**Occurence in the group**[5]	**Occurence in the whole genome**[6]	**Possible function**[7]
S10	Seeder			20 / 58	3002 / 19128	-
S11	Seeder			9 / 58	1720 / 19128	-
S12	Seeder			22 / 58	6173 / 19128	-
S13	Seeder			26 / 58	9564 / 19128	Matches to motif sequence TAGGGTTT, involved in translation regulation
S14	Seeder			19 / 58	4220 / 19128	-
S15	Seeder			11 / 58	1971 / 19128	-
S16	Seeder			36 / 58	9671 / 19128	-
S17	Seeder			40 / 58	9715 / 19128	-

S18	Seeder			21 / 58	6929 / 19128	-
S19	Seeder			18 / 58	4889 / 19128	-
S1	Seeder			12 / 58	2657 / 19128	-
S20	Seeder			10 / 58	3483 / 19128	-
S21	Seeder			16 / 58	4025 / 19128	-
S22	Seeder			11 / 58	1654 / 19128	-
S23	Seeder			39 / 58	10098 / 19128	-
S24	Seeder			24 / 58	5632 / 19128	-
S25	Seeder			13 / 58	3206 / 19128	-
S26	Seeder			9 / 58	1936 / 19128	-

S27	Seeder			19 / 58	3796 / 19128	-
S28	Seeder			10 / 58	2296 / 19128	-
S29	Seeder			12 / 58	2020 / 19128	-
S2	Seeder			40 / 58	9892 / 19128	-
S30	Seeder			13 / 58	3476 / 19128	-
S31	Seeder			11 / 58	2455 / 19128	-
S32	Seeder			85 / 58	20125 / 19128	-
S33	Seeder			20 / 58	5012 / 19128	-
S34	Seeder			14 / 58	2508 / 19128	-
S35	Seeder			19 / 58	5417 / 19128	-

S36	Seeder			29 / 58	11125 / 19128	-
S37	Seeder			19 / 58	4872 / 19128	-
S38	Seeder			13 / 58	2237 / 19128	-
S3	Seeder			17 / 58	2654 / 19128	-
S40	Seeder			19 / 58	4652 / 19128	-
S41	Seeder			30 / 58	7594 / 19128	-
S42	Seeder			19 / 58	6041 / 19128	-
S4	Seeder			27 / 58	7959 / 19128	-
S5	Seeder			19 / 58	6269 / 19128	-
S6	Seeder			29 / 58	8464 / 19128	-

S7	Seeder			33 / 58	13465 / 19128	-
S8	Seeder			40 / 58	13342 / 19128	-
W1	Weeder			15 / 58	2801 / 19128	-
M7	MEME			14 / 58	2114 / 19128	-
M1	MEME			134 / 58	37904 / 19128	-
M2	MEME			155 / 58	52185 / 19128	-
M3	MEME			99 / 58	35886 / 19128	-
M4	MEME			159 / 58	48468 / 19128	-
M5	MEME			184 / 58	62806 / 19128	-
M6	MEME			165 / 58	49464 / 19128	-

[1]Name of the motif

[2]*de novo* discovery software that was used to locate the motif

[3]Forward sequence of the motif; Motif logo representing the occurence of a specific nucleotide at a respective position. The x-axis represents the position of a nucleotide and y-axis represents the amount of information in bits

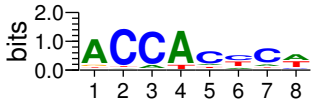
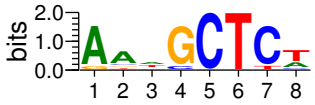
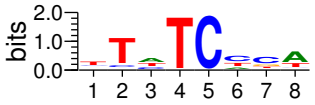
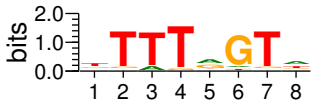
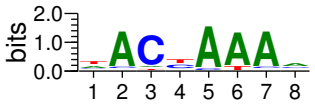
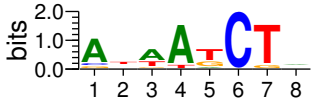
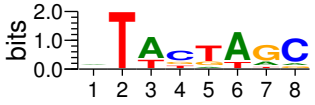
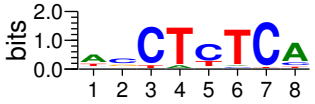
[4]Reverse sequence of the motif; Motif logo representing the occurence of a specific nucleotide at a respective position.The x-axis represents the position of a nucleotide and y-axis represents the amount of information in bits

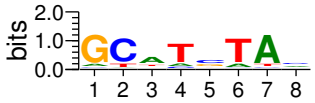
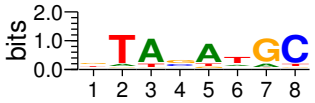
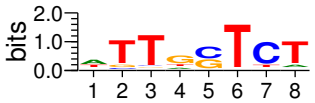
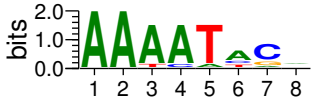
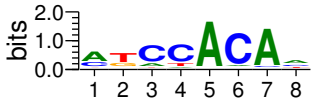
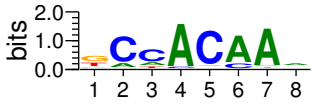
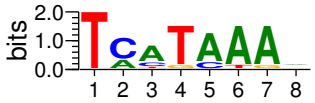
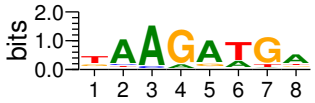
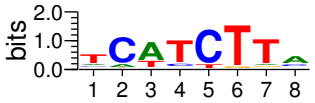
[5]Number of 5' UTRs containing the motif in the specific group. Note: Motif might occur more than once per sequence

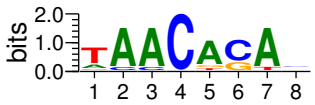
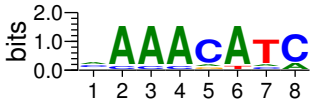
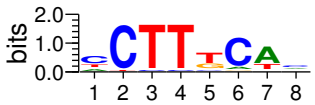
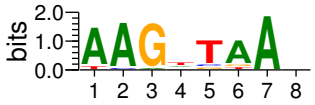
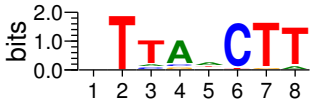
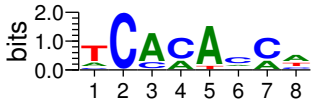
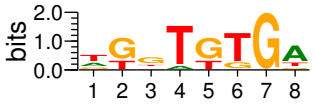
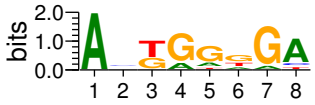
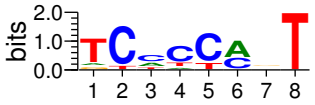
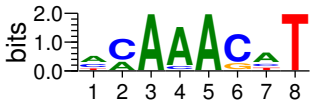
[6]Number of 5' UTRs containing the motif in the whole Arabidopsis genome. Note: Motif might occur more than once per sequence

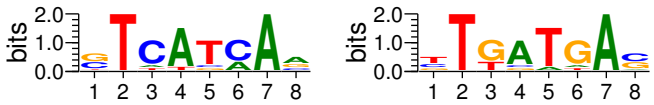
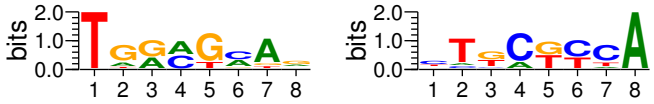
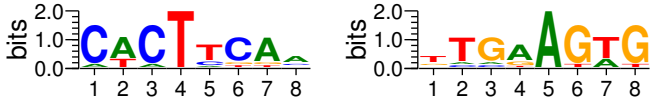
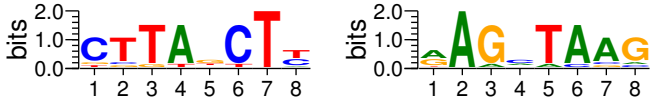
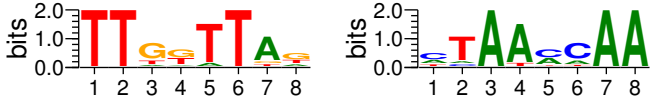
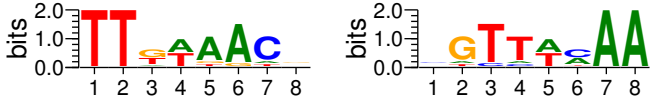
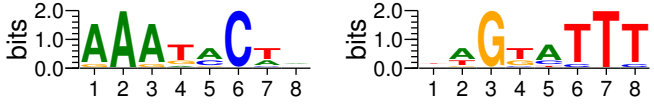
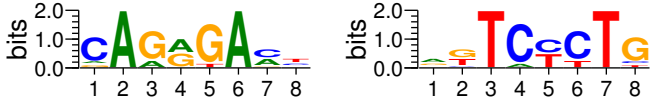
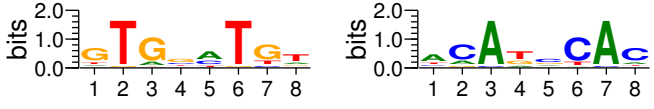
[7]Function of the motif obtained from literature

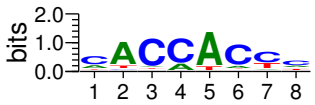
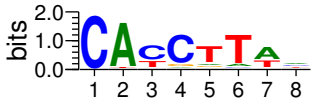
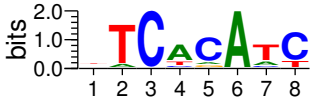
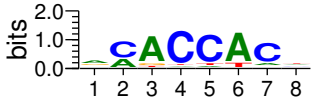
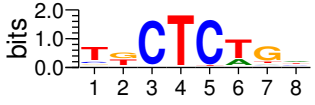
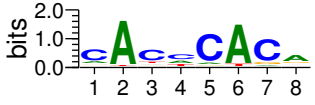
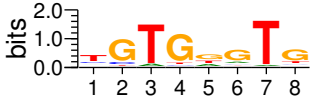
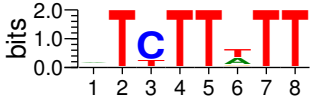
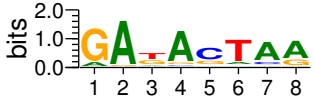
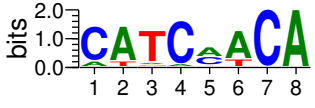
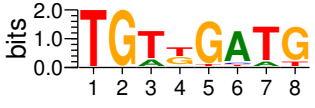
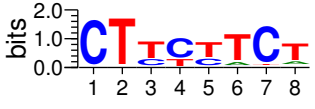
Supplemental Table 11: *De novo* motifs of length 8 found in the 5' UTRs of genes that belong to YYdown group

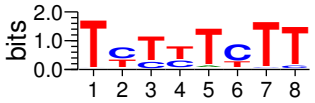
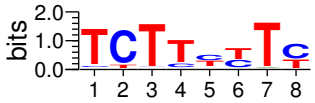
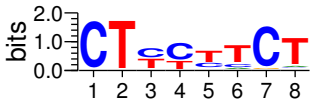
***De novo* motif**						
Motif ID[1]	**Software**[2]	**Forward**[3]	**Reverse**[4]	**Occurence in the group**[5]	**Occurence in the whole genome**[6]	**Possible function**[7]
S0	Seeder			13 / 22	4134 / 19128	-
S10	Seeder			17 / 22	4152 / 19128	-
S11	Seeder			42 / 22	39501 / 19128	-
S13	Seeder			25 / 22	14796 / 19128	-
S15	Seeder			24 / 22	16745 / 19128	-
S16	Seeder			7 / 22	1970 / 19128	-
S17	Seeder			21 / 22	10278 / 19128	-
S18	Seeder			9 / 22	1717 / 19128	-
S19	Seeder			11 / 22	2974 / 19128	-

S1	Seeder			9 / 22	2539 / 19128	-
S20	Seeder			8 / 22	2451 / 19128	-
S21	Seeder			11 / 22	4494 / 19128	-
S22	Seeder			26 / 22	18892 / 19128	-
S23	Seeder			13 / 22	9649 / 19128	-
S24	Seeder			22 / 22	10516 / 19128	-
S25	Seeder			12 / 22	6362 / 19128	-
S26	Seeder			10 / 22	2723 / 19128	-
S27	Seeder			13 / 22	6975 / 19128	-
S28	Seeder			40 / 22	25074 / 19128	-

S29	Seeder			13 / 22	6007 / 19128	-
S31	Seeder			9 / 22	2220 / 19128	Matches to motif sequence TAGGGTTT, involved in translation regulation
S32	Seeder			16 / 22	9203 / 19128	-
S33	Seeder			26 / 22	18392 / 19128	-
S34	Seeder			16 / 22	6289 / 19128	-
S35	Seeder			12 / 22	3171 / 19128	-
S36	Seeder			11 / 22	5218 / 19128	-
S37	Seeder			17 / 22	3439 / 19128	-
S38	Seeder			11 / 22	4365 / 19128	-

S39	Seeder			10 / 22	3460 / 19128	-
S3	Seeder			11 / 22	4559 / 19128	-
S40	Seeder			9 / 22	3664 / 19128	-
S41	Seeder			6 / 22	2344 / 19128	-
S42	Seeder			8 / 22	2437 / 19128	-
S43	Seeder			6 / 22	2756 / 19128	-
S44	Seeder			19 / 22	10360 / 19128	Matches to motif sequence TAGGGTTT, involved in translation regulation
S45	Seeder			7 / 22	3029 / 19128	-
S47	Seeder			10 / 22	3551 / 19128	-

S48	Seeder			12 / 22	2761 / 19128	-
S49	Seeder			16 / 22	4400 / 19128	-
S4	Seeder			13 / 22	4352 / 19128	-
S5	Seeder			21 / 22	6168 / 19128	-
S6	Seeder			23 / 22	11719 / 19128	-
S7	Seeder			12 / 22	4853 / 19128	-
S8	Seeder			20 / 22	9777 / 19128	-
S9	Seeder			15 / 22	7304 / 19128	-
W1	Weeder			14 / 22	3157 / 19128	-
M1	MEME			59 / 22	42584 / 19128	-

M2	MEME			95 / 22	67585 / 19128	-
M3	MEME			63 / 22	45269 / 19128	-
M4	MEME			65 / 22	42790 / 19128	-

[1]Name of the motif

[2]*de novo* discovery software that was used to locate the motif

[3]Forward sequence of the motif; Motif logo representing the occurence of a specific nucleotide at a respective position. The x-axis represents the position of a nucleotide and y-axis represents the amount of information in bits

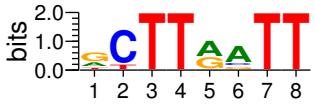
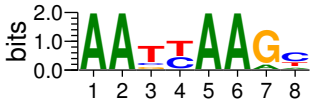
[4]Reverse sequence of the motif; Motif logo representing the occurence of a specific nucleotide at a respective position.The x-axis represents the position of a nucleotide and y-axis represents the amount of information in bits

[5]Number of 5' UTRs containing the motif in the specific group. Note: Motif might occur more than once per sequence

[6]Number of 5' UTRs containing the motif in the whole Arabidopsis genome. Note: Motif might occur more than once per sequence

[7]Function of the motif obtained from literature

Supplemental Table 12: *De novo* motifs of length 8 found in the 5' UTRs of genes that belong to YYup group

***De novo* motif**						
Motif ID[1]	**Software**[2]	**Forward**[3]	**Reverse**[4]	**Occurence in the group**[5]	**Occurence in the whole genome**[6]	**Possible function**[7]
W1	Weeder			6 / 6	1886 / 19128	-

[1]Name of the motif

[2]*de novo* discovery software that was used to locate the motif

[3]Forward sequence of the motif; Motif logo representing the occurence of a specific nucleotide at a respective position. The x-axis represents the position of a nucleotide and y-axis represents the amount of information in bits

[4]Reverse sequence of the motif; Motif logo representing the occurence of a specific nucleotide at a respective position.The x-axis represents the position of a nucleotide and y-axis represents the amount of information in bits

[5]Number of 5' UTRs containing the motif in the specific group. Note: Motif might occur more than once per sequence

[6]Number of 5' UTRs containing the motif in the whole Arabidopsis genome. Note: Motif might occur more than once per sequence

[7]Function of the motif obtained from literature