2013

THE ROLE OF POLYADENYLATION IN SEED GERMINATION

Liuyin Ma
University of Kentucky, lyma223@yahoo.com

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained and attached hereto needed written permission statements(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine).

I hereby grant to The University of Kentucky and its agents the non-exclusive license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless a preapproved embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's dissertation including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Liuyin Ma, Student

Dr. Arthur G. Hunt, Major Professor

Dr. Arthur G. Hunt, Director of Graduate Studies

THE ROLE OF POLYADENYLATION IN SEED GERMINATION

DISSERTATION

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy in the
College of Agriculture
at the University of Kentucky

By
Liuyin Ma

Lexington, Kentucky

Director: Dr. Arthur G. Hunt, Professor of Plant Physiology

Lexington, Kentucky

2013

ABSTRACT OF DISSERTATION

THE ROLE OF POLYADENYLATION IN SEED GERMINATION

Seed germination has many impacts on the uses of seeds, and is an important subject for study. Seed germination is regulated at both transcriptional and post-transcriptional levels. Therefore, it is important to study how polyadenylation regulates gene expression during seed germination. To this end, a modified Illumina GAIIx sequencing protocol (described in Chapter Two) was developed that allows deep coverage of poly(A) site position and distribution.

Alternative polyadenylation (APA) regulates gene expression by choosing one potential poly(A) site on a precursor RNA consequentially shortening/lengthening the mRNA relative to other possible sites. To further explore this phenomenon, genes affected by APA during seed germination and other developmental stages were identified (Chapter Three). These genes were categorized based on the location of poly(A) sites. Several genes were chosen to demonstrate how APA, especially that occurring in the coding regions and 5' untranslated regions, might down regulate gene expression by generating truncated transcripts.

In animal oocytes, maternally-derived mRNAs are stored with short poly(A) tails and reactivated by the cytoplasmic polyadenylation complex. It has been reported that seeds also contain stored mRNAs. Moreover, germination and its completion are less sensitive to *de novo* transcription inhibitors than to poly(A) polymerase inhibitors. Together, these considerations suggest that stored RNA without or with a short poly(A) tail (stored, unadenylated RNA) may be present in dry seed and function in seed germination upon reactivation by cytoplasmic polyadenylation. To further explore this, in Chapter Four, mRNA polyadenylation was studied through the course of germination using a combination of transcriptional inhibitors and the modified sequencing protocol described in Chapter Two. 273 putative stored, unadenylated RNAs were identified. Gene ontology analysis revealed that genes whose products are involved in translation are overrepresented; these genes encode 21 60S- and 10 40S-ribosomal proteins. These results indicate that transcripts whose products are involved in translation might be a major component of the stored, unadenylated RNA pool and, more importantly, translation might be

the first cellular process to be activated during seed germination.

KEYWORDS: Stored RNA, Alternative polyadenylation, Seed germination, Illumina sequencing, *Arabidopsis thaliana*

Liuyin Ma
Student's Signature

April 23, 2013
Date

THE ROLE OF POLYADENYLATION IN SEED GERMINATION

By

Liuyin Ma

Arthur G. Hunt
Director of Dissertation

Arthur G. Hunt
Director of Graduate Studies

April 23, 2013

ACKNOWLEDGMENTS

This dissertation work was carried out under the guidance of my major advisor Dr. Arthur G. Hunt. This dissertation has been crafted and shaped under Dr. Hunt's tireless patience and direction. I would also like to thank my co-advisor Dr. Bruce Downie for his patience, advice and knowledge of seed biology as well as his editorial comments. I am appreciative of the guidance and advice of Drs. Sharyn Perry and Daniel Noonan while serving on my committee. Sincerely thanks to all committee members including Dr. Rebecca McCulley.

I would like to express my gratitude to the present and past members of Hunt lab, Carol Von Lanken, Dr. Pratap Pati, Patrick Thomas, Jason Briones, Morgan Taylor, Dr. Bobby Gaffney and Dr. Lavanya Dampanaboina. I am proud to have worked with them. My special appreciation is given to Carol Von Lanken and I have truly enjoyed the technical support, friendship and stimulating discussion about American music, history and art. I would like to acknowledge Dr. Chappell for providing the supercomputer that I used for my data analysis as well as allowing me to join their discussion group. I would like to thank all of my colleagues in Dr. Chappell's lab.

I would like to thank my mother Yanxia Chen, my father Zhixi Ma and my brother Chenyang Ma for their constant emotional support in my life. I am grateful for the opportunity to pursue my doctorate by the supporting stipend from the China Scholarship Counsel.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter One: Literature review

1.1 Nuclear polyadenylation in eukaryotes

1.1.1 Background: The definition of polyadenylation and overview

In eukaryotes, precursor messenger RNAs (pre-mRNA) are transcribed from DNA by RNA polymerase II (Pol II) in a continuous fashion such that the RNA sequence differs from the sense (non-template) DNA strand only in the substitution of uridine in RNA for the deoxythymidine in DNA and the oxidation state of the 2' carbon of ribose. However, this precursor is processed in the nucleus to a typical mature messenger RNA (Figure 1.1). Processing entails the addition of a cap situated at the 5' end of the 5' untranslated region (5' UTR), the removal of introns (when present) by the spliceosome complex from UTRs and the protein coding sequence (CDS), and the pre-mRNA cleavage and addition, at the site of cleavage, of a tract of adenosine monophosphates (a Poly(A) tail) to the 3' untranslated region (3' UTR) (1). Therefore, to complete mRNA synthesis and become a translatable mature mRNA, the pre-mRNA has to complete the three main steps of mRNA processing: 5'-7-methyl guanine cap addition, exon/intron splicing, and 3' end polyadenylation (1). One of the critical steps for the messenger RNA maturation process is polyadenylation, a non-template assisted, linear addition of approximately 200 adenosine monophosphate molecules to the 3'-hydroxyl of the terminating mRNA nucleotide (1).

Almost all nuclear encoded, eukaryotic pre-mRNAs undergo polyadenylation during maturation except the replication-dependent histone mRNAs in metazoans (1).

Polyadenylation occurs in two steps and is directed by the interaction of the pre-mRNA with the polyadenylation machinery (1). The first step involves an ATP-dependent, pre-mRNA endonucleolytic cleavage to remove the specific fragments located downstream of the poly(A) site (2). This step is followed by the addition of a poly(A) tract to the upstream 3'-OH of the cleavage product (1, 2).

1.1.2 Significance and function of polyadenylation

Polyadenylation is important as the location of the poly(A) site defines the end of a transcript; its incorrect definition may lead to the production of truncated protein or RNA (2). One function of polyadenylation lies in regulating the production of distinct mRNA isoforms through selecting different poly(A) sites, which is termed alternative polyadenylation (APA) (2). APA may be used to control the inclusion or exclusion of a specific segment of RNA sequence such as a microRNA target site and thus control gene expression at the post-transcriptional level (2, 3). The polyadenylation process is also tightly associated with the termination of transcription as the polyadenylation signal indicates the termination of transcription by DNA dependent RNA polymerase II (2, 4). Polyadenylation competes with other events such as transcription initiation, capping and splicing (2, 5, 6). For example, splicing and polyadenylation may compete for the same intron (2, 6-8). The poly(A) tail of the mature mRNA is essential for the function of mRNA including: 1) mRNA localization as mRNA without a poly(A) tail loses its localization priority; 2) nuclear export as the poly(A) tail is one of the recognition signals for the RNA export complex; 3) stability control as a long poly(A) tail protects the mRNA

through the poly(A) binding protein, a short poly(A) tail being vulnerable to 3' → 5' mRNA degradation enzymes; 4) translation initiation as poly(A) binding proteins, bound to the poly(A) tail, are critical for forming the translation initiation complex; and 5) some RNAs such as ribosomal RNA (rRNA) and some mRNAs (particularly in the organelles) are polyadenylated as part of a quality-control degradation process (2, 9, 10).

1.1.3 Nuclear polyadenylation signal cis-elements

Polyadenylation requires multiple defined sequence elements within the 3' end of the pre-mRNA (9, 10). Multi-subunit protein factors required for cleavage and polyadenylation assemble onto these signal elements and function as the polyadenylation machinery complex (9, 10). Although the tripartite arrangement of one A-rich element, one or two more U-rich elements and cleavage sites are similar among animal, yeast and plant pre-mRNA, the sequence and location of the polyadenylation signal elements are different in different eukaryotes (9, 10).

In animals, multiple polyadenylation signal elements are required to define the exact poly(A) cleavage site of pre-mRNA. The three core elements include a highly conserved A-rich element consisting of the hexanucleotide sequence A(A/U)UAAA, an U/GU-rich downstream element (DSE), and the cleavage site of polyadenylation (Figure 1.2)(9, 10). In addition to A(A/U)UAAA and the DSE, a U-rich upstream sequence element (USE) and a G-rich auxiliary downstream element (aux DSE) have been identified in cellular mRNA (Figure 1.2)(9, 10). These additional cis-elements

are functional in regulating the efficiency of 3' processing by providing binding sites for regulatory- and 3' end processing-factors (10). The AAUAAA or AUUAAA polyadenylation signal element is located approximately 40 nucleotides (nt) upstream of the cleavage/polyadenylation site (10). AAUAAA or AUUAAA is found in approximately 53 to 58% and around 15 to 17% of human (*Homo sapiens*) poly(A) sites, respectively. A more variable GU-rich element (DSE; Fig. 1.2A) and an additional U-rich element downstream from GU-rich element are located less than 40 nt downstream of the cleavage/polyadenylation site (10, 11). A distal downstream G-rich element is typically located >30nt from the poly(A) site (Fig. 1.2A). In addition, a putative C-rich element located 40 to 100nt downstream from the poly(A) site has been reported using a bioinformatics analysis (10, 11). No significant element has been identified at the cleavage site despite a bias toward CA dinucleotides found in vitro (10, 11).

The polyadenylation signals in yeast are significantly different from those in animals due to the lack of conservation (apparently tolerant of significant variation) of the A(A/U)UAAA element (10, 11). Instead, yeast includes the degenerate A-rich positioning element (PE) in which either AAUAAA or AAAAAA are elements located 10-30nt upstream of the poly(A) site (9-11). There is an UA-rich efficiency element (EE) with UAUAUA being the optimal sequence located 25-75 nt upstream of the poly(A) site (Figure 1.2B) (9-11). The yeast pre-mRNA also includes both U-rich elements upstream (UUE) and downstream (DUE) of the polyadenylation/cleavage site and in close proximity to it (9-11). The cleavage/polyadenylation site itself is

defined by a pyrimidine followed by multiple adenosines Py(A)n (Figure 1.2B) (9-11).

The animal poly(A) sites are not functional in plant cells according to a study in the mid-1980s (12) indicating that poly(A) signals in plants may differ from those in animals. In the early 1990s, detailed studies of poly(A) signals established a general plant poly(A) signal element (13). In plants, the three core polyadenylation signal elements include a near upstream element (NUE), a far upstream element (14) and a cleavage element (CE) (Figure 1.2C) (2). The NUE is an A-rich signal element that is functionally equivalent to the dominant AAUAAA signal element in animals and is located 10-40nt upstream of the cleavage site (2). The canonical AAUAAA is the best signal concerning the frequency of correct pre-mRNA cleavage and polyadenylation (15) but this stereotypical signal can only be found in approximately 10% of the Arabidopsis (*Arabidopsis thaliana*) or rice (*Oryza sativa*) genes (2). The FUE is a 60-120nts, U-rich signal element located further upstream from the NUE (2). Compared with the NUE, the signal and the size of nucleotide sequence patterns of FUEs are more diverse and variable (2). No highly conserved sequence patterns are found in Arabidopsis or rice but a clear trend of U-enrichment in this region is observed for FUEs (2). The next polyadenylation element is CE, which includes the cleavage site and its surrounding regions (2). In plants, there is a YA (CA or UA) dinucleotide located at the cleavage site (2). The complicated transition of preferred nucleotides (U->A->C-rich) in the regions surrounding the cleavage site is observed in both Arabidopsis and rice (2, 11).

1.1.4 3’ end processing factors

In animals, multiple polyadenylation factors have been identified. These factors include the cleavage and polyadenylation specificity factor (CPSF), cleavage stimulatory factor (CstF), animal cleavage factor I (CFIm), animal cleavage factor II (CFIIm), poly(A) polymerase (PAP), and poly(A)-binding proteins (PABPs). CPSF, CstF, CFIm and CFIIm are multi-subunit protein complexes (9, 10) (Appendix 1.1). All of these factors, except PABP, contribute to the pre-mRNA endonucleolytic cleavage reaction (10). However, only CPSF and PAP are required in the reaction for the addition of a poly(A) tail (10). PABPs are required for neither endonucleolytic cleavage nor the polyadenylation reaction (10). However, PABP activates PAP and controls poly(A) tail length by regulating the interaction between CPSF and PAP (16).

In yeast, the 3’ end polyadenylation complex consists of several polypeptide subunits which include yeast cleavage and polyadenylation factor (CPF) and yeast cleavage factor I and II (CFIy and CFIIy) (17). The CPF contains subunits that are homologous to the animal CPSF subunits. The CPF itself consists of cleavage factor II (CF II), polyadenylation factor I (PF I) and additional factors such as polyadenylation factor subunit 2 (Pfs2), suppressor of *sua7-1* (transcription factor TFIIB) clone 2 (Ssu72), mutant Pcf11 (protein 1 of cleavage and polyadenylation factor I) extragenic suppressor 1 (Mpe1), and yeast catalytic subunit of type 1 serine/threonine protein phosphatase (Glc7) (1, 18). Rna15p, the yeast homolog of CstF64, associates with the A-rich PE (19). Only the Cleavage Factors, CFIA, CFIB and

CFII are required in the reconstitution of the pre-mRNA cleavage reaction *in vitro*. CPF, CFIA, CFIB and Pab1p are required for polyadenylation based on in vitro assays (19).

In plants, polyadenylation factors are encoded by homologs of human or yeast polyadenylation factors except for animal cleavage factor I and heterogeneous nuclear ribonucleoprotein 1 (Hrp, a yeast RNA binding protein 1) and both of which are missing in plants (9). Thus, the plant polyadenylation complex consists of several cleavage and polyadenylation specificity factors (CPSF), cleavage stimulatory factors (CstF), poly(A) polymerases (PAP), poly(A)-binding proteins (PABP) and other proteins such as flowering locus Y (FY), factor interacting with poly(A) polymerase (Fip1), and symplekin (SYM5) (20) (Appendix 1.1). CPSF100 serves as the core of the CPSF complex and CPSF30 is linked to the CstF complex via its interaction with Fip1, and Fip1 further mediates the interaction between CstF and other polyadenylation factors (9, 21). The interactions between CPSF30 and other CPSF subunits are different from those in other eukaryotes (9, 22). CPSF73 II and FY are involved in plant female gamete development and flowering regulation, respectively (9, 23). Some Arabidopsis polyadenylation factors are encoded by modest gene families and differ in that respect from human or yeast polyadenylation factors that are encoded by single genes (20). Unlike the yeast and human polyadenylation factors, which have essential functions, some plant polyadenylation factors affect only specific biological functions, and CPSF30, is one such example as mutants of this gene are not lethal (8, 9, 24).

1.1.4.1 Cleavage and polyadenylation stimulatory factor (CPSF) Subunits

Both cleavage and polyadenylation require CPSF. In animals, CPSF recognizes the near upstream element (NUE, A(A/U)UAAA; Figure 1.2) and functions in assembling other components of the 3' processing complex. Originally, the purified CPSF complex was thought to be comprised of CPSF subunits: 160, 100, 73 and 30 (25, 26). However, symplekin (SYMPK), a human factor interacting with poly(A) polymerase (hFip1) and the WD repeat-containing protein (Wdr33) have been added to the list of CPSF components (27-29). The CPSF complex has functions in RNA binding, protein-protein interactions, and the catalysis of cleavage. Even though none of the CPSF subunits have a canonical RNA Recognition Motif (RRM), CPSF recognizes A(A/U)UAAA elements specifically (30, 31). CPSF160 is thought to be the main CPSF subunit interacting with the A(A/U)UAAA element based on UV-cross linking (32) and *in vitro* RNA-binding assays (33). The central domain of Cft1p, the yeast CPSF160 homolog, is reported to be responsible for the RNA-binding activity of the protein (34). CPSF30 (35) and hFip1 (29) are also involved in RNA binding. The zinc finger domain of CPSF30 and the C-terminal arginine-rich domain of hFip1 are responsible for binding to the RNA at preferred U-rich sequences (29, 35). The CPSF30 and hFip1 may be involved in recognizing the USE sequences and in stabilizing the binding of CPSF160 to A(A/U)UAAA elements (10).

CPSF recruits other 3' processing complex components through direct physical interactions (10). Both CPSF and CstF can be isolated in a pre-assembled complex indicating an association among each other, and CPSF160, hFip1 and symplekin are

responsible for mediating this association (10, 28, 29, 33). Pfs2p, the yeast homolog of Wdr33, interacts with yeast homologs of mammalian CPSF and CstF complexes, indicating the Wrd33 may also be involved in bridging the CPSF and CstF complexes in animals (36). These interactions help the binding of CPSF and CstF to A(A/U)UAAA and the DSE, respectively (10). After endonucleolytic cleavage, CPSF remains bound to the A(A/U)UAAA element and anchors PAP to the pre-mRNA for polyadenylation through interactions among CPSF160, hFip1 and PAP (10). CPSF73 is the subunit responsible for the endonucleolytic cleavage step (37, 38).

In yeast, Cft1p/Yhh1p (CPSF160) binds to RNA directly, near the A-rich cleavage sites and in a different manner than its animal homolog, CPSF160, which binds to the A(A/U)UAAA element (34). Brr5p/Ysh1p (CPSF73) may be involved in pre-mRNA binding and endonuclease cleavage (19, 39, 40). Brr5p/Ysh1p interacts with Clp1p and may act as a bridge between CPF and CFII (19, 41, 42). Cft2p/Ydh1p (CPSF100) interacts with many subunits of CPF (Cft1p/Yhh1p, Brr5p/Ysh1p, Pta1p, Pfs2p, Ssu72p, YDL094c), Pcf11p of CFIA, and the CTD of DNA dependent RNA polymerase II (41). Mutation of Cft2p/Ydh1p disrupts both cleavage and polyadenylation *in vitro* and *in vivo* where it is lethal (41) indicating that it has a function in maintaining yeast cell viability (43). The cleavage reaction is not affected but the polyadenylation reaction is abolished after deletion of the last CCCH zinc finger motif in the yeast mutant *yth1-1* (44) indicating that the Yth1p is involved in the polyadenylation process. Fip1p interacts with both Pap1p and Yth1p. The

amino acids 80 to 105 of Fip1p are responsible for interacting with Pap1p, and amino acids 206 to 220 are responsible for interacting with Yth1p (45).

The *Arabidopsis* CPSF complex (AtCPSF) consists of AtCPSF100 (At5g23880), AtCPSF30 (At1g30460), AtCPSF73-I (At1g61010), AtCPSF73-II (At2g01730), AtCPSF160 (At5g51660), AtFip1 (At5g58040), AtFY (At5g13480) and *Arabidopsis* symplekin (At5g01400, At1g27590/At1g27595). AtCPSF100 is an essential gene as mutants deficient in AtCPSF100 are embryo lethal (46). In addition, an AtCPSF100 point mutant is deficient in transcription termination and/or polyadenylation (47). AtCPSF100 can form a homodimer (21). AtCPSF100 interacts with AtCPSF160, AtCPSF73-I, Arabidopsis symplekin and FY as those factors co-purified with FLAG-tagged AtCPSF100 (47). Moreover, results from immunoprecipitation assays of CPSF100 using nuclear extracts (8) as well as two-hybrid and in vitro assays (48) indicate that AtCPSF100 may interact with AtCPSF30 (20). In addition, a recent proteomic study based on a TAP-tagged assay revealed that AtCPSF100 interacts with AtCPSF30 and AtCPSF73-II (21). Overall, AtCPSF100 serves as a core and associates with all other CPSF subunits except AtFip1 (V) (21).

AtCPSF160 and AtCPSF100 are stably associated with AtFY *in vivo* (49). AtCPSF160 and AtFY interact with both AtCPSF73-I and AtCPSF73-II (21). Arabidopsis has two AtCPSF73 genes that encode CPSF73-like proteins: AtCPSF73-I and AtCPSF73-II. Both AtCPSF73 genes are essential genes as knockout or knockdown mutants are lethal (23, 48). Plants hemizygous for the wild-type AtCPSF73-II gene were defective

in female gametogenesis (23) while overexpression of AtCPSF73-I (48) altered male gametogenesis. Both AtCPSF73 proteins interact with AtCPSF100, AtCPSF160 and AtFY (21). Both AtCPSF73-I and AtCPSF73-II interact with themselves to form homo-dimers (e.g. AtCPSF73-I-AtCPSF73-I) (21). In addition, AtCPSF73-I interacts with CLP-similar protein 3 (AtCLPS3), an ortholog of human polyadenylation factor CLP1 while AtCPSF73-II interacts with AtCPSF30 (21). The homolog of AtCPSF73-I protein in animals is thought to be involved in the cleavage step of mRNA during 3' end formation (37, 38) while analogous functions of the Arabidopsis CPSF73 protein are not clear.

In Arabidopsis, CPSF30 protein is encoded by a complex gene whose transcripts are alternatively spliced to generate transcript encoding the AtCPSF30 protein or a transcript encoding a larger protein with a longer C-terminal end than that of AtCPSF30 protein (8). The AtCPSF30 can form homodimers (21). The AtCPSF30 interacts with AtCPSF100, AtCPSF160, AtCPSF73-II and AtCLPS3(21). The AtCPSF30 possesses three CCCH zinc finger motifs while its homologs in animal and yeast possess five such motifs (8, 50). The AtCPSF30 is an RNA binding protein and the N-terminal zinc finger motif is responsible for this activity (50). The AtCPSF30 is also an endonuclease and the C-terminal zinc finger motif is responsible for this activity. The endonuclease reaction produces a 3'-OH group that is suitable substrate for poly(A) polymerase (50). Thus, CPSF30 interacts with other CPSF subunits as well as other polyadenylation complexes and sits at the center of protein-protein interactions (17, 21). The CPSF30 deficient mutant is more tolerant to oxidative

stresses than the wild-type plant (24). It is thought that RNA binding and nuclease activity of AtCPSF30 is affected by calmodulin and sulfhydryl reagents (8, 51), respectively. It is also thought that one of three zinc finger motifs is engaged in a dithiothreitol-sensitive disulfide bond (52). These studies reveal that AtCPSF30 is a possible mediator of regulated alternative polyadenylation as well as communication of the protein with calcium availability and cellular (cytoplasmic) redox status (53). Next-generation sequence proved that poly(A) site choice is changed in AtCPSF30 deficient mutants genome-wide (53).

AtFip1(V) (or FIPS5), possesses a conserved motif that is also found in the yeast Fip1 and hFip1 (54). AtFip1 interacts with AtPAP(IV), AtCstF77, AtCPSF30 and at least one AtPabN isoform (54).

AtFY, homologs of yeast Pfs2 protein, contains seven conserved WD repeats (55). The C-terminal domain of AtFY consists of repetitive proline- and glutamine-rich regions that are frequently associated with protein-protein interaction (20). AtFY is involved in the autonomous pathway of flowering time determination (49). The transient interaction of AtFY and Arabidopsis RNA-binding protein FCA down-regulates an endogenous floral repressor gene through general chromatin silencing (49). AtFY can homodimerize (21) as well as interact with AtCPSF100, AtCPSF160, AtCPSF73-I and AtCPSF73-II (21).

Plants possess at least one gene encoding proteins that are similar to symplekin and the yeast counterpart Pta1p. The plant symplekin isoform is associated with AtCPSF100, AtFY, AtCPSF160 and AtCPSF73-I indicating that symplekin is a plant polyadenylation factor subunit (20, 47).

1.1.4.2 Cleavage stimulatory factor (CstF) Subunits

In animals, CstF binds to the DSE and is required for endonucleolytic cleavage but not polyadenylation (10, 19, 56). The CstF complex consists of CstF77, CstF50 and either CstF64 or its paralog CstF64 τ (57, 58). CstF64/ τ are responsible for the RNA-binding activity of the CstF complex (57). The N-terminal RNA-recognition motif (RRM) domain of CstF64 binds with high affinity to DSE-like, GU-rich sequences (59). The intact CstF complex only weakly binds to polyadenylation sites based on observations from UV crosslinking and gel mobility shift assays and the stable association of CstF with DSEs requires the binding of CPSF to the A(A/U)UAAA element (26). Despite containing a similar structure to CstF64, CstF64 τ has a different RNA binding specificity (60). CstF64 τ is highly expressed in testis and thought to function in mediating testis-specific poly(A) site choice (58, 61). The proline-rich domain of CstF 77 interacts with both CstF64 and CstF50 as the scaffold of the CstF complex, though neither CstF64 nor CstF50 interact with each other (28). The HAT (half a tetratricopeptide repeat (TPR)) of CstF77 forms a homodimer (62, 63) and, together with a capacity for homodimerization exhibited by CstF50 (28), indicates that the CstF complex may function as a dimer in the pre-mRNA 3' processing complex. In order to facilitate co-operative binding of CPSF and CstF to

pre-mRNA, both CstF77 and CstF64 interact with the CPSF subunit (10). CstF50 interacts with the DNA dependent RNA Polymerase II C-terminal domain (CTD) and may function in recruiting the CstF to the transcription elongation complex (10, 64).

In yeast, The CFI complex consists of CFIA and CFIB (65, 66). CFIA is further classified into four subunits: Rna14, Rna15, Pcf11 and Clp1. Hrp1 is the only protein in CFIB (65, 66). The Rna 15p specifically binds to the A-rich PE of yeast pre-mRNA through its RRM (67). Ten HATs are identified in Rna14p, the yeast homolog of CstF-77 (66). HATs function in protein-protein interaction. The HAT C-domain regulates the dimer formation of Rna14p based on solution- and microscopy-assays (66). Rna14p and Rna15p interact with each other and form a heterotetramer based on electronic microscopy and analytical ultracentrifugation results (66).

In Arabidopsis, cleavage stimulatory factor (CstF) subunits consist of AtCstF77 (At1g17760), AtCstF64 (At1g71800) and AtCstF50 (At5g60940). AtCstF77 is encoded by one single-copy gene (68). Deletion mutants affecting the 13th exon of the gene encoding *AtCstF77* are viable while mutation of *AtCstF77* homologs in other organisms is lethal. In addition, the Arabidopsis mutant possesses a delayed flowering phenotype indicating that AtCstF77 may be involved in flowering control (68). Another mutant having a T-DNA insertion in the 5' end of the AtCstF77 gene possesses female gametophytic lethality (68). An Arabidopsis AtCstF64 mutant possesses reduced organ size, pale leaves and sterility indicating AtCstF64 functions in regulation of plant growth and development (68). The AtCstF77 interacts

physically with AtCstF64 through a proline-rich region in plants, just like its animal counterparts (68). AtCstF64 and AtCstF77 function in flowering time control by repressing the expression of the *FLOWERING LOCUS C* (*FLC*) gene (68). Further analysis proved that AtCstF64 and AtCstF77 are required for 3' processing of the *FLC* natural antisense transcript (*COOLAIR)* but not the sense transcripts of a specific RNA binding protein (FPA) (68). These processes trigger histone demethylase activity and result in localized chromatin remodeling during vernalization ultimately reducing *FLC* sense transcription (68). Unlike its animal counterpart, AtCstF50 are not co-immunoprecipitation with AtCstF77 suggested that they are not interact with each other (20, 69). The functions of AtCstF50 are not yet clear (20).

1.1.4.3 Poly(A) polymerase and Poly(A)-Binding Proteins (PABPs)

In animals, PAP consists of a highly conserved N-terminal nucleotidyltransferase catalytic domain, a RNA-binding domain and a C-terminal domain (CTD) with a bipartite nuclear localization signal (NLS) and is rich in serine/threonine residues (10, 19, 56). The CTD is the most regulated domain of the PAP and is subject to such post-translational regulation as hyperphosphorylation by the mitosis-promoting factor during mitosis which inhibits PAP and blocks protein synthesis (70). PAP can bind to the 14-3-3ε protein through the CTD domain and the 14-3-3ε protein inhibits PAP activity and its subcellular localization (71). The PAP CTD is acetylated by the cAMP response element-binding protein (CREB protein) to inhibit the association of CFIm25 and PAP while also inhibiting PAP nuclear localization (72).

Finally, the CTD of PAP can undergo sumoylation to stabilize PAP but inhibit its activity *in vitro* (73).

Neo-poly(A) polymerase (Neo-PAP) , Star-poly(A) polymerase (Star-PAP) and Testis-specific poly(A) polymerase (TPAP) are the three non-canonical nuclear PAPs (10). TPAP is specifically expressed in testes and located in both nuclear and cytoplasmic compartments (74). Neo-PAP is involved in the 3' end processing based on *in vitro* assay (75) but its in vivo function is still not characterized (10). Star-PAP consists of a Zinc finger domain and a RRM domain in the N-terminus, a split catalytic domain in the middle of the protein, and a PAP associated domain and a RS domain in the C-terminus (76). Star-PAP interacts with the subunits of CPSF and the type I phosphatidylinositol 4-phosphate 5-kinases (PIPKIα) to control 3' end pre-mRNA processing (76). Phosphatidylinositol-4,5-bisphosphate (PtdIns4,5P) directly regulates the activity of Star-PAP (76).

At least five PABPs exist in humans with PABPN1 located in the nucleus and the other four (PABPNC1,3-5) in the cytoplasm (77, 78). PABPN1 stimulates PAP activity and regulates poly(A) tail length by stabilizing the polyadenylation complex (79). PABP coats the entire poly(A) tail and is terminated by a poorly understood, but PABPN1-dependent mechanism, when it reaches ~250nts (79). Recently, it has been reported that PABPN1 suppresses the alternative polyadenylation by repressing the cleavage at weak proximal poly(A) sites (80). In the cytoplasm, PABPs have important functions in regulating translation through promoting the

interaction of the 5'- and 3' ends of the mRNA and stimulating translational initiation (77, 78).

In yeast, Pap1p is the poly(A) polymerase and is required for accurate pre-mRNA cleavage and polyadenylation (81). *PAP1* is an essential gene (82). Pap1p interacts with Fip1p in vitro (83). The Pab1p is the yeast poly(A) binding protein that binds to the poly(A) tail at the 3' end of mRNAs and is highly conserved (84). Pab1p is thought to be an important mediator of several functions of the poly(A) tail in mRNA biogenesis, stability and translation (85). Pab1p can copurify and interact with nuclear cleavage factor IA (CF IA) (85). Pab1p regulates both mRNA decay and translation (85). Pab1p can bind to eukaryotic initiation factor 4G (eIF4G) leading to the formation of a circular message (85) thus regulating translation. Pab1 is also required for the export of mRNA from nucleus to the cytosol (85).

In Arabidopsis, there are four genes in a small gene family that have been found encoding homologs of canonical nuclear poly(A) polymerases (20). One of these PAPs (AtPAP-III) is truncated at its C-terminus and lacks a nuclear localization signal (86). This AtPAP-III is located in the cytosol (86) and further studies reveal that this AtPAP-III is specifically expressed in pollen (17).

Eight putative poly(A) binding proteins (PABs) have been identified using a comparable evolutionary analysis in Arabidopsis (87). These PABs can be classified into four groups (87). Class I PABs consist of PAB3 and PAB5 and possess an

expression pattern limited to reproductive tissue (87). Class II PABs consist of PAB2, PAB4 and PAB8 and are highly and universally expressed (87). It has been reported that Class II PABs are required for replication of tulip mosaic virus (88). Class III PABs consist of PAB6 and PAB7 and have a restricted and weak expression pattern (87). Class IV PABs consist only of PAB1, which has a low and tissue specific expression pattern (87). Arabidopsis has three nuclear poly(A) binding proteins (AtPABN). At5g51120 encodes AtPABN1, a homolog of the human PABN1 protein. Mutants deficient in *AtPABN1* gene are lethal (Dr. Arthur Hunt, personal communication). The other two *AtPABNs* genes are At5g65260 and At5g10350. Their functions are not clear (17, 89).

1.1.4. 4 Other subunits

Animal cleavage factor I (CFIm) has an RNA-binding activity and is required only for the endonucleolytic cleavage step (10, 19). CFI_m functions as a heterodimer consisting of CFI_m25 and one of two structurally similar proteins, CFI_m59 or CFI_m68 (90, 91). UV crosslinking assays suggests that all CFI_m subunits are involved in RNA recognition (92). The Nudix domain of CFI_m25 functions as an RNA-binding domain and specifically recognizes the UGUAN motif (91). The RNA binding activity of the RNA-Recognition Motif (RRM) domains present on the CFI_m59 and 68 subunits, depend on the presence of CFI_m25 (93). CFI_m interacts with RNA to improve the 3' processing efficiency at canonical poly(A) signals but also functions as the primary RNA-binding factor for noncanonical UGUAN motif-containing pre-mRNA (94, 95). CF I_m25 is not only self-associating but also interacts with PAP and PABP1 (19). Due

to the C-terminal RS domains (arginine/serine dipeptide repeats), the CFI_m59 and 68 binds to other RS proteins (93, 96) and the CFI_m proteins are also components of the spliceosome (97).

Animal Cleavage Factor II_m ($CFII_m$) consists of pre-mRNA cleavage and polyadenylation factor II protein 11 (Pcf11) and human RNA 5'-kinase (hClp1) and is apparently required only for the pre-mRNA cleavage step (10). The Drosophila (*Drosophila melanogaster*) and yeast Pcf11 homologs are involved in both pre-mRNA processing and transcription termination (5). The hClp1 has an RNA 5' kinase activity (98, 99) but this activity may not be required for pre-mRNA 3' end processing as the yeast Clp1 protein does not have a detectable kinase activity (100).

In Arabidopsis, orthologs of animal CFI_m 68 and CFI_m 25 have been identified but the functions of these factors are not clear. In Arabidopsis, CLP1-Similar protein 3 (AtCLPS3) is an ortholog of human hCLP1. Mutants deficient in AtCLPS3 revealed that AtCLPS3 is essential for embryo development and female gametophyte transmission (101). The Pcf11-similar protein 4 (PCFS4) is an ortholog of human pre-mRNA cleavage complex 2 protein Pcf11 (Pcf11). The PCFS4 regulates alternative polyadenylation pre-mRNA of *FCA*, encoding an RNA-binding protein involved in flowering control that promotes flowering in the autonomous pathway (102).

1.2 Alternative polyadenylation in plants and animals

1.2.1 Definition and Classification of alternative polyadenylation

In eukaryotes, a large portion of genes have different mRNA isoforms formed by cleavage/polyadenylation at distinct sites, which is now known as alternative polyadenylation (103). Alternative polyadenylation can be defined as the following four types. In type I alternative polyadenylation, more than one polyadenylation site is present in the 3' untranslated region and with the same terminal exon (104). In this type of alternative polyadenylation, different RNA isoforms are produced with no effect on the encoded protein (104). However, it is thought that mRNA stability, translation ability, and other post-transcriptional events may be changed with this type of alternative polyadenylation (104). In type II alternative polyadenylation, at least one polyadenylation site is present in an upstream intron (104). The type II alternative polyadenylation may or may not produce different protein products depending on the stability of those transcripts generated by alternative polyadenylation in the intron (104). In type III alternative polyadenylation, at least one polyadenylation site is present in an upstream exon which comprises the CDS (104). This type of alternative polyadenylation may or may not have an in frame stop codon (104). In type IV alternative polyadenylation, at least one of the polyadenylation sites is present in the 5' untranslated region.

1.2.2 APA is an evolutionarily conserved mechanism for regulating gene expression in eukaryotes

APA was reported first in 1980 as the mRNAs of the Immunoglobulin M (IgM) gene encode different proteins (105-107) while the mRNAs of dihydrofolate reductase (DHFR) differ in their 3' UTR (108). In 1997, it was reported that the mRNAs of approximately 80 genes have APA (109). Genome-wide APA analyses were first conducted through bioinformatics analyses of the expressed sequence Tag (EST) data (110, 111). In 2008 and 2009, several studies were performed to detect the APA phenomena during T-cell activation, neuronal activity, and development, based on microarray techniques (112, 113). The use of high throughput sequencing based techniques led to studies of alternative polyadenylation genome-wide at an accelerated pace for all kinds of eukaryotic organisms (103). The RNA polyadenylation profiles for yeast (114), nematodes (*Caenorhabditis elegans*) (115-117), Drosophila (118), Arabidopsis (119, 120) and mammals (114, 121-125) have been characterized. These studies have concluded that approximately: 72.1% of yeast (114); 70% to 75% of Arabidopsis (119, 120); 40% of nematodes (115, 116); 54.3% of Drosophila (118); 55% of Zebrafish (126) and 70% mammals (124) genes undergo alternative polyadenylation.

1.2.3 APA in animals

In animals, alternative polyadenylation events occur in different development stages, tissues or cell types (103, 104). It is thought that the 3' UTR length influences the stability or translation ability of the transcripts (104). In animals, miRNA binding sites are normally located in the 3' UTR and miRNA can bind to those sites and regulate the mRNA expression through inhibiting mRNA translation or cleavage and degradation of mRNAs (104). It has been reported that long 3' UTRs contain more miRNA binding sites and AU-rich content (104). In addition, it has been reported that alternative polyadenylation leads to 3' UTR shortening during development of spermatocytes (127), proliferating T cells (112) and some cancer cell lines (3). On the other hand, it has also been reported that alternative polyadenylation leads to 3' UTR lengthening in ovulated oocytes and zygotes (128), developing mouse embryos (129) and neurological tissues (130). It is thought that the length of the 3' UTRs is generally shorter in rapidly proliferating cell lines than in primary animal tissues, indicating that the use of alternative polyadenylation to produce short 3' UTRs may correlate positively with proliferation (104). It has also been reported that alternative polyadenylation correlates positively with cancer cell formation through the alteration of miRNA mediated repression (3). The oncogene *HMGA2* can be activated through the loss of its miRNA target sites (131, 132). In addition, in mantle cell lymphomas, the shorter length 3' UTR in *Cyclin D1* mRNA leads to an increase in mRNA stability (133). After evaluating 23 genes, it was also reported that most of them, when expressed in cancer cell lines, are the shorter isoforms compared with those in normal tissues due to using proximal poly(A)

signals (3). Further studies revealed that one of the oncogenic transformation mechanisms is the loss of miRNA targeting sites through alternative polyadenylation in the mRNA oncogenes in order to avoid the miRNA mediated repression through alternative polyadenylation (3).

1.2.4 APA in plants

Alternative polyadenylation has been reported in plants. Multiple polyadenylation sites have been identified for genes encoding AGAMOUS and rbohA in Arabidopsis (134, 135). One typical example of how APA regulates the development of the plant is the APA regulation of flowering time control. FCA and FPA are RNA binding proteins and accelerate flowering time by suppressing the expression of *FLOWERING LOCUS C* (*FLC*) (136, 137). FLC represses flowering time by mediating the signals from the autonomous, vernalization and ambient temperature pathways (2). Basically, the expression level of *FLC* is regulated by APAs of sense transcripts of *FCA* and *FPA* as well as natural antisense transcripts of *FLC* (2). FCA and FPA play important roles in the APA of the *FLC* natural antisense transcript, *COOLAIR* (2). An AtCPSF30-deficient mutant has a different polyadenylation profile genome-wide compared with wild-type plants (53). The AtCPSF30 deficient mutants are more tolerant to oxidative stresses than wild-type plants (24) indicating that alternative polyadenylation may be relevant to oxidative stress. In addition, one study revealed that 10% of APA transcripts are differentially distributed among different developmental stages and tissues under environmental stress (138).

1.3 Stored RNA in plants and animals

1.3.1 Stored RNA in animals

In the 1960s, the presence of a temporarily inactive, stored mRNA in unfertilized eggs of animals was reported (139-143). Since then, the regulation of polyadenylation and translation of such stored mRNAs have been well characterized in the oocytes and early embryos of other animals (144-155). These maternal mRNAs have a short poly(A) tail of 20 to 40 nucleotides and are therefore translationally repressed (masked) and stored in the growing oocyte. During oocyte maturation or after fertilization, these repressed mRNAs are thought to regain a long poly(A) tail 80-250 residues in length and thus become translationally active. Those proteins are thought to be involved in meiotic maturation of the oocyte and to regulate the maternal to zygotic transition (156).

In animals, spermatogenesis is the process of male primordial germ cell differentiation through diploid cell proliferation and meiosis to form haploid cells which then developed into spermatozoa (157). Spermatogenesis yields the spermatozoa that consist of highly differentiated, transcriptionally inactive specialized cells retaining a minimal cytoplasm and a compact nucleus (158). As RNA synthesis ceases during mid-spermatogenesis, post transcriptional mRNA processing including polyadenylation and translation controls gene expression (157). Specific transcripts encoding protamines and transition nuclear proteins have been found stored as translationally inactive mRNAs (159-169). Those spermatozoal RNAs are RNAs synthesized prior to transcriptional arrest and thus

stored in a stable form in preparation for translation during the late stages of spermatogenesis. The spermatozoal RNAs are used as markers of male fertility status (158). The spermatozoal RNAs can be delivered to the oocyte at fertilization and may play an important role in early embryonic development (170). The paternal stored RNAs may also function in packaging the paternal genome (158).

1.3.1.1 Polyadenylation of stored RNA in animals

1.3.1.1.1 Polyadenylation of stored RNA in oogenesis and spermatogenesis

In animal oogenesis, the maternal stored mRNAs are regulated by controlling their poly(A) tail length; shorter poly(A) tails lead to translational inactivation while cytoplasmic polyadenylation reactivates these maternal, stored mRNA by poly(A) tail reacquisition (157). Cytoplasmic polyadenylation was discovered in the 1980s (156, 171). It is known that cytoplasmic polyadenylation is involved in oocyte maturation, mitotic cell cycle progression, cellular senescence and synaptic plasticity (171). The best example of its function is the regulation of translation by cytoplasmic polyadenylation elements (CPEs) in the 3' untranslated region of stored mRNAs which function in *Xenopus* oocyte maturation (156). During oocyte maturation in *Xenopus* and mouse, quiescent prophase I arrested oocytes transition to metaphase II concomitantly with translational activation of stored RNAs encoding key cell cycle regulators (such as c-Mos kinase and mitotic cyclins) through cytoplasmic polyadenylation of stored transcripts (156, 171-173).

1.3.1.1.2 Cytoplasmic polyadenylation cis-elements and RNA binding proteins

For stored mRNAs to regain their poly(A) tail through cytoplasmic polyadenylation, they must be recognized by the cytoplasmic polyadenylation machinery (156). In oocytes and early embryos of *Xenopus*, this machinery recognizes several cis-elements in the 3' UTR of mRNA (156). In *Xenopus*, at least four such cytoplasmic polyadenylation elements have been found (156). These elements include a C-rich element (127, 174), the U-rich embryonic element (UREE), polyadenylation response element (PRE)(154) and cytoplasmic polyadenylation elements (CPE)(175). The putative cytoplasmic polyadenylation factor poly(rC) binding protein 2 (PCBP2) binds to the C-rich element (174). The Elav (embryonic lethal, abnormal vision) related protein A, the ortholog of HuR (human antigen R) binds to the U-rich embryonic element (176, 177). The RNA binding protein Musashi binds to some polyadenylation response elements (PRE) (178).

The CPE is the best characterized cytoplasmic polyadenylation element by far as it is required for cytoplasmic polyadenylation of several groups of mRNAs during oocyte maturation and in the embryonic cell cycle (156, 179, 180). The CPE canonical sequence is U_5AU, but not all CPEs adhere to this motif and variations such as U_4AU (c-mos), $U_{4\text{-}5}A_2U$ (cyclin B1) and U_5A (c-Mos in mature *Xenopus* oocytes) are known to exist (153, 181-184). The cytoplasmic polyadenylation element binding protein (CPEB), an RNA binding protein, binds to CPE (156). CPEB consists of two RNA recognition motifs and a zinc finger region and all of these are required for the recognition of the CPE (156). The CPEB is required for cytoplasmic polyadenylation

as depletion of CPEB from an egg extract leads to the inhibition of polyadenylation, and injection of a CPEB antibody blocks polyadenylation in oocytes and embryos (156).

1.3.1.1.3 Symplekin

Symplekin is thought of as a scaffold protein involved in 3' end RNA processing and is found in nuclear complexes (156). Symplekin or its homologues are required for cleavage and polyadenylation in plants, yeast and vertebrates (43, 156, 185). In the *Xenopus* oocyte, symplekin is also found in the cytoplasm and associated with CPSF100 (156, 186). The symplekin antibody also precipitates CPEB from oocytes as well as inhibits cytoplasmic polyadenylation in oocytes suggesting that symplekin is involved in cytoplasmic polyadenylation (156, 187). Cleavage and polyadenylation specificity factor 100 (CPSF100) and 30 (CPSF30) are present in the *Xenopus* oocyte cytoplasm and reside in a cytoplasmic CPSF-related complex while CPSF73 is absent from the cytoplasmic CPSF complex (188).

1.3.1.1.4 The cytoplasmic poly(A) polymerase

Germline Development Defective-2 (Gld-2), a *Caenorhabditis elegans* germline determinant, was cloned in 2002 and encodes a cytoplasmic poly(A) polymerase (156, 189). The Gld-2 protein is a member of the DNA polymerase β nucleotidyl transferase family but it lacks the RNA binding domain (156, 190). The *Xenopus* Gld-2 protein (xGld-2) coimmunoprecipitates with symplekin in extracts from both mature and immature oocytes (156, 187). Also, the tagged xGld-2 binds to CPSF160

and CPEB in rabbit reticulocyte lysates (156, 187). Overexpression of *Gld-2* leads to an increase in CPE-dependent polyadenylation (187). This evidence indicates that xGld-2 is a cytoplasmic poly(A) polymerase that functions with the cytoplasmic CPSF complex in the *Xenopus* oocyte (156). Specifically, symplekin and Gld-2 seem to act as scaffold and anchored proteins, respectively, to associate with the cytoplasmic polyadenylation machinery by interacting with both CPEB and CPSF (157).

In mice, the testis-specific poly(A) polymerase (TPAP) has also been identified as a cytoplasmic poly(A) polymerase (157). TPAP is only expressed in the testis and its mRNA is most abundant around spermatids (157). TPAP has a high degree of identity (86%) with canonical PAPs and also has elements that are required for polyadenylation activity (157). No TPAP gene is found in the lower vertebrates including *Xenopus* (157). TPAP-deficient mice show spermatogenic arrest and an increase in apoptotic cells (157).

1.3.2 Stored RNAs present in plants

1.3.2.1 The history of stored RNA studies in plants

In 1965, it was reported that long-lived stored mRNAs exist in embryos of cotton based on a combination of transcriptional inhibitor and radioactive labeling experiment (191). Despite early skepticism (192) the presence of stored mRNA in the desiccated cells of plant seeds is now a firmly established fact (193, 194) and has been under study for almost 50 years (193, 195-199). This section will examine the

conclusions reached by a number of workers that led to the consensus that a stored transcriptome exists in mature, dehydrated seeds.

Studies comparing the protein synthetic capacity of subcellular fractions of peanut (*Arachis hypogaea*) or wheat (*Triticum* spp.) embryos suggested that translatable mRNA was limiting in the dry embryo and that its activation or formation during germination permitted protein synthesis (200). Subsequent studies indicated that stored mRNA and monoribosomes capable of assembling on this substrate to initiate protein synthesis were held separate in the cells of dry seed (201). Dure and Waters (1965) utilized Actinomycin D, a polypeptide antibiotic isolated from soil bacteria of the genus *Streptomyces* (202), to prevent *de novo* mRNA synthesis when working with cotton (*Gossypium* spp.) embryos. Actinomycin D has the ability to inhibit transcription through binding DNA at the transcription initiation complex to prevent elongation of the RNA chain by RNA polymerase (203). After the partial inhibition of RNA synthesis by Actinomycin D application, C^{14}-labeled leucine was still incorporated into soluble protein and there was no loss of polyribosomes (indicating persistent translation) during the first 16 hours of germination in cotton embryos. This suggested that the protein synthesis observed in this experiment comes from the messenger RNA pre-existing in the mature seed (191). In 1968, it was suggested that masked mRNA also exists in the dry wheat embryo and is activated upon germination, supporting early protein biosynthesis (204).

The use of inhibitors of transcription or translation to investigate whether mRNA, manufactured during the latter stages of seed development, persists during dehydration to be used for translation upon imbibition has been extensive (Appendix 1.2). However, while the mode of action of inhibitors of translation are distinct from that of transcription, the inhibitors of transcription can be divided into those that truly inhibit transcription and those that inhibit both transcription and subsequent post-transcriptional maturation of the mRNA, viz. the production of a poly(A) tail. This discrimination has been used to produce Appendix 1.2 in which the studies, and their conclusions concerning the necessity of transcription for the completion of germination, have been outlined. For example, cordycepin (3'-deoxyadenosine), an analogue of adenosine lacking the 3'-OH moiety on ribose (205), has been used in a number of studies (Appendix 1.2). Incorporation of 3' deoxyadenosine into RNA terminates subsequent elongation of the RNA, whether during transcription or post-transcriptional addition of the poly(A) tail (205, 206). This distinction is important because it has bearing on an important subtlety regarding the necessity of transcription for the completion of germination versus the necessity of re-adenylation of stored mRNA for the completion of germination (207). Certainly, the requirement for translation to permit the completion of germination has been confirmed on a host of occasions, in a variety of species (Appendix 1.2). The requirement for transcription versus the requirement for re-adenylation of stored mRNA is less clear (Appendix 1.2).

In 1976, using the transcriptional/polyadenylation inhibitor cordycepin or the transcriptional inhibitor actinomycin D, it was determined that cordycepin but not actinomycin D could inhibit seed germination of intact lettuce (*Lactuca sativa*) (208). Approximately 60% of ^{3}H-adenosine incorporation into mRNA was inhibited by cordycepin during the 12 hr incubation but actinomycin D had little effect on incorporation of tritiated adenosine into mRNA (208). The authors suggested that RNA synthesis might be essential for seed germination (208). In 1985, in vitro wheat germ translation assay results showed that translation products of polyadenylated RNA from dry embryonic axes of *Vigna unguiculata* are the same as those from axes after 2 hr imbibition but different from that of axes after 4 to 24 hr imbibition (209). However, those translation products remained unchanged during 4 to 24 hr imbibition (209). The investigators suggested that the stored poly(A)+ mRNA from dry embryonic axes directs the protein synthesis required for the onset of seed germination 0 to 4 hrs imbibition but are subsequently decimated, and that continued translation relies on *de novo* synthesized mRNA (209). In 1990, it was found that a stored mRNA in cotyledons of *Vigna unguiculata* seeds showed induced mRNA synthesis before germination in both mature and immature seeds (210). In 1991, poly(A)+ RNA was found in the cytoplasm of both unfertilized and fertilized eggs and early embryos in a fern *Marsilea vestita* based on in situ hybridization experiments using [^{3}H]polyuridylic acid as a probe (211). The authors of this study also suggested that poly(A)+ RNA is present in the mature sperm of this fern (211). In 1995, experiments were performed to study the synthesis of ribosomal proteins in maize (*Zea mays*) axes at the onset of germination and the origin of these

transcripts. Ribosomal proteins were observed in the presence of transcription inhibitors, suggesting the presence of stored ribosomal protein transcripts in the embryonic axes (212). In addition, immunoprecipitation of ribosomal proteins from in vitro translation products using transcript from embryonic axes revealed the presence of mRNAs encoding ribosomal proteins (212). Interestingly, only two of the studied mRNAs (encoding ribosomal proteins S4 and S6) seem to be stored as mature mRNA (212).

In 2004, a microarray analysis was performed to study the role of stored- and newly-synthesized mRNAs in seed germination of Arabidopsis using *transparent testa* mutants, the seed coat of which is highly permeable to the transcription inhibitor alpha-amanitin (213). The authors found that seeds can complete germination (radicle protrusion) in the presence of the inhibitor, while seedling growth was blocked in alpha-amanitin treated samples (213). However, seed germination was abolished by cycloheximide (213), a translational inhibitor which binds to the ribosome and inhibits eEF2- mediated translocation (214). They suggested that stored mRNAs are involved in seed germination while *de novo* synthesized mRNA can hasten, but is not necessary for, the completion of germination (213). In 2005, a microarray experiment was performed to identify stored mRNA in *Arabidopsis* dry seed (198). More than 12,000 stored mRNA species were detected (198). In 2011, it was reported that a significant increase in the *de novo* synthesis of ribosomal proteins between the 0–2 and 22–24 h germination

periods was found in maize embryonic axes, as indicated by the amount of [^{35}S]methionine incorporated (215).

1.3.2.2 Stored, unadenylated RNAs may be present in plants

It has been reported that stored mRNAs are polyadenylated during the onset of germination based on the following three experiments (207). First, actinomycin D can inhibit the total amount of $^{32}PO_4$ incorporated into poly(A)-containing mRNA by 62%. However, when partitioned, this inhibition was comprised of 70% for mRNA but only 30% into poly(A), indicating that actinomycin D inhibits the *de novo* RNA synthesis more than polyadenylation (207). Second, the poly(A) portion of mRNA-poly(A) has far more $^{32}PO_4$ and [^{3}H]adenosine incorporation and this trend is enhanced when cotyledons are incubated in actinomycin D (207) indicating that polyadenylation is a very important step in seed germination. The absolute amount of poly(A) mRNA accumulated in the early onset of germination is more than would be expected from 70% inhibition of mRNA production resulting from actinomycin D treatment (207). Therefore, they suggested that over 50% of the total mass of mRNA stored in dry seed is re-polyadenylated during the onset of germination (207).

In summary, there are many different papers that have worked with stored mRNAs in dry seed, to the point that the existence of stored mRNA is now taken for granted. However, there are only two publications that specifically indicate that stored, unadenylated mRNAs (stored mRNAs which have a short- (or no-) poly(A) tail) potentially exist in dry seed. These two publications are those by Walbot et al.

(1974) (193) and by Harris and Dure (1978) (207). It is necessary to revisit these studies as seed germination is abolished by the polyadenylation inhibitor cordycepin (207) and by translational inhibitors, indicating a possibility of re-polyadenylation and subsequent translation of the stored, unadenylated mRNAs during seed germination. Appendix 1.2 is built around the hypothesis, stated in Walbot et al. 1974 (193), that those inhibitors of transcription and polyadenylation usually have a greater repressive effect on seed germination than do those inhibitors of transcription that do not interfere with polyadenylation. This in turn suggests not only that deadenylated messenger RNA exists in the dry seed, but also that its re-adenylation and subsequent translation is an important requirement for the completion of seed germination. This is elegantly explained in Walbot et al. (193).

1.4 Summary

Alternative polyadenylation occurs in plants and may function in gene regulation. Therefore, it is of interest to identify the genes that are subject to alternative polyadenylation during seed germination and other developmental stages. To identify those genes, polyadenylation site choice has been studied genome-wide during germination. These studies revealed that alternative polyadenylation might act as a mechanism to down-regulate the gene expression in certain developmental stages such as seed germination stages but release such regulation in other developmental stages such as leaf.

Stored, unadenylated mRNAs may be present in dry seed. However, during seed germination, the stored, unadenylated mRNAs may regain their long poly(A) tail with the help of cytoplasmic polyadenylation. These polyadenylated mRNAs may be translated into proteins and these proteins may further function in seed germination. To identify these stored, unadenylated mRNA, gene expression was studied in dry seed and in seed imbibed in with or without a transcriptional inhibitor. The stored, unadenylated mRNA candidates have been identified and genes encoded ribosomal proteins were overrepresented in this gene list, suggesting that translation might be the first thing to be activated early in seed germination.

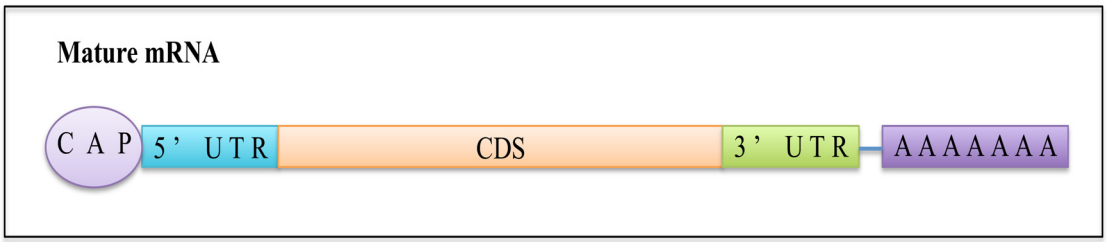


Figure 1.1 A typical structure of mature mRNA in Plants
The cartoon depicts 5 separate regions comprising a mature mRNA. CAP: A 7-methylguanosine attached from the 5' carbon by a triphosphate bridge to the 5' carbon of the terminal nucleotide of the mRNA. The cap interacts with the cap-binding complex. 5' UTR: 5' end untranslated region. CDS: coding region. 3' UTR: 3' untranslated region. A: adenosine. The poly(A) tail interacts with the poly(A) binding proteins (PABPs).

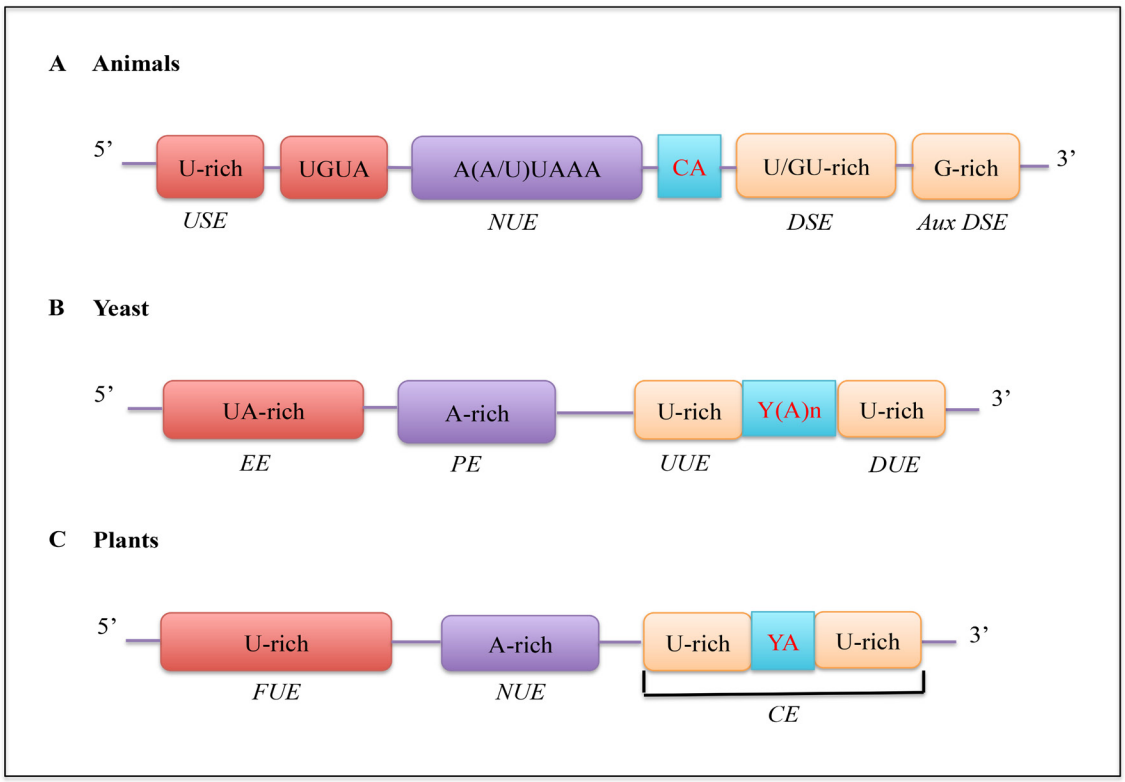


Figure 1.2 Polyadenylation signals of nuclear mRNA in animals, yeast and plants

Figure modified from Millevoi and Vagner, 2010 (9). The cartoons depict regions and motifs known to be of importance in the recognition and cleavage of the pre-mRNA which dictates the site of polyadenylation, A) in animals. USE: upstream sequence element. DSE: downstream sequence element. Aux DSE: auxiliary downstream sequence element. B) In Yeast which has the following sequence designations, EE: efficiency element; PE: positioning element; UUE: upstream U-rich element of the cleavage site; DUE: downstream U-rich element of cleavage site. $Y(A)_n$: a U or C followed by multiple adenosines. C) In plants where the motifs/regions are referred to as, FUE: far upstream element; NUE: near upstream element; CE: cleavage element; YA: predominant dinucleotide located at the poly (A) or cleavage site and where Y=U or C. The 'A' is the last nucleotide before poly(A) addition.

Chapter Two: A protocol for the genome-wide characterizations of polyadenylation sites

2.1 Introduction

An important aspect of the annotation of eukaryotic genomes is the accurate description of 3' UTRs and their associated poly(A) sites. Plant poly(A) site databases have been assembled from the analysis of expressed sequence tags (ESTs) (216, 217). However, these datasets are not built to specifically identify poly(A) sites and they fail to capture a majority of poly(A) sites. To better study poly(A) site position and distribution on a genome-wide basis, a protocol designed to specifically query the mRNA-poly(A) junction [called polyadenylated tags, or PATs hereafter] was developed and used to study poly(A) site choice in Arabidopsis. Combined with bioinformatics tools, the tags generated from this strategy were used to study how polyadenylation controls gene expression at different developmental stages in Arabidopsis. Gene expression analyses showed that this protocol is highly reproducible.

2.2 Results

2.2.1 Polyadenylated Tag (PAT) preparation

Figure 2.1 presents an overview of the 3' end polyadenylated sequencing tag preparation procedure developed in the course of this research. Basically, total RNA was isolated (218) and further purified using DNase-treatment to eliminate genomic DNA contamination. RNA was fragmented by a zinc ion-catalyzed (zinc chloride) nucleic acid alkylation and fragmentation mechanism (219) in order to generate 3'

end sequencing tags of short length (200bp to 400bp) that are suitable for the Illumina GAIIx sequencing platform. To selectively retrieve the polyadenylated transcripts, oligo d(T) beads (New England Biolabs, Ipswich, MA, USA) were used for affinity purification; this step reduces or eliminates ribosomal RNAs and tRNAs that together represent approximately 95% of total RNA (220). To add the Illumina sequencing adapter to fragmented and polyadenylated RNA, cDNA was produced using an anchored oligo d(T) primer with an Illumina-compatible sequence at its 5' end.

To attach an Illumina-compatible sequence to the other end of the cDNA, an oligonucleotide primed, template-switching, mechanism was used. This employs the propensity of Moloney Murine Leukemia Virus-derived reverse transcriptase (Clontech) to add short oligo-dC tracts to the 3' end of cDNA (221). Accordingly, an anchored oligo-dG primer with a suitable Illumina-compatible sequence was included in the reverse transcription reactions, such that template-switching due to binding of this primer to the oligo-dC tracts added the 5' end Illumina sequencing adapter to the first strand cDNA.

For some experiments (Chapter 3), these cDNAs were amplified for 10-15 cycles using Phire Hot Start II DNA polymerase (Thermal Scientific). The PCR products were separated on agarose gels, purified, and submitted for sequencing (referred to as the old method). In other instances (Chapter 4), these cDNAs were further size selected using Agencourt Ampure XP beads (Beckman Coulter, Brea, CA, USA) to

remove artifacts of reverse transcription (referred to as the new method), especially a 120bp artifact generated from dimerization of both 5' end and 3' end Illumina sequencing adapters. The cDNAs were then amplified using Phire Hot Start II DNA polymerase (Thermal Scientific) with a very low PCR cycle number in order to avoid a PCR-generated skewing of transcript abundance in so far as possible. An S1 nuclease reaction was performed to purge the amplicons of heteroduplexes, an important possible PCR artifact (222). The S1 nuclease-treated PCR products were separated on agarose gels run at low voltage in order to thoroughly separate the 120bp primer artifact from the tags. The tags with a size range from 250bp to 400bp (Figure 2.2) were excised from the gel and purified. These tags were further amplified in a low cycle PCR reaction and a second round of agarose gel size selection performed. The resulting tags were sent for sequencing. The Illumina GAII high-throughput sequencing platform at Ohio State University was used to perform the sequencing reactions. The sequencing output was processed using CLC genomic Workbench software as described in the following sections.

To better evaluate the quality of the PATs obtained using these techniques, this chapter focused on wild-type leaf tags. Eight wild-type leaf tags were divided into two groups. The first six wild-type leaf replicates (wt leaf R1 to R6) were generated using the old method (Table 2.1). The last two wild-type leaf replicates (wt leaf R7 and R8) were generated using the new method (Table 2.1). Twelve PAT datasets for wild-type seed germination experiment (discussed in Chapter 3) were generated by the old method (Table 2.1). Twenty PAT datasets, generated by the new method,

were analyzed to identify stored mRNAs (discussed in Chapter4) (Table 2.1). Overall, forty PAT datasets were generated through either the old method (18 out of 40) or the new method (22 out of 40).

2.2.2 The PATs map to the reference database

To evaluate the quality of the PATs, they were mapped to the Arabidopsis genome using the CLC Genomics Workbench software package (see Methods). More than 34% of PATs mapped to the Arabidopsis TAIR10 genome database (Table 2.1). However, the leaf replicates using the old method (wt leaf R1 to R6) had a relatively lower mapping percentage (from 34.71% to 54.24%) to the genome database compared with that of the leaf replicates prepared using the new method (wt leaf R7 and R8 from 83.94% to 84.09%). These results suggest that purification with AMPure beads and S1 nuclease reactions might reduce contamination by artifacts. Similar to the mapping results for the genome, 73.24% and 72.03% of PATs mapped to 3' UTR in the leaf R7 and R8, respectively from the new method while the leaf R1 to 6 have a mapping percentage ranging from 16.70% to 47.20% (Table 2.1). In addition to artifact contamination, it is possible that ribosomal RNAs (rRNAs) may occur in these lower percentage 3' UTR mapping replicates due to reduced efficiency of oligo d(T) selection. Therefore, all of the tags were mapped to the ribosomal RNA reference database in Arabidopsis. The mapping percentages to ribosomal RNAs in those lower percentage 3' UTR mapping datasets were high (Table 2.1) suggesting that ribosomal RNA contamination was present in those replicates. Overall, the new method may

decrease the artifact contamination. However, the ribosomal RNA contamination exists in replicates generated from both samples.

2.2.3 The PATs map to the poly(A) sites

To further evaluate the quality of the PATs, they were mapped to the Arabidopsis TAIR10 genome reference database. Browser tracks were built to illustrate the locations of PATs within their respective genes for leaf replicates made by both the old method (wt leaf R1 and R2) and the new method (wt leaf R7 and R8) as shown in Figure 2.3 and Figure 2.4. Figure 2.3 depicts the PATs that mapped to the genomic region of 340000 to 370000 nts on Chromosome 3. If a gene is transcribed from left to right in the diagram, the mapped PATs should be colored red. If a gene is transcribed from right to left in the diagram, the mapped PATs should be green in color. It is thought that most of the poly(A) sites are located in the 3' UTR of Arabidopsis genes (119) and the mapping results coincide with this assumption as most of the PATs from replicates generated by both the old and new method in figure 2.3 mapped to the 3' end of the various Arabidopsis genes (Figure 2.3). Closer inspection of the genomic region between 13830000 to 13847000 nts on Chromosome 4 reveals a similar pattern, with most PATs mapped to the 3' end of the Arabidopsis genes in this region (Figure 2.4). Overall, the PAT mapping results concur with the EST database–poly(A) site placement in replicates from both the new and the old method suggesting that these PAT datasets were reproducible.

2.2.4 The PATs are suitable for gene expression analysis

To further evaluate the quality of PATs, RNA-seq analysis was performed. The PAT abundance was used as a measure of expression level. Basically, the PATs were mapped to the Arabidopsis genomic reference database in a strand-specific pattern. The results were then normalized and expression represented in terms of tags per million (TPM). Different samples were then compared using x=y plots , so as to measure and visualize, respectively, the correlation between two replicates. Figure 2.5 shows a typical scatter plot for wild-type leaf replicates 7(x) and 8(y). The Pearson correlation coefficient was 0.95, indicating that expression values are very reproducible.

To illustrate the variation seen in all of the wt leaf PAT samples (in addition to the seed samples that are the subject of chapters 3 and 4), a Principle Component Analysis (PCA) was conducted and presented as a correlation scatter plot of the two principal components (Figure 2.6). Examining cluster 1 only, open circles in light blue (wt leaf from the new method) are clustered with each other as well as the open orange circles (wt leaf from the old method). In addition, the wild-type replicates from both the old (open orange circles) and new methods (light blue open circles) cluster together in one sector suggesting that replicates from both the old and the new method are reproducible. Overall, these results show that replicates made through the tag protocol are reproducible and suitable for gene expression analysis.

2.3 Discussion

2.3.1 Artifacts present in PATs

In some replicates, such as wild-type leaf replicate 2, a high percentage of PATs did not map to the Arabidopsis genome (Table 2.1). Those PATs failing to map to the database may result from a variety of artifacts such as primer dimerization (223). Another potential source of contamination comes from ribosomal RNA. Some samples do have ribosomal RNA contamination (Table 2.1). Both the failure of oligo d(T) selection to eliminate rRNA and/or the non-specific binding of rRNA to oligo d(T) magnetic beads may lead to rRNA contamination and decrease the mapping percentage of PATs to 3' UTRs. It is reported that incubation of RNAs with oligo d(T) magnetic beads at a higher temperature (50 °C, 5 minutes) instead of room temperature may help to remove the non-specific binding of ribosomal RNA (224). Another possibility for rRNA contamination is that, before degradation in the cell, the truncated rRNA can acquire a short poly(A) tract (225).

2.3.2 Poly(A) site determination using PATs is not affected by PCR artifacts or rRNA contamination

Although PCR artifacts or rRNA contamination exists in the PATs, closer inspection of the genome map of the PATs revealed that poly(A) site determination coincided well with the canonical poly(A) sites identified from expressed sequence tag (EST) databases (Figure 2.3 and Figure 2.4). These results reveal that most of the PATs mapped to 3' UTR regions of expressed genes and hence, poly(A) site determination was not affected by PCR artifacts or rRNA contamination. The

robust nature of the data set in defining PATs was probably due to the fact that the artifacts were not mapped to the reference genome and thus had less influence on PAT 3' UTR mapping than anticipated. Although rRNA contamination may occur, the rRNA tags have been excluded by mapping method and thus did not affect the mapping of PATs to other genes.

2.3.3 Sample reproducibility is not affected by an abundance of PCR artifacts or rRNA contamination

The artifacts or rRNA contamination existing in some PAT samples may possibly influence tag quality and downstream polyadenylation analyses. However, such was not the case. The fact is the artifacts will not map to the reference database and the rRNA contamination is excluded by the mapping method employed. Thus, when PAT relative abundance was used to represent gene expression level and the repeatability of this metric was assessed, good agreement between the different replicates was seen (Figure 2.6). This was true for replicate PATs made with or without Ampure bead purification or S1 nuclease digestion (Figure 2.6), and for samples with high quantities of unmapped- or rRNA-contaminated-PATs (Figure 2.6). Overall, tag quality is not affected by an abundance of artifacts or rRNA contamination and the protocol was successful even without the Ampure bead purification or S1 nuclease digestion.

2.4 Methods and material

2.4.1 RNA isolation and clean up

Total RNA was isolated from 0.1g to 0.3g *Arabidopsis thaliana* tissues using methods described in Chang *et al* (218). DNase treatment was performed using RNase-free DNase I (Thermo Scientific Catalog # EN0521) following the manufacture's protocol. Total RNA was purified after DNase treatment using a kit (RNeasy Plant Mini Kit, Qiagen, Valencia, CA, USA).

2.4.2 RNA fragmentation

Between 1-5μg total RNA was diluted to 25μL. A 1/10 volume of RNA fragmentation buffer (100mM $ZnCl_2$ in 100mM Tris-HCl, pH 7.0) was added to the total RNA. The sample was incubated for 5-15 minutes at 70°C. The reaction was terminated by adding 1/10 volume of RNA fragmentation stop buffer (0.5M EDTA, pH 8.0) to the sample. The sample was next incubated at room temperature for 5 min before the fragmented RNA was purified using a kit (Qiagen).

2.4.3 Poly(A) RNA enrichment

The fragmented RNA was heated to 65°C for 5 minutes to disrupt secondary structures, and then placed on ice for later usage. Meanwhile, 15μL of oligo (dT) beads (New England Biolabs, Ipswich, MA, USA) were aliquoted into nuclease-free microcentrifuge tubes. The beads were washed twice with 100μL of Binding Buffer (20mM Tris-HCl pH 7.5, 1.0M LiCl, and 2mM EDTA) and the supernatant removed using a 6-Tube Magnetic Separation Rack (New England Biolabs, Ipswich, MA, USA).

The beads were resuspended in 50µL of Binding Buffer (20mM Tris-HCl pH 7.5, 1.0M LiCl and 2mM EDTA), and added to the fragmented RNA. The beads were mixed gently, and the tubes were allowed to stand at room temperature for 5 minutes before the removal of the supernatant using the Magnetic Separation Rack (New England Biolabs, Ipswich, MA, USA). The beads were washed twice with 100µL of Bead Washing Buffer (10mM Tris-HCl pH 7.5, 0.15M LiCl and 1mM EDTA). The supernatant was removed from the beads, and 10µL of Elution Buffer (10mM Tris-HCl pH 7.5) added to them and heated to 80°C for 2 minutes to elute the adherent mRNA. Approximately 9µL of mRNA was usually recovered.

2.4.4 cDNA synthesis

To the 9µL of fragmented, poly(A)-enriched mRNA, 1µL of a 100µM stock of primers combining oligo d(T) with the right Illumina adapter sequence was added along with 1µL of a 100µM stock of the left Illumina adapter. The solution was mixed, heated to 65°C for 5 min to disrupt the secondary structure, chilled on ice for 2 minutes, and this heating and chilling was repeated once.

Next, reagents were added to the mRNA in the following order: SMARTScribe™ 5X 1st strand buffer (5µL), 10mM dNTPs (2.5µL), 100mM DTT (1µL), and RNase Inhibitor (New England Biolabs; 0.5µL). The mixture was brought to 42°C in a thermocycler with a heated lid (Minicycler PTC-150), and 1µL SMARTScribe™ Reverse Transcriptase was added and the reaction incubated for 2 hr at 42°C. At the end of 2 hr, the mixture was heated to 70°C for 5 min to inactivate the

enzymes. One µL each of RNase H (New England Biolabs, Ipswich, MA, USA) and RNase A/T1 (Thermal Scientific) were added to the tube. The reaction was incubated at 37°C for 1 hr. The cDNA was purified using a kit (QIAquick PCR Purification Kit) following the manufacturer's protocol.

2.4.5 Agencourt Ampure XP beads size selection

Agencourt® Ampure®XP beads (Beckman Coulter, Brea, CA, USA) were added to the cDNA (an 1.8:1 v/v ratio of beads to cDNA), incubated at room temperature for 20 minutes, the beads separated from the suspension using a magnetic rack and the supernatant discarded. The beads were washed twice with fresh 70% ethanol before being air-dried at room temperature for 15 minutes. Water (25µL) was added to the beads, and they were resuspended and incubated for at least 1 min. The beads were removed by magnetic separation. The supernatant was retained for the next step.

2.4.6 PCR Amplification

One µL cDNA (after AMPure beads) was used to perform a low cycle PCR (15 cycles) using the Phire PCR system (Thermal Scientific). Reactions (25µL total volume) contained the following: 5X Phire Hot Start II polymerase buffer (5µL), 10mM dNTPs (0.5µL), 10µM Primer 1 (PE-PCR1; 1µL), 10µM Primer 2 (PE-PCR2; 1µL), cDNA (1µL), Phire Hot Start II DNA polymerase (0.5µL), and nuclease-free water to 25µL. The PCR reaction was performed using the program: one cycle of enzyme activation (5 min at 98°C), followed by from 10 to 15 cycles (depending on

the experiment) of denaturation (15 sec at 98°C), annealing (15 sec at 60°C), and extension (30 seconds at 72°C). PCR was conducted for the number of cycles indicated in the Results. After PCR, the amplicons were purified and eluted for the next step (e.g. QIAquick PCR Purification Kit).

2.4.7 S1 nuclease reaction

The PCR product from the prior step was used to perform S1 nuclease reaction (Thermal Scientific). Reactions contained: 5X Reaction Buffer for S1 nuclease (10μL), S1 nuclease (Thermal Scientific; 0.1μL), PCR product (30μL), and nuclease-free water to 50μL. The S1 reaction was incubated at room temperature for 30 minutes before being purified and eluted using a kit (e.g. Qiagen QIAquick PCR Purification Kit column).

2.4.8 Size-selection by Agarose gel electrophoresis

The S1-treated PCR products were loaded onto a 1.5% (w/v) agarose gel and run at a very low voltage (between 25 and 50 V) for 3 hours. The gel containing amplicons within a size range of 250-400bp was excised and the PCR amplicons extracted from the gel using a kit (e.g. QIAquick Gel Extraction Kit). A 1μL aliquot of the purified PCR products was used to perform another round of low-cycle-number PCR amplification as described above and these were again size-selected through agarose gel and recovered from the gel using a kit (Qiagen). The purified PCR amplicons were the final tags delivered to the facility at the Ohio Agriculture

Research and Development Center (OARDC) Ohio State University, Wooster, OH for sequencing on an Illumina GAIIx sequencing platform.

2.4.9 Bioinformatics processing of PATs

PATs were analyzed using the CLC Genomics Workbench (abbreviated "CLC"). The input of the data analysis pipeline was the "raw" DNA sequences in FASTQ format returned by the sequencing center. The FASTQ files were imported into the ".clc" file format with CLC. Each lane of Illumina sequencing had more than one sample and the sequences from one specific sample contained a specific identifier (the barcode identifiers were the Linker (NN) + Barcode (XXXTTTT)). Therefore, each of the barcodes was distinctly recognized and the sequences that contained them were extracted into tissue- and treatment-specific bins during input. The end result was an extracted CLC file with all the sequences containing one specific bar code; in the process, the bar code itself was removed (trimmed) from the 5' end of each tag. To better perform downstream analyses, the oligo-dT and the PE-PCR sequences were also trimmed with the "Trim sequence" tool. In addition, tags were discarded if they were shorter than 20nt. The trimmed tags, in their appropriate bins, were retained for further analysis. For some purposes, the reversed complemented sequences for these tags were generated using the Geneious software package version 5.6 (Biomatters, Auckland, New Zealand).

PATs were mapped to the reference database using the read-mapping tool in CLC. The mapping was performed with the stringency options of "fraction" set at 0.9

("Fraction" refers to the minimum length fraction of a read that must match the reference sequence; setting a value at 0.9 means that at least 90% of the read needs to match the reference) and "similarity" set at 0.7 (a "Similarity" value of 70% dictates that the reads must have at least 70% identity with the reference sequence in order to be included in the final mapping.). After mapping the tags to the whole genome reference database, mapping results were converted to genome browser tracks for easier display.

2.4.10 Expression level analysis of PATs

The RNA-Seq analysis was performed using the tool of the same name in CLC. PATs were mapped to the annotated TAIR10 genome reference database; for this, the 3' UTRs of genome reference genes were extended to the downstream 120 flanking residues as described in (119). The entire tag mapping operation was performed using the stringency criteria of "fraction" at 0.9 and "similarity" at 0.7 with the strand specific alignment and expression based on unique reads. The RNA-seq data was used to measure and analyze expression levels. For this, read counts were normalized based on the total gene reads such that expression levels were represented as tags per million. Multi-group comparisons were performed based on these calculated expression values, and the degrees of correlation between different samples estimated with scatter plots. Multiple-group comparisons were performed by the "Principle Component Analysis" tools in CLC and correlations were calculated and then graphically represented by the scatter plot dots.

Table 2.1 The mapping distribution of all PATs to genome, 3' UTR, and rRNAs using CLC genome workbench

A table of general statistics describes the TAG datasets used in this thesis. The headings of the columns refer to various parameters and the rows contain numbers describing the data set for the sample and replication named on the far left of each row. 3' UTR: 3' untranslated region. rRNA: ribosomal RNAs. Total tag number: total number of tags after trimming the Illumina sequence adaptor each Illumina read represents a tag. Map to genome: number of tags mapped to the Arabidopsis reference database (TAIR website). Percentage of genome tags: percentage of tags mapped to the genome. Map to 3' UTR: number of tags mapped to the Arabidopsis 3' UTR reference database (TAIR website). Percentage of 3' UTR tags: percentage of tags out of the total that mapped to the 3' UTR. Map to rRNAs: the number of tags that mapped to the Arabidopsis rRNA reference database (TAIR website). Percentage of rRNA tags: percentage of tags out of the total that mapped to rRNA. wt: wild-type. R: replicates. W: data from seeds imbibed in water. A: data from seeds imbibed in alpha-amanitin. O: PATs generated by the old method. N: PATs generated by the new method. *tt2*: *transparent testa 2*, a mutant described in Chapter Four.

Sample name	Total tag number	Map to genome	Percentage of genome tags (%)	Map to 3' UTR	Percentage of 3' UTR tags (%)	Map to rRNAs	Percentage of rRNA tags (%)	Methods
wt leaf R1	5401559	2329299	43.12	1147045	21.24	735731	13.61	O
wt leaf R2	5366655	1862945	34.71	896164	16.70	54707	10.21	O
wt leaf R3	1489608	529749	35.56	392422	26.34	8009	0.54	O
wt leaf R4	3625226	1966494	54.24	1711063	47.20	118177	3.26	O
wt leaf R5	5580059	2252808	40.37	1009936	18.10	754526	13.52	O
wt leaf R6	2465522	674606	27.36	502554	20.38	9030	0.37	O
wt leaf R7	3690262	3097426	83.94	2702859	73.24	35194	0.95	N
wt leaf R8	2693378	2264888	84.09	1940003	72.03	29902	1.11	N
wt 0 R1	2714976	1410560	51.95	1205126	44.39	22472	0.83	O
wt 0 R2	712815	505561	70.92	442296	62.05	4346	0.61	O
wt 0 R5	7278302	3711914	51.00	3293535	45.25	126882	1.74	O
wt 12 R1	2666918	1338845	50.20	1026014	38.47	30812	1.16	O
wt 12 R2	5569438	2781831	49.95	2340986	42.03	53481	0.96	O
wt 24 R3	2352105	1279574	54.40	962856	40.94	23011	0.98	O
wt 24 R4	2362190	1317202	55.76	1088498	46.08	21374	0.90	O
wt 36 R2	2170319	1107887	51.05	852005	39.26	31393	1.45	O
wt 36 R4	1783651	821891	46.08	658938	36.94	16849	0.94	O
wt 48 R1	1814431	802305	44.22	674850	37.19	17900	0.99	O
wt 48 R2	1069393	479355	44.82	350156	32.74	10658	1.00	O
wt 48 R4	2003382	825953	41.23	644612	32.18	26245	1.31	O
***tt2* 0 R1**	3843750	2534494	65.94	244015	6.35	1993286	51.86	N
***tt2* 0 R2**	965054	727077	75.34	396330	41.07	186504	19.33	N

Table 2.1 (continued)

Sample name	Total tag number	Map to genome	Percentage of genome tags (%)	Map to 3' UTR	Percentage of 3' UTR tags (%)	Map to rRNAs	Percentage of rRNA tags (%)	Methods
***tt2* 0 R3**	1868146	1175740	62.94	450901	24.14	481251	25.76	N
***tt2* 12w R1**	2020776	1310924	64.87	887354	43.91	235473	11.65	N
***tt2* 12w R2**	2495440	1817372	72.83	1351780	54.17	149849	6.00	N
***tt2* 12w R3**	1480597	962427	65.00	651587	44.01	125835	8.50	N

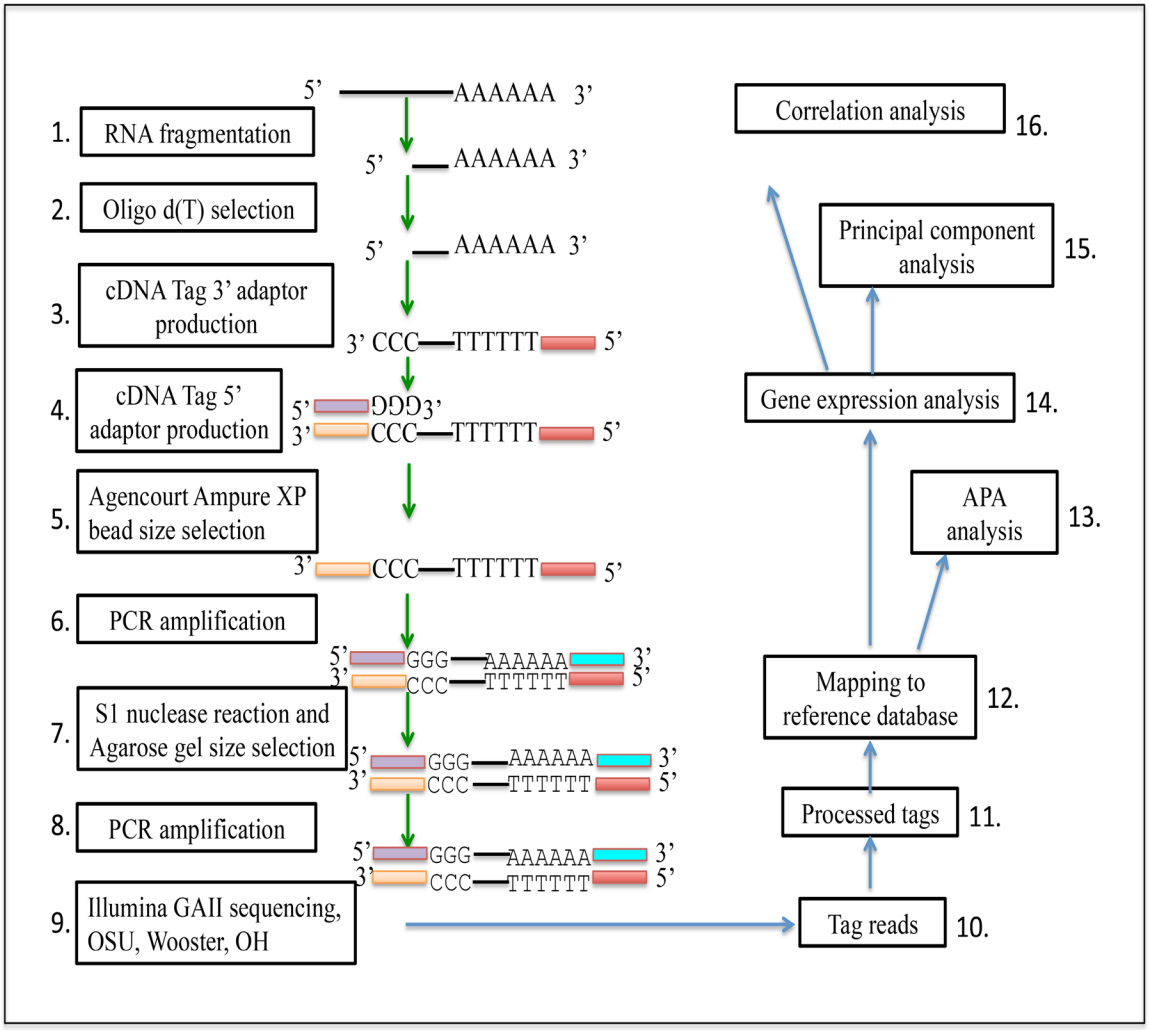


Figure 2.1 Overview of the 3' end polyadenylated sequencing tag preparation protocol

Following RNA isolation, 1) metal ion based cleavage and 2) oligo d(T) selection of polyadenylated mRNA fragments was performed. Illumina sequencing sites were added to the cDNA during reverse transcription by using, 3) a modified oligo d(T) primer and 4) a second, template switching primer. 5) First strand purification was performed using a kit. The purified first strand was used as 6) template for 10-15 rounds PCR amplification, 7) at which time the poly(A) mRNA was degraded, and amplicons of a defined size selected by 7) agarose gel electrophoresis. Upon recovery from the gel, the amplicons underwent 8) and additional, limited cycle amplification, another round of agarose gel fractionation and purification before being sent for 9) sequencing. 10) The tags were first 11) processed, then 12) mapped to the reference genome allowing both 13) alternative polyadenylation analysis and 14) gene expression analysis. These data were then further analyzed using expression analysis tools.

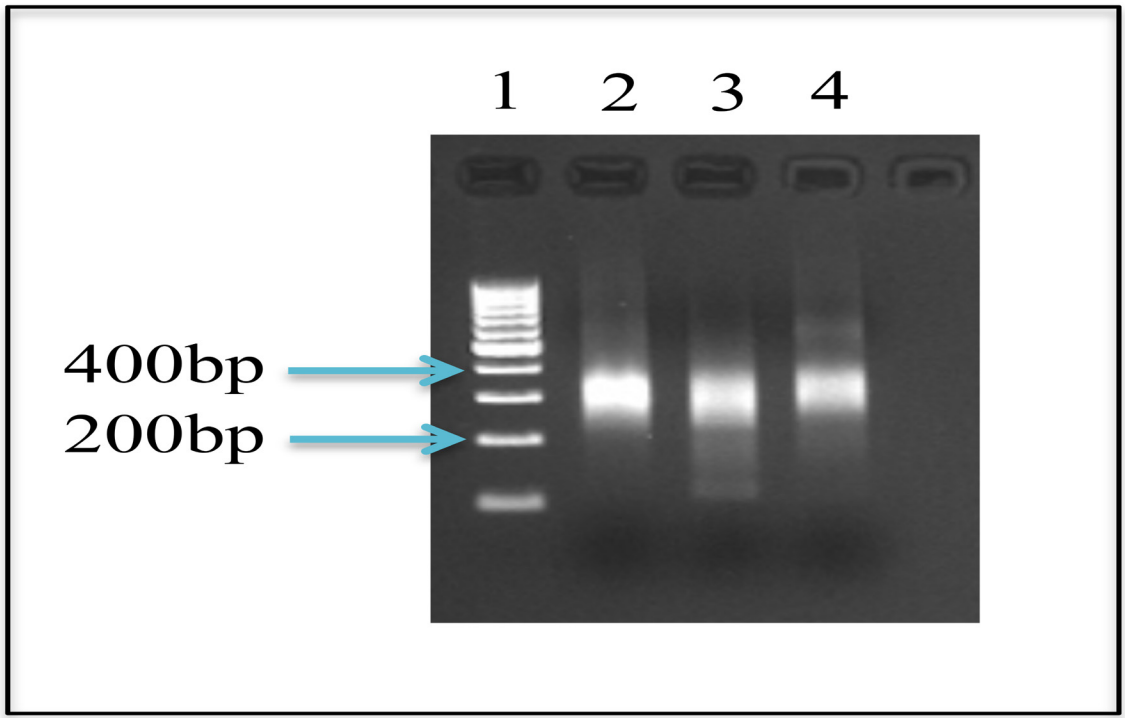


Figure 2.2 Agarose gel electrophoresis for polyadenylated tags
1.5% agarose gel was used to perform the analysis. The electrophoresis was run at a constant voltage of 50 Volts. Lane 1: 1kb DNA ladder; Lane 2: tags made from wild type leaf replicate 7; Lane 3: tags made from *tt2* (*transparent testa 2*) 72hr, water imbibed seeds from replicate 1 (Detailed information about *tt2* and alpha-amanitin will be discussed in Chapter 4); Lane 4: tags made from *tt2* 48hr, alpha-amanitin imbibed seeds in replicate 3. Note the polyadenylated tag abundance in the size range of 200 to 400bp in lane 2 to lane 4.

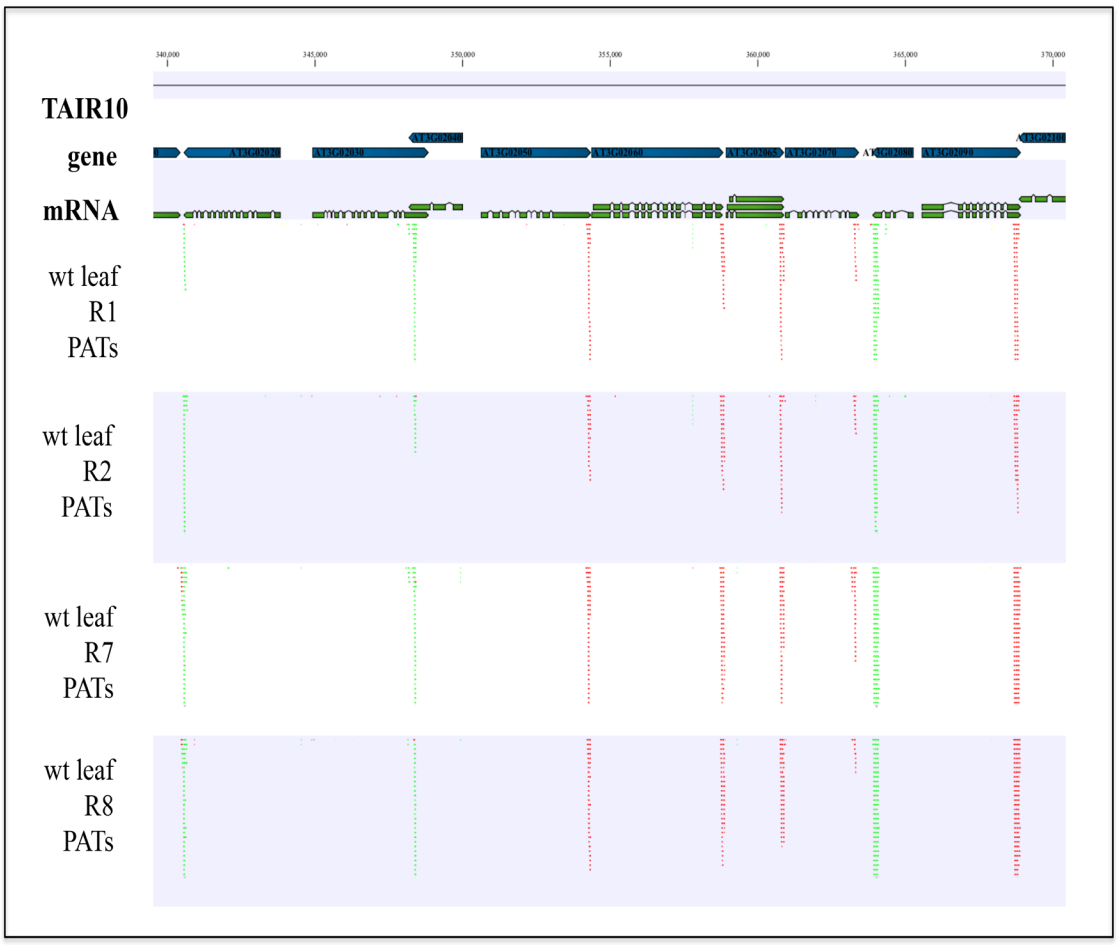


Figure 2.3 Tags mapped to the poly(A) site in the chromosome 3 genomic region

A graphical depiction of the location of the PATs when mapped to the Arabidopsis genome, annotated to show the gene structure as predicted by the expressed sequence tag database. TAIR10: Arabidopsis expressed sequence tag database. Gene: gene in the Arabidopsis reference database (TAIR10). wt leaf: wild-type leaf. R: replicate. PATs: polyadenylated tags. Each red or green line represents a read. The position of red or green columns represents the location of reads. The number of lines represents the expression level. The gray zones beneath the red polyadenylated tags indicate that many more tags than are shown also map to the respective genomic position. If a gene is transcribed from left to right in the diagram, the mapped PATs are colored red. If a gene is transcribed from right to left in the diagram, the mapped PATs are green in color.

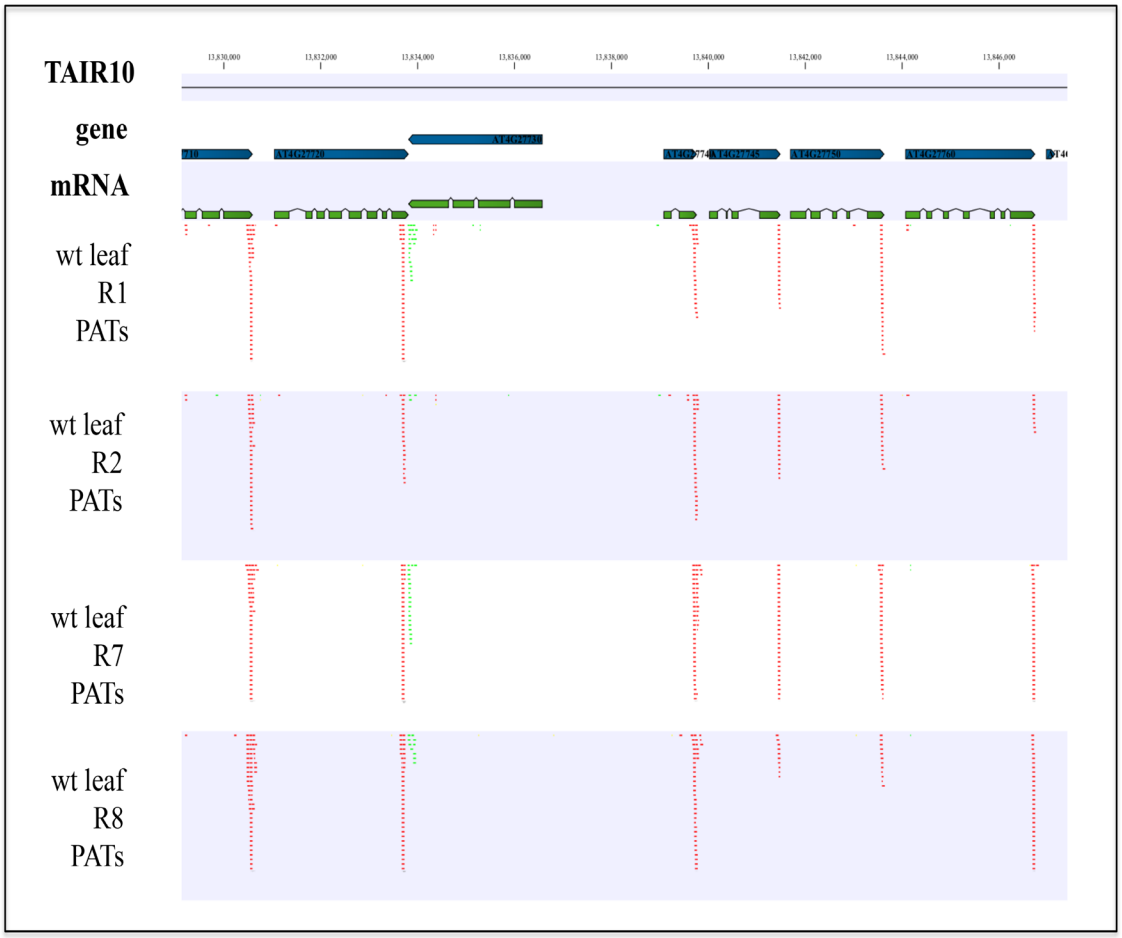


Figure 2.4 Tags mapped to the poly(A) site in the chromosome 4 genomic region

A graphical depiction of the location of the PATs when mapped to the Arabidopsis genome, annotated to show the gene structure as predicted by the expressed sequence tag database. TAIR10: Arabidopsis expressed sequence tag database. Gene: gene in the Arabidopsis reference database (TAIR10). wt leaf: wild-type leaf. R: replicate. PATs: polyadenylated tags. Each red or green line represents a read. The position of red or green columns represents the location of reads. The number of lines represents the expression level. The gray zones beneath the red polyadenylated tags indicate that many more tags than are shown also map to the respective genomic position. If a gene is transcribed from left to right in the diagram, the mapped PATs are colored red. If a gene is transcribed from right to left in the diagram, the mapped PATs are green in color.

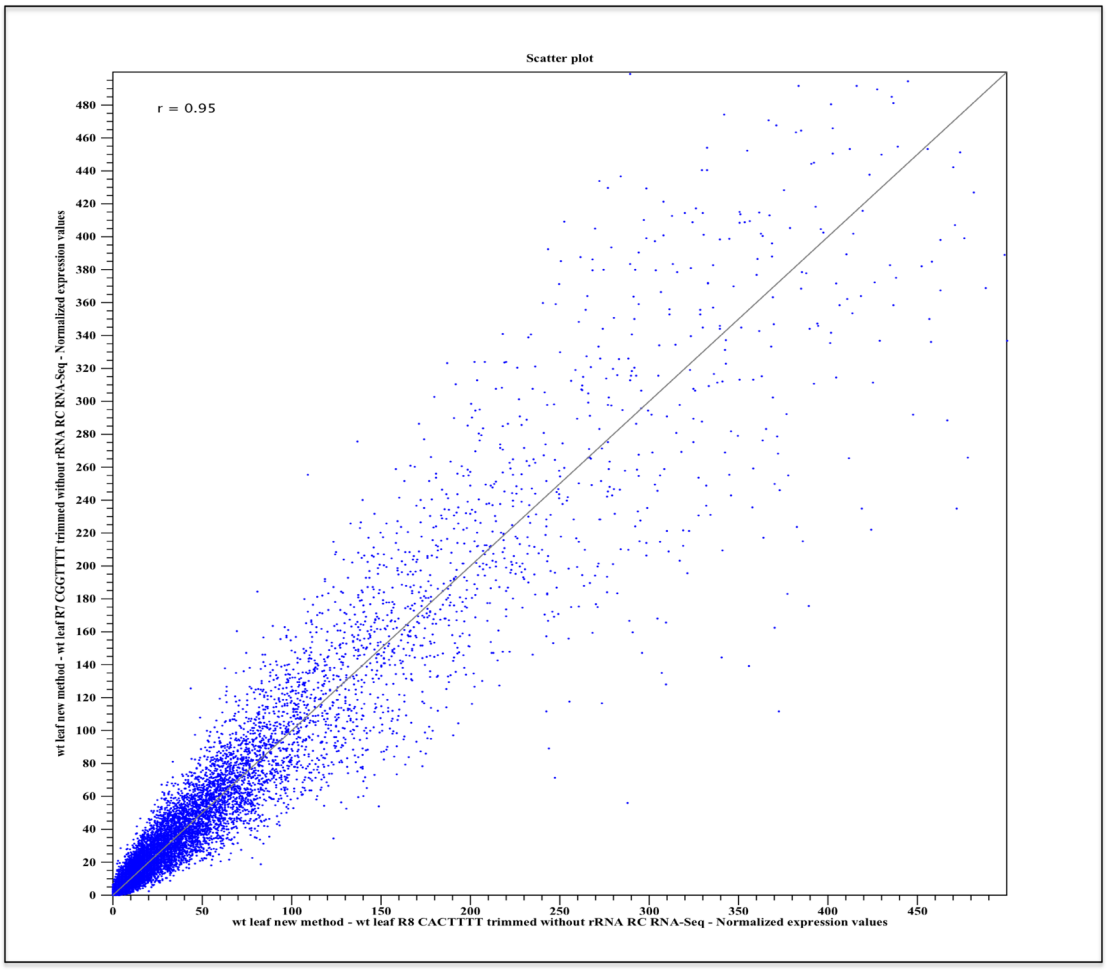


Figure 2.5 The scatter plot for two wild-type leaf biological replicates
wt leaf R7: abscissa. wt leaf R8: ordinate. Each of the blue dots (x, y) represent the normalized tags expression level with x= the normalized expression level in wild-type leaf replicate 7, y= the normalized expression level in wild-type leaf replicate 8. wt: wild-type. Both replicates were prepared using the new method. R: replicate. Note the seven characters "bar code" associated with each replication, which provides information on the genotype, tissue, and treatment from which the RNA was extracted. These were trimmed from the sequence prior to mapping or analysis. Without rRNA: these PATs were without rRNA contamination. RNA-seq: the method used to analyze the gene expression values in CLC. Normalized expression values: the expression values of each gene normalized according to the total number of PATs in each replicate. Pearson correlation coefficient represents the overall variability between replicates. A high Pearson correlation coefficient represents a high reproducibility between replicates.

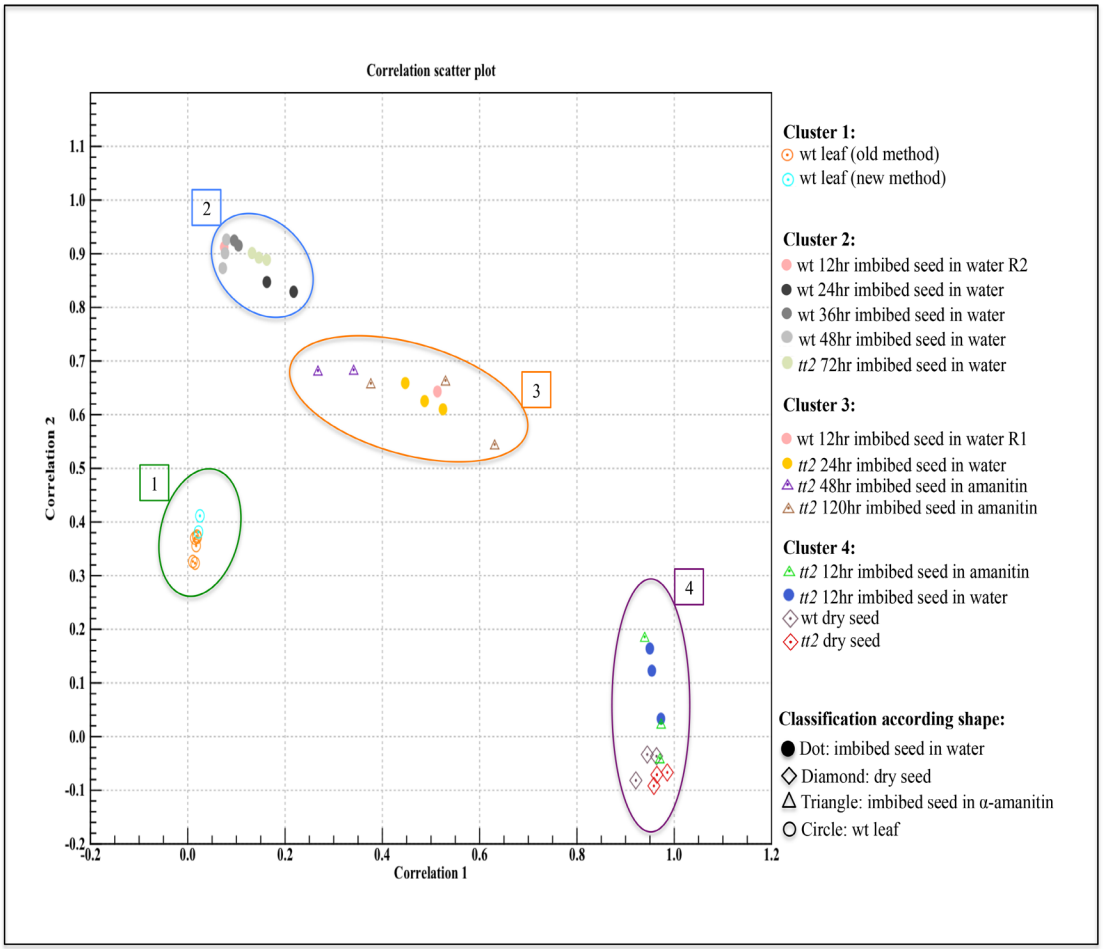


Figure 2.6 A correlation scatter plot generated by Principle Component Analysis for all replicates and treatments

Shapes in the same color represent each replicate generated from independent biological replications. The Dots represent water imbibed seed samples. The open diamonds represent dry seed samples. The open triangles represent samples of seeds imbibed in alpha-amanitin. The open circles represent wt leaf samples. R: replicate. The datasets from both seed germination and alpha-amanitin will be discussed in Chapters 3 and 4, respectively.

Chapter Three: The role of polyadenylation in seed germination: identifying genes producing alternatively polyadenylated mRNA.

3.1 Introduction

As one of the key features for the propagation of plant species, seed germination is important from both economic and ecological aspects (226). Seed germination is a critical phase in the plant life cycle because of its high vulnerability to pathogen infection, biotic stress, abiotic stress and wounding (226). Seed germination is a complex process in which gene expression is regulated at both transcriptional and post-transcriptional levels (226). One post-transcriptional event is polyadenylation of precursor mRNA to generate mature mRNA (see Chapter 1). Polyadenylation regulates gene expression through a process that chooses one of several potential poly(A) sites carried on a precursor RNA (227). The consequence, called alternative polyadenylation (APA), functions to control mRNA contexts (at RNA processing steps through loss of, or changes to, the exonic contents of mRNAs) resulting in variations in the protein-coding potential and thus generates different RNA isoforms (227). APA can also affect mRNA expression which is controlled by RNA regulatory elements (227).

In animals, it has been reported that the length of the mRNA 3' untranslated region (3' UTR) varies under different biological conditions through changing APA patterns (113, 130, 228). For example, mRNAs expressed in brain tissues have a longer 3' UTR than they do when expressed in other tissue types because a more distal poly(A) site is chosen than the stereotypical (more proximal) site used in the

other tissue types (130). However, poly(A) sites chosen in testes tend to result in a shorter 3' UTR than elsewhere resulting from preference for the more proximal poly(A) site in this tissue (130, 228). 3' UTRs gradually lengthen via APA during mouse embryonic development (129). However, the 3' UTR tends to be shorter in rapidly proliferating cells (112). Moreover, mRNAs expressed in transformed cells apparently have shorter 3' UTRs than those in untransformed cells (3). The expression of shorter mRNA isoforms (i.e. a shorter 3' UTR on the same length coding region; CDS) of some proto-oncogenes led to far more oncogenic transformation than that from the full-length mRNA isoform (i.e. with a longer 3' UTR on the same length CDS) (3). One explanation for these observations of shorter 3' UTRs associated with more rapid transcriptional activity is that the shorter 3' UTR results in the loss of microRNA recognition sequences found in the longer 3' UTRs leading to a loss of post-transcriptional regulation of mRNA accumulation and hence, run-away transcription and accumulation of mRNA species possessing the shorter 3' UTR. With consequent loss of microRNA-mediated repression in the 3' UTR, the shorter mRNA isoforms generated by APA in transformed cells exhibit increased stability and typically produce ten-fold more protein than untransformed cells (3).

In plants, it has been reported that several genes can produce mRNAs terminating in one of two or three different poly(A) sites, thus generating significant heterogeneity (229). Fourteen distinct poly(A) sites were identified for one gene in one case (230). A recent large-scale sequencing study, focused on the mRNA-

poly(A) junction in leaf or dry seed mRNA of Arabidopsis, used next generation sequencing to confirm that more than 74% of Arabidopsis genes have two or more poly(A) clusters (119). The microRNA390 targeting site of the gene encoding trans-acting small interfering RNA (TAS3a, AT3G17185) is located between the two poly(A) sites of TAS3a in WT leaf (227) providing a glimpse of possible gene regulation by APA in plants. Another interesting observation is that 113 genes exhibit different poly(A) site choice depending on if the transcript arises in the leaf or in seeds (119). In addition, 10% of APA transcripts are differentially distributed among different developmental stages and tissues under environmental stress (231). Shen et al. (231) also report that one seedling stage shows a biased usage whereby there is an increase of intron, exon, and 5' UTR APA (231).

These studies in animals and plants raise interesting questions concerning the role APA may play in controlling gene expression. In particular, the report of differential poly(A) site choice in seeds and leaves raises the possibility that APA may play important roles in seed germination. To test this possibility, it was necessary to identify those genes capable of alternatively polyadenylating their transcripts among seed germination stages as well as dry seed and the leaf to provide some bioinformatic evidence of the control of gene regulation by APA. Using a modified RNA-seq strategy (see Chapter 2) as well as bioinformatics tools (53), global poly(A) site choice during seed germination was evaluated. In addition, genes whose transcripts possess different poly(A) sites during seed germination stages or compared to those produced in leaves or dry seeds have

been identified. The results of these experiments show that 5' UTR APA may down-regulate the production of full-length mRNAs by generating 5' UTR mapped-transcripts from an unknown mechanism in one developmental stage, but such regulation is released in other developmental stages in order to generate full-length mRNAs. The coding region (CDS) APA may also down-regulate the production of full-length mRNA through generating more truncated nonstop RNAs by CDS APA in one developmental stage but generating full-length mRNAs in other stages. Overall, APA might be involved in regulating the production of full-length mRNAs or transcripts, from the same gene, destined for rapid turn over.

3.2 Results

3.2.1 Preparation and Characterization of polyadenylated cDNA tags

To study Arabidopsis genes with different poly(A) sites during seed germination on a genome-wide basis, a seed germination experiment was performed using *Arabidopsis thaliana* (Col.). Seeds were imbibed on top of water-saturated filter paper in Petri dishes for 12hr, 24hr, 36hr, or 48hr, by which time the seeds had completed germination (seed Methods and Materials). Total RNA was isolated from seeds at those germination stages as well as from dry seed and the leaves of young seedlings. Short cDNA tags that include the mRNA-poly(A) site junction (called Poly(A) Tags, or PATs) were prepared and sequenced as described in Chapter 2.

To measure the variability among different PAT samples, gene expression was assessed as described in Chapter 2; the results are shown in Figure 2.6. Four sample clusters (numbered 1-4 in Figure 2.6) were apparent in the correlation scatter plot. The dry seed (gray diamond) was located in the cluster 4 (Figure 2.6). The seed germination stage samples such as the 24hr (black dot), 36hr (dark gray dot) and 48hr (gray dot) imbibed seeds were located in the cluster 2 (Figure 2.6). Interesting, the replicates from 12hr-imbibed seeds (pink dot) were located in two clusters with replicate 1 located in cluster 3 and replicate 2 located in cluster 4 and similar results were observed from Pearson correlation coefficient analysis (data not show). However, replicates from other samples (see Cluster 2 and 4) were located in the same cluster. Overall, these results show that the variability amongst replicates from the same time point is low, and that there is a change in the gene expression program in the imbibed samples compared with the dry seed. These data therefore support the conclusion that the PATs are accurate predictors of gene expression. The 12hr-imbibed seeds is a stage during intense alterations in biological processes as the seed switches from an anabolic to a catabolic metabolism thus leading to dramatic alterations in the gene expression program over a very short period of time resulting in large variations in the transcriptome, even between biological replicates.

3.2.2 Genome-Wide Characterization of Poly(A) Site Distributions

To test whether the distribution of PATs changed during seed germination or between tissues (seeds or leaves), the genomic distribution of PATs was determined. Between 77.14% and 85.98% of PATs mapped to annotated 3' UTRs (Table 3.1). Between 6.73% and 12.08% of PATs mapped to protein-coding regions (CDS) (Table 3.1). Between 2.77% and 4.92% of PATs mapped to introns (Table 3.1), while between 2.04% and 6.77% of PATs mapped to the 5' UTR. Among the samples, wild-type 24hr and 48hr imbibed seeds have fewer PATs mapping to the 3' UTR with 79.13% and 77.14% respectively, while they possess a greater number of PATs mapping to the CDS, with 12.08% and 11.17%, respectively (Table 3.1). Compared to wild-type dry seed and the leaf which had 2.04% and 2.45% of PATs mapping to the 5' UTR, the wild-type seed germination stages (24hr, 36hr and 48hr imbibed seed) have a greater number of PATs mapped to the 5' UTR with 4.84%, 6.23% and 6.77%, respectively (Table 3.1). These results suggested that there is no significant shift of polyadenylation from 3' UTR to other genomic regions.

3.2.3 The genome-wide distribution of poly(A) sites located in the 3' UTR does not change during seed germination or in the leaf

To evaluate whether there are more subtle shifts in poly(A) site choice during germination, a more accurate analysis of poly(A) site choice was performed using an approach described in Thomas et al. (53). This software focuses on 3' UTR–situated sites reflecting the fact that the overwhelming majority of PATs map to 3' UTRs (53). One peculiarity of poly(A) site choice in plants was that it varies both

locally (with the actual site of polyadenylation shifting by only a few nucleotides) and on a larger scale with many tens of nucleotides separating the alternative polyadenylation sites (Figure 3.1A). The local alterations in polyadenylation site are referred to as a single cluster of PATs while there can be two or more such polyadenylation site clusters (PACs) per gene (Figure 3.1A). The program measures relative poly(A) site distribution on a gene-by-gene basis and allows one to summarize results genome-wide using other software, such as Excel (Figure 3.1B).

To represent the variability of poly(A) site distribution genome-wide, the resulting set of values was grouped in increments of 0.05 and the running sums of numbers of genes whose values fall within a given increment plotted (Figure 3.1C). As explained in Thomas et al. (53) and in the methods section, plots made from poly(A) site profile data sets that contain considerable poly(A) site differences yield curves that are shifted to the right while plots from data sets that contain largely similar poly(A) site selections, genome-wide, have curves shifted to the left (Figure 3.1C). The comparison of poly(A) sites in wild-type leaves and a mutant (*oxt6)* deficient in the polyadenylation factor subunit CPSF30 (taken from ref. 54) was used as an example to illustrate these plots (Figure 3.1C). Comparisons of different members of replicate data sets for the wild type dry seed as well as all germination stages are largely superimposable demonstrating the reproducibility of the PATs method (Figure 3.2A). This is in marked contrast to the comparison of the wild-type-and *oxt6*-leaf, which produces a curve, shifted to the right.

Individual replicates from dry seed and from the various time points were compared and the curves plotted (Figures 3.2 A, B). As shown in Fig. 3.2 A, B, these plots were superimposable on those obtained by Thomas et al. for comparisons of biological replicates. This result corroborates that obtained from the gene expression analysis and is indicative of a high degree of reproducibility in the various samples.

The curves representing the comparison between the germination stage averages was also superimposable on the curve obtained when comparing replicates from the same biological sample (Figure 3.2C). This indicates that the genome wide differences in poly(A) site choice in the 3' UTR between germination stages is very limited. These results indicate that, for poly(A) sites located in 3' UTRs, poly(A) site choice does not change appreciably during seed germination.

3.2.4 Identification of genes exhibiting alternative poly(A) site choice in 3' UTRs during germination or seedling development

It has been reported that 113 Arabidopsis genes exhibited different poly(A) site usage between wild type dry seed and the leaf (119). Therefore, it was possible that a relatively small number of genes might exhibit developmental stage-specific changes in poly(A) site choice among seed germination stages and/or between specific seed developmental stages and the leaf. To identify these candidates, genes with poly(A) metric values (Figure 3.1B) larger than 0.2 metric values as described in the preceding section and summarized in Figure 3.1B, were considered as a

potential APA candidates. There were 455 such genes identified. Of these 455 genes, 59 high confidence genes were selected based on closer visual inspection using the genome browser utility of CLC (Appendix 3.1). One example, shown in Figure 3.3, affects the gene AT4G14300, which encodes a RNA-binding protein, which can interact with the Arabidopsis nuclear import receptor TRANSPORTIN1 (232). This gene has two 3' UTR poly(A) sites (Figure 3.3). The transcripts generated by the distal poly(A) site were abundant in dry seed (Figure 3.3). The transcripts generated by the proximal poly(A) site were abundant during seed germination stages (24hr, 36hr and 48hr imbibed seed) and in the leaf (Figure 3.3).

3.2.5 Identification of genes, expressed during germination or in the leaf, exhibiting alternative poly(A) site choice within the 5' UTR

More than 15% of all PATs map to genomic regions other than the 3' UTR (Table 3.1), raising the possibility of stage-specific poly(A) site choice involving sites in these regions. A systematic visual examination of the genes producing transcripts with altered polyadenylation in these regions of the gene was performed using the genome browser functions of the CLC Genomics Workbench. 303 genes were identified as having more than 50% of their PATs located in the 5' UTR in at least one germination stage, in dry seeds, or in leaves. Internal priming of tags from A-rich regions in the mRNA has been reported as a common problem resulting in contamination of PATs. This is because, if there are long stretches of adenosines in the internal region of RNAs, these sites will be treated as poly(A) sites since primers containing oligo d(T) can recognize those long stretches of adenosines and treat the

sites as legitimate poly(A) sites. For 5' UTR poly(A) sites, it is also possible that they are generated from transcripts transcribed from nearby genes. Therefore, if the nearest upstream gene was in an antisense orientation or in the same orientation with the gene but more than 500 nt away and no long stretches of adenosine were observed after the poly(A) site, it was accepted that the 5' UTR poly(A) site of this gene was a legitimate polyadenylation site. With these criteria in mind, a closer inspection ruled out internal priming and transcription from nearby genes for 19 genes which changed their poly(A) sites from the 5' UTR site to another (non-5' UTR) site in the gene, or *vice versa*, at different development stages (Appendix 3.2). Most of these candidates had 5' UTR poly(A) sites in germination stages but shifted to 3' UTR poly(A) sites in leaf. For example, AT1G13460 has three poly(A) sites, one in the 5' UTR and two located in the 3' UTR (Figure 3.4). In the 24hr, 36hr and 48hr imbibed seeds; the dominant poly(A) site is located in the 5' UTR. However, the proximal 3' UTR site is the dominant site in the leaf (Figure 3.4). Therefore, the product of this gene (the B'theta subunit of PROTEIN PHOSPHATASE 2A) is predicted to accumulate in the leaf but be less abundant or not present in seeds during the different stages of germination.

Another example, AT1G70230, encodes ALTERED XYLOGLUCAN4 (AXY4) (233), is shown in Figure 3.5. This gene mainly has three poly(A) sites: the 5' UTR site and two major 3' UTR sites (Figure 3.5). Transcripts from this gene were not present in dry seeds. The transcripts of *AXY4* with 5' UTR poly(A) sites were predominate in all germination stages (24hr, 36hr, and 48hr imbibed seed) (Figure 3.5). The

transcripts of *AXY4* using the two 3' UTR poly(A) sites were dominant in the leaf (Figure 3.5). Therefore, it was reasonable to predict that this gene was down-regulated during seed germination stages due to the usage of the 5' UTR poly(A) site resulting in the lack of full-length transcript.

AT4G00430, encoded a PLASMA MEMBRANE INTRINSIC PROTEIN1;4 (PIP1;4 and is depicted in Figure 3.6. PIP1;4 transcripts have four major poly(A) sites: the 5' UTR poly(A) site and three 3' UTR poly(A) site (Figure 3.6). The short transcripts of PIP1;4, generated by the 5' UTR poly(A) site, were dominant in germination stages (24hr, 36hr and 48hr imbibed seed). The full-length transcripts of PIP1;4, generated by the proximal and middle 3' UTR poly(A) sites, were dominant in the leaf (Figure 3.6).

3.2.6 Identification of genes, expressed during germination or in the leaf, exhibiting alternative poly(A) site choice within the introns

321 genes were identified that had more than 50% of their PATs located in introns in at least one of the stages under study. Closer inspection, which ruled out internal priming, showed that only seven genes showed changes in their poly(A) sites during different development stages (Appendix 3.3). One example, shown in Figure 3.7, affects the gene AT1G06630, which encodes a F-box/RNI-like superfamily protein. This gene has two major poly(A) sites (Figure 3.7). The transcripts generated by the intronic poly(A) site were abundant in wild-type dry seed, 24hr imbibed seeds and

leaf (Figure 3.7). The transcripts generated by the 3' UTR poly(A) site were also abundant in the leaf (Figure 3.7).

3.2.7 Identification of genes, expressed during germination or in the leaf, exhibiting alternative poly(A) site choice within the coding region (CDS)

There were 993 genes identified as having more than 50% of their PATs located in the CDS in at least one stage under study. Closer inspection which rule out internal priming showed that 73 of these genes had differences in their poly(A) sites in the different stages examined (Appendix 3.4).

One example, the AT3G48670 gene, encoded INVOLVED IN DE NOVO2 (IDN2) (Figure 3.8). The IDN2 gene has two major poly(A) sites: the 3' UTR poly(A) site and the coding region (CDS) poly(A) sites (Figure 3.8). In the 24hr and 36hr imbibed seeds, the dominant transcripts of *IDN2* were those truncated transcript generated by CDS APA (Figure 3.8). However, the full-length transcripts of IDN2, generated by the 3' UTR site, were dominant in the leaf (Figure 3.8).

Another interesting example of CDS APA was AT3G14980, encoding INCREASED DNA METHYLATION1 (IDM1) (Figure 3.9). This gene contained three major poly(A) sites: two of them located in the coding region and one located in the 3' UTR (Figure 3.9). The truncated transcripts generated by the proximal CDS poly(A) site dominated in the 36hr imbibed seeds (Figure 3.9). The full-length transcripts generated by the 3' UTR poly(A) site were more abundant in the leaf (Figure 3.9).

Both proximally truncated- and full-length-transcripts occurred in 24hr-imbibed seed but the full-length transcripts were more abundant (Figure 3.9).

A third example of CDS APA involved AT2G22125. The AT2G22125 gene encodes a CELLULOSE SYNTHASE INTERACTIVE PROTEIN1 (CSI1) (234). This gene has two major poly(A) sites: the CDS poly(A) site and the 3' UTR poly(A) site (Fig. 3.10). The truncated transcripts generated from the CDS poly(A) site were consistently abundant in all germination stages (24hr, 36hr and 48hr imbibed seed) and in the leaf (Figure 3.10). However, more 3' UTR full-length transcripts were generated in the leaf (Figure 3.10).

3.3 Discussion

3.3.1 Stage-specific APA is not genome-wide, but involves a small number of genes

As most PATs located to the 3' UTR, we tried to determine whether the global 3' UTR poly(A) sites changed during seed germination, or not, by using software and methods described in Thomas et al. (53). Changes to the 3' UTR APA sites, among different developmental stages, were not a global event as expected (Figure 3.2). However, consistent with a previous report (Wu et al., 2011 (235)), dozens of genes were found to change their poly(A) site among different developmental stages (Appendix 3.1, 3.2, 3.3 and 3.4). Therefore, while not widespread, APA events may affect the expression of numerous genes during germination and development.

More than 77% of PATs mapped to the 3' UTR indicating that most PATs are located in the 3' UTR during seed germination stages (24hr, 36hr and 48hr imbibed seed) (Table 3.1) Compared with other germination stages, the 36hr-imbibed seed had fewer PATs (6.73%) mapping to the coding region (compared with 12.08% and 11.17% in 24hr- and 48hr-imbibed seeds, respectively) (Table 3.1). This result suggests that certain mechanisms may act in 36hr-imbibed seeds to decrease the coding region polyadenylation. For other seed germination stages, dry seed, and leaves, approximately 10% of PATs mapped to the coding region (Table 3.1) as reported previously (119). Interestingly, compared with the leaf and dry seeds (less than 2.45%), seed germination stages had more PATs that mapped to the 5' UTR (more than 4.84%) (Table 3.1) indicating that polyadenylation in the 5' UTR might be a way to regulate gene expression during seed germination.

3.3.2 5' APA and possible regulation mechanisms

There were 19 genes that had differentially-utilized 5' UTR poly(A) sites in at least one developmental stage (Appendix 3.2). The 5' UTR poly(A) sites of these genes did not result from internal priming or from upstream genes flanking genes in the same orientation. Therefore, where did these transcripts come from? Recently, a new class of human RNAs called PROMoter uPstream Transcripts (PROMPTs) were discovered, produced upstream of promoter region of coding genes in mammals by all three mammalian DNA dependent RNA polymerases (RNAPI, II, and III) (236). Studies of PROMPTs generated by RNAP II revealed that these transcripts have a 3' end poly(A) tail and 5' cap structure (236) (Figure 3.11A). They were largely

nuclear RNAs and rapidly degraded by exosome (236). They were estimated to have relatively small sizes ranging from 200 to 600 nts (236). RNAP II was associated with PROMPT-transcribing DNA (236). Moreover, RNAPII occupancy of the upstream PROMPT region increases under conditions, which reduce the coding gene transcription (236) (Figure 3.11A). These results lead to a possible explanation for the transcripts, which terminate in the 5' UTR (5' UTR APA transcripts) in seed germination stages. These transcripts generated by 5' UTR APA may be the PROMPTs. If true, the production of these transcripts may lead to a reduction of transcripts that encode a functional protein.

In addition, a recent study using a new genome-wide nuclear run-on assay (PRO-seq) mapped RNAP II with base pair resolution (237). This method can detect the RNAP II pausing efficiently (237). The highest density of PRO-seq signals/ reads mapped to within +30 (+, downstream) to +60 nucleotides from the transcription start site (TSS)(237) (Figure 3.11B). It may be that transcripts derived from the promoter-associated RNA polymerase II pausing may become polyadenylated in Arabidopsis. This would also be a mechanism for down regulating the production of full-length mRNAs.

However, PROMPTs are transcribed far upstream (0.5kb-2.5kb) from transcription start site and the length of PROMPTs are thought to be less than 600 nt (236) suggesting that the PROMPTs may not terminate in the 5' UTR. Moreover, the short transcript generated from Promoter-associated PoI II Pausing may not have a

poly(A) tail and will be very short as most of PoI II pausings are between +30 to +60 nt after the transcription start site. Overall, it is possible that the transcripts generated by 5' UTR APA in plants may not result from the two possible mechanisms described from the animal system and they may be generated by an unknown mechanism. However, regardless of the mechanism, it seems likely that this is a mechanism for down regulating the production of full-length transcripts.

3.3.3 Intronic APA may regulate the production of different RNA isoforms

More than 2% of PATs mapped to the intron. The AT1G06630 encoded a F-box/RNI-like superfamily protein the gene of which can produce three different RNA isoforms (Figure 3.7). Each of these RNA isoforms has the potential to translate into different proteins based on the Arabidopsis EST database mRNA and CDS annotation (Figure 3.7). Interestingly, intronic APA regulates the production of the shortest and longest RNA isoforms in dry seed, seed germination stages and leaf. The shortest RNA isoform was generated in dry seed, 24hr imbibed seed and the leaf. However, the leaf sample contained both the shortest and the longest RNA isoforms. These results suggested that the intronic APA might regulate the production of different RNA isoforms or proteins.

3.3.4 CDS APA acts as a mechanism down regulation for gene expression especially during seed germination

Approximately 10% of PATs mapped to the CDS, and these transcripts will almost always lack a stop codon (119). These "nonstop RNAs" are degraded by the nonstop RNA decay pathway and become translationally repressed (238). Thus, this mode of APA may be an additional down regulation mechanism for gene expression in seed germination stages.

One interesting example of CDS APA transcripts was produced from the gene, *INVOLVED IN DE NOVO2* (*IDN2*, AT3G48670), the product of which is an RNA-binding protein that is involved in transcriptional silencing through the RNA dependent DNA methylation (RdDM) pathway (239, 240). Interestingly, the full-length transcripts of *IDN2* were abundant in the leaf while the truncated transcripts of *IDN2*, generated by usage of the CDS poly(A) site, were abundant in 24hr and 36hr imbibed seeds (Figure 3.8). This indicates that *IDN2* expression might be down regulated in these two germination stages but released from such regulation in the leaf. Another interesting example involves transcripts produced from *INCREASED DNA METHYLATION1* (*IDM1*, AT3G14980), the protein of which is involved in preventing DNA hypermethylation and transcriptional silencing (241). This protein acts as a "counterpart" of IDN2 as IDN2 is involved in DNA methylation. Interestingly, like those for *IDN2*, the full-length of *IDM1* transcripts are abundant in the leaf while the truncated *IDM1* transcripts are abundant in 36hr imbibed seeds (Figure 3.9). Both *IDM1* and *IDN2* were downregulated or not expressed in 36hr and

48hr imbibed seed germination stages. This may indicate that the RdDM pathway might not be active in those stages as *IDN2* is required for RdDM. The CDS APA might be one mechanism by which the RdDM- and demethylation-pathways are regulated through production of either the full-length- or truncated-transcripts.

3.3.5 A model to describe the potential functions of APA among different developmental stages in plants

FLC represses flowering time while its expression level is regulated by APAs of sense transcripts of *FCA* and *FPA* as well as natural antisense transcripts of *FLC* (2). An AtCPSF30-deficient mutant has a different polyadenylation profile genome-wide compared with wild-type plants (68). APA might also be involved in down-regulating gene expression (Figure 3.12) by generating the truncated transcripts possibly represented by the 5' UTR and CDS APA examples shown. APA might also be involved in regulating tissue specific gene expression (Figure 3.11) by generating truncated nonstop transcripts in one stage but generating full-length transcripts with stop codons in another stage just like regulating the expression of *AXY4* genes in the leaf but not during seed germination stages (Figure 3.5). The APA might also function in regulating translation (Figure 3.11). In animals, it has reported that 3' UTR APA leads to inclusion or exclusion of some RNA regulatory cis-elements such as miRNA targeting sites (104)(Chapter 1). The shortening of the 3' UTR by 3' UTR APA has been observed during development of spermatocytes, proliferating T cells and some cancer cell lines (104)(Chapter 1). The lengthening of the 3' UTR by 3'

UTR APA has been observed in ovulated oocytes, developing mouse embryos and neurological tissues (104)(Chapter 1). The shorter 3' UTR may avoid mRNA repression by loss of miRNA targeting sites by shifting 3' UTR APA in cancer cells (3)(Chapter 1). Although the location of miRNA targeting sites may be different between animals and plants (242), it is possible that the plant might also present such a regulatory mechanism.

3.4 Methods and Material

3.4.1 Seed germination experiment

Approximately 0.10g of wild-type *Arabidopsis thaliana* (Col.) seeds were placed on the top of two layers of Whatman No.1 filter paper in a BD Falcon Bacteriological Petri dish (standard style dish 100 x 15 mm) wetted with distilled water. The seeds were next subjected to 4°C for 3 d to alleviate dormancy before transfer to 25°C. Germination was allowed to proceed for 12hr, 24hr, 36hr or 48hr. RNA was isolated from imbibed or dry seeds or from seedling leaves as described in Chapter 2.

3.4.2 3' UTR alternative polyadenylation assay

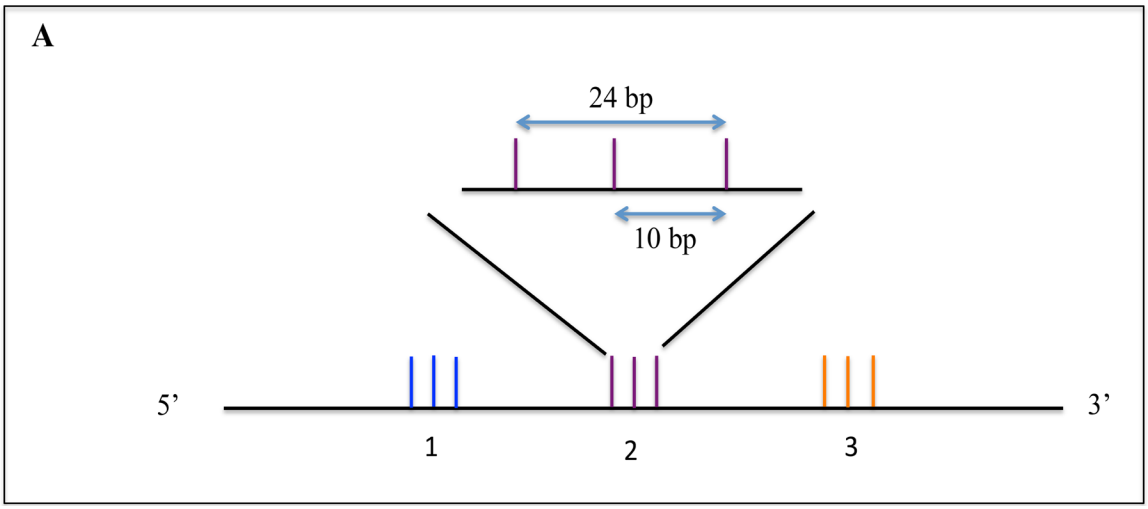
To assay poly(A) site choice *in vivo*, and to compare different replicates or biological samples (tissues, mutants, treatment), tags were mapped to a dataset that consisted of the TAIR10 3' UTR sequences (3' UTRs for all annotated Arabidopsis genes) in which each sequence was extended by 500nts downstream of the TAIR-established annotation as described in (53). This was done to take into account the fact that the Arabidopsis genome annotation is incomplete, and that mRNAs in some mutants

may be extended downstream from “normal” poly(A) sites. In the mapping step, genomic positions adjacent to tracts of 6 or more A’s are masked to eliminate internal RT priming artifacts from inclusion in the data set. The program default costs for mismatches, deletions, and insertions were used, the fractional length of the tag that must match the reference was set to 0.9, and the similarity of a tag for the reference was set to 0.7. The mapping was performed using CLC. The mapping files were exported as .sam files that were used for pairwise comparisons using the Java program described by Thomas et al. (53). The output for this was further analyzed with Excel much as described in Thomas et al. (53).

Table 3.1 Distribution of PAT in different gene regions

Genomic regions as defined in the TAIR10 database. As explained by Wu *et al*. (2011) (235), the 3' UTR were extended by 120 nucleotides. [a], total number of curated poly(A) site tags that map to the respective genomic regions. [b], percentage of total PATs that fall within the indicated regions. Wild-type dry seed, leaf and different imbibed seed germination stages are calculated separately.

Developmental stages	Region	3' UTR	CDS	Intron	5' UTR
Dry seed	PAT No[a].	4473866	516653	145326	106712
	PAT(%)[b]	85.34	9.85	2.77	2.04
24hr imbibed seed	PAT No.	2051330	313151	102417	125415
	PAT(%)	79.13	12.08	3.95	4.84
36hr imbibed seed	PAT No.	1509108	123578	88734	114348
	PAT(%)	82.21	6.73	4.83	6.23
48hr imbibed seed	PAT No.	1317216	190740	84082	115613
	PAT(%)	77.14	11.17	4.92	6.77
Leaf	PAT No.	4646320	460597	164650	132498
	PAT(%)	85.98	8.52	3.05	2.45



B

		f@1	f@2	f@3	metric
Case 1	wt leaf	0.95	0.05	0	0.9
	oxt6 leaf	0.05	0.95	0	
Case 2	wt leaf	0.4	0.2	0.4	0.2
	oxt6 leaf	0.2	0.4	0.4	
Case 3	wt leaf	0.5	0.5	0	1.0
	oxt6 leaf	0	0	1.0	

$$\text{Case 1: metric value} = \frac{|0.95-0.05|+|0.05-0.95|+|0-0|}{2}$$

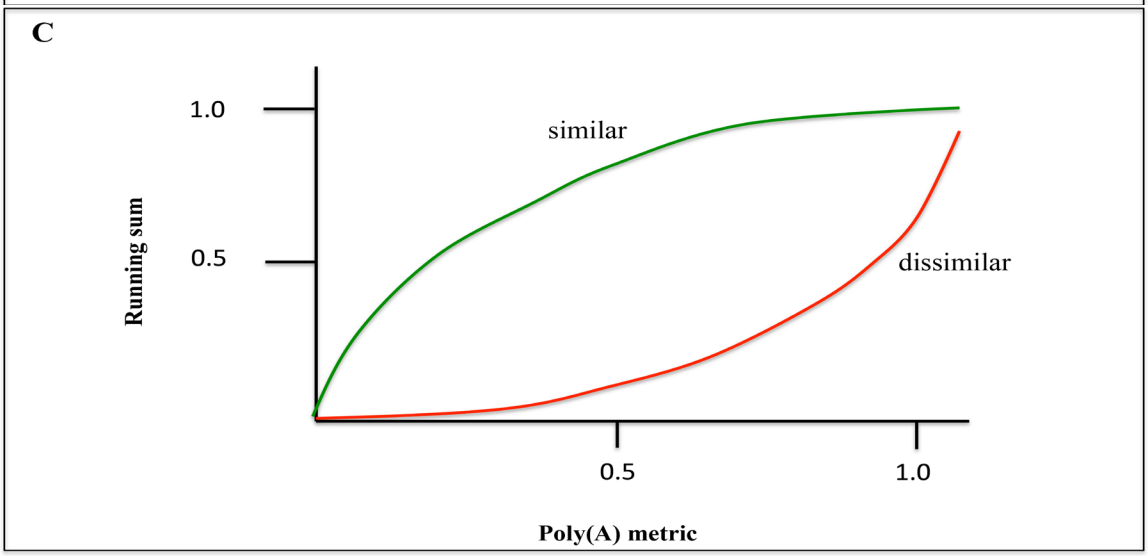


Figure 3.1 Strategy for assessing poly(A) site choice between different replicates/samples

Figure 3.1 (continued)

Figure modified from Thomas *et al*., 2012 (53)

(A) Illustration of clustering of closely situated poly(A) sites. This is referred to as poly(A) clusters or PACs. The bottom line: the hypothetical reference sequence depicted from 5' to 3'. There are three clusters depicted in the figure and these clusters were defined by the existence of PATs that end at the indicated positions. Cluster two was expanded and presented above the larger figure. The clustering poly(A) sites were defined by the following criteria: the maximum distance between distinct sites is set to 10 nucleotides; the span of a single cluster is a maximum of 24 nucleotides.

(B) Hypothetical results were used to describe how to calculate metric values for further analysis. Three cases exemplify how the metric value is calculated based on the frequency of PAT distribution at different PACs. f@1: frequency or percentage of PATs at PAC 1. Metric value: the fraction of all tags mapped to one of two clusters in the reference sequence is calculated. From this, the absolute values of the differences between the two data sets (here, the wild type [wt] and mutant *oxt6* is calculated and the result is divided by two to generate the metric value. The metric value represent the poly(A) site choice difference. The greater it is, the bigger the difference the change in the poly(A) site cluster.

(C) An illustration of two hypothetical outcomes. For making such a plot, the set of metrics (from 0 to 1) for a data set are divided by 20 steps of an increment of 0.05. The number of genes falling into each of these metrics values was counted. The running sum (normalized so that the final value is 1.0) is then calculated and plotted as shown. If the two data sets are similar regarding poly(A) site choice, more genes (higher running sum) fall into the small metric values. On the other hand, if two data sets are different, more genes (higher running sum) fall into the large metric values.

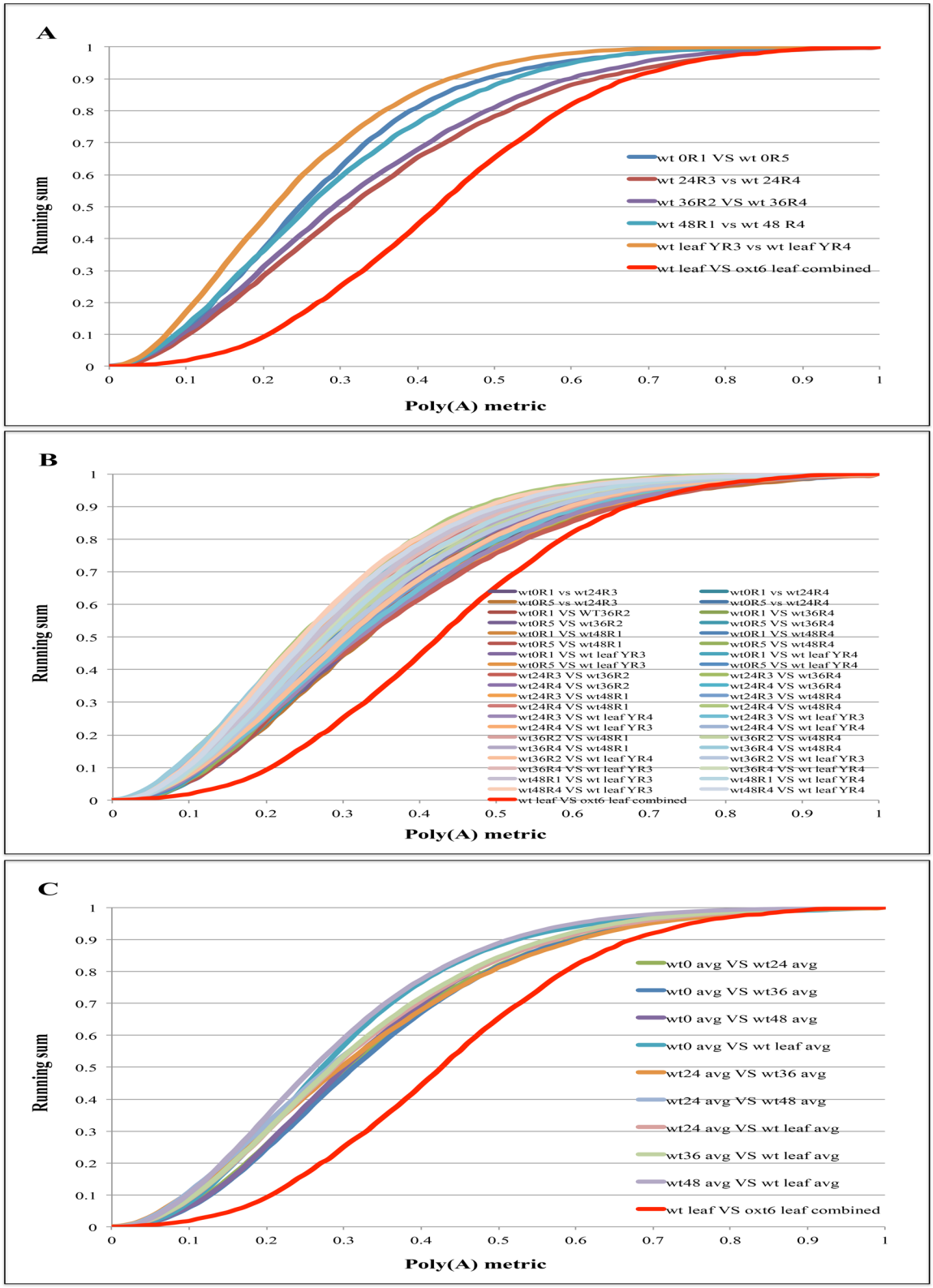


Figure 3.2 Plots of Pairwise comparisons of data sets from dry seed, leaf and seed germination stages

Figure 3.2 (continued)

A graphic depiction of the similarity among replicates in polyadenylation tag placement.

(A) Comparison of the replicates from the same stages. The stages being compared are listed in the legend (eg. wt0 R1 VS wt0 R5) denoting the pairwise comparison between wild-type dry seeds for replications 1 versus 5.

(B) Results of comparison among replicates from all stages (eg.wt0 R1 VS wt24 R3) denote the pairwise comparisons that were made.

(C) Comparison of the average values (the replicates averaged) from different tissues, (seeds versus leaves) and stages during germination on water. The stages being compared are listed in the legend (for example, average wt0 VS average wt24) denoting the pairwise comparison between the mean wild-type dry seed values versus the mean wt seeds imbibed on water for 24 hours values wt: wild-type. 0: dry seed. R: replicate. YR: young leaf Replicate (YR3 = wt leaf R7; YR4= wt leaf R8). 24: 24hr imbibed seed. 36: 36hr imbibed seed. 48: 48hr imbibed seed. Avg: average.

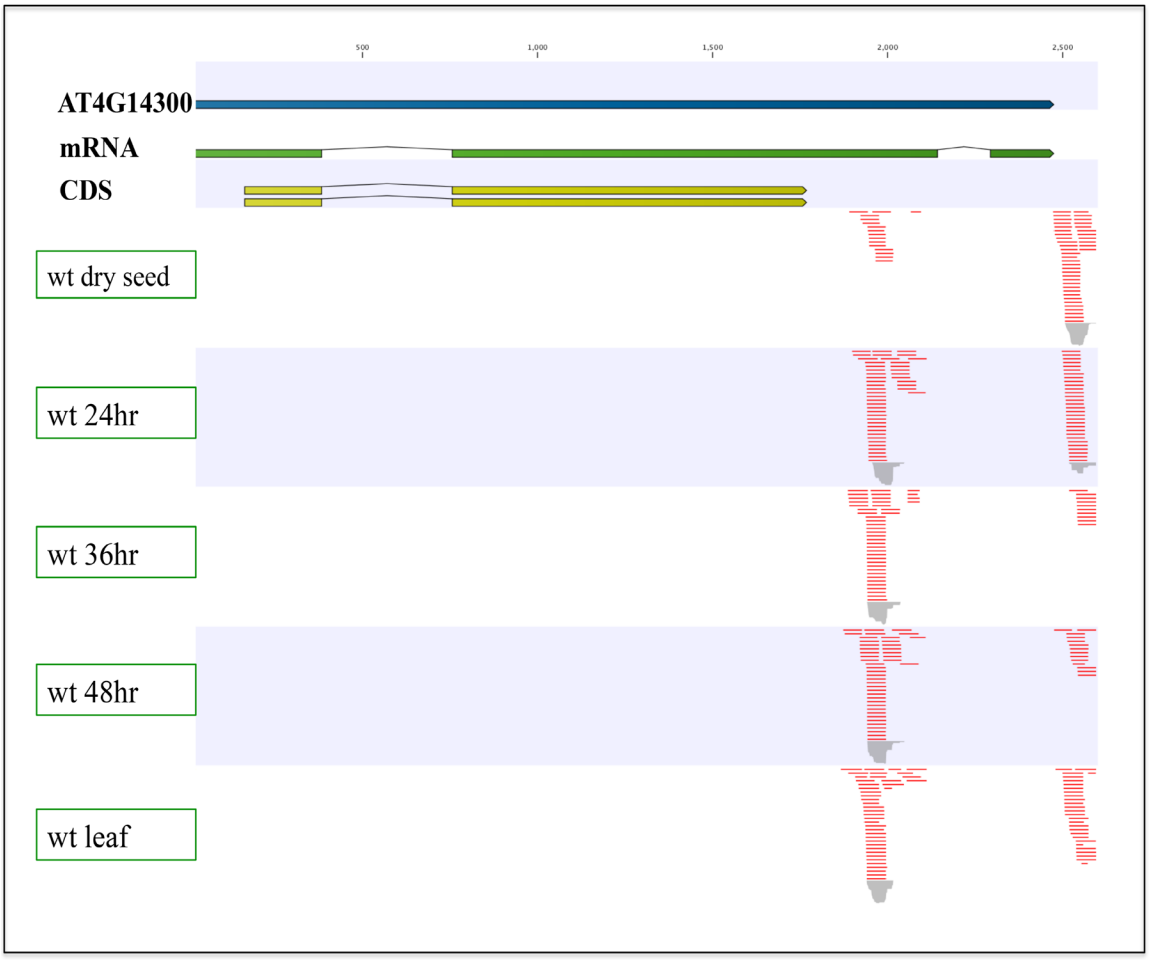


Figure 3.3 The 3' UTR APA during seed germination: AT4G14300
A graphical depiction of the position and frequency of use of two 3' UTR APA sites in AT4G14300 in different tissues and germination stages, 24hr: 24hr imbibed seeds. 36hr: 36hr imbibed seeds. 48hr: 48hr imbibed seeds. Each of the red lines represents a read. The position of the red columns represents the location of the reads. The number of lines represents the expression level. The gray zones beneath the red polyadenylated tags indicate that many more tags than are shown also map to the respective genomic positions. CDS: coding region.

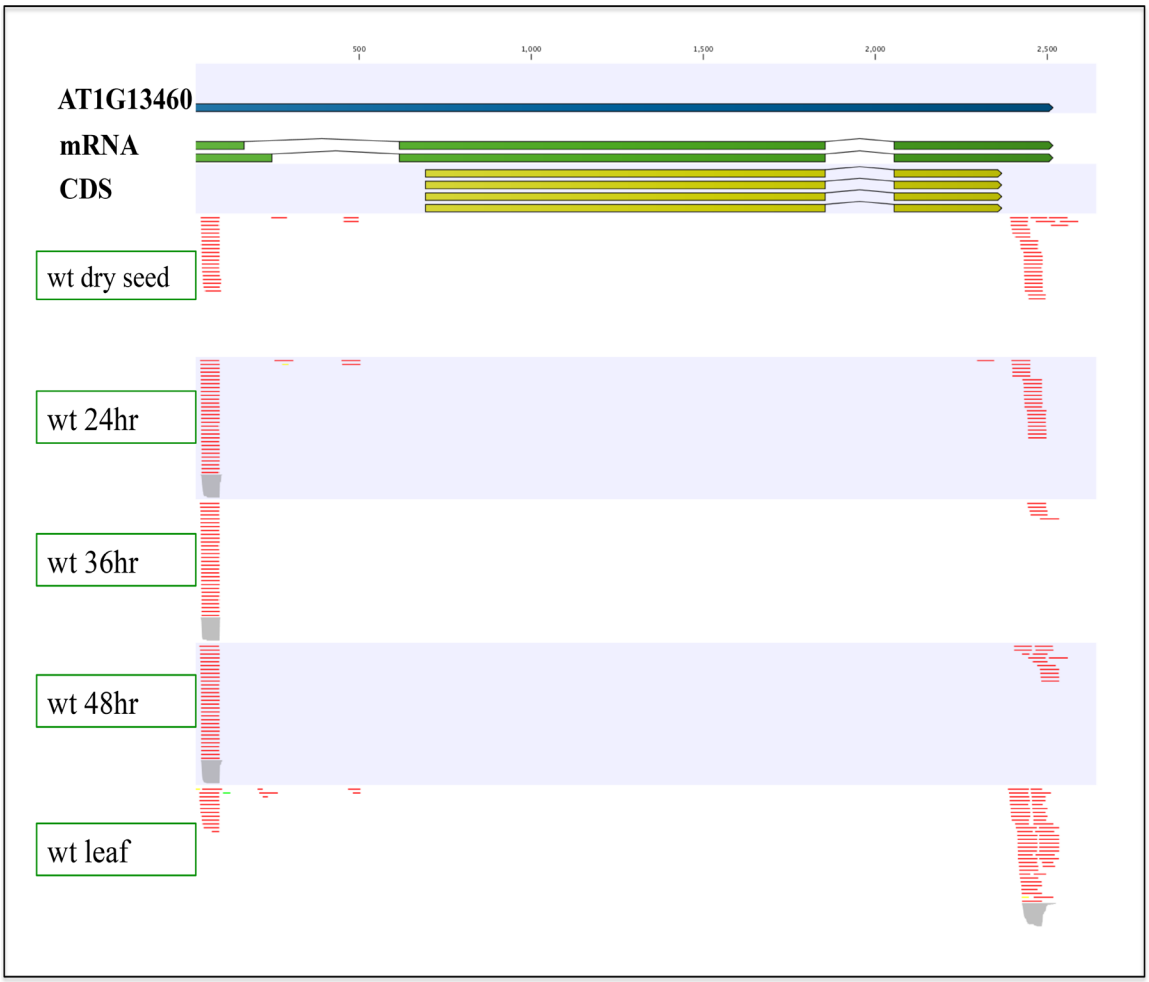


Figure 3.4 The 5' UTR APA during seed germination: AT1G13460
A graphical depiction of the position and frequency of use of 5' UTR versus two 3' UTR APA sites in AT1G13460 in different tissues and germination stages, 24hr: 24hr imbibed seeds. 36hr: 36hr imbibed seeds. 48hr: 48hr imbibed seeds. Each of the red lines represents a read. The position of the red columns represents the location of the reads. The number of lines represents the expression level. The gray zones beneath the red polyadenylated tags indicate that many more tags than are shown also map to the respective genomic positions. CDS: coding region.

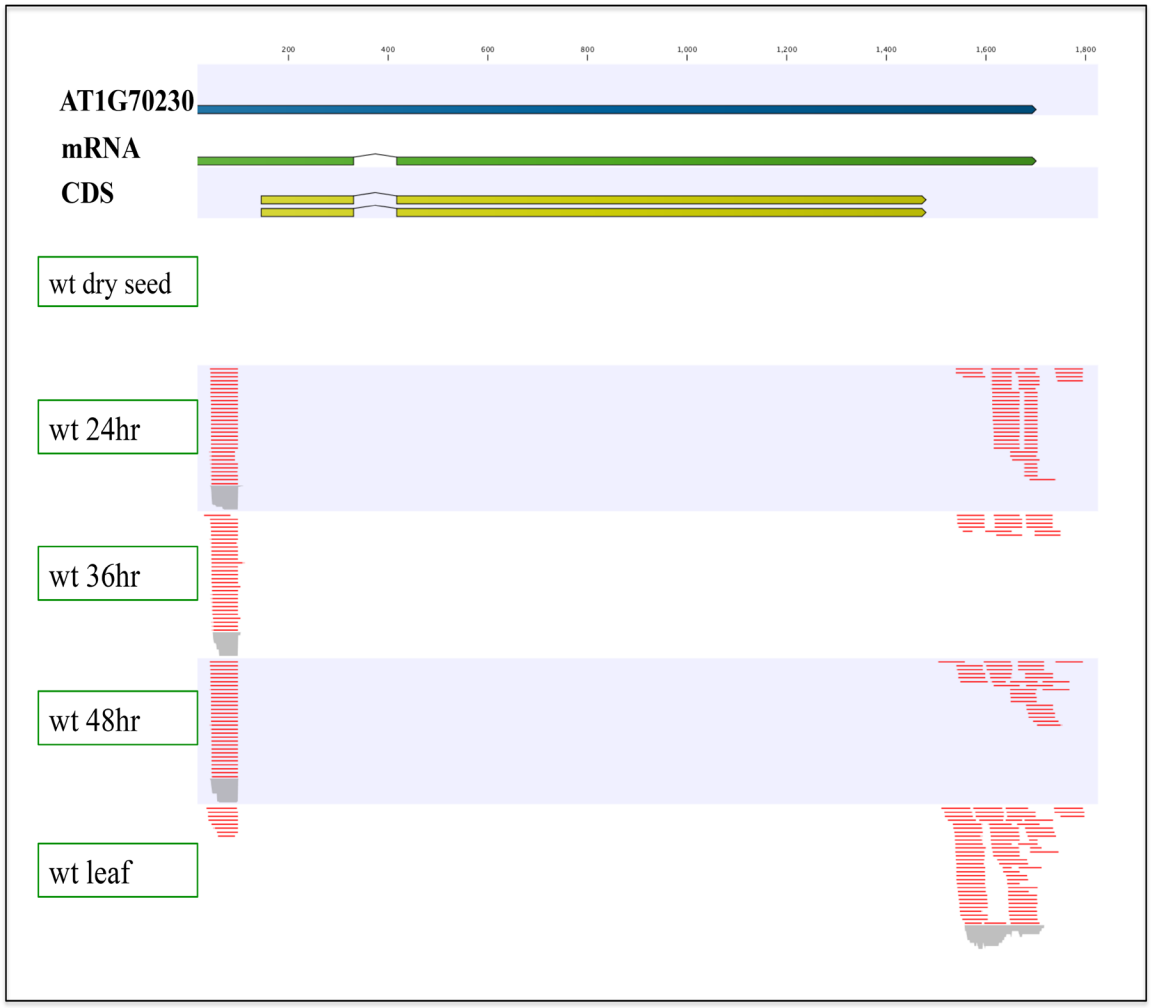


Figure 3.5 The 5' UTR APA during seed germination: AT1G70230
A graphical depiction of the position and frequency of use of 5' UTR versus two 3' UTR APA sites in AT1G70230 in different tissues and germination stages, 24hr: 24hr imbibed seeds. 36hr: 36hr imbibed seeds. 48hr: 48hr imbibed seeds. Each of the red lines represents a read. The position of the red columns represents the location of the reads. The number of lines represents the expression level. The gray zones beneath the red polyadenylated tags indicate that many more tags than are shown also map to the respective genomic positions.

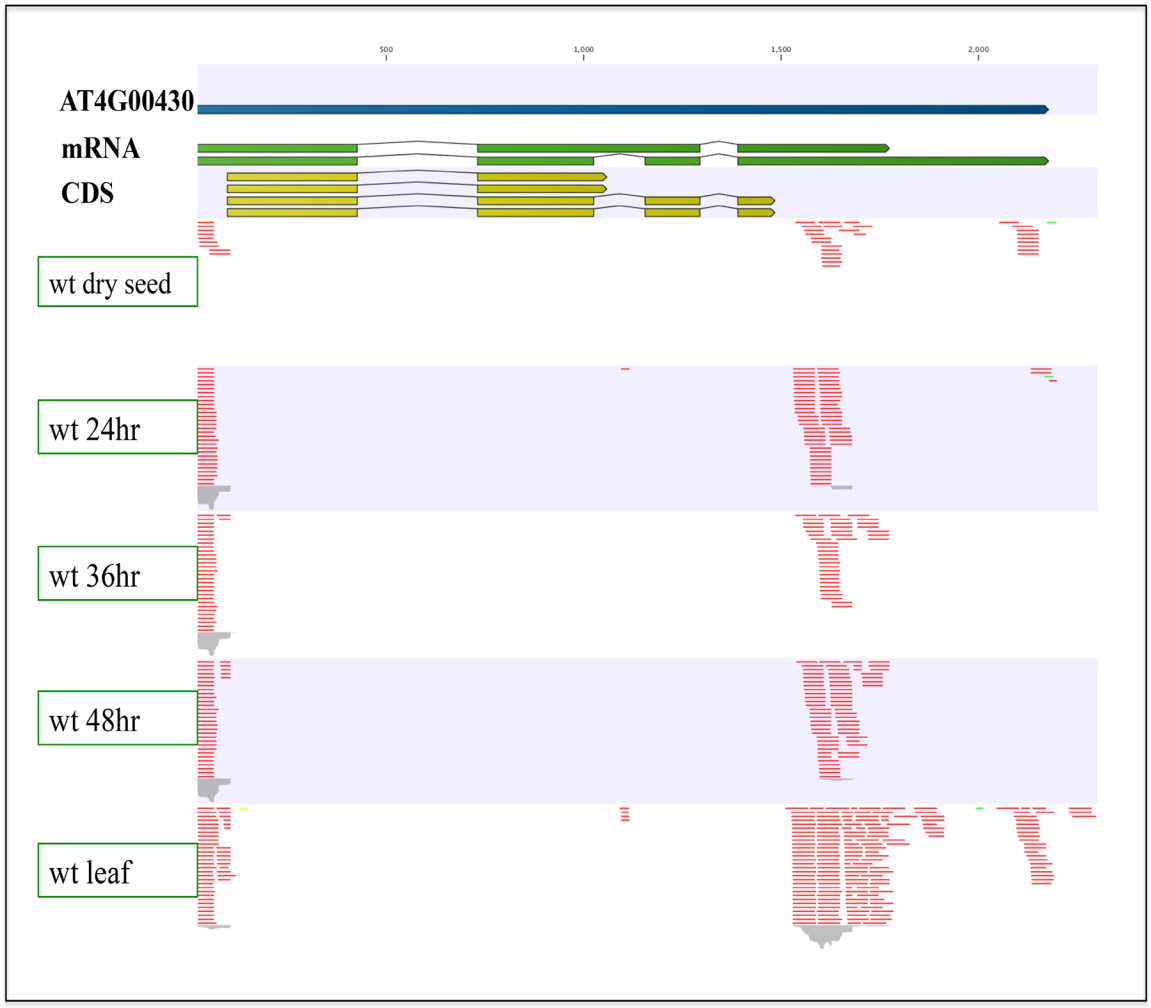


Figure 3.6 The 5' UTR APA during seed germination: AT4G00430
A graphical depiction of the position and frequency of use of 5' UTR versus 3' UTR APA sites in AT4G00430 in different tissues and germination stages, 24hr: 24hr imbibed seeds. 36hr: 36hr imbibed seeds. 48hr: 48hr imbibed seeds. Each of the red lines represents a read. The position of the red columns represents the location of the reads. The number of lines represents the expression level. The gray zones beneath the red polyadenylated tags indicate that many more tags than are shown also map to the respective genomic positions.

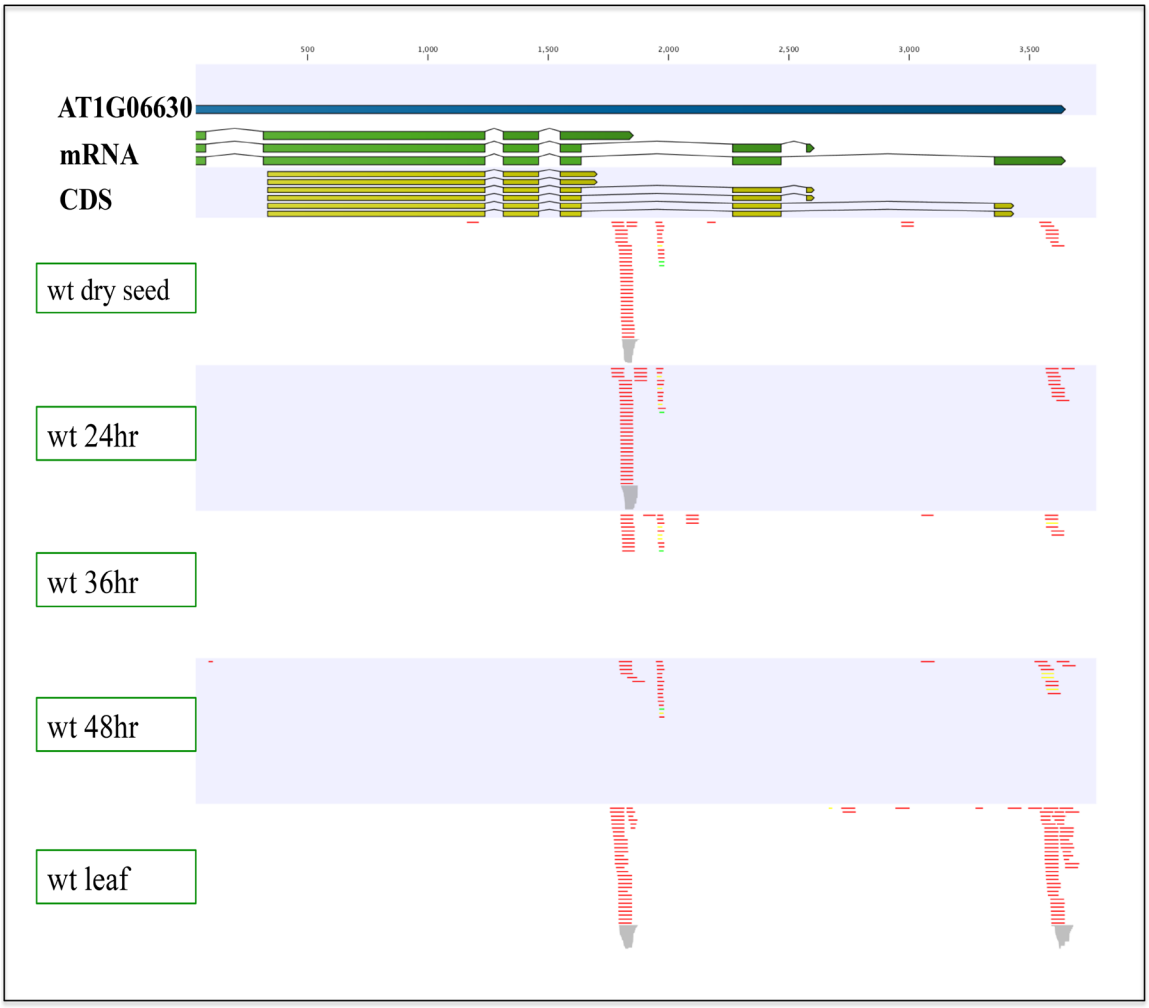


Figure 3.7 The intronic APA during seed germination: AT1G06630
A graphical depiction of the position and frequency of use of intron versus 3' UTR APA sites in AT1G06630 in different tissues and germination stages, 24hr: 24hr imbibed seeds. 36hr: 36hr imbibed seeds. 48hr: 48hr imbibed seeds. Each of the red lines represents a read. The position of the red columns represents the location of the reads. The number of lines represents the expression level. The gray zones beneath the red polyadenylated tags indicate that many more tags than are shown also map to the respective genomic positions.

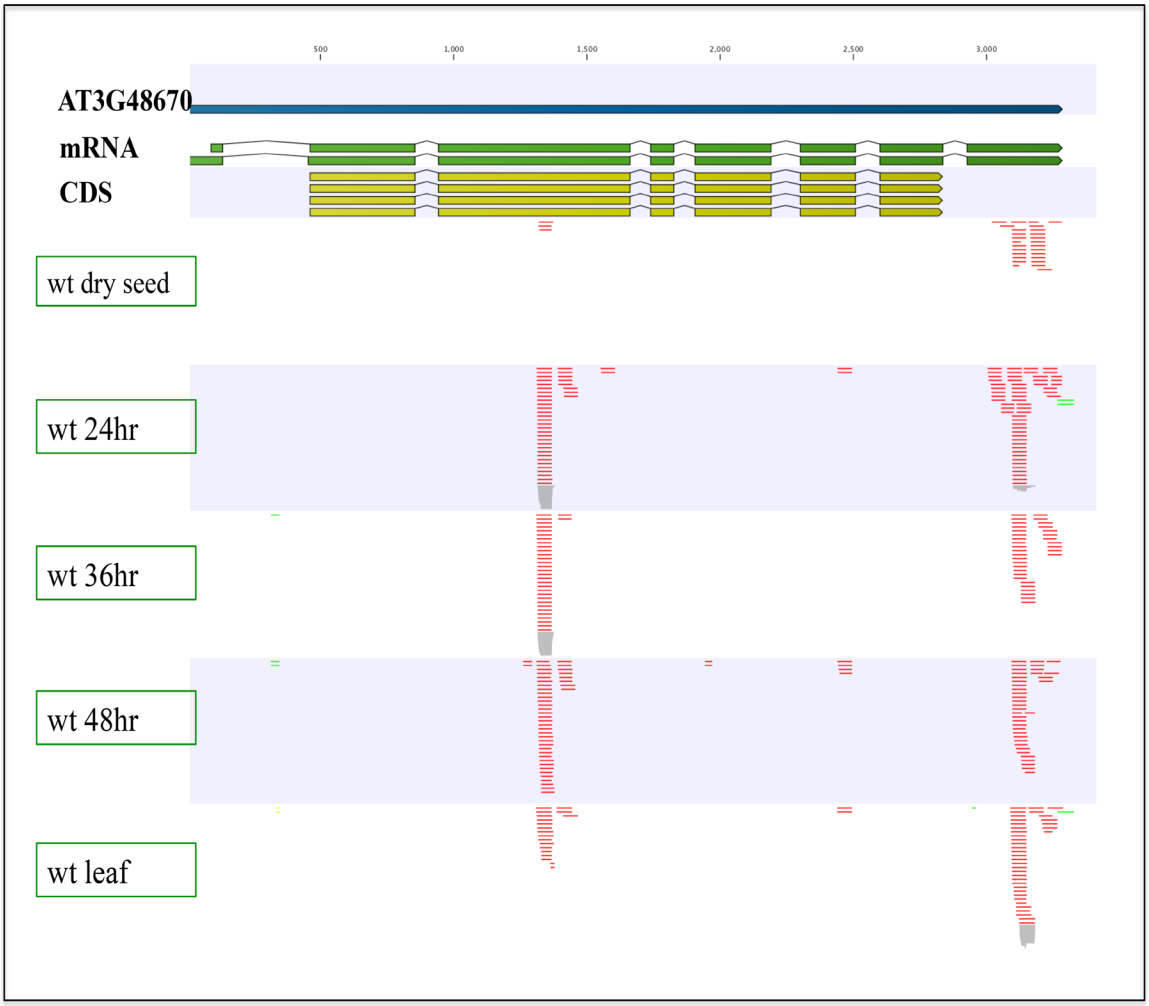


Figure 3.8 The CDS APA during seed germination: AT3G48670

A graphical depiction of the position and frequency of use of CDS versus 3' UTR APA sites in AT3G48670 in different tissues and germination stages, 24hr: 24hr imbibed seeds. 36hr: 36hr imbibed seeds. 48hr: 48hr imbibed seeds. Each of the red lines represents a read. The position of the red columns represents the location of the reads. The number of lines represents the expression level. The gray zones beneath the red polyadenylated tags indicate that many more tags than are shown also map to the respective genomic positions.

Figure 3.9 The CDS APA during seed germination: AT3G14980
A graphical depiction of the position and frequency of use of CDS versus 3' UTR APA sites in AT3G14980 in different tissues and germination stages, 24hr: 24hr imbibed seeds. 36hr: 36hr imbibed seeds. 48hr: 48hr imbibed seeds. Each of the red lines represents a read. The position of the red columns represents the location of the reads. The number of lines represents the expression level. The gray zones beneath the red polyadenylated tags indicate that many more tags than are shown also map to the respective genomic positions.

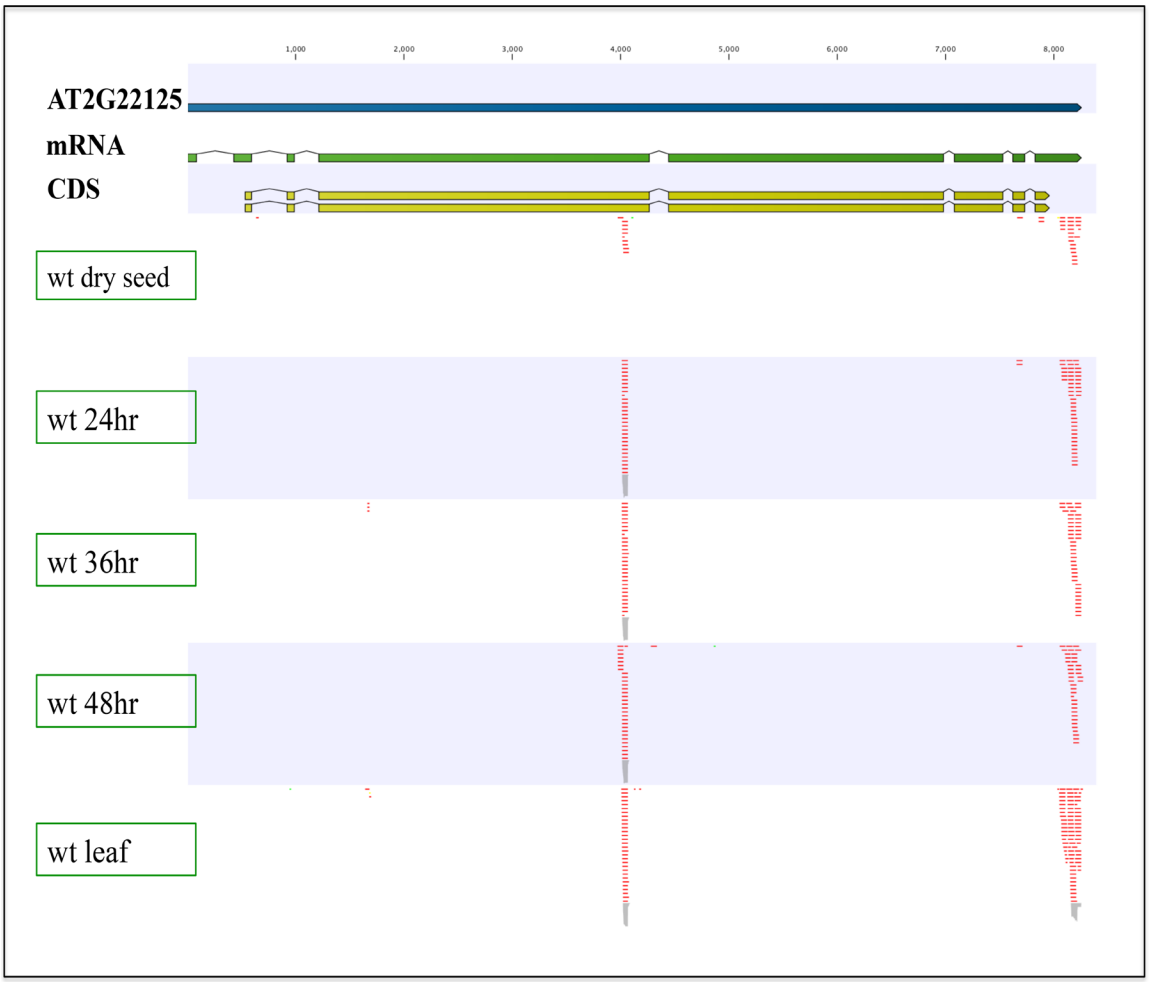


Figure 3.10 The CDS APA during seed germination: AT2G22125

A graphical depiction of the position and frequency of use of CDS versus 3' UTR APA sites in AT2G22125 in different tissues and germination stages, 24hr: 24hr imbibed seeds. 36hr: 36hr imbibed seeds. 48hr: 48hr imbibed seeds. Each of the red lines represents a read. The position of the red columns represents the location of the reads. The number of lines represents the expression level. The gray zones beneath the red polyadenylated tags indicate that many more tags than are shown also map to the respective genomic positions.

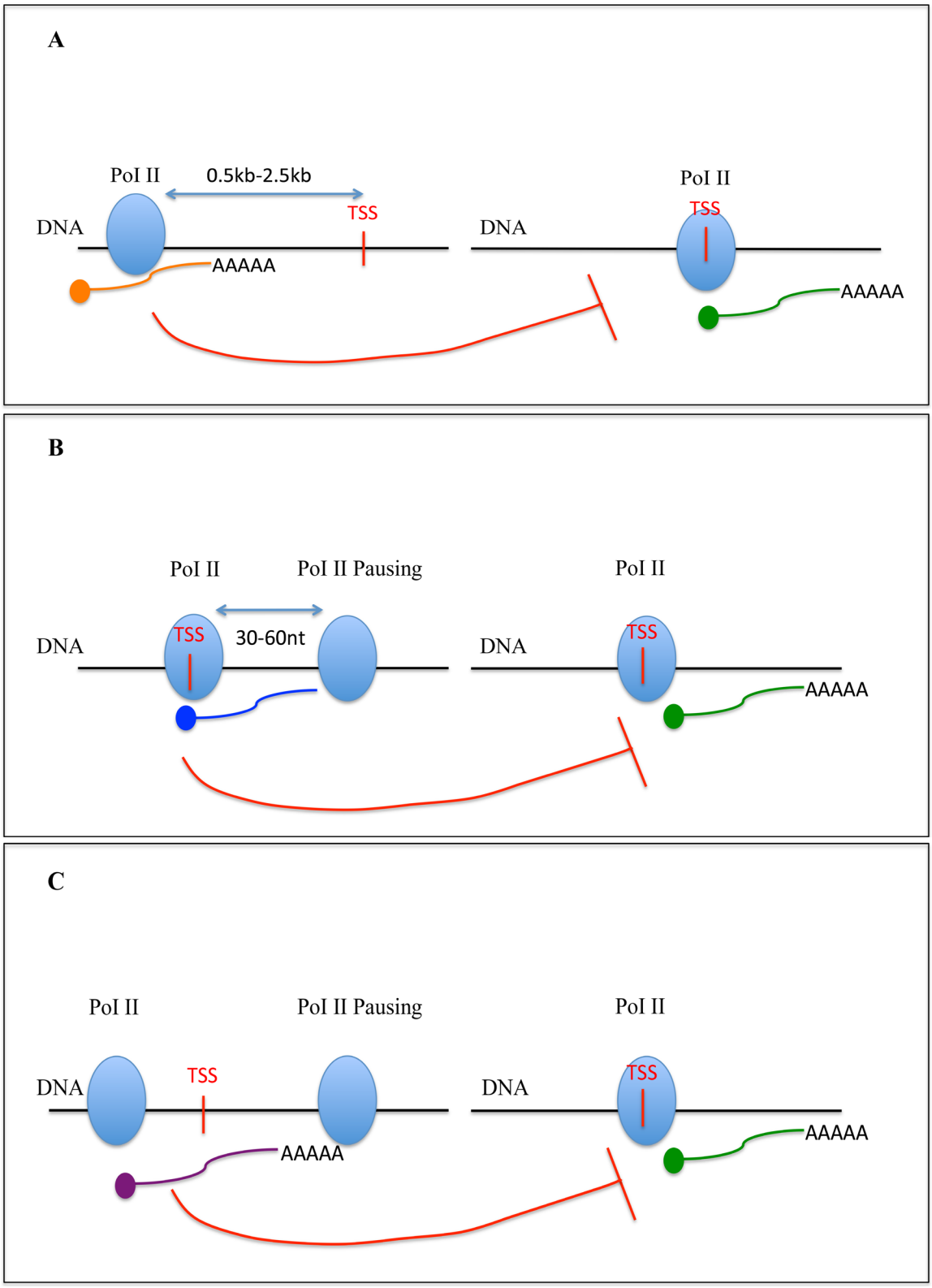


Figure 3.11 The possible mechanism explaining a function for transcripts terminated within the 5' UTR

Figure 3.11 (continued)

(A) In animal systems, a means of down regulating gene expression has been recently described (PROMoter uPstream Transcripts (PROMPTs)). This is depicted in the figure. DNA: chromosomal DNA. PoI II: DNA-dependent RNA polymerase II. TSS: transcription start site. Orange colored transcript: Promoter upstream transcripts transcribed from 0.5kb-2.5kb upstream of the transcription start site of the coding gene. Green-colored transcript: mRNA transcribed for this coding gene. The red colored line: repression.

(B) Another explanation might be that the 5' UTR APA arise from DNA dependent RNA Polymerase II pausing during pre-mRNA transcription, leading to aborted transcript polyadenylation and subsequent repression of transcription. The transcripts produced by PoI II pausing in animals are known to undergo such processing and downregulate production of full-length transcripts. DNA: chromosomal DNA. PoI II: DNA-dependent RNA polymerase II. TSS: transcription start site. Blue colored transcript: Short transcript produced from the coding gene but abortive elongation after 30 to 60nt from transcription start site. Green-colored transcript: mRNA transcribed for this coding gene. The red colored line: repression.

(C) In plants, it is possible that the 5' UTR APA transcripts commence just prior to the TSS and proceed through it but terminate in the 5' UTR and these abortive transcripts become polyadenylated, repressing full-length transcript production from this gene in the process. DNA: chromosomal DNA. PoI II: DNA-dependent RNA polymerase II. TSS: transcription start site. Dark purple colored transcript: Transcript mapped to the 5 'UTR of the coding sequence. Green-colored transcript: mRNA transcribed for this coding gene. The red colored line: repression.

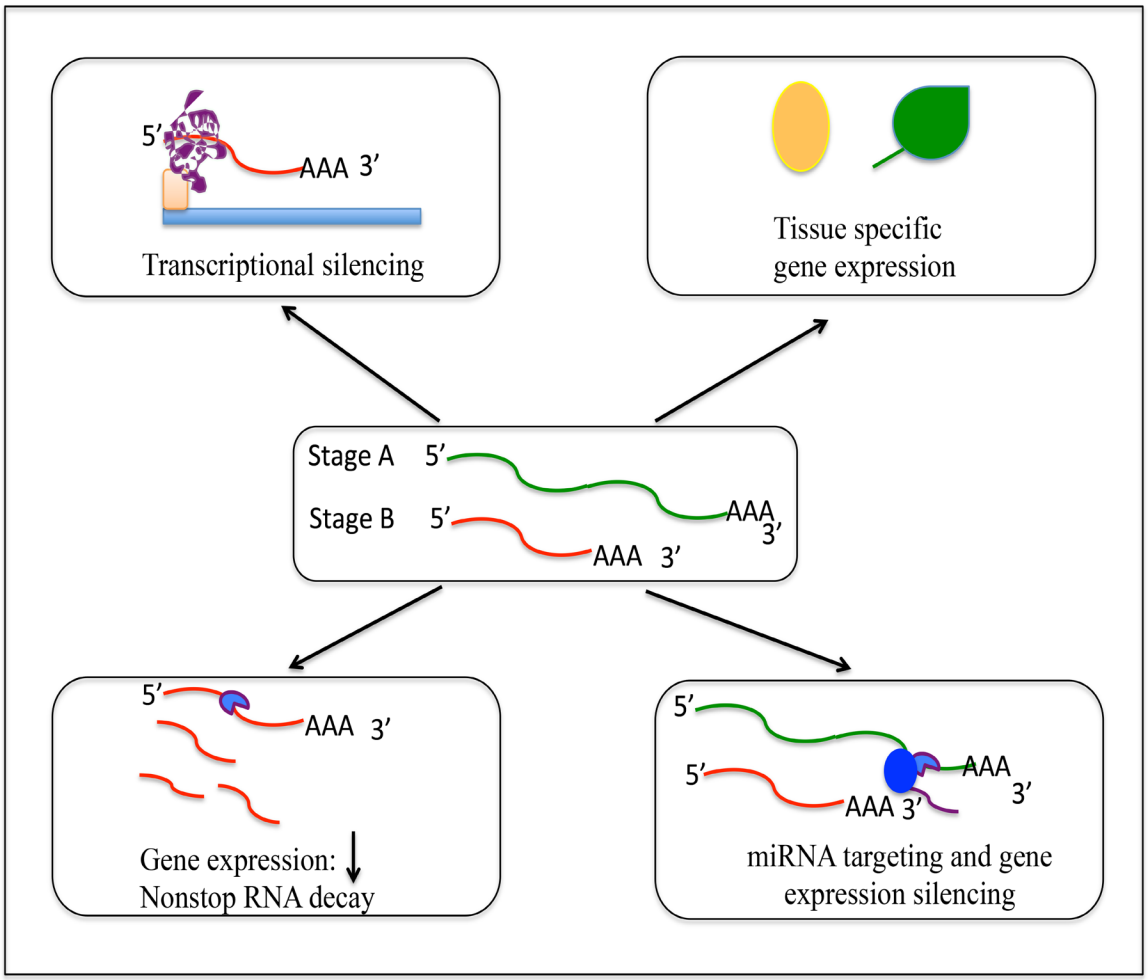


Figure 3.12 A model describing the potential functions of APA among different developmental stages in plants

Chapter Four: The role of polyadenylation in seed germination: defining the trans(crypto)me

4.1 Introduction

Approximately 70% of our food comes from seeds (243). The growing global population is projected to increase annual cereal consumption for food alone by a billion metric tons in the next 30 years (244). This, and more recent demands for seed as biofuel feedstock, continues to push the capacity of humanity to produce seeds sufficient to meet our ever increasing needs (245). Thus, the requirement to understand all aspects of seed biology is increasingly important (246). Seed germination, a complex developmental process under the control of many genes (247), is one of those aspects. A more sophisticated understanding of the mechanisms occurring during germination would lead to applications for improving the establishment of crop plants and preventing preharvest sprouting (248).

Many groups have reported that seeds contain a sizeable and diverse population of stored mRNA which can be the "source" of new protein production before nuclear transcription becomes active in the early germination stages (198, 207, 209-212, 215, 249). In addition, it has been reported that seed germination is sensitive to inhibitors of polyadenylation (cordycepin) and translational (cycloheximide) but less sensitive to transcription inhibitors such as actinomycin D and alpha-amanitin (208, 250) (Chapter 1, Appendix 1.2) indicating that seeds may complete germination without the help of *de novo* synthesized RNA but do require mRNA

polyadenylation and subsequent translation. Harris and Dure (207) raised the possibility that stored unadenylated mRNAs (RNAs with a poly(A) tail less than 15 nucleotides) exist in dry seed and can be repolyadenylated by cytoplasmic poly(A) polymerase(s) in the early stages of germination in cotton.

The polyadenylation of mRNAs is usually thought of as a process that takes place in the nucleus. It occurs during the maturation of pre-mRNA (see Chapter 1) and acts as a gatekeeper, preventing the export of pre-mRNA to the cytoplasm until it has been properly processed (2, 85). However, mRNA polyadenylation also occurs in the cytosol (156, 171). Indeed, the polyadenylation of specific mRNAs in the cytoplasm plays important roles in regulating gene expression during Xenopus oogenesis and early development, stages when the nucleus is not "active" (156, 171). During oocyte development, a population of maternal mRNAs is synthesized and stored for later use in particular development stages. These maternal mRNAs typically have short poly(A) tails (20 to 40 nucleotides) and are translationally repressed (156). Upon oocyte maturation or following fertilization, these masked mRNAs are elongated by cytoplasmic poly(A) polymerase(s) to regain their long poly(A) tail (80-250 residues) and the mRNAs are thus activated for translation (156). The cytoplasmic polyadenylation resulting in activation of those stored mRNAs, which encode key cell cycle regulators in Xenopus (171), is a process that is essential for meiotic maturation and early development.

For these reasons, it is important and interesting to identify stored unadenylated mRNAs (mRNAs typically have short poly(A) tails less than 25nt) in Arabidopsis and demonstrate the roles (if any) their translated products might have in seed germination. Using a combination of transcriptional inhibitors and a modified RNA-seq strategy (see Chapter 2), 273 putative stored, unadenylated mRNAs have been identified. Gene ontology results show that many of these stored, unadenylated mRNA candidates can generate proteins that function in translation. These results indicate that translation might be the first cellular process that is upregulated during seed germination and the transcripts encoding ribosomal proteins might be one of the major components of the stored unadenylated RNA pool. Also identified were 210 putative *de novo* synthesized mRNAs. Genes from this list were enriched in transcripts producing cell wall-, transporters-, toxin- and stress related- proteins. 547 putative stored, degraded mRNAs genes were identified. Transcripts encoding proteins involved in lipid storage and stress responses are overrepresented in this gene list.

4.2 Results

4.2.1 The completion of seed germination in 100μM alpha-amanitin

To identify stored, unadenylated RNAs in Arabidopsis seed, a seed germination experiment was performed in the presence or absence of alpha-amanitin. Alpha-amanitin is an amatoxin produced by the toadstool *Amanita phalloideswe* (251). Alpha-amanitin inhibits RNA synthesis catalyzed by DNA-dependent RNA polymerase II through blocking translocation by binding to the bridge helix (251-

255). Because alpha-amanitin is a relatively large molecule (Mr 919.0), it might have difficulty accessing the embryo that is covered by a seed coat composed of five cell layers in the integuments (256). To circumvent this, the *transparent testa2-5* (*tt2-5*, SALK_005260) mutant allele in the Columbia background was used, similar to the approach described in Rajjou *et al*. (213). This mutant features a permeable seed coat because the third cell layer lacks proanthocyanidin (256-258). It was expected that, in seeds imbibed in alpha-amanitin where *de novo* transcription is lacking, polyadenylated transcripts detected only following imbibition ("absent" in mature dry seeds) would have to be repolyadenylated from mRNAs that were stored in a unadenylated form in the dry seed.

Most seeds imbibed in 100μM alpha-amanitin completed germination but did so more slowly than they did in water. Accordingly, for these studies, seeds imbibed in 100μM alpha-amanitin were allowed to proceed for 12-, 48-, or 120-hours, with the 120 hr time point being sufficient for the seeds to complete germination. Seeds imbibed in water were allowed to proceed for 12, 24, or 72 hours, the last time of which is sufficient to complete germination. RNA was isolated from these samples and poly(A) tags (PATs) prepared and sequenced as described in Chapter 2.

The sequencing output was processed, mapped to the Arabidopsis reference genome and analyzed using CLC Genomics Workbench (see Methods and Materials of Chapter 2). The initial analysis involved determination of gene expression levels and comparisons of the individual sequencing samples using the RNA-seq tool in

CLC. To better study variability amongst the different replicates and developmental stages, a Principle Component Analysis was conducted (much as described in Chapters 2 and 3); for this the *tt2*-5 samples were compared with the wild-type seed samples described in Chapter 3 and the wild-type leaf samples described in Chapter 2. When this was done, the treatments and developmental stages could be resolved into three clusters (Figure 2.6). The dry seed samples all clustered in one sector along with the 12hr samples with or without alpha-amanitin (Figure 2.6). The 36hr and later water samples cluster in the opposite sector (Figure 2.6). The other alpha–amanitin time points are scattered in the middle with some of the replicates for the 24hr water time points (Figure 2.6). These results show that the PAT protocol was generally very reproducible as the replicates from same sample are clustering with each other. In addition, they indicate that gene expression in the presence of alpha-amanitin proceeds to a certain stage, similar to that seen in the 24hr imbibed seeds in water, and then ceases further progress as 120hr imbibed seed in alpha-amanitin are still clustered with 48hr imbibed seed in alpha-amanitin and 24hr imbibed seeds in water. Overall, as most of *tt2* water samples group with wild-type water samples and *tt2* dry seeds group with wild-type seeds, it would appear that *tt2* seeds are similar to the wild-type seeds.

4.2.2 Genome-wide characterization of poly(A) site distribution in the samples from germination stages

To study global aspects of poly(A) site choice, the genome-wide distribution of PATs was determined using CLC Genomic Workbench. The genome-wide distribution of *tt2* PATs was more diverse than expected (Table 4.1). The percentage of PATs from the *tt2* dry seed and *tt2* seed germination stages that mapped to the 3' UTR (67.87% and 80.94%) was lower than that seen in the wild-type (77.14% to 85.98% of which mapped to the 3' UTR; Chapter 3 Table 3.1). Additionally, in *tt2* seeds that had been imbibed for 48hr in alpha-amanitin, only 31.27% of the PATs mapped to the 3' UTR (Table 4.1). One reason for this disparity is because between 10 and 23% of all PATs in the *tt2* samples mapped to protein-coding regions (CDS in Table 4.1). This is in contrast to what was seen in the wild-type (between 6.73% and 12.08% of PATs mapping to the CDS; Chapter 3 Table 3.1). In addition, between 3.37% and 15.77% of PATs in samples from *tt2* dry seed and *tt2* germination stages mapped to introns with the exception of *tt2* 48hr imbibed seeds in alpha-amanitin, in which approximately 44% of PATs mapped to introns (Table 4.1). As it is not possible to directly map to the intron region reference database of Arabidopsis in CLC, the PATs mapping to ribosomal RNAs may be erroneously categorized as intron mapped PATs. Further inspection revealed that ribosomal RNAs are contaminating the *tt2* 48hr imbibed seeds in alpha-amanitin. The percentage of PATs mapping to introns in *tt2* was somewhat higher than seen in the wild-type (Chapter 3, Table 3.1). In contrast, fewer *tt2*-derived PATs mapped to the 5' UTR than for wild-type (Tables 3.1 and 4.1). These results suggest a different distribution of PATs mapping between

tt2 and WT seeds; this may be a result of the different germination conditions between seeds submerged to water (*tt2*) and seeds germinated on wetted filter paper (245).

4.2.3 Gene expression analysis to identify putative stored, unadenylated RNAs

To identify genes whose mRNAs are in a stored, unadenylated form in dry seed, a detailed expression analysis was performed using the gene expression results from PAT abundance. If a gene had a normalized expression level of 10 tags per million (tpm) or greater in any of the *tt2* germination stages or *tt2* dry seed, the gene was defined as an expressed gene. Using this criterion, 14,648 genes were identified as expressed in at least one stage of germination or development. It was thought that in alpha-amanitin treated samples; newly appearing polyadenylated transcripts (<10 tpm in dry seed, >10 tpm after imbibition in alpha-amanitin) would have to be repolyadenylated from those transcripts that were stored in a unadenylated form in dry seed. However, in seeds imbibed in water, newly appearing polyadenylated transcripts (<10 tpm in dry seed, >10 tpm after imbibition in water) could come from both those stored in a unadenylated form in dry seed and also from mRNAs newly synthesized by RNA polymerase II. Thus, putative stored, unadenylated RNAs should have <10 tpm in dry seed and >10 tpm in both alpha-amanitin treated- and water treated-samples.

Based on these criteria, genes that showed more than a five-fold increase in expression level between *tt2* dry seed and any germination stage (in both water

and alpha-amanitin treated seed) were identified. This list was filtered to retain only genes with expression levels greater than 20 tpm in at least one of the alpha-amanitin treated samples. There were 898 genes that fit these criteria. From this list, genes with expression differences between samples germinated in water and alpha-amanitin that were less than two-fold were selected. Finally, from this latter list, genes with expression values of at least 20 tpm in one or more stages were identified. This resulted in a list of 273 genes that were designated as "high confidence, stored, unadenylated RNA candidates" (Appendix 4.1).

This list was assessed by visual inspection and by using gene ontology analysis. The high confidence, putative stored, unadenylated genes were significantly enriched for genes that encode ribosomal proteins and other translation-related proteins (Figure 4.1). In particular, 39 genes that encode cytoplasmic ribosomal proteins were present in the gene list of high confidence stored, unadenylated RNAs (Appendix 4.2). Eukaryotic 80S ribosomes consist of four ribosomal RNAs and 80 proteins present in either the small (40S) or large (60S) ribosomal subunits (259). There are 249 genes identified in Arabidopsis encoding 80 cytoplasmic, ribosomal proteins (32 small subunit and 48 large subunit proteins) (259). This is because a small gene family encodes each of the ribosomal proteins. Of the 39 ribosomal protein genes that encode stored, unadenylated mRNA, 28 encode large (60S) ribosomal subunit proteins (RPLs) (Appendix 4.2). These 28 genes encode 21 different RPLs, about 44% of the total complement of RPLs (Appendix 4.2). There are 11 of the ribosomal protein genes that yield stored, unadenylated mRNA

encoding small subunit (40S) ribosomal proteins (RPSs) (Appendix 4.2). Those 11 genes encode 10 different RPSs, about 32% of the total complement of RPSs (Appendix 4.2).

To better assess the expression level of these RNAs during the seed developmental and germination stages, the relative expression level of globular-, heart-, torpedo-, or cotyledon-stage embryos, and dry- or 24hr imbibed-seed (Appendix 4.3), was obtained from the Arabidopsis eFP Browser at bar.utoronto.ca (260). The gene expression levels were calculated for the 39 ribosomal protein genes from these six stages. Expression levels were normalized based on the average gene expression level across all tissues (Appendix 4.4) and plotted using the six developmental stages as the abscissa and normalized expression level as the ordinate in order to determine whether there were any trends/patterns in expression level that occurred within these stages (Figure 4.2). For most genes, the normalized expression level of the 39 ribosomal protein genes dramatically decreased in cotyledon stage embryos and the dry seed, with the dry seed having the lowest value which dramatically increased in 24hr imbibed seeds (Figure 4.2).

To further analyze the possible location of these ribosomal protein transcripts, their relative expression level in different tissues such as the embryo proper, cellularized endosperm, chalazal endosperm, chalazal seed coat and general seed coat (Appendix 4.5) was obtained from the Arabidopsis eFP Browser at bar.utoronto.ca (260). The average gene expression level for those 39 genes from

the five tissues mentioned above was calculated in the linear cotyledon stage. Based on the prior data (Figure 4.2), this stage would still retain the poly(A) tract on these messenger RNAs as it is prior to their precipitous decline. The expression level was normalized based on the average gene expression level (Appendix 4.6) in order to ascertain whether there were any trends/patterns in gene expression of the ribosomal protein genes in different tissues at the same developmental stage. A plot was constructed using the five tissues as the abscissa and the normalized gene expression value as the ordinate (Figure 4.3). Most of the genes encoding ribosomal proteins were expressed in the embryo of the linear cotyledon stage indicating that these ribosomal protein transcripts are potentially stored in the tissues of the embryo (Figure 4.3).

Overall, these results indicate that transcripts encoding ribosomal proteins are a major component of the stored, unadenylated RNA pool, and that translation might be a major priority for rapid activation early during seed germination.

4.2.4 Identification of putative *de novo* synthesized mRNAs

To better understand the population of RNAs that are newly synthesized in the nucleus, gene expression analysis was performed to identify putative *de novo* synthesized mRNAs. By definition, *de novo* synthesized mRNAs are newly synthesized in the nucleus, plastid or mitochondria. Thus, a typical *de novo* synthesized mRNA should have <10 tpm in dry seed and in all alpha-amanitin-treated samples, but have >10 tpm in water treated samples. To generate lists of

genes that encode *de novo* synthesized mRNAs, genes with the following properties were identified: the 48hr amanitin/24hr water and 120hr amanitin/72hr water expression ratios are less than 1.0; the *tt2* dry seed/ *tt2* 12hr water, *tt2* dry seed/ *tt2* 24hr water, or *tt2* dry seed/*tt2* 72hr water ratios are less than 1.0; and the expression level at any stage (other than dry seed) after the commencement of germination of *tt2* seed in water was greater than 20 tpm. This yielded a list of 1288 genes transcribing *de novo* synthesized mRNAs. To further identify higher-confidence *de novo* synthesized mRNA candidates, the expression ratio limit was set at 0.5; using this filter, 210 high confidence genes transcribing *de novo* synthesized mRNAs were identified (Appendix 4.7). A gene ontology analysis of the list of high confidence genes producing *de novo* synthesized transcripts showed enrichment for those genes whose products were responsive to inorganic substances, or involved in toxin-related functions [transporters, oxidative stress responsive proteins] and cell wall related-functions (Figure 4.4). Further analysis revealed that fifteen transporters are responsible for metal ion, ion and drug translocation (Appendix 4.8). Moreover, twenty-five genes have been identified to respond to inorganic substance such as toxins (e.g. herbicides), metal ions (e.g. cadmium ions) and nitric oxide as well as oxidative stresses (Appendix 4.9). Six GLUTATHIONE S-TRANSFERASES (GSTs) were present in this gene list. Furthermore, six genes have been identified to be involved in cell wall related functions (Appendix 4.10).

4.2.5 Other classes of mRNAs seen in dry seed

To determine additional classes of the reservoir of stored mRNAs genome-wide, gene expression analysis was performed to identify putative stored, degraded mRNAs. To generate lists of genes that encode such mRNAs, those that showed at least a five-fold decrease in expression in any germination stages (either in water or alpha-amanitin) when compared with *tt2* dry seed were identified. Genes fitting this criteria and with transcript abundance levels of at least 20 tpm in *tt2*-and WT-dry seed (Ch. 3) were thus identified. Genes on this list that showed a comparable decrease in expression in the WT germination study were then selected. In this way, 547 genes were identified as high confidence stored, degraded mRNA candidates (Appendix 4.11).

This list was subjected to gene ontology analysis and found to be enriched for genes encoding proteins involved in lipid storage, lipid localization and abiotic stress responses (Figure 4.5). Further analysis revealed that 18 genes related to lipid storage or localization functions were present in this list (Appendix 4.12). Eight are Oleosin family proteins and five are seed storage proteins of the albumin family that are also listed as lipid transporters. Seventeen out of these 18 genes are included on the ATH1 microarray allowing a survey of their transcript abundance during seed development. The relative expression level (transcript abundance in dry seed) of these genes in globular-, heart-, torpedo-, or cotyledon-stage embryos, and dry- or 24hr imbibed-seed for these lipid storage and localization genes was obtained from the Arabidopsis eFP Browser at bar.utoronto.ca. These expression levels were

normalized based on the average gene expression level across all six stages (Appendix 4.13) and plotted using the six developmental stages as the abscissa and normalized expression level as the ordinate (Figure 4.6). For most genes, the normalized expression level of these 17 lipid storage and lipid localization genes was high in the cotyledonary stage embryo and in the dry seed but dramatically decreased in 24hr imbibed seeds. These results corroborate those obtained with the PAT-based expression analysis results.

There were 34 genes listed in Appendix 4.6 that were also identified as producing products that were involved in abiotic stress responses, such as responses to heat, temperature stimulus, light intensity and oxidative stress (Appendix 4.14). Twelve of these genes encode heat shock related-proteins; e.g. the HSP20-like chaperones are abundant in this list. Generally, the gene expression analysis using the microarray data from the Arabidopsis eFP Browser at bar.utoronto.ca (Appendix 4.15, 4.16) indicates a decrease in transcript abundance for this category (abiotic stress related) of stored, degraded mRNAs from dry seed to 24hr imbibed seeds (Figure 4.7) in a pattern similar to that of the lipid storage and localization gene list.

Overall, the results presented above indicate that the PAT data are similar to those generated from microarray data.

4.3 Discussion

4.3.1 Ribosomal protein RNAs might be major components of the stored, unadenylated RNA pool

Translation plays important roles during seed germination (213) (Appendix 1.2). Reduced seed germination rates and seed viability have been observed a mutant deficient in eukaryotic initiation factor 4G (261). Moreover, translation inhibitors abolish the completion of seed germination (207, 208, 213, 250). It is reasonable to assume that transcripts encoding proteins involved in translation accumulate in dry seeds and are translated early during seed germination to rebuild the translation machinery as soon as possible.

In plants, it has been reported that large (L3, L16) subunit ribosomal proteins are not present as mature mRNAs in the Maize embryonic axis (212). We identified 273, high confidence, stored, unadenylated mRNA candidates (Appendix Table 4.1). The mRNAs encoding proteins involved in translation, including many ribosomal proteins, are overrepresented in this gene list (Figure 4.1). Further analysis revealed that 39 genes were “translation related” (Appendix Table 4.2). Additional analysis showed that these 39 genes encode 31 different ribosomal proteins. Approximately 38.75% (31/80) of the total ribosomal proteins required to decorate a functional ribosome (Table 4.2) were present in unadenylated form and stored in dry seed. These results indicate that ribosomal protein transcripts might be major components of the stored, unadenylated RNA pool.

The normalized values of stored, unadenylated transcripts encoding ribosomal proteins decreased from the torpedo to cotyledon-stage of embryogenesis and again from cotyledon-stage to the dry seed stage, but increased in 24hr imbibed seeds (Figure 4.5). Because microarray data is generated using polyadenylated RNA, the unadenylated RNAs are not recovered by oligo d(T) selection and not labeled using oligo d(T) and thus, although they may be present, they would register no (or very low) expression values. Therefore, it is possible that the decrease in gene expression of those RNAs in cotyledon-stage embryos and dry seeds is a cryptic event resulting from deadenylation of the polyadenylated form of those RNAs. The cytoplasmic de-adenylase, polyadenine ribonucleases (PARN) is responsible for deadenylation in animals (156). AtPARN, the Arabidopsis homolog of animal PARN has been identified (262). Therefore, AtPARN might be involved in deadenylation of some ribosomal protein transcripts during seed maturation and after-ripening. During early seed germination, these stored, unadenylated ribosomal protein mRNAs may be quickly repolyadenylated by an unknown plant cytoplasmic polyadenylation mechanism akin to that in animal systems (see Chapter 1). This would lead to the translation of these messages encoding ribosomal proteins in order to build the translation machinery. Further tissue localization analysis revealed that most of these ribosomal protein genes are expressed in the embryo in the linear cotyledon stage (Figure 4.3) and this result was coincident with an early report that two large (L3, L16) subunit ribosomal proteins are not present as mature RNAs in the Maize embryonic axis (212).

Therefore, it is possible that Arabidopsis may undergo similar events as documented in animal systems.

4.3.2 The *de novo* mRNA candidates and their possible functions/mechanisms

Although alpha-amanitin treatment delays the radical protrusion of seeds, the *tt2* seed do complete seed germination after 120hr imbibition in 100μM alpha-amanitin (Dr. Pratap Kumar Pati, pers. Comm. 2012) indicating that *de novo* transcription is not required for seed germination. Similar observations have been reported (213). However, gene expression in the 48hr and 120hr imbibed seeds in alpha-amanitin resembles that seen in 24hr imbibed seeds in water (Figure 2.6), indicating that mRNAs from *de novo* transcription may function in later germination stages. Therefore, the *de novo* mRNAs may have fewer functions in early seed germination stage but have some functions in later germination stages.

Transporters, inorganic substance related proteins and cell wall related functions were overrepresented in the *de novo* mRNAs candidate gene lists (Figure 4.4). Further analysis revealed that these transporters are responsible for zinc, iron, nitrate and copper transport. Translocation of these metal ions might play important roles in seedling establishment or late seed germination stages. Glutathione S-transferases (GSTs) are encoded by another prominent set of *de novo*-synthesized mRNAs. GSTs are induced by abiotic stresses, suggesting that *de novo*-synthesized mRNAs might reflect abiotic stress regulation in seedling establishment or later seed germination stages. Xyloglucan endotransglucosylase/

hydrolases (XTH) are also over-represented in the *de novo* mRNA candidate gene list, suggesting that *de novo* mRNAs might regulate cell elongation and expansion in seedling establishment or late seed germination stages.

4.3.3 The stored degraded mRNA candidates and their possible functions/mechanisms

There were 547 genes encoding mRNAs that are stored in dry seed and are apparently degraded upon imbibition (Appendix 4.11). mRNAs that encode lipid storage, lipid localization and abiotic stress responsive proteins were abundant in this set of genes. Detailed studies for these lipid storage or localization proteins revealed that most of them were oleosins and the 2S seed storage albumin1 (SESA1) (Appendix 4.12). Oleosins serve to control the size of oil-bodies through the prevention of coalescence during dehydration of seeds and pollen (263). Mutants deficient in Oleosin 1 lead to the formation unusually large oil bodies, which alter the accumulation of lipid and proteins and cause delay of seed germination (264). It was reported that the oil bodies experience cytoplasmic compression as water potential decreases during the seed maturation stage especially late maturation stages as water is withdrawn from the cytoplasm (264, 265).

Another group of stored degraded mRNAs encoded stress responsive proteins. Heat shock related proteins, especially HSP20-like chaperones, are the major component of this group (Appendix 4.14). In addition, At5g05410, a gene encoding dehydration-responsive element binding protein 2 (266), is in this group. During the late seed

maturation stage, dehydration occurs and leads to the decrease of moisture content on a fresh weight basis (267). At this stage, the cytoplasm is condensed and proteins aggregate (267). Small heat shock proteins (HSPs) were reported to accumulate late in seed development and may be involved in mitigating the aggregation effect by acting as molecular chaperones to stabilize large complexes (267).

Therefore, it is interesting to ask why these classes of mRNAs are degraded during seed germination instead of earlier, during seed maturation. Since Oleosins and HSPs function in controlling the size of oil bodies (264) and protect other proteins (267) respectively during the desiccation of seeds in the late seed maturation stage, mRNAs from those genes are needed during this stage to produce these proteins. However, with desiccation, the seed contains very low moisture content on a fresh weight basis (approximately 5 to 10% in orthodox seed) (267) and it is possible that the mRNA degradation pathway has been "shut down" before mRNAs that encode these proteins have been degraded. Therefore, those mRNAs are retained in the dry seed and are degraded after the mRNA degradation pathway is "re-activated" upon imbibition.

4.3.4 Summary

In summary, the results reported here lead to the identification of 273 stored, unadenylated mRNA candidates, 210 *de novo* mRNA candidates, and 547 stored, degraded mRNA candidates. The major component of stored unadenylated mRNAs was that encoding ribosomal proteins. These stored, unadenylated mRNAs may

accumulate in seed developmental stages and the poly(A) tail of these mRNAs may be shortened by de-adenylase and stored in the dry seed. Because they lack a poly(A) tail, they are undetectable by techniques that use poly(A) selection. Therefore, they apparently possess lower expression levels in dry seed in both our PAT data and microarray data. However, they are quickly activated by polyadenylation (presumably in the cytoplasm) and may help to restore the translational apparatus during early seed germination.

The major components of stored, degraded RNAs were mRNAs encoding Oleosins, lipid transporters and stress responsive proteins such as heat shock proteins. They are transcribed during late embryogenesis because the proteins they encode were functional in late seed development when dehydration occurs and during storage in the dry state, and these transcripts remain and are detected in mature, dehydrated seeds. However, after seed maturation, these mRNAs may not be degraded because the mRNA degradation system may shut down when seed moisture content is 5% fresh weight, because there is insufficient free water to permit hydrolysis and these mRNAs remain intact in the dry seed. However, upon imbibition, mRNA degradation becomes active again and these mRNAs are degraded.

The major components of *de novo* mRNAs were metal ion responsive transporters, abiotic stress responsive proteins such as Glutathione S-transferases (GSTs), and cell wall related proteins such as XTH. Arabidopsis seed can finish germination in

the presence of transcription inhibitors, indicating that *de novo* mRNA encode proteins that may have no function in seed germination *sensu stricto* but do contribute to the speed with which this event is completed. Because metal ion responsive transporters, GSTs and XTH are abundant in the *de novo* mRNA gene list, and certain metals are known cofactors for proteins of the photosynthetic apparatus, it is possible that they are involved in seedling development.

Overall, the following model is described (Figure 4.8). The stored, degraded mRNAs and stored, unadenylated mRNAs accumulated during seed development stages. During the late seed maturation stages, protein encoded by stored degraded mRNA function in oil body formation and protein protection as dehydration occurs. In addition, stored, unadenylated mRNAs are deadenylated (and not degraded) at this time. This model predicts that the mRNA degradation pathway is shut down as moisture content decreases leading to lower free water availability for hydrolysis. During the early seed germination stages following imbibition, the stored, unadenylated mRNAs are re-polyadenylated to allow rapid synthesis of ribosomal proteins to restore the translation machinery to full activity. At the same time, the mRNA degradation pathway is reactivated due to greater water availability; leading to the degradation of so-called stored, degraded mRNAs. Following imbibition, but before the completion of seed germination, the transcription system was active and *de novo* mRNA was made. These mRNAs accumulated and may function in late germination stages and in seedling development.

4.4 Methods and material

4.4.1 Seed germination experiment

A *transparent testa* mutant (SALK_005260; *tt2-5*) in the *Arabidopsis thaliana* (Col.) background was used in this experiment. The seeds were first subjected to 4°C for 3 d to alleviate dormancy before transfer to 25°C. Approximately 0.15g of seeds were placed in 1.5mL of water or water containing 100μM alpha-amanitin in one well of a 24FB well TC plate (Sarstedt). The plates were incubated at 25°C in a shaker at 180-rpm speed under 24hr constant light conditions with a light intensity of ~ 8 $\mu mole \bullet m^{-2} \bullet Sec^{-1}$ for the indicated periods of time (0-120 hrs). RNA was isolated from imbibed seeds and dry *tt2* seeds as described in Chapter 2. Using these RNA samples, poly(A) tags were prepared and sequenced, again as described in Chapter

Table 4.1 Distribution of PAT in different gene regions

Genomic regions as defined in the TAIR10 database. As explained by Wu *et al*. (2011) (235), the 3'UTR were extended by 120 nucleotides. [a]: total number of curated poly(A) site tags that map to the respective genomic regions. [b]: percentage of total PATs that fall within the indicated regions. *tt2* dry seed, and different imbibed seed germination stages under both water (W) and alpha-amanitin (A) are calculated separately.

Sample name	Regions	3' UTR	CDS	Intron	5' UTR
tt2 dry seed	PAT No.[a]	855942	215376	98311	15814
	PAT(%)[b]	72.21	18.17	8.29	1.33
tt2 12W	PAT No.	2933304	482114	163553	83929
	PAT(%)	80.08	13.16	4.47	2.29
tt2 24W	PAT No.	1401098	314642	325602	23020
	PAT(%)	67.87	15.24	15.77	1.11
tt2 72W	PAT No.	4347302	573045	668716	114401
	PAT(%)	76.22	10.05	11.72	2.01
tt2 12A	PAT No.	4926039	816365	205281	138005
	PAT(%)	80.94	13.41	3.37	2.27
tt2 48A	PAT No.	506627	375627	715278	22602
	PAT(%)	31.27	23.18	44.15	1.4
tt2 120A	PAT No.	2830057	425335	492055	93307
	PAT(%)	73.68	11.07	12.81	2.43

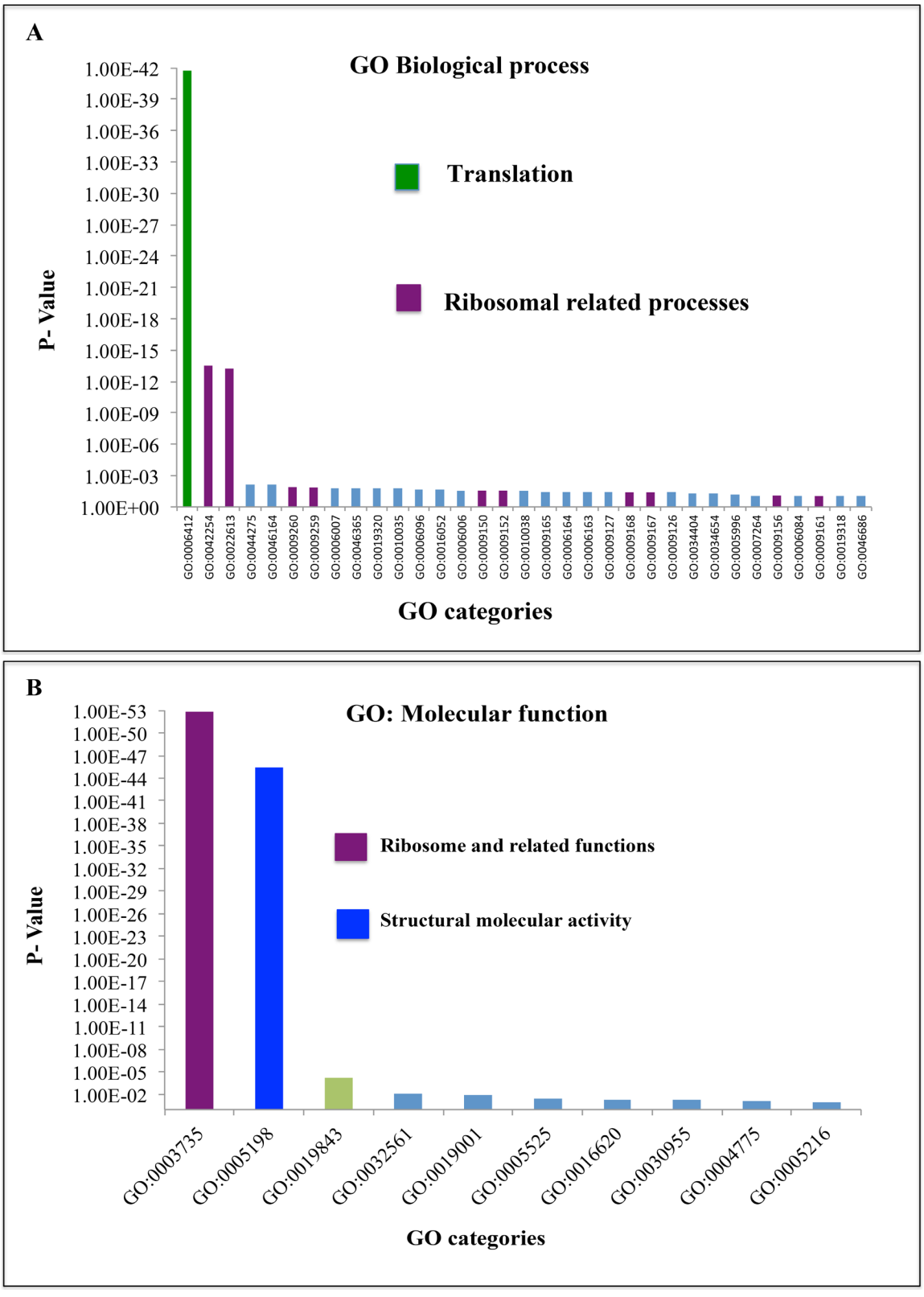


Figure 4.1 The stored, unadenylated RNA, high confidence candidate genes analyzed by Gene Ontology

Figure 4.1 (continued)

(A) Abundance of ribosomal and translation related proteins based on GO Biological processes. GO:0006412 = translation; GO:0044275 = cellular carbohydrate catabolic process; GO:0009259 = ribonucleotide metabolic process; GO:0019320 = hexose catabolic process; GO:0016052 = carbohydrate catabolic process; GO:0009152 = purine ribonucleotide biosynthetic process; GO:0006164 = purine nucleotide biosynthetic process; GO:0009168 = purine ribonucleoside monophosphate biosynthetic process; GO:0034404 = nucleobase-containing small molecule biosynthetic process; GO:0007264 = small GTPase mediated signal transduction; GO:0009161 = nucleoside monophosphate metabolic process; GO:0003735 = structural constituent of ribosome; GO:0005198 = structural molecule activity; GO:0019843 = rRNA binding; GO:0032561 = guanyl ribonucleotide binding; GO:0019001 = guanyl nucleotide binding; GO:0005525 = GTP binding; GO:0016620 = oxidoreductase activity, acting on the aldehyde or oxo group of donors, NAD or NADP as acceptor; GO:0030955 = potassium ion binding; GO:0004775 = succinate-CoA ligase (ADP-forming) activity; GO:0005216 = ion channel activity.

(B) The GO molecular function of the stored, unadenylated gene products. (B) GO:0003735 = structural constituent of ribosome ; GO:0005198 = structural molecule activity ; GO:0019843 = rRNA binding ; GO:0032561 = guanyl ribonucleotide binding ; GO:0019001 = guanyl nucleotide binding ; GO:0005525 = GTP binding ; GO:0016620 = oxidoreductase activity, acting on the aldehyde or oxo group of donors, NAD or NADP as acceptor ; GO:0030955 = potassium ion binding ; GO:0004775 = succinate-CoA ligase (ADP-forming) activity ; GO:0005216 = ion channel activity.

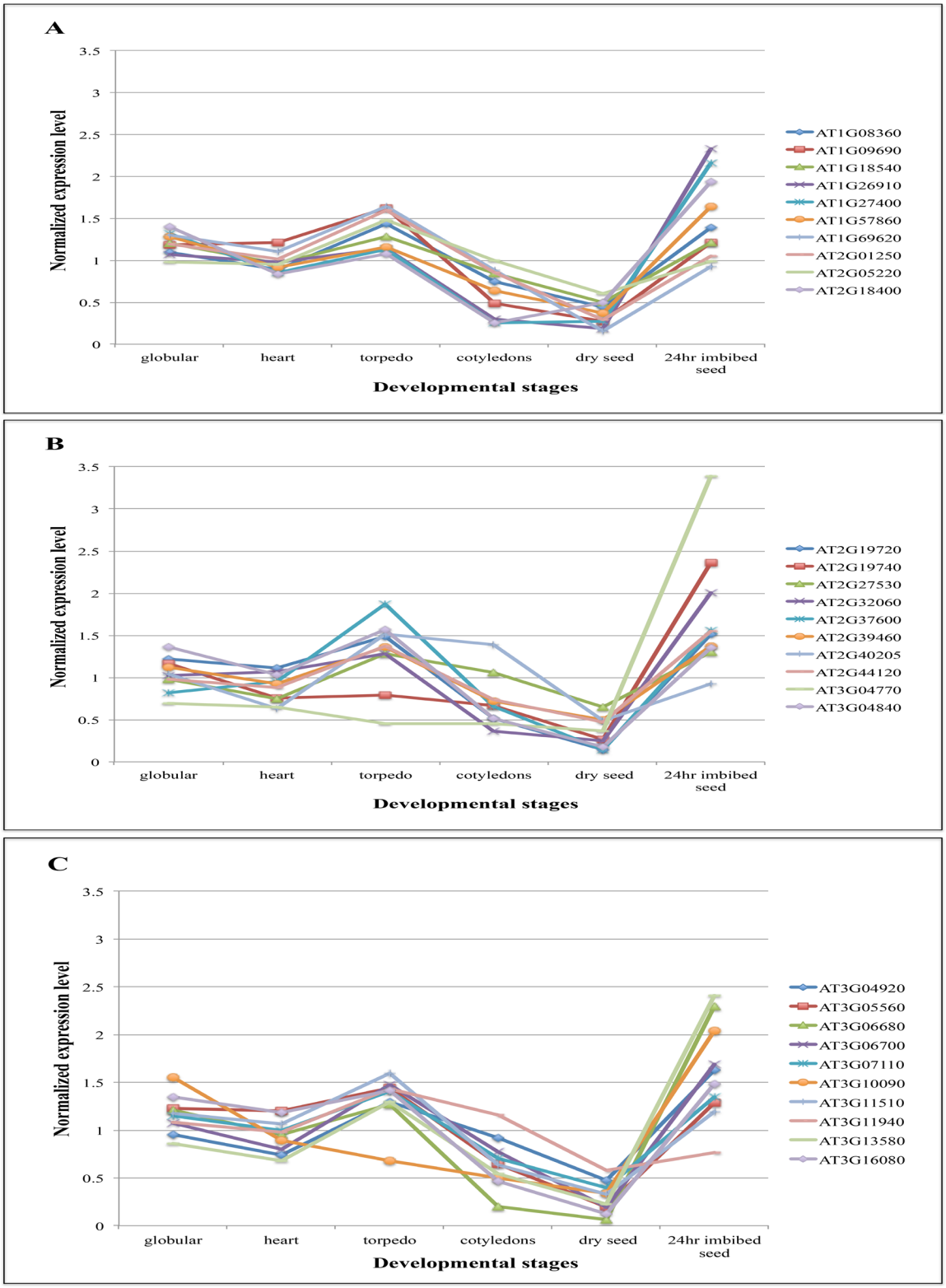


Figure 4.2 The normalized gene expression level for ribosomal protein genes from different developmental stages

Figure 4.2 (continued)

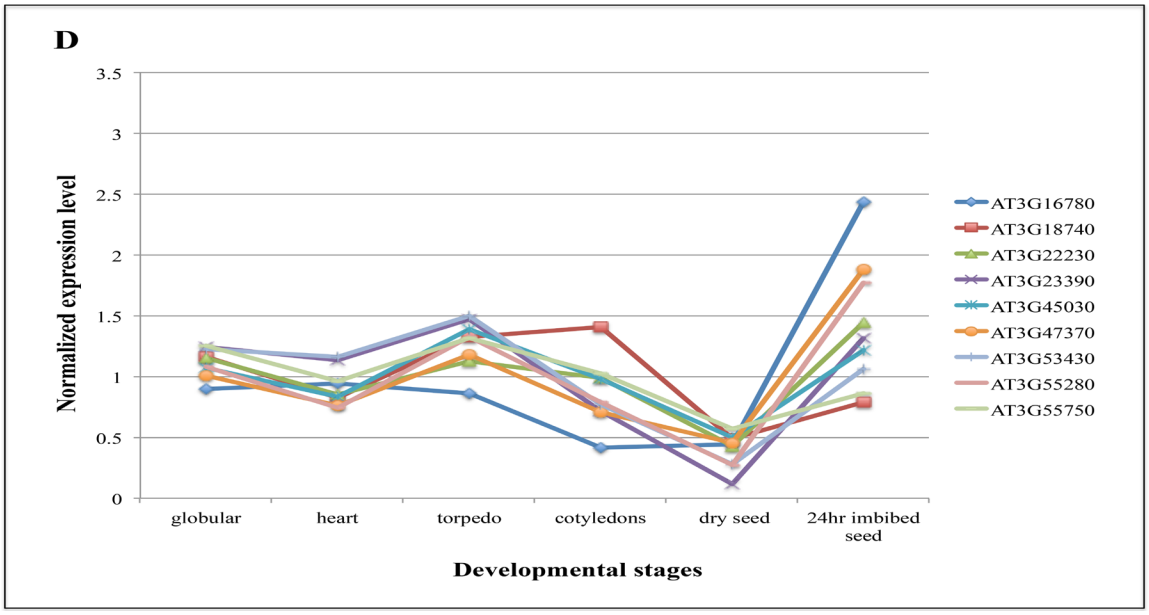


Figure 4.2 The normalized gene expression level for ribosomal protein genes from different developmental stages (continued).

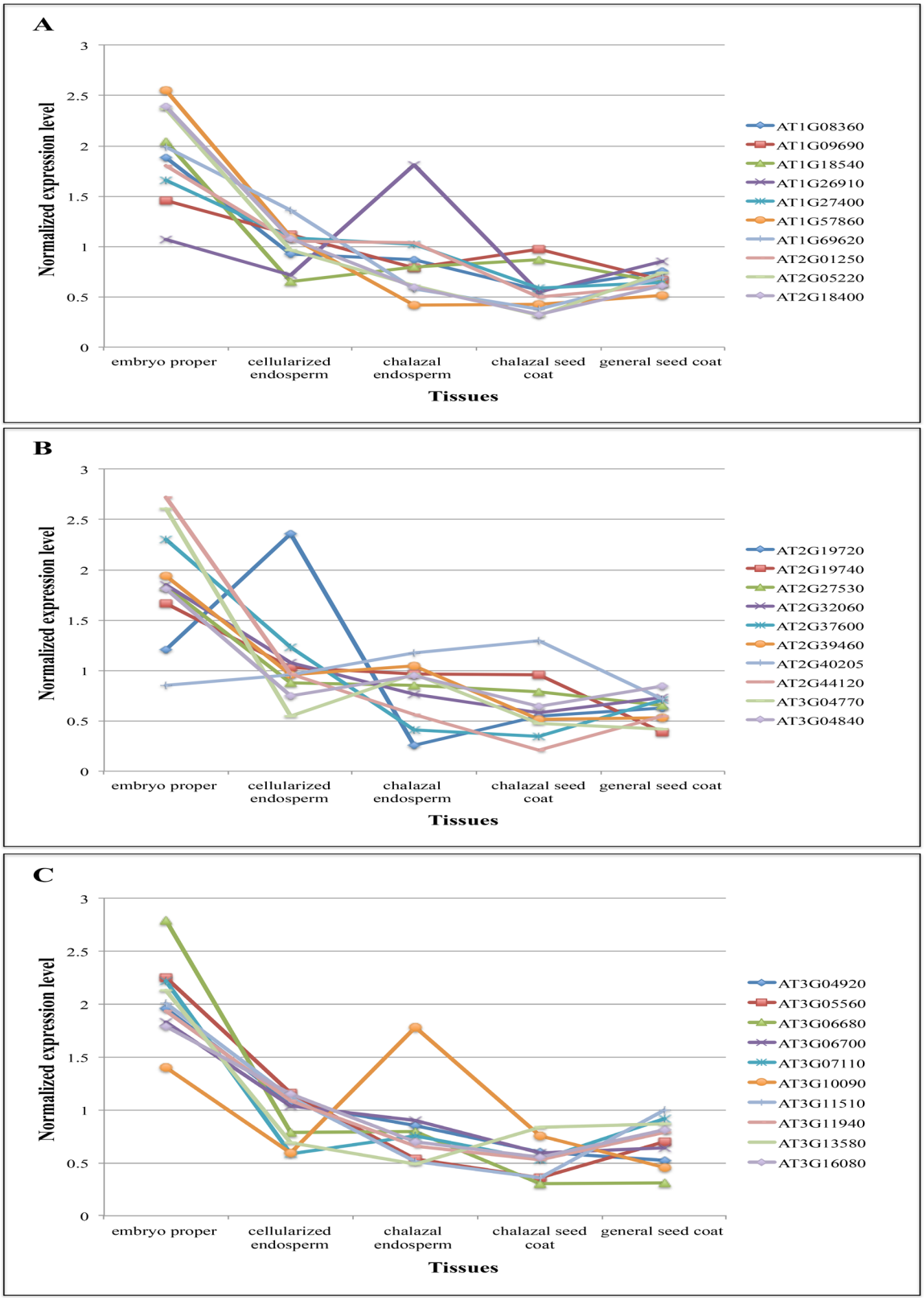


Figure 4.3 The localization of stored ribosomal protein transcripts in the linear cotyledon stage

Figure 4.3 (continued)

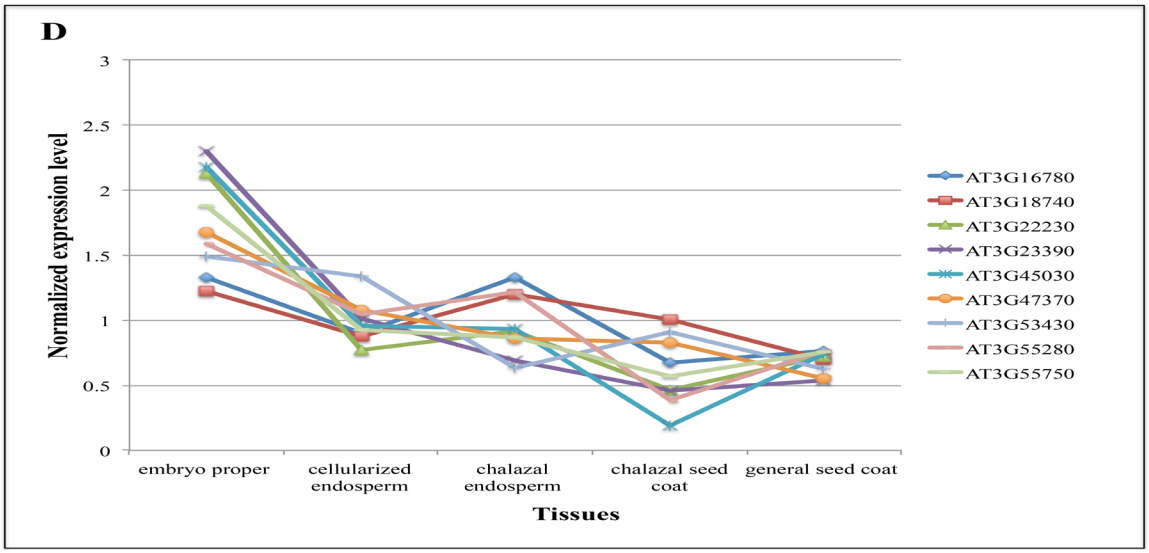


Figure 4.3 The localization of stored ribosomal protein transcripts in the linear cotyledon stage. Tissue types are presented along the abscissa (continued).

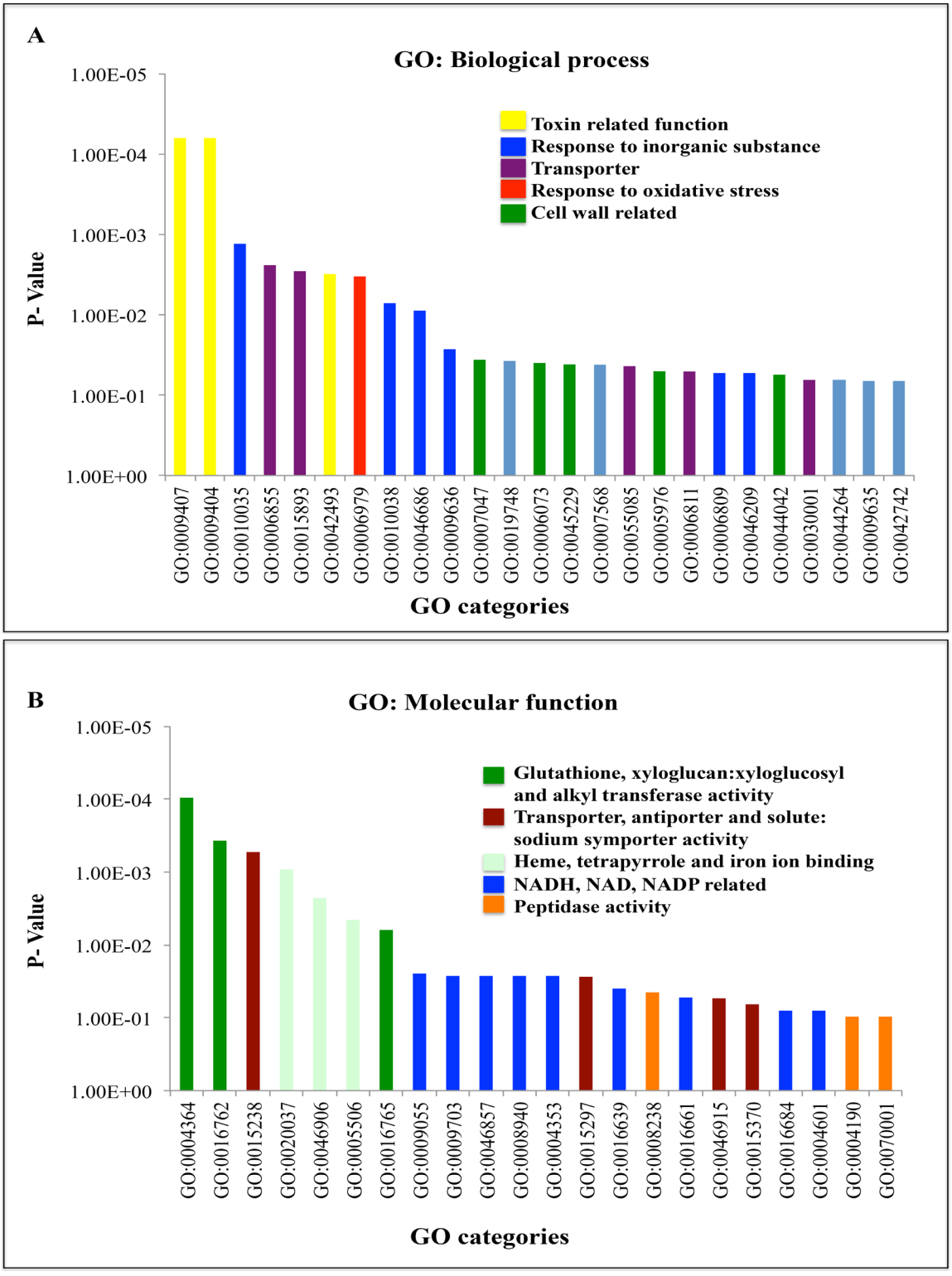


Figure 4.4 The high confidence candidate, *de novo* RNA genes analyzed by Gene Ontology

Figure 4.4 (continued)

(A) The GO biological process results demonstrated an abundance of genes whose products have toxin related function, respond to inorganic substances, are transporters, are oxidative stress responsive proteins and cell wall related proteins. GO terms for Figure 4.4A:GO:0009407 = toxin catabolic process; GO:0009404 = toxin metabolic process; GO:0010035 = response to inorganic substance; GO:0006855 = multidrug transport; GO:0042493 = response to drug; GO:0006979 = response to oxidative stress; GO:0010038 = response to metal ion; GO:0046686 = response to cadmium ion; GO:0009636 = response to toxin; GO:0007047 = cell wall organization; GO:0019748 = secondary metabolic process; GO:0006073 = cellular glucan metabolic process; GO:0045229 = external encapsulating structure organization; GO:0007568 = aging; GO:0055085 = transmembrane transport; GO:0005976 = polysaccharide metabolic process; GO:0006811 = ion transport; GO:0006809 = nitric oxide biosynthetic process; GO:0046209 = nitric oxide metabolic process; GO:0044042 = glucan metabolic process; GO:0030001 = metal ion transport; GO:0044264 = cellular polysaccharide metabolic process; GO:0009635 = response to herbicide; GO:0042742 = defense response to bacterium.

(B) The GO molecular functions showed an abundance of transcripts encoding cell wall related proteins, transporters and stress responsive proteins. GO term: GO:0004364 = glutathione transferase activity; GO:0016762 = xyloglucan:xyloglucosyl transferase activity; GO:0015238 = drug transporter activity; GO:0020037 = heme binding; GO:0046906 = tetrapyrrole binding; GO:0005506 = iron ion binding; GO:0016765 = transferase activity, transferring alkyl or aryl (other than methyl) groups; GO:0009055 = electron carrier activity; GO:0009703 = nitrate reductase (NADH) activity; GO:0046857 = oxidoreductase activity, acting on other nitrogenous compounds as donors, with NAD or NADP as acceptor; GO:0008940 = nitrate reductase activity; GO:0004353 = glutamate dehydrogenase [NAD(P)+] activity; GO:0015297 = antiporter activity; GO:0016639 = oxidoreductase activity, acting on the CH-NH2 group of donors, NAD or NADP as acceptor; GO:0008238 = exopeptidase activity; GO:0016661 = oxidoreductase activity, acting on other nitrogenous compounds as donors; GO:0046915 = transition metal ion transmembrane transporter activity; GO:0015370 = solute:sodium symporter activity; GO:0016684 = oxidoreductase activity, acting on peroxide as acceptor; GO:0004601 = peroxidase activity; GO:0004190 = aspartic-type endopeptidase activity; GO:0070001 = aspartic-type peptidase activity.

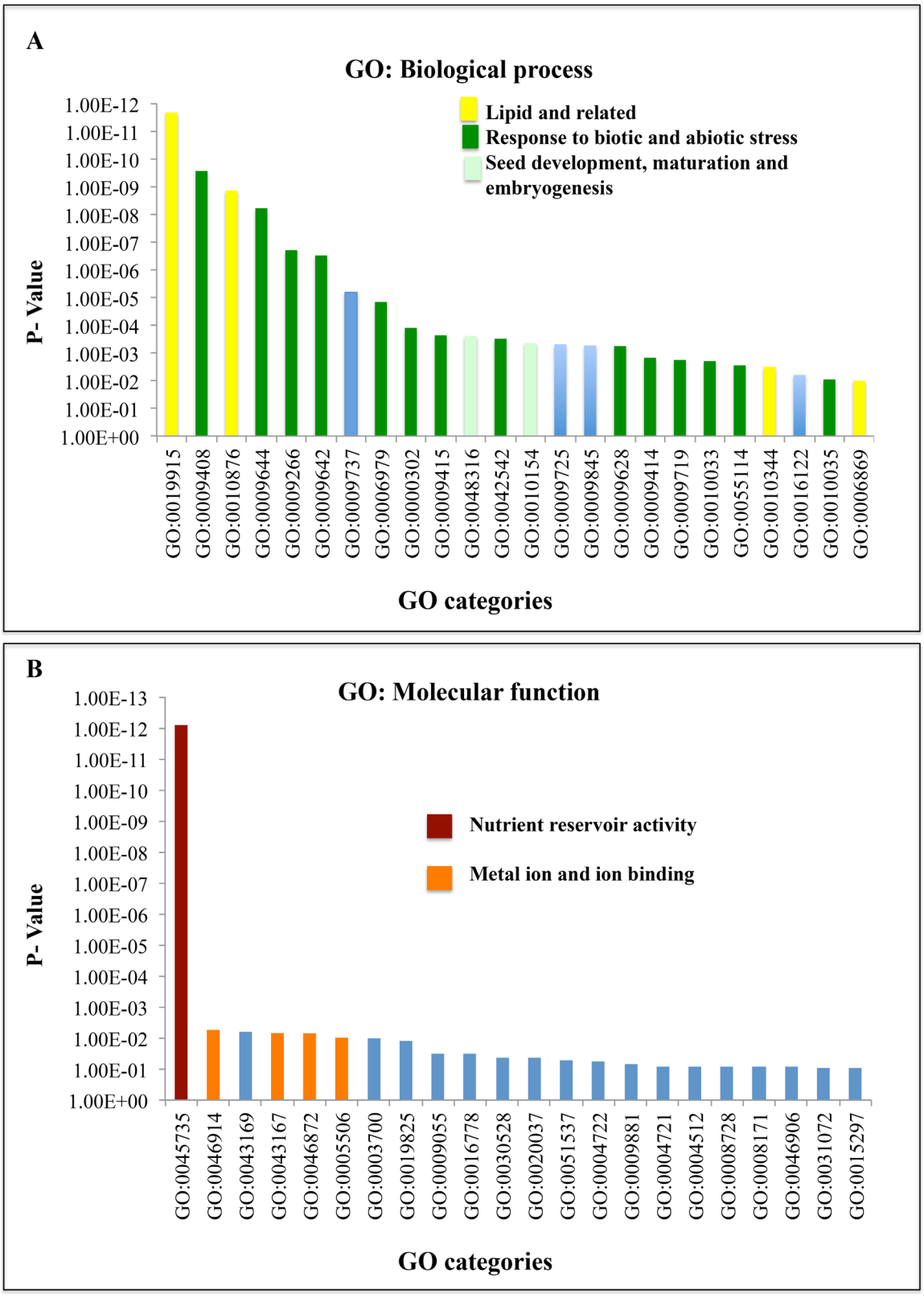


Figure 4.5 The stored, degraded high confidence candidate RNA genes analyzed by Gene Ontology

Figure 4.5 (continued)

(A) The GO Biological process analysis showed an abundance of transcripts encoding proteins involved in lipid storage and stress responses. GO:0019915 = lipid storage; GO:0009408 = response to heat; GO:0010876 = lipid localization; GO:0009644 = response to high light intensity; GO:0009266 = response to temperature stimulus; GO:0009642 = response to light intensity; GO:0009737 = response to abscisic acid stimulus; GO:0006979 = response to oxidative stress; GO:0000302 = response to reactive oxygen species; GO:0009415 = response to water; GO:0048316 = seed development; GO:0042542 = response to hydrogen peroxide; GO:0010154 = fruit development; GO:0009725 = response to hormone stimulus; GO:0009845 = seed germination; GO:0009628 = response to abiotic stimulus; GO:0009414 = response to water deprivation; GO:0009719 = response to endogenous stimulus; GO:0010033 = response to organic substance; GO:0055114 = oxidation reduction; GO:0010344 = seed oilbody biogenesis; GO:0016122 = xanthophyll metabolic process; GO:0010035 = response to inorganic substance; GO:0006869 = lipid transport.

(B) When analyzed by GO molecular function, the stored, degraded mRNAs showed an abundance of transcripts encoding proteins with a nutrient reservoir activity or capable of binding ions. GO term: GO:0045735 = nutrient reservoir activity; GO:0046914 = transition metal ion binding; GO:0043169 = cation binding; GO:0043167 = ion binding; GO:0046872 = metal ion binding; GO:0005506 = iron ion binding; GO:0003700 = transcription factor activity; GO:0019825 = oxygen binding; GO:0009055 = electron carrier activity; GO:0016778 = diphosphotransferase activity; GO:0030528 = transcription regulator activity; GO:0020037 = heme binding; GO:0051537 = 2 iron, 2 sulfur cluster binding; GO:0004722 = protein serine/threonine phosphatase activity; GO:0009881 = photoreceptor activity; GO:0004721 = phosphoprotein phosphatase activity; GO:0004512 = inositol-3-phosphate synthase activity; GO:0008728 = GTP diphosphokinase activity; GO:0008171 = O-methyltransferase activity; GO:0046906 = tetrapyrrole binding; GO:0031072 = heat shock protein binding; GO:0015297 = antiporter activity.

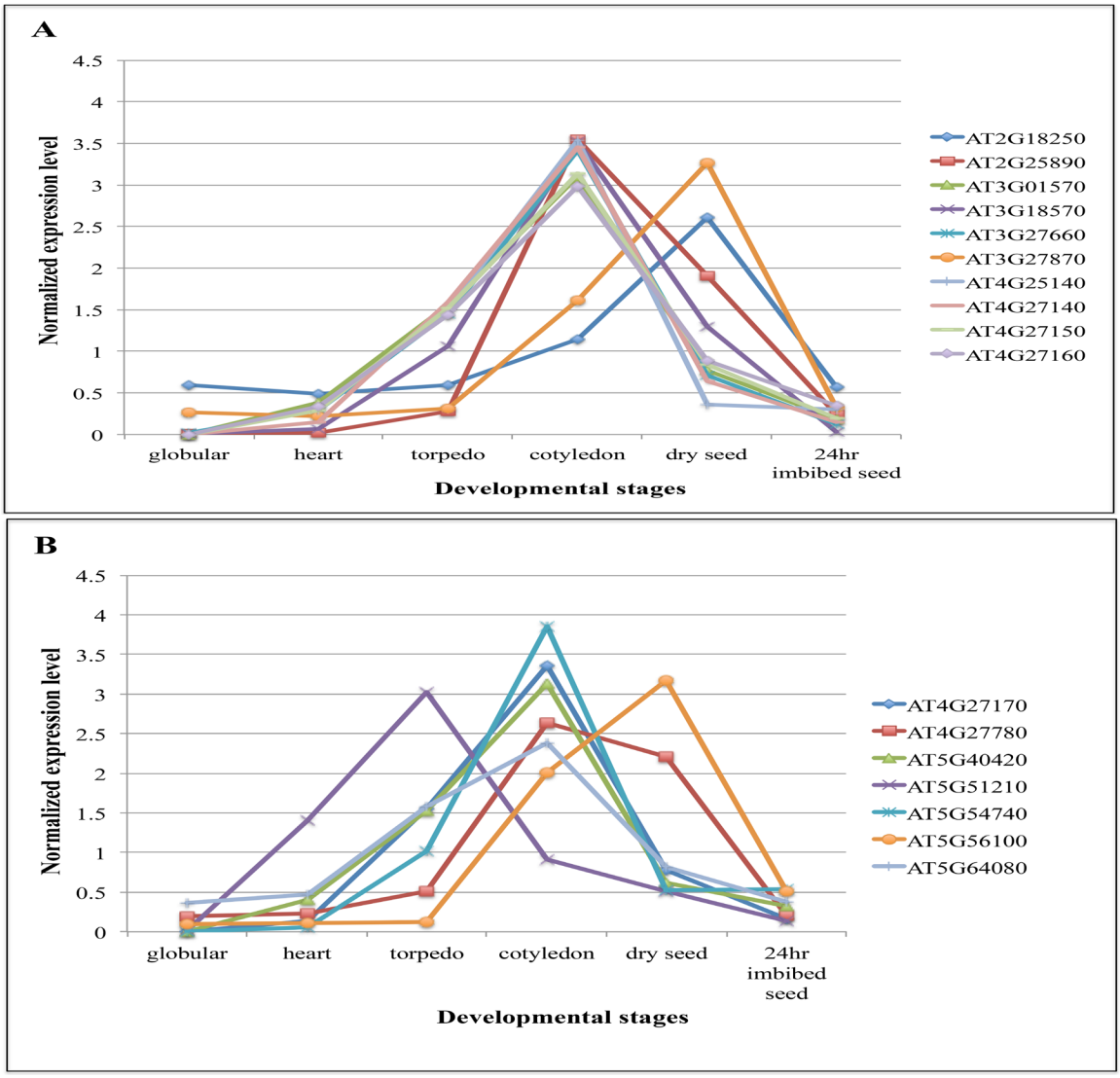


Figure 4.6 The normalized gene expression level for lipid storage or localization genes in different developmental stages

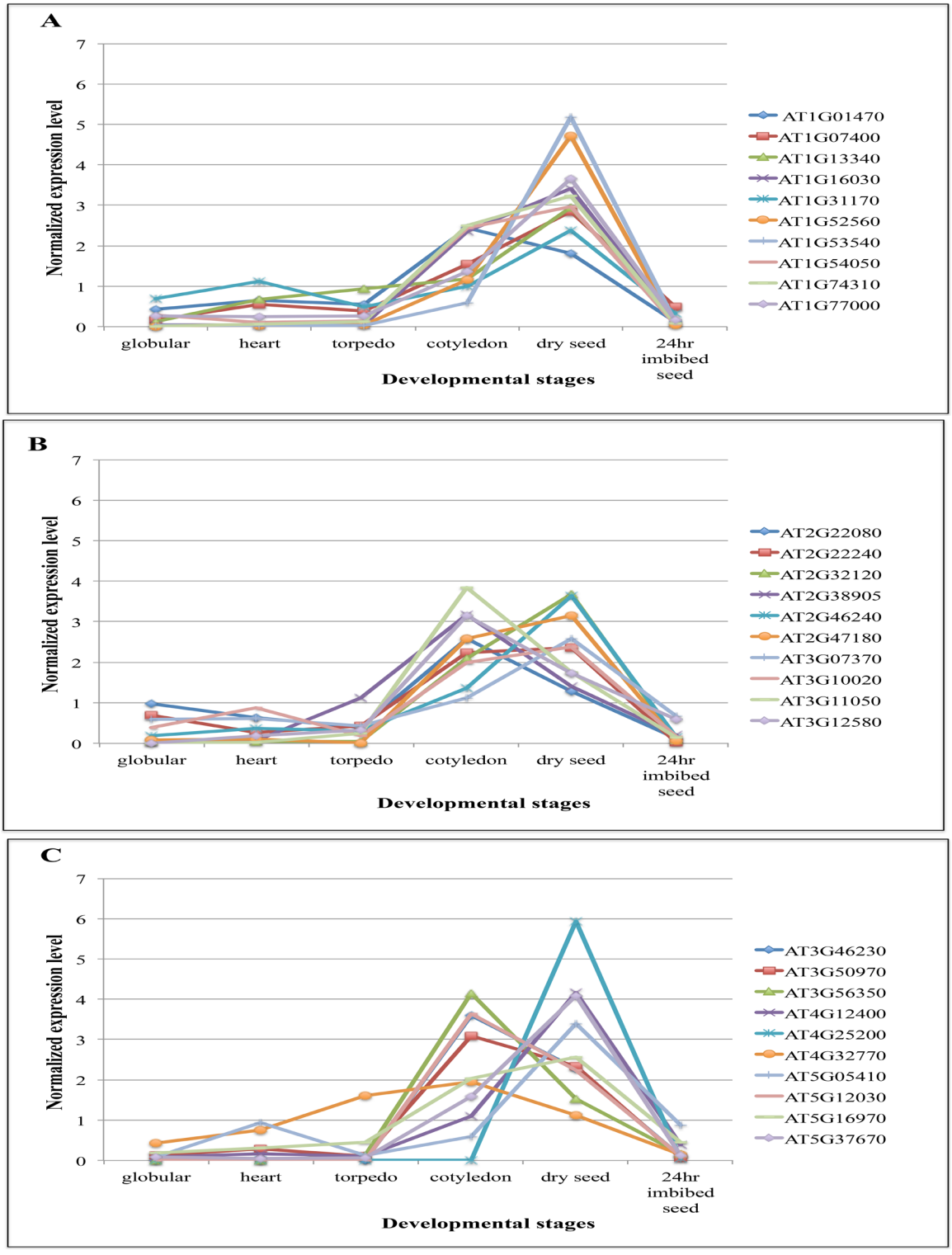


Figure 4.7 The normalized gene expression level for stress responsive genes from different developmental stages

Figure 4.7 (continued)

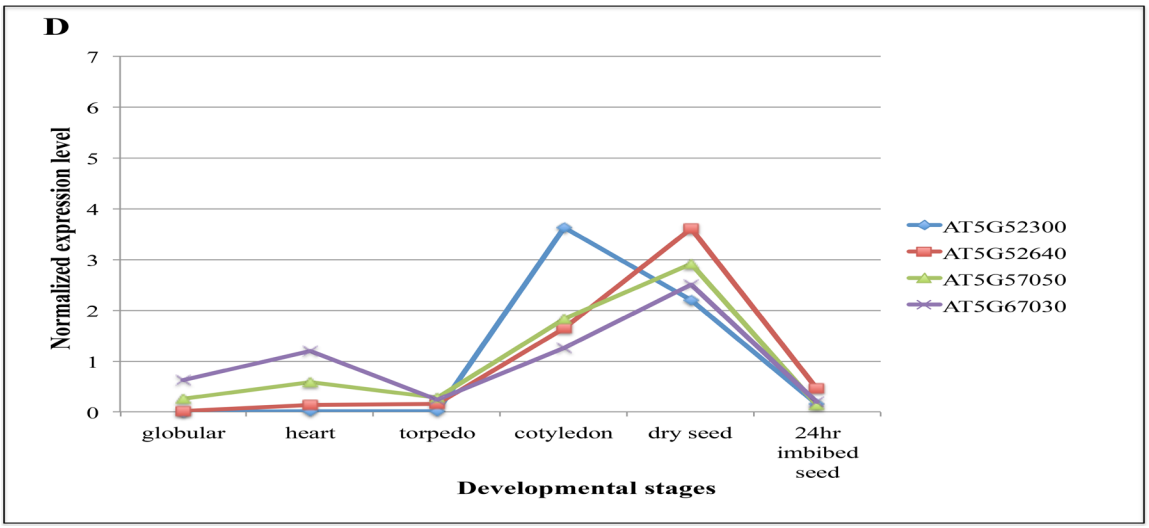


Figure 4.7 The normalized gene expression level for stress responsive genes from different developmental stages (continued).

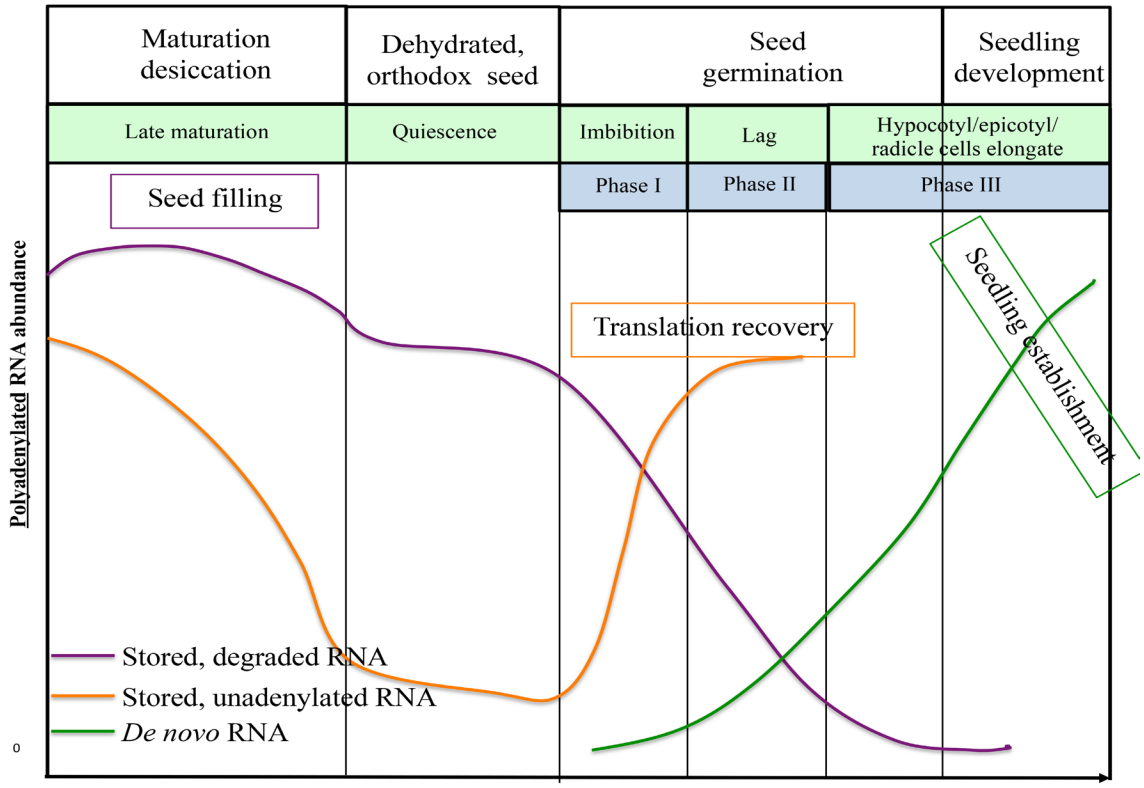


Figure 4.8 A model describing the abundance of three groups of RNAs in different developmental stages

A model describing the abundance of stored, unadenylated RNAs, stored, degraded RNAs and de novo RNAs in late maturation, quiescence and germination of orthodox seeds. The orange line, stored, unadenylated RNAs; Purple line, stored, degraded RNA; Green line, *de novo* RNAs.

Chapter Five: Future prospects

5.1 Alternative polyadenylation amongst different developmental stages

Although APA amongst different developmental stages is not a global event, the genes subject to alternative polyadenylation in different seed germination stages, dry seed and leaves have been identified.

19 genes have been identified producing untranslatable transcripts by choosing a 5' UTR poly(A) site in dry and germinating seeds. This may down-regulate the expression of these genes by decreasing the production of full-length mRNA. However, such down-regulation of gene expression seems to be reduced later in development by the decreasing of the transcripts generated by the 5' UTR poly(A) site and increasing of the production of full-length mRNA.

Future experiments may be performed to identify the mechanism of 5' UTR polyadenylation. 5' and 3' RACE (Rapid amplification of cDNA 3' ends) can be used to confirm these transcripts and to establish where the locations of transcription start sites are. After that, chromatin immunoprecipitation assay for DNA dependent RNA polymerase II may be performed to identify whether PoI II is associated with the region where the transcription start site of those transcripts located. The other experiment is to evaluate the protein expression level of those genes by either Western blot or mass spectrometry (MS). Bioinformatics tools also can be used to identify whether there are any conserved cis-element located in the promoter region of these genes.

73 genes have been identified that encode nonstop RNAs by choosing the coding region poly(A) sites in at least one developmental stage. This may be a mechanism to down-regulate the gene expression by decreasing the production of full-length mRNAs. The poly(A) sites of coding region APA are usually followed by an AG-rich region that is different from the canonical plant polyadenylation signal. Therefore, the identification of the coding region polyadenylation cis-elements and machinery, if they are different with the canonical ones, will be the main focus of future studyies. Bioinformatics tools will be used to identify whether there are any other conserved cis-element exist in genes affected by coding region APA. Since transcripts generating by coding region polyadenylation are nonstop RNAs, the protein products of those candidates may be evaluated.

The 3' UTR length of many mRNAs may be longer or shorter in different developmental stages. It will be interesting to evaluate the levels of protein encoded by these genes performed amongst different developmental stages. Western blot or MS may be used for this.

5.2 Three classes of RNAs in the seed

Stored, unadenylated mRNA, stored, degraded mRNA, and *de novo* synthesized mRNA have been identified and may have function in different developmental stages.

Genes encoding ribosomal proteins are a major component of those that produce stored, unadenylated mRNAs. Therefore, translation may be the first thing to be upregulated in the early seed germination stages. Future research will be focused on evaluating those stored, unadenylated mRNA and their function in translation recovery. The 3' RACE can be used to evaluate the unadenylated and polyadenylated mRNA abundance. Immunoblotting blot or MS may be performed to assess the ribosomal protein expression level in dry seed or imbibed seeds. A mutant deficient of Arabidopsis homolog of poly(A) ribonuclease (AtPARN) may be used to reduce deadenylation of these mRNAs during seed development. If the seed germination is not completed in *AtPARN* mutants, this may suggest that stored, unadenylated mRNAs play important roles in seed germination.

Seed germination is the most critic stage in plant life cycle. Manipulation of these stored, unadenylated mRNAs might be used to promote the seed germination. The genes producing those stored, unadenylated mRNA may be overexpressed in seed developmental stages so that more stored, unadenylated mRNA are present in dry seed. These stored, unadenylated mRNA may be repolyadenylated as soon as

possible during the early seed germination stage and this may promote the speed of seed germination.

mRNAs that are stored in dry seed and degraded early upon imbibition have also been identified. Genes encoding oleosin and lipid transporters were overrepresented in this gene list. Since those mRNAs may function in seed filling, it would be interesting to overexpression or knockout the respective genes to evaluate how seeds perform. Bioinformatics analysis may also be performed to identify conserved cis-elments present in those mRNA that may be responsible for the expression of this group of genes.

mRNAs synthesized *de novo* upon imbibition have been identified. The genes encoding transporters, cell wall related proteins, and stress responsive proteins were overrepresented in this gene lists. These results suggested that de novo mRNA may function in seedling development. The 3' RACE can be used to evaluate the RNA abundance of those candidate genes in seed germination stages and seedling development. Protein levels may also be evaluated. Since those mRNAs may function in seedling establishment, it would be interesting to overexpress or knockout these genes and evaluate how seed germination and seedling establishment are affected.

APPENDICES

Appendix 1.1. A Table compares subunits in different organisms (20, 227).

Factor or enzyme	Mammalian protein	Yeast protein	Arabidopsis protein	Arabidopsis gene
CPSF				
	CPSF160	Yhh1p	CPSF160	At5g51660
	CPSF100	Ydh1p	CPSF100	At5g23880
	CPSF73	Ysh1p	CPSF73-I	At1g61010
			CPSF73-II	At2g01730
	CPSF30	Yth1p	CPSF30	At1g30460
	Wdr33	Pfs2p	FY	AT5G13480
CstF				
	CstF77	Rna14p	CstF77	At1g17760
	CstF64	Rna15p	CstF64	AT1G71800
	CstF64 τ	Pti1p?		
	CstF50		CstF50	At5g60940
CF Im				
	CF Im68			At1g13190
				At5g55670
	CF Im59			
	CF Im25		CFIS1	At4g29820
			CFIS2	At4g25550
CF Iim				
	hCLP1	Clp1p	CLPS3	At3g04680
			CLPS5	At5g39930
	hPCF11	Pcf11p		At2g36480
			PCFS5	At5G43620
			PCFS1	At1g66500
			PCFS4	At4g04885
Poly(A) polymerase				
	PAP	Pap1p	PAPS1	At1g17980
			PAPS2	At2g25850
			PAPS3	At3g06560
			PAPS4	At4g32850
Other proteins				
	hFip1	Fip1p	FIPS5	At5g58040
			FIPS3	At3g66652
	Symplekin	Pta1p		At5g01400
				At1g27595/At1g27590

Appendix 1.2: A summary of inhibitor treatment experimental results in plants

Cd: Cordycepin, MP: 6-methylpurine, T: Tagetin, A: Alpha-amanitin, AD: Actinomycin D, Cd: Cordycepin, Ch: Cycloheximide, PM: Puromycin.

Inhibits both transcription and polyadenylation		**Inhibits transcription**				**Inhibits translation**	
Cd	**MP**	**T**	**A**	**AD**	**Cd**	**Ch**	**PM**
Whole seed application							
Chenopodium bonus-henricus Totally prevented completion of germination. (268)			*Chenopodium bonus-henricus* Totally prevented completion of germination. (268)				
Cotton (*Gossypium sp.*) Totally inhibited (193)				Cotton (*Gossypium sp.*) not inhibited. (193)	Cotton (*Gossypium sp.*) not inhibited. (193)		
Lettuce (*Lactuca sativa*) Totally inhibited. (208)				Lettuce (*Lactuca sativa*) Partially inhibited. (208)			
		Arabidopsis thaliana [*tt2-1*] Delays do not prevent. (269)					
				Rice (*Oryza sativa*) Little inhibition (20%?). (270)		Rice (*Oryza sativa*) Totally inhibited. (270)	
Wild Oat (*Avena fatua*) Totally inhibited. (195)							
			Arabidopsis thaliana [*tt2-1*] Delays do not prevent. (213)			*Arabidopsis thaliana* [*tt2-1*] Totally prevented. (213)	
			Oryza sativa Delays do not prevent. (271)				
						Arabidopsis	

Appendix 1.2 (continued)

Inhibits both transcription and polyadenylation		Inhibits transcription				Inhibits translation	
Cd	**MP**	**T**	**A**	**AD**	**Cd**	**Ch**	**PM**
						thaliana [*tt2-1*] Totally prevented. (272)	
			Kentucky bluegrass (*Poa pratensis*) does not prevent. (273)	Kentucky bluegrass (*Poa pratensis*) does not prevent. (273)		Kentucky bluegrass (*Poa pratensis*) delayed/reduced. (273)	
			White clover (*Trifolium repens*) Delayed but did not reduce. (273)	White clover (*Trifolium repens*) does not prevent. (273)		White clover (*Trifolium repens*) delayed/reduced. (273)	
	Lettuce totally inhibited. (274)						
Applied to excised seed part							
			Cress (*Lepidium sativum*) delayed autolysis of endosperm caps. (275)			Cress (*Lepidium sativum*) prevented autolysis of endosperm caps. (275)	
			Wheat (*Triticum aestivum*) Totally inhibited excised embryo water uptake (251)	Wheat (*Triticum aestivum*) Partially inhibited excised embryo water uptake (251)		Wheat (*Triticum aestivum*) Totally inhibited excised embryo water uptake (251)	
				Mung Bean (*Phaseolus Vulgaris*) excised embryos did not stop water uptake. (276)			Mung Bean (*Phaseolus Vulgaris*) excised embryos did not completely stop water uptake. (276)
	Lettuce totally inhibite						

Appendix 1.2 (continued)

Inhibits both transcription and polyadenylation		Inhibits transcription				Inhibits translation	
Cd	**MP**	**T**	**A**	**AD**	**Cd**	**Ch**	**PM**
	d. (274)						

Appendix 3.1: The 59 high confidence genes capable of 3' UTR APA

Descriptions of the location of PA sites as "proximal", "middle" (if present), or "distal" assumes a 3' UTR location and is in reference to the termination codon. A site described as "middle" is with reference to the occurrence of two additional major PA sites, one upstream, the other downstream of the middle site.

3' UTR APA Candidates	Function	Description
AT1G04080	PRP39	This gene has two poly(A) sites. The more proximal is dominant in dry seed and 48hr imbibed seed. However, the downstream site is dominant in 24-, and 36-hr imbibed seeds.
AT1G04430	S-adenosyl-L-methionine-dependent methyltransferases superfamily protein	This gene has three main poly(A) sites. The middle site is dominant. In 36-, and 48-hr imbibed seeds, there are more tags located in the up-, and down-stream sites compared with the leaf.
AT1G06890	Nucleotide/sugar transporter family protein	This gene has three poly(A) sites. The dominant one in dry seeds and the leaf is the most proximal. However, the middle site is most dominant during germination stages.
AT1G07430	AIP1, AKT1 INTERACTING PROTEIN PHOSPHATASE 1, ATAIP1, HAI2, HIGHLY ABA-INDUCED PP2C GENE 2	This gene has three major poly(A) sites. One located in the CDS, the other two in the 3' UTR. The dominant site in the leaf is in the CDS, followed by the distal 3' UTR site. The proximal 3' UTR site is not used in the leaf. However, the proximal 3' UTR site is used in dry seed and 24hr imbibed seed.
AT1G12500	Nucleotide-sugar transporter family protein	This gene has two poly(A) sites. The 48hr imbibed seed have more tags located in the proximal site. The distal site predominates in the other stages.
AT1G15930	Ribosomal protein L7Ae/L30e/S12e/Gadd45 family protein	WT 24hr imbibed seed utilize the proximal site predominately relative to the WT dry seed.
AT1G18650	PDCB3, PLASMODESMATA CALLOSE-BINDING PROTEIN 3	This gene has two poly(A) sites located in the 3' UTR. The leaf use both sites compared with the germination stages (especially 24hr imbibed seed), which uses the proximal site exclusively.
AT1G22360	ATUGT85A2, UDP-GLUCOSYL	The distal 3' UTR site dominates in dry seed with more

Appendix 3.1 (continued)

3' UTR APA Candidates	Function	Description
	TRANSFERASE 85A2, UGT85A2	proximal sites used in other stages.
AT1G25570	Di-glucose binding protein with Leucine-rich repeat domain	The distal site is dominant in 36hr and 48hr imbibed seed. The proximal site is dominant in the leaf.
AT1G31835	Unknown protein	The proximal site is dominant in dry seed while the distal site is dominant in 24 imbibed seed.
AT1G52420	UDP-Glycosyltransferase superfamily protein	The dominant poly(A) site in the leaf is proximal to that in 36hr and 48hr imbibed seed.
AT1G56220	Dormancy/auxin associated family protein	This gene has at least two main poly(A) sites. Both sites are equivalent in dry seed and the leaf. However, during seed germination stages, the more proximal site dominates.
AT1G61890	MATE efflux family protein	The proximal site is dominant in seed germination. However, both sites are utilized in the leaf where the distal site dominates.
AT1G67480	Galactose oxidase/kelch repeat superfamily protein	The dominant poly(A) site gradually changes during germination from the more proximal- (dry seed, 24hr and 36hr imbibed seeds) to the more distal-site (48hr imbibed seed and the leaf).
AT1G77740	PHOSPHATIDYLINOSITOL-4-PHOSPHATE 5-KINASE 2, PIP5K2	The dominant site gradually changes from the more proximal (24hr, 36hr and 48hr imbibed seed) to the more distal (leaf).
AT2G05710	ACO3, ACONITASE 3	This gene has four poly(A) sites. There are three additional poly(A) sites present during seed germination stages (24hr, 36hr and 48hr imbibed seed) compared with dry seed and leaf that only have the dominant poly(A) site.
AT2G20610	ABERRANT LATERAL ROOT FORMATION 1, ALF1, HLS3, HOOKLESS 3, ROOTY, ROOTY 1, RTY, RTY1, SUPERROOT 1, SUR1	This gene has two main poly(A) sites. The more distal site dominates in 36hr imbibed seed but the proximal site is dominant in dry seeds.
AT2G21410	VACUOLAR PROTON ATPASE A2, VHA-A2	The dominant sites changes from the more distal (24hr, 36hr and 48hr imbibed seed) to the more

Appendix 3.1 (continued)

3' UTR APA Candidates	Function	Description
		proximal in the leaf.
AT2G26890	GRAVITROPISM DEFECTIVE 2, GRV2, KAM2, KATAMARI2	The dominant poly(A) site is proximal in dry seed but the more distal site at other stages.
AT2G40000	ARABIDOPSIS ORTHOLOG OF SUGAR BEET HS1 PRO-1 2, ATHSPRO2, HSPRO2, ORTHOLOG OF SUGAR BEET HS1 PRO-1 2	There are two dominant poly(A) sites in dry seed but only the distal site is abundant in 24hr imbibed seed.
AT3G02350	GALACTURONOSYLTRANSFERASE 9, GAUT9	There are more tags situated at the proximal site in samples from germination stages compared with those of the leaf.
AT3G02520	GENERAL REGULATORY FACTOR 7, GF14 NU, GRF7	The dominant site is located distally in 48hr imbibed seed but proximately in the other stages.
AT3G03960	TCP-1/cpn60 chaperonin family protein	The dominant site is located distally in 36hr, 48hr imbibed seed comparing with that of the leaf.
AT3G05060	SAR DNA-binding protein, putative, strong similarity to SAR DNA-binding protein-1 (Pisum sativum)	The dominant sites are located in coding region. However, tags and sites located in the 3' UTR decreases as germination goes on. (High in 24hr imbibed seed but low in 48hr imbibed seed and leaf.
AT3G16850	Pectin lyase-like superfamily protein	The dominant poly(A) site in dry seed is proximal to that utilized in the leaf.
AT3G18215	Protein of unknown function, DUF599	The dominant poly(A) site in 24hr imbibed seed is distal to that utilized by the leaf
AT3G27380	SDH2-1, SUCCINATE DEHYDROGENASE 2-1	The proximal poly(A) site is utilized in 36hr and 48hr imbibed seed but not in any other stage or the leaf.
AT3G28430	Unknown protein	The dominant site is proximal in 36hr and 48hr imbibed seed but not in the leaf.
AT3G50060	MYB DOMAIN PROTEIN 77, MYB77	The dry seed sample has at least one more proximal, dominant site compared with the 36hr imbibed seed sample
AT3G50980	DEHYDRIN XERO 1, XERO1	This gene has at least three main 3' UTR poly(A) sites. The middle and distal sites are gradually abandoned in favor of the proximal 3' UTR site during germination. This gene also has a CDS poly(A) site that is also

Appendix 3.1 (continued)

3' UTR APA Candidates	Function	Description
		gradually deserted during germination. The expression level of this gene also decreased as germination progresses.
AT3G51950	Zinc finger (CCCH-type) family protein / RNA recognition motif (RRM)-containing protein	The dominant site is located proximally in 36hr and 48hr imbibed seed but located distally in the leaf.
AT3G52230	Unknown protein	This gene has four main poly(A) sites in the 3' UTR of dry seed. But it only has two poly(A) sites in the 3' UTR in other stages. In dry seed, three out of four sites are abundant. In 24 imbibed seeds, the two downstream sites predominate.
AT3G53900	PYRIMIDINE R, PYRR, UPP, URACIL PHOSPHORIBOSYLTRANSFERASE	The dominant site is located distally in 36hr and 48hr imbibed seed but proximally in the leaf.
AT3G62290	ADP-RIBOSYLATION FACTOR A1E, ARFA1E, ATARFA1E	The distal site is more dominant than the proximal in dry seed and germination stages but not in the leaf in which the proximal site is dominant.
AT4G04885	PCF11P-SIMILAR PROTEIN 4, PCFS4	In the leaf, the dominant site is the middle with very few tags from proximal and distal sites. However, all three sites are abundant in 24hr-imbibed seed.
AT4G13930	SERINE HYDROXYMETHYLTRANSFERASE 4, SHM4	3' UTR dominant site of dry seed is proximal to that in all other stages (24, 36, 48 imbibed seed and leaf)
AT4G14300	RNA-binding (RRM/RBD/RNP motifs) family protein	The dominant poly(A) site of dry seed is located distal with that in other germination stages and leaf samples.
AT4G19600	CYCT1;4	This gene has two main poly(A) sites but the dominant site in 24hr imbibed seed is the distal site. In dry seed, the proximal site prevails.
AT4G21960	PRXR1	The number of poly(A) sites increases as germination progresses, from a single PA site in dry seed to as many as six in both 48hr imbibed seed and the leaf.
AT4G37120	SMP2, SWELLMAP 2	The dominant poly(A) site gradually transitions from the more distal to the more proximal from dry seed,

Appendix 3.1 (continued)

3' UTR APA Candidates	Function	Description
		through the seed germination stages, to the leaf.
AT4G38130	ARABIDOPSIS HISTONE DEACETYLASE 1, ARABIDOPSIS HISTONE DEACETYLASE 19, ATHD1, ATHDA19, HD1, HDA1, HDA19, HISTONE DEACETYLASE 1, HISTONE DEACETYLASE 19, RPD3A	Dry seed have an additional, proximal poly(A) site but all other stages do not, having but a single site.
AT5G01530	LHCB4.1, LIGHT HARVESTING COMPLEX PHOTOSYSTEM II	The poly(A) site numbers increase from two in the dry seed, through three during the seed germination stages to the leaf, which has four.
AT5G03280	ATEIN2, CKR1, CYTOKININ RESISTANT 1, EIN2, ENHANCED RESPONSE TO ABA3, ERA3, ETHYLENE INSENSITIVE 2, ORE2, ORE3, ORESARA 2, ORESARA 3, PIR2	This gene has three main poly(A) sites. Dry seed and 24hr imbibed seeds use the proximal and distal sites equally regularly and the middle site infrequently. In all other germination stages and the leaf, the proximal site predominates.
AT5G03730	ATCTR1, CONSTITUTIVE TRIPLE RESPONSE 1, CTR1, SIS1, SUGAR-INSENSITIVE 1	The dominant poly(A) site in dry seed is distal to that used during seed germination and in the leaf.
AT5G06390	FASCICLIN-LIKE ARABINOGALACTAN PROTEIN 17 PRECURSOR, FLA17	Only one dominant poly(A) site in dry seed and located proximal. The other stages have three major poly(A) site and the middle site is dominant.
AT5G06760	ATLEA4-5, LATE EMBRYOGENESIS ABUNDANT 4-5, LEA4-5	This gene has at least four main poly(A) sites in the 3' UTR. The second most proximal site is less abundant in 36hr imbibed seeds compared with dry seed. The expression of this gene goes down as germination progresses. It is highly expressed in dry seed but nearly silent in 48hr imbibed seeds and leaf.
AT5G07360	Amidase family protein	The 3' UTR of dry seed has two additional, proximal and middle sites that are used predominately relative to all other stages and the leaf in which a single, distal site is used.
AT5G09260	VACUOLAR PROTEIN SORTING-ASSOCIATED PROTEIN 20.2, VPS20.2	The dominant site is distal during germination stages but this site is not used in the leaf. The dominant site is

Appendix 3.1 (continued)

3' UTR APA Candidates	Function	Description
		the proximal one in the leaf.
AT5G14105	Unknown protein	The dominant poly(A) site in 24hr through 48hr imbibed seed located distal is different with that of dry seed and Leaf which located proximal.
AT5G16650	Chaperone DnaJ-domain superfamily protein	This gene has two main poly(A) sites. The 36hr imbibed seed uses the distal site predominantly but the dry seed uses the proximal site predominantly..
AT5G17560	BolA-like family protein;	The 3' UTR dominant site in dry seed is proximal to that used in 24hr imbibed seeds.
AT5G22000	RHF2A, RING-H2 GROUP F2A	This gene has two poly(A) sites. The dominant poly(A) site located in proximal in dry seed and germination stages. However, The distal poly(A) site dominant in leaf.
AT5G32450	RNA binding (RRM/RBD/RNP motifs) family protein	This gene has two poly(A) sites: the proximal one and distal one. The dominant poly(A) site located distal in seed germination stages. However, the proximal poly(A) site dominant in leaf.
AT5G39760	ATHB23, HB23, HOMEOBOX PROTEIN 23, ZHD10, ZINC FINGER HOMEODOMAIN 10	This gene has three poly(A) sites. The dominant poly(A) site is the middle site in germination stages. However, the proximal one is dominant in leaf.
AT5G57930	ACCUMULATION OF PHOTOSYSTEM ONE 2, APO2, EMB1629, EMBRYO DEFECTIVE 1629	This gene has two poly(A) sites. The dominant poly(A) site located distal in seed germination stages. However, the proximal poly(A) site dominant in leaf.
AT5G60790	ABCF1, ARABIDOPSIS THALIANA GENERAL CONTROL NON-REPRESSIBLE 1, ATGCN1, ATP-BINDING CASSETTE F1, GCN1, GENERAL CONTROL NON-REPRESSIBLE 1	This gene has two major poly(A) sites The dominant poly(A) site located distal in 48hr imbibed seed. The dominant poly(A) site located proximal in leaf.
AT5G64300	ARABIDOPSIS THALIANA GTP CYCLOHYDROLASE II, ARABIDOPSIS THALIANA RIBOFLAVIN A1, ATGCH, ATRIBA1, GCH, GTP CYCLOHYDROLASE	The distal poly(A) site is used in all stages. The proximal poly(A) site first increases in prevalence as dry seed transitions through the seed germination stages and then decreases in the transition into the

Appendix 3.1 (continued)

3' UTR APA Candidates	Function	Description
	II, RED FLUORESCENT IN DARKNESS 1, RFD1, RIBA1, RIBOFLAVIN A1	seedling leaf.
AT5G64740	CELLULOSE SYNTHASE 6, CESA6, E112, ISOXABEN RESISTANT 2, IXR2, PRC1, PROCUSTE 1	The dominant site transitions from the more proximal to the more distal as seed germination stages transition into the seedling leaf.
AT5G65110	ACX2, ACYL-COA OXIDASE 2, ATACX2	The proximal poly(A) site is discarded as seed germination stages transition into the seedling leaf.

Appendix 3.2: The 19 genes capable of 5’ UTR APA

5’ UTR APA	Functions	Description
AT1G13460	Encodes protein phosphatase 2A (PP2A) B'theta subunit. Targeted to peroxisomes.	Upstream gene is located in another direction, thus, the poly(A) site in 5’ UTR is coming from this gene. This gene has two poly(A) sites. One located in the 5' UTR, the other located in the 3' UTR. The 5' UTR site is the dominant site in the dry seed and seed germination samples (24hr, 36hr and 48hr imbibed seed). However, in the leaf samples, the dominant poly(A) site located in 3' UTR indicating that this gene probably down regulated during seed germination.
AT1G13930	Involved in response to salt stress. Knockout mutants are hypersensitive to salt stress.	This gene has two poly(A) sites. One located in the 5' UTR, the other located in the 3' UTR. The abundant poly(A) site is 5' UTR in the seed germination samples (24hr, 36hr, 48hr imbibed seed). However, The 3' UTR poly(A) site is the most abundance one in the leaf indicating that this gene probably down regulated during seed germination.
AT1G35190	2-oxoglutarate (2OG) and Fe(II)-dependent oxygenase superfamily protein;	This gene has 5' UTR and 3' UTR poly(A) sites. The 5' UTR poly(A) site is dominant in 24hr and 36hr imbibed seed. However, the 3' UTR poly(A) site is dominant in dry seed, leaf and 48hr imbibed seed.
AT1G70230	ALTERED XYLOGLUCAN 4, AXY4, TBL27, TRICHOME BIREFRINGENCE-LIKE 27	The 5' UTR poly(A) site is dominant in germination stages (24hr, 36hr and 48hr imbibed seed). The 3' UTR poly(A) site is dominant in leaf.
AT3G10520	AHB2, ARABIDOPSIS HEMOGLOBIN 2, ARATH GLB2, ATGLB2, GLB2, HAEMOGLOBIN 2, HB2, HEMOGLOBIN 2, NON-SYMBIOTIC HAEMOGLOBIN 2, NSHB2	The 5’ UTR poly(A) site dominant in germination stages. The 3' UTR poly(A) site is dominant in leaf.
AT3G17780	Unknown protein	The 5' UTR poly(A) site is dominant in germination stages. The 3' UTR poly(A) site is dominant in WT leaf.
AT3G28210	PMZ, SAP12, STRESS-ASSOCIATED PROTEIN 12	The 5' UTR poly A) site is dominant in germination stages and dry seed. The 3' UTR poly(A) site dominant in WT leaf.
AT3G51770	ARABIDOPSIS ETHYLENE OVERPRODUCER 1, ATEOL1, ETHYLENE OVERPRODUCER 1, ETO1	The 5' UTR poly(A) site is dominant in germination stages and WT leaf. The 3' UTR poly(A) site is dominant in dry seed.
AT3G55850	LAF3, LAF3 ISF1, LAF3 ISF2, LAF3 ISOFORM 2, LONG AFTER FAR-RED 3, LONG AFTER FAR-RED 3 ISOFORM 1	The 5' UTR poly(A) site is dominant in dry seed. The 3' UTR poly(A) site is dominant in WT leaf.
AT4G00430	PIP1;4, PIP1E, PLASMA MEMBRANE INTRINSIC PROTEIN 1;4, PLASMA MEMBRANE INTRINSIC PROTEIN 1E, TMP-C, TRANSMEMBRANE PROTEIN C	The 5' UTR poly(A) site is dominant in germination stages. The 3' UTR poly(A) site is dominant in WT leaf.
AT4G00720	ASKTHETA, ATSK32, SHAGGY-LIKE PROTEIN KINASE 32, SHAGGY-LIKE PROTEIN KINASE THETA, SK32	The 5' UTR poly(A) site is dominant in germination stages. The 3' UTR poly(A) site is dominant in both WT leaf and dry seed.
AT4G07990	Chaperone DnaJ-domain superfamily protein	The 5' UTR poly(A) site is dominant in germination stages. The 3' UTR poly(A) site is dominant in WT leaf and dry seed.
AT4G20890	TUB9, TUBULIN BETA-9 CHAIN	The 5' UTR poly(A) site is dominant in germination stages. The 3’ UTR poly(A) site is dominant in WT leaf.
AT4G32530	ATPase, F0/V0 complex, subunit C protein	The 5' UTR poly(A) site is dominant in germination stages and dry seed. The 3' UTR poly(A) site is dominant in WT leaf.
AT4G33080	AGC (cAMP-dependent, cGMP-dependent and protein kinase C) kinase family protein	The 5' UTR poly(A) site is dominant in germination stages. The 3' UTR poly(A) site is dominant in WT leaf.
AT5G01750	Unknown protein	The 5' UTR poly(A) site is dominant in germination stages. The 3' UTR poly(A) site is dominant in WT leaf.
AT5G06680	ARABIDOPSIS THALIANA GAMMA TUBULIN COMPLEX PROTEIN 3, ATGCP3, ATSPC98, GAMMA TUBULIN COMPLEX PROTEIN 3, GCP3, SPC98, SPINDLE POLE BODY COMPONENT 98	The 5' UTR poly(A) site is dominant in germination stages. The 3' UTR poly(A) site is dominant in both WT leaf and dry seed.
AT5G59613	Unknown protein	The 5' UTR poly(A) site is dominant in germination stages. The 3' UTR poly(A) site is dominant in both WT leaf and dry

Appendix 3.2 (continued)

5’ UTR APA	Functions	Description
		seed.
AT5G65480	Unknown protein	The 5' UTR poly(A) site is dominant in germination stages. The 3' UTR poly(A) site is dominant in both WT leaf and dry seed.

Appendix 3.3: The seven genes capable of intronic APA

Intronic APA candidates	Functions	Description
AT1G06630	F-box/RNI-like superfamily protein	WT leaf has both intronic and 3' UTR poly(A) sites. Other stages only have the 3' UTR one.
AT1G29465	Unknown protein	This gene have seven poly(A) sites. Four of them are intronic poly(A) sites (proximal to distal order, name them number 1-4). The 3' UTR has one poly(A) site. WT leaf are abundant in both number 2 intronic and 3' UTR poly(A) sites. Other stages are mostly abundant in the number 2 intronic poly(A) site. The number 3 intronic site also abundant in dry seed.
AT1G58210	EMBRYO DEFECTIVE 1674	Both dry seed and germination stages have more tags located in upstream intronic poly(A) site; WT leaf has more tags located in downstream intronic poly(A) site.
AT2G23040	Unknown protein	The germination stages have the intronic poly(A) site, but not the WT leaf.
AT3G27330	Zinc finger (C3HC4-type RING finger) family protein	The dry seed and 24hr-imbibed seed have a downstream intronic poly(A) site. The WT leaf has an upstream intronic poly(A) site.
AT4G16530	Family of unknown function (DUF577)	The intronic poly(A) site dominates in the dry seed and in germination stages. However, the 3' UTR poly(A) site is dominant in the wt leaf.
AT4G31980	Unknown protein	The second intronic poly(A) site is dominant in the WT leaf. However, in other stages, both sites exist.

Appendix 3.4: The 73 genes capable of coding region APA

Coding region APA gene list	Functions	Description
AT1G03530	ATNAF1, NAF1, NUCLEAR ASSEMBLY FACTOR 1	The 3' UTR poly(A) site is dominant in dry seed. However, the coding region poly(A) site is dominant in germination stages and WT leaf.
AT1G06720	P-loop containing nucleoside triphosphate hydrolases superfamily protein	The coding region poly(A) site is dominant in germination stages. However, the 3' UTR poly(A) site is dominant in WT leaf.
AT1G07140	SIRANBP	The 3' UTR poly(A) site is dominant in WT leaf. However, the coding region poly(A) site is dominant in germination stages.
AT1G12830	Unknown protein	Two coding region dominant poly(A) site: the upstream one is dominant in all stages. However, the downstream one dominant in WT leaf but not germination stages
AT1G13350	Protein kinase superfamily protein	The dominant poly(A) site located in coding region in germination stages. However, the 3' UTR poly(A) site is dominant in WT leaf.
AT1G22400	ARABIDOPSIS THALIANA UDP-GLUCOSYL TRANSFERASE 85A1, ATUGT85A1, UGT85A1	The 3' UTR poly(A) site is dominant in both WT leaf and dry seed. However, the coding region poly(A) site is dominant in germination stages.
AT1G22530	PATELLIN 2, PATL2	The coding region poly(A) site is dominant in all stages. However, there are more tags located in 3' UTR poly(A) site in WT leaf and 24hr imbibed seed comparing with other stages.
AT1G26270	Phosphatidylinositol 3- and 4-kinase family protein	The coding region poly(A) site is dominant in germination stages. The 3' UTR poly(A) site is dominant in WT leaf.
AT1G28190	Unknown protein	The dominant poly(A) site located in coding region in germination stages. However, both 3' UTR and CDS poly(A) sites are dominant in WT leaf.
AT1G31930	EXTRA-LARGE GTP-BINDING PROTEIN 3, XLG3	The 3' UTR poly(A) site dominant in WT leaf. However, the coding region poly(A) site is dominant in germination stages.
AT1G31970	STRESS RESPONSE SUPPRESSOR 1, STRS1	The dominant coding region poly(A) site in WT leaf is different with that in other stages.
AT1G32810	RING/FYVE/PHD zinc finger superfamily protein	The dominant coding region poly(A) sites is different between germination stages and WT leaf. The WT dry seed has a similar dominant pattern as germination stages
AT1G56660	Unknown protein	Massive coding region poly(A) sites occur in all stages and the dominant sites are varied in every stage.
AT1G64330	Myosin heavy chain-related	The 3' UTR poly(A) site is dominant in both WT leaf and dry seed. However, the coding region poly(A) site is dominant in germination stages.
AT1G65280	DNAJ heat shock N-terminal domain-containing protein	The coding region dominant poly(A) site in 24hr-imbibed seed is different with other stages.
AT1G66760	MATE efflux family protein	The 3' UTR poly(A) site is dominant in WT leaf. However, the coding region poly(A) site is dominant in germination stages.
AT1G67230	CROWDED NUCLEI 1, CRWN1, LINC1, LITTLE NUCLEI1	The dominant coding region poly(A) site in 36hr and 48hr imbibed seed is different with that in dry seed, 24hr imbibed seed and WT leaf.
AT1G67785	Unknown protein	The 3' UTR poly(A) site is dominant in WT leaf. However, the coding region poly(A) site is dominant in germination stages.
AT1G68790	CROWDED NUCLEI 3, CRWN3, LINC3, LITTLE NUCLEI3	The 3' UTR poly(A) site is dominant in dry seed. However, the coding region poly(A) site is dominant in germination stages.
AT1G70200	RNA-binding (RRM/RBD/RNP motifs) family protein	The dominant coding region poly(A) sites are diversely changed among germination stages and WT leaf.

Appendix 3.4 (continued)

Coding region APA gene list	Functions	Description
AT1G73960	TAF2, TBP-ASSOCIATED FACTOR 2	The upstream coding region poly(A) site is dominant in germination stages and dry seed. However, the downstream one is dominant in WT leaf.
AT1G76180	EARLY RESPONSE TO DEHYDRATION 14, ERD14	The dominant poly(A) site in WT leaf located in 3' UTR. However, both 36hr and 48hr imbibed seed have more tags located in coding region poly(A) site.
AT1G77260	S-adenosyl-L-methionine-dependent methyltransferases superfamily protein	The 3' UTR poly(A) site is dominant in WT leaf. However, the coding region poly(A) site is dominant in germination stages.
AT2G02160	CCCH-type zinc finger family protein	The coding region poly(A) site is dominant in germination stages. However, the 3' UTR poly(A) site is dominant in WT leaf.
AT2G03150	EMB1579, EMBRYO DEFECTIVE 1579	The 3' UTR poly(A) site is dominant in WT leaf. However, the coding region poly(A) site is dominant in germination stages.
AT2G16940	Splicing factor, CC1-like	The coding region poly(A) site is dominant in all stages. However, More tags located in 3' UTR in WT leaf comparing with other stages.
AT2G22100	RNA-binding (RRM/RBD/RNP motifs) family protein	The dominant poly(A) site was changed between WT leaf and WT 24hr imbibed seed.
AT2G22125	Ethylene-responsive nuclear protein / ethylene-regulated nuclear protein (ERT2)	More tags located in 3' UTR sites in WT dry seed, 48hr imbibed seed and WT leaf. However, coding region poly(A) site is dominant in all stages.
AT2G22795	Unknown protein	More tags located in 3' UTR poly(A) site of WT leaf than other stages. All germination stages and WT leaf have many dominant poly(A) sites
AT2G26460	SMU2, SUPPRESSORS OF MEC-8 AND UNC-52 2	The coding region poly(A) site is dominant in both WT leaf and 24hr imbibed seed comparing with other stages which have a equal distribution of coding region and 3' UTR poly(A) site.
AT2G28510	Dof-type zinc finger DNA-binding family protein;	The 3' UTR poly(A) site is dominant in WT leaf. The coding region poly(A) site is dominant in 48hr-imbibed seed. Both coding region and 3' UTR poly(A) sites are dominant in 24hr and 36hr imbibed seed.
AT2G29210	Splicing factor PWI domain-containing protein	The 3' UTR poly(A) site is dominant in dry seed. However, the coding region poly(A) site is dominant in germination stages and WT leaf.
AT2G33250	Unknown protein	The 3' UTR poly(A) site is dominant in both WT leaf and 48 imbibed seed. However, the coding region poly(A) site is dominant in 36hr-imbibed seed.
AT2G42190	Unknown protein	The 3' UTR poly(A) site is dominant in WT leaf. However, the coding region poly(A) site is dominant in germination stages.
AT2G42560	Late embryogenesis abundant domain-containing protein / LEA domain-containing protein	Massive coding region poly(A) sites in dry seed than germination stages and WT leaf.
AT2G46550	Unknown protein	The coding region poly(A) site is dominant in germination stages and dry seed. However, 3' UTR poly(A) site is dominated in WT leaf.
AT3G10040	Sequence-specific DNA binding transcription factors	More tags located in coding region in 36hr-imbibed seed than that of 24hr and 48hr imbibed seed. This gene only expressed during seed germination stages. A dominant 3' UTR poly(A) site has been identified in all germination stages.
AT3G12860	NOP56-like pre RNA processing ribonucleoprotein	The coding region poly(A) site is dominant in WT leaf which is different with other stages
AT3G13480	Unknown protein	There is one more dominant coding region poly(A) site in WT leaf than other stages.
AT3G14980	IDM1, INCREASED DNA METHYLATION 1, REPRESSOR OF SILENCING 4, ROS4	The 3' UTR poly(A) site is dominant in WT leaf. However, the coding region poly(A) site is

Appendix 3.4 (continued)

Coding region APA gene list	Functions	Description
		dominant in germination stages.
AT3G25840	Protein kinase superfamily protein	The coding region poly(A) site is dominant during seed germination stages and dry seed are different with that of WT leaf.
AT3G29075	Glycine-rich protein	The coding region poly(A) sites are changed between germination stages and WT leaf.
AT3G47060	FTSH PROTEASE 7, FTSH7	The 3' UTR poly(A) site is dominant in dry seed. However, the coding region poly(A) site is dominant in germination stages and WT leaf.
AT3G48670	IDN2, INVOLVED IN DE NOVO 2, RDM12, RNA-DIRECTED DNA METHYLATION 12	The 3' UTR poly(A) site is dominant in WT leaf. However, the coding region poly(A) site is dominant in germination stages.
AT3G49601	Unknown protein	The 3' UTR poly(A) site is dominant in dry seed. However, the coding region poly(A) site is dominant in germination stages. Both 3' UTR and coding region sites are dominant in WT leaf.
AT3G52280	GENERAL TRANSCRIPTION FACTOR GROUP E6, GTE6	The 3' UTR poly(A) site is dominant in dry seed, 24hr imbibed seed and WT leaf. However, the coding region poly(A) site is dominant in 36hr and 48hr imbibed seed
AT3G57000	Nucleolar essential protein-related; CONTAINS InterPro DOMAIN/s: Ribosomal biogenesis, methyltransferase	The 3' UTR poly(A) site is dominant in WT leaf and dry seed. However, the coding region poly(A) site is dominant in germination stages.
AT3G58050	Unknown protein	The 3' UTR poly(A) site is dominant in WT leaf. However, the coding region poly(A) site is dominant in germination stages.
AT3G58840	PEROXISOMAL AND MITOCHONDRIAL DIVISION FACTOR 1, PMD1	The dominant poly(A) sites in WT leaf and 48hr imbibed seed located in 3' UTR. However, The dominant site in 36hr-imbibed seed located in coding region sites. Both coding region and 3' UTR poly(A) sites are dominant in 24hr-imbibed seed.
AT4G02510	ATTOC159, PLASTID PROTEIN IMPORT 2, PPI2, TOC159, TOC160, TOC86, TRANSLOCON AT THE OUTER ENVELOPE MEMBRANE OF CHLOROPLASTS 159, TRANSLOCON AT THE OUTER ENVELOPE MEMBRANE OF CHLOROPLASTS 160, TRANSLOCON AT THE OUTER ENVELOPE MEMBRANE OF CHLOROPLASTS 86	The 3' UTR poly(A) site is dominant in both WT leaf and dry seed. However, the coding region poly(A) site is dominant in germination stages.
AT4G08310	Unknown protein	The coding region poly(A) site is dominant in 36hr-imbibed seed. However, the 3' UTR poly(A) site is dominant in other stages.
AT4G11100	Unknown protein	The 3' UTR poly(A) site is dominant in WT leaf. However, the coding region poly(A) site is dominant in germination stages.
AT4G11560	Bromo-adjacent homology (BAH) domain-containing protein	The 3' UTR poly(A) site is dominant in dry seed. However, the coding region poly(A) site is dominant in germination stages. Both sites are dominant in WT leaf.
AT4G11740	SAY1	The 3' UTR poly(A) site is dominant in dry seed. However, the coding region poly(A) site is dominant in germination stages and WT leaf.
AT4G16630	DEA (D/H)-box RNA helicase family protein	The coding region poly(A) site is dominant in germination stages and dry seed. However, 3' UTR poly(A) site is dominated in WT leaf.
AT4G27120	CONTAINS InterPro DOMAIN/s: DDRGK domain	The 3' UTR poly(A) site is dominant in WT leaf. However, the coding region poly(A) site is dominant in germination stages.
AT4G27500	PPI1, PROTON PUMP INTERACTOR 1	More 3' UTR tags in WT leaf and dry seed comparing with that in germination stages.
AT5G03380	Heavy metal transport/detoxification superfamily protein	The dominant coding region poly(A) sites are changed in different developmental stages.

Appendix 3.4 (continued)

Coding region APA gene list	Functions	Description
		Basically, WT 24hr and 36hr imbibed seed have a dominant coding region site. Other stages have another dominate stages. Dry seed have more tags located in 3' UTR than other stages.
AT5G10910	MraW methylase family protein	Both coding region and 3' UTR poly(A) site are dominant during germination stages. However, only the 3' UTR poly(A) site is dominant in WT leaf.
AT5G12410	THUMP domain-containing protein	The dominant poly(A) sites are diversity changed in all stages. The 3' UTR poly(A) site is dominant in both WT leaf and 48hr imbibed seed. However, the dominant site of 36hr-imbibed seed is totally different with other stages. So does 24hr imbibed seed.
AT5G16730		The coding region poly(A) site is dominant in germination stages. The 3' UTR poly(A) site is dominant in WT leaf.
AT5G18570	ATOBGC, ATOBGL, CHLOROPLASTIC SAR1, CPSAR1, EMB269, EMB3138, EMBRYO DEFECTIVE 269, EMBRYO DEFECTIVE 3138, OBG-LIKE PROTEIN	The 3' UTR poly(A) site is dominant in WT leaf. However, the coding region poly(A) site is dominant in germination stages.
AT5G20050	Protein kinase superfamily protein	The 3' UTR poly(A) site is dominant in WT leaf. The coding region poly(A) site is dominant in both WT 24hr and 48hr imbibed seed. However, both coding region and 3' UTR poly(A) sites are dominant in 36hr-imbibed seed.
AT5G20610	Unknown protein	The 3' UTR poly(A) site is dominant in WT leaf and dry seed. However, the coding region poly(A) site is dominant in germination stages.
AT5G23420	HIGH-MOBILITY GROUP BOX 6, HMGB6	The 3' UTR poly(A) site is dominant in WT leaf and 36hr imbibed seed. However, there are no much difference between 3' UTR and coding region poly(A) site.
AT5G38720	Unknown protein	The dominant coding region poly(A) site in WT leaf is different with that in germination stages.
AT5G43560	TRAF-like superfamily protein	The 3' UTR poly(A) site is dominant in dry seed. However, the coding region poly(A) site is dominant in germination stages and WT leaf.
AT5G49400	zinc knuckle (CCHC-type) family protein;	The 3' UTR poly(A) site is dominant in WT leaf. However, the coding region poly(A) site is dominant in germination stages.
AT5G50740	Heavy metal transport/detoxification superfamily protein	The 3' UTR poly(A) site is dominant in WT leaf. However, the coding region poly(A) site is dominant in germination stages.
AT5G54500	FLAVODOXIN-LIKE QUINONE REDUCTASE 1, FQR1	The short isoform RNA was dominant in WT leaf and 48hr imbibed seed. However, the longer RNA isoform was dominant in 24hr and 36hr imbibed seed.
AT5G58590	RAN BINDING PROTEIN 1, RANBP1	WT leaf has more tags located in 3' UTR site than other stages. All stages have coding region poly(A) site
AT5G60030	Unknown protein	Many coding region poly(A) sites occur. And the dominant poly(A) sites are diversity in germination stages and WT leaf.
AT5G65900	DEA(D/H)-box RNA helicase family protein	The dominant coding region poly(A) site in dry seed is different with other stages.

Appendix 4.1: High confidence genes list for stored, unadenylated mRNA.

Gene name	Annotation
AT1G01220	GHMP kinase-related
AT1G01490	Heavy-metal-associated domain-containing protein
AT1G03230	Extracellular dermal glycoprotein, putative / EDGP, putative
AT1G04410	Malate dehydrogenase, cytosolic, putative
AT1G07610	MT1C (metallothionein 1C)
AT1G07790	HTB1; DNA binding
AT1G07930	Elongation factor 1-alpha / EF-1-alpha
AT1G08360	60S ribosomal protein L10A (RPL10aA)
AT1G09570	PHYA (PHYTOCHROME A); G-protein coupled photoreceptor/ signal transducer
AT1G09690	60S ribosomal protein L21 (RPL21C)
AT1G10030	ERG28 (ARABIDOPSIS HOMOLOG OF YEAST ERGOSTEROL28)
AT1G10630	ATARFA1F; GTP binding / phospholipase activator/ protein binding
AT1G11660	Heat shock protein, putative
AT1G11680	CYP51G1 (CYTOCHROME P450 51); oxygen binding
AT1G14620	DECOY (endoxyloglucan transferase A2)
AT1G16920	RAB11 (ARABIDOPSIS RAB GTPASE HOMOLOG A1B); GTP binding
AT1G17200	Integral membrane family protein
AT1G18540	60S ribosomal protein L6 (RPL6A)
AT1G21980	ATPIP5K1 (ARABIDOPSIS THALIANA 1-PHOSPHATIDYLINOSITOL-4-PHOSPHATE 5-KINASE 1); 1-phosphatidylinositol-4-phosphate 5-kinase
AT1G22780	PFL (POINTED FIRST LEAVES); structural constituent of ribosome
AT1G25520	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G68650.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO42107.1); contains InterPro domain Protein of unknown function UPF0016; (InterPro:IPR001727)
AT1G26340	B5 #6 (cytochrome b5 family protein #6); heme binding / transition metal ion binding
AT1G26910	60S ribosomal protein L10 (RPL10B)
AT1G27400	60S ribosomal protein L17 (RPL17A)
AT1G27970	NTF2B (NUCLEAR TRANSPORT FACTOR 2B); Ran GTPase binding / protein transporter
AT1G28510	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT3G58150.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO66379.1); contains InterPro domain Optic atrophy 3-like (InterPro:IPR010754)
AT1G28580	GDSL-motif lipase, putative
AT1G28650	Lipase, putative
AT1G28660	Lipase, putative
AT1G29980	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT2G34510.1); similar to unknown [Populus trichocarpa] (GB:ABK95079.1); contains InterPro domain Protein of unknown function DUF642 (InterPro:IPR006946); contains InterPro domain Galactose-binding like (InterPro:IPR008979)
AT1G30270	CIPK23 (CBL-INTERACTING PROTEIN KINASE 23); kinase
AT1G31180	3-isopropylmalate dehydrogenase, chloroplast, putative
AT1G32090	Early-responsive to dehydration protein-related / ERD protein-related
AT1G48140	Dolichol-phosphate mannosyltransferase-related
AT1G48630	Guanine nucleotide-binding family protein / activated protein kinase C receptor, putative / RACK, putative
AT1G54690	G-H2AX/GAMMA-H2AX/H2AXB/HTA3; DNA binding
AT1G54730	Sugar transporter, putative

Appendix 4.1 (continued)

Gene name	Annotation
AT1G55920	AtSerat2;1 (SERINE ACETYLTRANSFERASE 1)
AT1G56210	Copper chaperone (CCH)-related
AT1G56450	PBG1 (20S proteasome beta subunit G1); peptidase
AT1G57860	60S ribosomal protein L21
AT1G58380	XW6; structural constituent of ribosome
AT1G60090	glycosyl hydrolase family 1 protein
AT1G61260	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G11220.1); similar to unknown [Populus trichocarpa] (GB:ABK92540.1); contains InterPro domain Protein of unknown function DUF761, plant (InterPro:IPR008480)
AT1G63660	GMP synthase (glutamine-hydrolyzing), putative / glutamine amidotransferase, putative
AT1G64490	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G42060.1); contains domain DEK C-terminal domain (SSF109715)
AT1G64740	TUA1 (ALPHA-1 TUBULIN)
AT1G66470	Basic helix-loop-helix (bHLH) family protein
AT1G69530	ATEXPA1 (ARABIDOPSIS THALIANA EXPANSIN A1)
AT1G69620	RPL34 (RIBOSOMAL PROTEIN L34); structural constituent of ribosome
AT1G70480	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G23560.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO66084.1); contains InterPro domain Protein of unknown function DUF220 (InterPro:IPR003863)
AT1G70770	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G23170.1); similar to hypothetical protein [Vitis vinifera] (GB:CAN67931.1); contains domain PTHR13448 (PTHR13448)
AT1G71010	Phosphatidylinositol-4-phosphate 5-kinase family protein
AT1G71860	PTP1 (PROTEIN TYROSINE PHOSPHATASE 1)
AT1G73220	ATOCT1 (ARABIDOPSIS THALIANA ORGANIC CATION/CARNITINE TRANSPORTER1); carbohydrate transmembrane transporter/ carnitine transporter/ transporter
AT1G74510	Kelch repeat-containing F-box family protein
AT1G74960	FAB1 (FATTY ACID BIOSYNTHESIS 1); fatty-acid synthase
AT1G75240	ATHB33 (ARABIDOPSIS THALIANA HOMEOBOX PROTEIN 33); DNA binding / transcription factor
AT1G75270	DHAR2; glutathione dehydrogenase (ascorbate)
AT1G77350	Similar to unnamed protein product [Vitis vinifera] (GB:CAO47891.1)
AT1G77480	nucellin protein, putative
AT1G78080	RAP2.4 (related to AP2 4); DNA binding / transcription factor
AT1G79260	Identical to Uncharacterized protein At1g79260 [Arabidopsis Thaliana] (GB:O64527); similar to hypothetical protein [Vitis vinifera] (GB:CAN83082.1); contains InterPro domain Region of unknown function DUF1794 (InterPro:IPR014878)
AT1G79550	PGK (PHOSPHOGLYCERATE KINASE)
AT1G80270	DNA-binding protein, putative
AT1G80460	NHO1 (NONHOST RESISTANCE TO P. S. PHASEOLICOLA 1); carbohydrate kinase
AT2G01250	60S ribosomal protein L7 (RPL7B)
AT2G05220	40S ribosomal protein S17 (RPS17B)
AT2G16430	ATPAP10/PAP10; protein serine/threonine phosphatase
AT2G18400	Ribosomal protein L6 family protein
AT2G18910	Hydroxyproline-rich glycoprotein family protein
AT2G19670	Protein arginine N-methyltransferase, putative
AT2G19720	RPS15AB (ribosomal protein S15A B); structural constituent of ribosome
AT2G19740	60S ribosomal protein L31 (RPL31A)

Appendix 4.1 (continued)

Gene name	Annotation
AT2G20360	Binding / catalytic/ coenzyme binding
AT2G22360	DNAJ heat shock family protein
AT2G24765	ARF3/ARL1/ATARL1 (ADP-RIBOSYLATION FACTOR 3); protein binding
AT2G26250	FDH (FIDDLEHEAD); acyltransferase
AT2G27500	glycosyl hydrolase family 17 protein
AT2G27530	60S ribosomal protein L10A (RPL10aB)
AT2G27840	HDT4 (histone deacetylase 13)
AT2G31680	AtRABA5d (Arabidopsis Rab GTPase homolog A5d); GTP binding
AT2G31740	Methyltransferase
AT2G31750	UGT74D1 (UDP-GLUCOSYL TRANSFERASE 74D1); UDP-glycosyltransferase/ abscisic acid glucosyltransferase/ transferase, transferring glycosyl groups / transferase, transferring hexosyl groups
AT2G32060	40S ribosomal protein S12 (RPS12C)
AT2G32730	26S proteasome regulatory subunit, putative
AT2G37600	60S ribosomal protein L36 (RPL36A)
AT2G38310	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G05440.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO48777.1); contains InterPro domain Bet v I allergen; (InterPro:IPR000916); contains InterPro domain Streptomyces cyclase/dehydrase (InterPro:IPR005031)
AT2G38700	MVD1 (mevalonate diphosphate decarboxylase 1)
AT2G39460	ATRPL23A (RIBOSOMAL PROTEIN L23A); RNA binding / structural constituent of ribosome
AT2G39870	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT3G55690.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO69095.1)
AT2G39890	ProT1 (PROLINE TRANSPORTER 1); amino acid transmembrane transporter
AT2G40205	60S ribosomal protein L41 (RPL41C)
AT2G40220	ABI4 (ABA INSENSITIVE 4); DNA binding / transcription factor
AT2G41420	Proline-rich family protein
AT2G41560	ACA4 (AUTO-INHIBITED CA(2+)-ATPASE, ISOFORM 4); calcium-transporting ATPase/ calmodulin binding
AT2G41650	Unknown protein
AT2G42500	PP2A-4 (protein phosphatase 2A-4); protein serine/threonine phosphatase
AT2G42770	peroxisomal membrane 22 kDa family protein
AT2G43130	ARA4 (Arabidopsis Rab GTPase homolog A5c); GTP binding
AT2G44120	60S ribosomal protein L7 (RPL7C)
AT2G45740	PEX11D
AT2G46450	ATCNGC12 (cyclic nucleotide gated channel 12); cyclic nucleotide binding / ion channel
AT3G01280	porin, putative
AT3G02550	LBD41 (LOB DOMAIN-CONTAINING PROTEIN 41)
AT3G03960	chaperonin, putative
AT3G04120	GAPC (GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE C SUBUNIT); glyceraldehyde-3-phosphate dehydrogenase
AT3G04120	GAPC (GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE C SUBUNIT); glyceraldehyde-3-phosphate dehydrogenase
AT3G04120	GAPC (GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE C SUBUNIT); glyceraldehyde-3-phosphate dehydrogenase
AT3G04120	GAPC (GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE C SUBUNIT); glyceraldehyde-3-phosphate dehydrogenase
AT3G04120	GAPC (GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE C SUBUNIT); glyceraldehyde-3-phosphate dehydrogenase
AT3G04120	GAPC (GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE C SUBUNIT); glyceraldehyde-3-phosphate

Appendix 4.1 (continued)

Gene name	Annotation
	dehydrogenase
AT3G04770	RPSAB (40S RIBOSOMAL PROTEIN SA B); structural constituent of ribosome
AT3G04840	40S ribosomal protein S3A (RPS3aA)
AT3G04920	40S ribosomal protein S24 (RPS24A)
AT3G05560	60S ribosomal protein L22-2 (RPL22B)
AT3G06350	EMB3004/MEE32 (EMBRYO DEFECTIVE 3004); 3-dehydroquinate dehydratase/ NADP binding / binding / catalytic/ shikimate 5-dehydrogenase
AT3G06650	ACLB-1 (ATP-citrate lyase B-1)
AT3G06680	60S ribosomal protein L29 (RPL29B)
AT3G06700	60S ribosomal protein L29 (RPL29A)
AT3G07110	60S ribosomal protein L13A (RPL13aA)
AT3G07330	ATCSLC06 (Cellulose synthase-like C6); transferase, transferring glycosyl groups
AT3G07810	Heterogeneous nuclear ribonucleoprotein, putative / hnRNP, putative
AT3G07950	Rhomboid protein-related
AT3G08590	2,3-biphosphoglycerate-independent phosphoglycerate mutase, putative / phosphoglyceromutase, putative
AT3G10090	40S ribosomal protein S28 (RPS28A)
AT3G10270	DNA topoisomerase, ATP-hydrolyzing, putative / DNA topoisomerase II, putative / DNA gyrase, putative]
AT3G10520	AHB2 (NON-SYMBIOTIC HAEMOGLOBIN 2)
AT3G11510	40S ribosomal protein S14 (RPS14B)
AT3G11940	ATRPS5A (RIBOSOMAL PROTEIN 5A); structural constituent of ribosome
AT3G13580	60S ribosomal protein L7 (RPL7D)
AT3G13920	EIF4A1 (eukaryotic translation initiation factor 4A-1); ATP-dependent helicase
AT3G14415	(S)-2-hydroxy-acid oxidase, peroxisomal, putative
AT3G16080	60S ribosomal protein L37 (RPL37C)
AT3G16780	60S ribosomal protein L19 (RPL19B)
AT3G16870	Zinc finger (GATA type) family protein
AT3G17390	MTO3 (S-adenosylmethionine synthase 3); methionine adenosyltransferase
AT3G17790	ATACP5 (acid phosphatase 5); acid phosphatase/ protein serine/threonine phosphatase
AT3G18740	60S ribosomal protein L30 (RPL30C)
AT3G18940	Clast3-related
AT3G20330	Aspartate carabmoyltransferase, chloroplast / aspartate transcarbamylase / ATCase (PYRB)
AT3G20820	leucine-rich repeat family protein
AT3G21090	ABC transporter family protein
AT3G22230	60S ribosomal protein L27 (RPL27B)
AT3G23390	60S ribosomal protein L36a/L44 (RPL36aA)
AT3G24240	Leucine-rich repeat transmembrane protein kinase, putative
AT3G26618	ERF1-3 (EUKARYOTIC RELEASE FACTOR 1-3); translation release
AT3G45030	40S ribosomal protein S20 (RPS20A)
AT3G45310	Cysteine proteinase, putative
AT3G46440	UXS5 (UDP-Xyl synthase 5); catalytic
AT3G46560	TIM9 (EMBRYO DEFECTIVE 2474); P-P-bond-hydrolysis-driven protein transmembrane transporter

Appendix 4.1 (continued)

Gene name	Annotation
AT3G47370	40S ribosomal protein S20 (RPS20B)
AT3G49010	ATBBC1 (breast basic conserved 1); structural constituent of ribosome
AT3G49730	Pentatricopeptide (PPR) repeat-containing protein
AT3G52450	U-box domain-containing protein
AT3G52930	Fructose-bisphosphate aldolase, putative
AT3G53020	STV1 (SHORT VALVE1); structural constituent of ribosome
AT3G53430	60S ribosomal protein L12 (RPL12B)
AT3G55010	ATPURM/PUR5; phosphoribosylformylglycinamidine cyclo-ligase
AT3G55280	60S ribosomal protein L23A (RPL23aB)
AT3G55420	Similar to hypothetical protein [Vitis vinifera] (GB:CAN83699.1)
AT3G55750	60S ribosomal protein L35a (RPL35aD)
AT3G56340	40S ribosomal protein S26 (RPS26C)
AT3G57450	Similar to unnamed protein product [Vitis vinifera] (GB:CAO40798.1)
AT3G57610	ATPURA; adenylosuccinate synthase
AT3G59540	60S ribosomal protein L38 (RPL38B)
AT3G60245	60S ribosomal protein L37a (RPL37aC)
AT3G60770	40S ribosomal protein S13 (RPS13A)
AT3G62120	tRNA synthetase class II (G, H, P and S) family protein
AT3G62250	UBQ5 (UBIQUITIN 5); protein binding
AT3G62400	Unknown protein
AT3G62530	PBS lyase HEAT-like repeat-containing protein
AT3G66658	ALDH22a1 (ALDEHYDE DEHYDROGENASE 22A1); 3-chloroallyl aldehyde dehydrogenase
AT4G00860	ATOZI1 (ARABIDOPSIS THALIANA OZONE-INDUCED PROTEIN 1)
AT4G01610	cathepsin B-like cysteine protease, putative
AT4G02230	60S ribosomal protein L19 (RPL19C)
AT4G05400	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT4G21140.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO39711.1)
AT4G09320	NDPK1 (nucleoside diphosphate kinase 1); ATP binding / nucleoside diphosphate kinase
AT4G12420	SKU5 (skewed 5); copper ion binding
AT4G12600	Ribosomal protein L7Ae/L30e/S12e/Gadd45 family protein
AT4G13170	60S ribosomal protein L13A (RPL13aC)
AT4G15800	RALFL33 (RALF-LIKE 33)
AT4G16720	60S ribosomal protein L15 (RPL15A)
AT4G17190	FPS2 (FARNESYL DIPHOSPHATE SYNTHASE 2); dimethylallyltranstransferase/ geranyltranstransferase
AT4G17390	60S ribosomal protein L15 (RPL15B)
AT4G17890	AGD8 (ARF-GAP DOMAIN 8); DNA binding
AT4G21850	Methionine sulfoxide reductase domain-containing protein / SeIR domain-containing protein
AT4G22000	Similar to hypothetical protein [Vitis vinifera] (GB:CAN76661.1)
AT4G22310	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT4G14695.1); similar to unknown [Populus trichocarpa] (GB:ABK93494.1); contains InterPro domain Protein of unknown function UPF0041 (InterPro:IPR005336)
AT4G23885	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G24165.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO69543.1)

Appendix 4.1 (continued)

Gene name	Annotation
AT4G25740	40S ribosomal protein S10 (RPS10A)
AT4G26210	Mitochondrial ATP synthase g subunit family protein
AT4G26230	60S ribosomal protein L31 (RPL31B)
AT4G27730	ATOPT6 (oligopeptide transporter 6); oligopeptide transporter
AT4G29410	60S ribosomal protein L28 (RPL28C)
AT4G29870	Similar to membrane protein, putative [Arabidopsis thaliana] (TAIR:AT2G19340.2); similar to unnamed protein product [Vitis vinifera] (GB:CAO43189.1); contains domain PTHR13160 (PTHR13160); contains domain PTHR13160:SF2 (PTHR13160:SF2)
AT4G30190	AHA2 (Arabidopsis H(+)-ATPase 2); ATPase
AT4G30360	ATCNGC17 (cyclic nucleotide gated channel 17); calmodulin binding / cyclic nucleotide binding / ion channel
AT4G30470	cinnamoyl-CoA reductase-related
AT4G31210	DNA topoisomerase family protein
AT4G31700	RPS6 (RIBOSOMAL PROTEIN S6); structural constituent of ribosome
AT4G31790	diphthine synthase, putative (DPH5)
AT4G33865	40S ribosomal protein S29 (RPS29C)
AT4G34670	40S ribosomal protein S3A (RPS3aB)
AT4G35000	APX3 (ASCORBATE PEROXIDASE 3); L-ascorbate peroxidase
AT4G36130	60S ribosomal protein L8 (RPL8C)
AT4G36860	Zinc ion binding
AT4G36890	IRX14 (IRREGULAR XYLEM 14); transferase, transferring glycosyl groups / xylosyltransferase
AT4G39400	BRI1 (BRASSINOSTEROID INSENSITIVE 1); kinase
AT5G01040	LAC8 (laccase 8); copper ion binding / oxidoreductase
AT5G01870	Lipid transfer protein, putative
AT5G02450	60S ribosomal protein L36 (RPL36C)
AT5G02610	60S ribosomal protein L35 (RPL35D)
AT5G03850	40S ribosomal protein S28 (RPS28B)
AT5G05370	Ubiquinol-cytochrome C reductase complex ubiquinone-binding protein, putative / ubiquinol-cytochrome C reductase complex 8.2 kDa protein, putative
AT5G10360	EMB3010 (EMBRYO DEFECTIVE 3010); structural constituent of ribosome
AT5G11000	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT2G25200.1); similar to hypothetical protein [Vitis vinifera] (GB:CAN69699.1); contains InterPro domain Protein of unknown function DUF868, plant (InterPro:IPR008586)
AT5G11520	ASP3 (ASPARTATE AMINOTRANSFERASE 3)
AT5G11560	Catalytic
AT5G11710	(EPSIN1); binding
AT5G12220	Las1-like family protein
AT5G13890	Similar to unknown [Populus trichocarpa] (GB:ABK92746.1); contains InterPro domain Protein of unknown function DUF716 (InterPro:IPR006904)
AT5G14040	Mitochondrial phosphate transporter
AT5G14430	Dehydration-responsive protein-related
AT5G15320	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT3G01130.1); similar to unknown [Populus trichocarpa] (GB:ABK94588.1)
AT5G15350	Plastocyanin-like domain-containing protein
AT5G16190	ATCSLA11 (Cellulose synthase-like A11); transferase, transferring glycosyl groups
AT5G17560	BolA-like family protein

Appendix 4.1 (continued)

Gene name	Annotation
AT5G19780	TUA5 (tubulin alpha-5)
AT5G20180	Ribosomal protein L36 family protein
AT5G20400	Oxidoreductase, 2OG-Fe(II) oxygenase family protein
AT5G22500	Acyl CoA reductase, putative / male-sterility protein, putative
AT5G23250	Succinyl-CoA ligase (GDP-forming) alpha-chain, mitochondrial, putative / succinyl-CoA synthetase, alpha chain, putative / SCS-alpha, putative
AT5G24300	ATSS1/SSI (STARCH SYNTHASE I); transferase, transferring glycosyl groups
AT5G37850	SOS4 (SALT OVERLY SENSITIVE 4); kinase/ pyridoxal kinase
AT5G39890	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G15120.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO14912.1); contains InterPro domain Protein of unknown function DUF1637 (InterPro:IPR012864)
AT5G39990	glycosyltransferase family 14 protein / core-2/I-branching enzyme family protein
AT5G41670	6-phosphogluconate dehydrogenase family protein
AT5G44710	Similar to unnamed protein product [Vitis vinifera] (GB:CAO41922.1); contains InterPro domain Ribosomal protein S27, mitochondrial (InterPro:IPR013219)
AT5G45600	GAS41 (TBP-ASSOCIATED FACTOR 14B)
AT5G45620	26S proteasome regulatory subunit, putative (RPN9)
AT5G45775	60S ribosomal protein L11 (RPL11D)
AT5G47455	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT4G17310.1); similar to transcription factor [Arabidopsis thaliana] (TAIR:AT4G17310.2); similar to hypothetical protein 25.t00029 [Brassica oleracea] (GB:ABD64980.1)
AT5G47930	40S ribosomal protein S27 (RPS27D)
AT5G53160	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT4G27920.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO69376.1); similar to hypothetical protein [Vitis vinifera] (GB:CAN64668.1); contains InterPro domain Bet v I allergen; (InterPro:IPR000916)
AT5G54500	FQR1 (FLAVODOXIN-LIKE QUINONE REDUCTASE 1)
AT5G56350	Pyruvate kinase, putative
AT5G56360	Calmodulin-binding protein
AT5G56710	60S ribosomal protein L31 (RPL31C)
AT5G58420	40S ribosomal protein S4 (RPS4D)
AT5G59380	MBD6 (methyl-CpG-binding domain 6); DNA binding
AT5G62390	ATBAG7 (ARABIDOPSIS THALIANA BCL-2-ASSOCIATED ATHANOGENE 7); calmodulin binding
AT5G63150	Contains InterPro domain Protein of unknown function DUF1713, mitochondria (InterPro:IPR013177)
AT5G65640	BHLH093 (BETA HLH PROTEIN 93); DNA binding / transcription factor
AT5G66200	Armadillo/beta-catenin repeat family protein
AT5G66920	SKS17 (SKU5 Similar 17); copper ion binding / oxidoreductase
AT5G67260	CYCD3;2 (CYCLIN D3;2); cyclin-dependent protein kinase
No microarray hit: AT1G10040; AT1G14860; AT1G52780; AT2G04170; AT2G33510; AT2G39960; AT2G43490; AT3G14870; AT3G56705; AT4G03415; AT4G16141; AT4G16380; AT4G27490; AT4G40042; AT5G56795	

Appendix 4.2: The 39 genes that encode cytoplasmic ribosomal proteins in the gene list of high confidence stored, unadenylated RNAs

Subunits name	Gene name	Protein name	MATDB AGI gene name
Large (60S) ribosomal subunit	RPL6A	L6	AT1G18540
	RPL6 family protein	L6	AT2G18400
	RPL7B	L7	AT2G01250
	RPL7C	L7	AT2G44120
	RPL7D	L7	AT3G13580
	RPL10aA	L10a	AT1G08360
	RPL10aB	L10a	AT2G27530
	RPL10B	L10	AT1G26910
	RPL12B	L12	AT3G53430
	RPL13aA	L13a	AT3G07110
	RPL17A	L17	AT1G27400
	RPL19B	L19	AT3G16780
	RPL21C	L21	AT1G09690
	RPL21-2	L21	AT1G57860
	RPL22B	L22	AT3G05560
	RpL23aA	L23a	AT2G39460
	RPL23aB	L23a	AT3G55280
	RPL27B	L27	AT3G22230
	RPL29A	L29	AT3G06700
	RPL29B	L29	AT3G06680
	RPL30C	L30	AT3G18740
	RPL31A	L31	AT2G19740
	RPL34B	L34	AT1G69620
	RPL35aD	L35a	AT3G55750
	RPL36A	L36	AT2G37600
	RPL36aA	L36a	AT3G23390
	RPL37C	L37	AT3G16080
	RPL41C	L41	AT2G40205
Small (40S) ribosomal subunits	RPS3aA	S3a	AT3G04840
	ATRPS5A	S5	AT3G11940
	RPS12C	S12	AT2G32060
	RPS14B	S14	AT3G11510
	RPS15A B	S15a	AT2G19720
	RPS17B	S17	AT2G05220
	RPS20A	S20	AT3G45030
	RPS20B	S20	AT3G47370
	RPS24A	S24	AT3G04920

Appendix 4.2 (continued)

Subunits name	Gene name	Protein name	MATDB AGI gene name
	RPS28A	S28	AT3G10090
	RPSAB	Sa	AT3G04770

Appendix 4.3: The relative expression level of stored ribosomal protein mRNA in globular-, heart-, torpedo-, or cotyledon-stage embryos, and dry- or 24hr imbibed-seed

Gene name	Globular	Heart	Torpedo	Cotyledons	Dry seed	24hr imbibed seed
AT1G08360	1049	859	1383	714	421	1339
AT1G09690	1337	1365	1830	553	307	1363
AT1G18540	1311	1048	1395	924	543	1318
AT1G26910	277	254	296	78	48	604
AT1G27400	919	595	789	176	192	1502
AT1G57860	483	344	437	240	138	618
AT1G69620	2327	1977	2947	1574	277	1659
AT2G01250	1868	1587	2496	1334	455	1636
AT2G05220	1565	1526	2362	1599	959	1581
AT2G18400	204	121	157	37	72	281
AT2G19720	59	54	72	25	7	73
AT2G19740	361	233	245	205	81	731
AT2G27530	1154	881	1512	1248	769	1534
AT2G32060	768	806	963	276	187	1509
AT2G37600	205	238	469	165	36	392
AT2G39460	1168	968	1421	756	520	1432
AT2G40205	662	400	960	879	309	588
AT2G44120	1072	962	1496	807	514	1696
AT3G04770	99	92	64	64	52	481
AT3G04840	1709	1297	1972	644	217	1704
AT3G04920	636	492	869	613	317	1087
AT3G05560	1444	1419	1697	762	230	1521
AT3G06680	516	407	541	85	26	977
AT3G06700	696	515	957	500	125	1092
AT3G07110	1267	1095	1548	770	435	1479
AT3G10090	567	328	247	184	122	748
AT3G11510	1557	1403	2110	841	429	1579
AT3G11940	2395	2171	3158	2569	1281	1691
AT3G13580	322	254	487	202	83	901
AT3G16080	1252	1102	1314	430	111	1378
AT3G16780	336	353	321	154	167	909
AT3G18740	2389	1700	2739	2893	1011	1622
AT3G22230	683	506	668	586	255	858
AT3G23390	1511	1374	1790	867	142	1594
AT3G45030	1455	1132	1886	1332	679	1653
AT3G47370	632	482	742	446	287	1186
AT3G53430	1694	1610	2073	1068	387	1477

Appendix 4.3 (continued)

Gene name	Globular	Heart	Torpedo	Cotyledons	Dry seed	24hr imbibed seed
AT3G55280	751	517	921	544	190	1231
AT3G55750	2114	1620	2232	1738	966	1461

Appendix 4.4: The normalized expression level of stored ribosomal protein mRNA in globular-, heart-, torpedo-, or cotyledon-stage embryos, and dry- or 24hr imbibed-seed

Gene name	Globular	Heart	Torpedo	Cotyledons	Dry seed	24hr imbibed seed
AT1G08360	1.091760624	0.894015611	1.439375542	0.743104944	0.438161318	1.39358196
AT1G09690	1.187564767	1.212435233	1.62546262	0.49119171	0.272686899	1.210658771
AT1G18540	1.202936229	0.961614926	1.280012234	0.847836061	0.498241321	1.209359229
AT1G26910	1.06743738	0.978805395	1.140655106	0.300578035	0.184971098	2.327552987
AT1G27400	1.321351546	0.855499641	1.134435658	0.253055356	0.276060388	2.159597412
AT1G57860	1.282300885	0.913274336	1.160176991	0.637168142	0.366371681	1.640707965
AT1G69620	1.297463061	1.102313911	1.643155841	0.877613605	0.154446613	0.92500697
AT2G01250	1.195392491	1.015571672	1.597269625	0.853668942	0.291168942	1.046928328
AT2G05220	0.978940784	0.954545455	1.477481234	1.000208507	0.599874896	0.988949124
AT2G18400	1.403669725	0.832568807	1.080275229	0.254587156	0.495412844	1.933486239
AT2G19720	1.220689655	1.117241379	1.489655172	0.517241379	0.144827586	1.510344828
AT2G19740	1.167025862	0.753232759	0.792025862	0.662715517	0.261853448	2.363146552
AT2G27530	0.975486052	0.744716822	1.278106509	1.054945055	0.650042265	1.296703297
AT2G32060	1.021956088	1.072521623	1.281437126	0.367265469	0.248835662	2.007984032
AT2G37600	0.817275748	0.948837209	1.869767442	0.657807309	0.143521595	1.562790698
AT2G39460	1.118595371	0.927055068	1.360893855	0.724022346	0.498004789	1.371428571
AT2G40205	1.045813586	0.631911532	1.516587678	1.388625592	0.488151659	0.928909953
AT2G44120	0.982434703	0.881625172	1.371009623	0.739575378	0.471055445	1.554299679
AT3G04770	0.697183099	0.647887324	0.450704225	0.450704225	0.366197183	3.387323944
AT3G04840	1.359406072	1.031685006	1.568606655	0.512263025	0.172610367	1.355428874
AT3G04920	0.950672646	0.735426009	1.298953662	0.916292975	0.473841555	1.624813154
AT3G05560	1.224939912	1.203732504	1.439558886	0.64640181	0.195108158	1.29025873
AT3G06680	1.213166144	0.956896552	1.271943574	0.19984326	0.061128527	2.297021944
AT3G06700	1.074903475	0.795366795	1.477992278	0.772200772	0.193050193	1.686486486
AT3G07110	1.152866242	0.996360328	1.40855323	0.700636943	0.395814377	1.345768881
AT3G10090	1.549180328	0.896174863	0.674863388	0.50273224	0.333333333	2.043715847
AT3G11510	1.179694406	1.063013007	1.598686703	0.637201667	0.325041041	1.196363177
AT3G11940	1.083301922	0.981982661	1.428420656	1.162005277	0.579419525	0.764869959
AT3G13580	0.859048466	0.677634504	1.299244108	0.538906181	0.221431747	2.403734993
AT3G16080	1.344549848	1.183461607	1.411132987	0.46178629	0.119205298	1.47986397
AT3G16780	0.9	0.945535714	0.859821429	0.4125	0.447321429	2.434821429
AT3G18740	1.160271977	0.825643516	1.330257407	1.405050996	0.491015056	0.787761049
AT3G22230	1.152418448	0.853768279	1.127109111	0.988751406	0.430258718	1.447694038
AT3G23390	1.245671888	1.132728772	1.475680132	0.714756801	0.117065128	1.314097279
AT3G45030	1.072876982	0.834705665	1.390684527	0.982180165	0.500675925	1.218876736
AT3G47370	1.004503311	0.766092715	1.179337748	0.708874172	0.45615894	1.885033113
AT3G53430	1.223251896	1.162594777	1.496931039	0.771211939	0.279456012	1.066554339
AT3G55280	1.084737602	0.74675012	1.330284064	0.785748676	0.27443428	1.778045258

Appendix 4.4 (continued)

Gene name	Globular	Heart	Torpedo	Cotyledons	Dry seed	24hr imbibed seed
AT3G55750	1.251998816	0.959431448	1.321883328	1.029315961	0.572105419	0.865265028

Appendix 4.5: The possible location of these ribosomal protein transcripts, their relative expression level in the linear cotyledon stage

Gene name	Embryo proper	Cellularized endosperm	Chalazal endosperm	Chalazal seed coat	General seed coat
AT1G08360	3877.52	1908.25	1787.78	1166.88	1560.64
AT1G09690	5842.15	4479.57	3166.59	3907.93	2663.02
AT1G18540	1580.92	501.99	614.13	669.06	497.36
AT1G26910	206.87	138.49	349.92	105.26	165.37
AT1G27400	1609.66	1059.32	994.23	569.09	626.94
AT1G57860	1850.67	793.85	301.48	311.61	370.94
AT1G69620	3878.05	2659.39	1131.62	732.35	1350.52
AT2G01250	2227.59	1297.63	1282.9	612.72	754.93
AT2G05220	2352.81	961.42	608.09	314.03	737.19
AT2G18400	299.88	134.94	75.07	40.62	76.46
AT2G19720	104.16	203.29	21.87	46.97	53.89
AT2G19740	1285.44	791.16	742.66	738.23	298.57
AT2G27530	2890.9	1381.71	1340.05	1247.8	1029.07
AT2G32060	2726.17	1581.92	1118.42	852.91	1072.41
AT2G37600	525.49	281.18	92.69	78.32	161.48
AT2G39460	2940.88	1453.44	1584.44	778.66	807.07
AT2G40205	10768.48	12087.65	14769.05	16345.05	9010.63
AT2G44120	2489.25	885.07	514.88	189.99	497.31
AT3G04770	100.26	21.02	37.04	18.13	15.97
AT3G04840	1370.12	565.15	719.39	484.26	637
AT3G04920	1197.73	651.02	519.24	366.01	320.19
AT3G05560	4171.11	2146.39	998.44	660.08	1289.46
AT3G06680	1889.65	534.15	539.21	205.17	210.9
AT3G06700	5678.36	3227.97	2788.68	1836.73	1992.03
AT3G07110	833.3	219.28	284.09	198.8	344.98
AT3G10090	2470.52	1049.17	3139	1328.32	805.25
AT3G11510	1026.73	571.9	261.23	182.18	511.74
AT3G11940	5916.42	3330.97	2015.88	1607.37	2397.25
AT3G13580	497.72	160.96	113.85	195.35	202.52
AT3G16080	3987.69	2575.94	1557.07	1227.05	1810.75
AT3G16780	1430.05	970.84	1436.62	726.75	824.75
AT3G18740	9091.34	6524.35	8961.14	7462.32	5181.39
AT3G22230	1037.18	374.16	446.94	225.98	350.29
AT3G23390	3519.33	1546.9	1051.27	702.07	821.63
AT3G45030	2886.13	1272.84	1231.21	253.01	977.4
AT3G47370	1564.16	1007.42	801.99	773.72	512.72
AT3G53430	652.25	586.57	278.85	396.48	272.37
AT3G55280	1133.26	749.48	869.99	276.04	543.02

Appendix 4.5 (continued)

Gene name	Embryo proper	Cellularized endosperm	Chalazal endosperm	Chalazal seed coat	General seed coat
AT3G55750	3877.52	1908.25	1787.78	1166.88	1560.64

Appendix 4.6: The possible location of these ribosomal protein transcripts, their normalized expression level in the linear cotyledon stage

Gene name	Embryo proper	Cellularized endosperm	Chalazal endosperm	Chalazal seed coat	General seed coat
AT1G08360	1.882095743	0.926238731	0.867764223	0.566387764	0.75751354
AT1G09690	1.456222712	1.116584061	0.789308778	0.974096253	0.663788196
AT1G18540	2.045989864	0.649663773	0.794792751	0.865881878	0.643671735
AT1G26910	1.070855463	0.716888737	1.811348883	0.544874781	0.856032135
AT1G27400	1.656287815	1.090005845	1.02303035	0.585575111	0.64510088
AT1G57860	2.550150887	1.093894255	0.41542765	0.429386394	0.511140814
AT1G69620	1.988349998	1.363519837	0.580203098	0.375489775	0.692437292
AT2G01250	1.80349171	1.050581547	1.038655908	0.496067697	0.611203137
AT2G05220	2.365327312	0.966534903	0.611325133	0.315700688	0.741111965
AT2G18400	2.391501986	1.076128044	0.598672983	0.323938944	0.609758043
AT2G19720	1.210656004	2.36284811	0.254195918	0.54593426	0.626365707
AT2G19740	1.666779044	1.02586578	0.962977754	0.957233549	0.387143872
AT2G27530	1.832111672	0.875660527	0.849258448	0.790794889	0.652174464
AT2G32060	1.854075788	1.075868185	0.760640548	0.58006646	0.729349019
AT2G37600	2.306480213	1.234154991	0.406834861	0.34376207	0.708767864
AT2G39460	1.94387196	0.960699267	1.047288052	0.514681095	0.533459625
AT2G40205	0.854900997	0.959628846	1.172503043	1.29762042	0.715346694
AT2G44120	2.719600131	0.966972577	0.562525948	0.207571288	0.543330056
AT3G04770	2.605238541	0.546201019	0.962477913	0.471104875	0.414977653
AT3G04840	1.81428632	0.748360664	0.95260228	0.641247696	0.84350304
AT3G04920	1.960798117	1.065781762	0.850045348	0.599193239	0.524181534
AT3G05560	2.250887164	1.158272426	0.538795616	0.356203888	0.695840906
AT3G06680	2.796101306	0.790377854	0.797865099	0.303588551	0.31206719
AT3G06700	1.828924288	1.039686236	0.898196765	0.591586322	0.641606388
AT3G07110	2.215693052	0.583051929	0.755377702	0.528596878	0.917280438
AT3G10090	1.404940254	0.596644094	1.785092798	0.755391674	0.45793118
AT3G11510	2.010216228	1.11971274	0.511457526	0.35668695	1.001926556
AT3G11940	1.937536883	1.090841629	0.660169807	0.526389043	0.785062638
AT3G13580	2.126281613	0.687628161	0.48637218	0.834543746	0.865174299
AT3G16080	1.786839629	1.154250123	0.697705785	0.549827486	0.811376977
AT3G16780	1.326820696	0.900759138	1.332916435	0.67428897	0.765214761
AT3G18740	1.221279971	0.876444834	1.203789628	1.002446499	0.696039069
AT3G22230	2.130126717	0.768437699	0.917910908	0.464110411	0.719414265
AT3G23390	2.302864733	1.012210124	0.68789588	0.459397739	0.537631524
AT3G45030	2.179662236	0.961273844	0.929834048	0.191078137	0.738151736
AT3G47370	1.67827966	1.080920427	0.860502445	0.830169892	0.550127575
AT3G53430	1.491525346	1.341332345	0.637657099	0.906646177	0.622839032
AT3G55280	1.586403456	1.049165824	1.217862752	0.386416895	0.760151073

Appendix 4.6 (continued)

Gene name	Embryo proper	Cellularized endosperm	Chalazal endosperm	Chalazal seed coat	General seed coat
AT3G55750	1.882095743	0.926238731	0.867764223	0.566387764	0.75751354

Appendix 4.7: high confidence gene lists for *de novo* synthesized mRNAs

Gene name	Annotation
AT1G02930	ATGSTF6 (EARLY RESPONSIVE TO DEHYDRATION 11); glutathione transferase
AT1G03220	Extracellular dermal glycoprotein, putative / EDGP, putative
AT1G05000	Tyrosine specific protein phosphatase family protein
AT1G05385	Photosystem II 11 kDa protein-related
AT1G05560	UGT1 (UDP-glucosyl transferase 75B1); UDP-glycosyltransferase/ transferase, transferring glycosyl groups
AT1G07690	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G54950.1)
AT1G08230	Amino acid transporter family protein
AT1G12860	Basic helix-loop-helix (bHLH) family protein / F-box family protein
AT1G13080	CYP71B2 (CYTOCHROME P450 71B2); oxygen binding
AT1G13380	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT4G27435.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO42142.1); contains InterPro domain Protein of unknown function DUF1218 (InterPro:IPR009606)
AT1G14300	Binding
AT1G14870	Uncharacterized protein
AT1G16460	ATRDH2 (ARABIDOPSIS THALIANA RHODANESE HOMOLOGUE 2); thiosulfate sulfurtransferase
AT1G17020	SRG1 (SENESCENCE-RELATED GENE 1); oxidoreductase, acting on paired donors, with incorporation or reduction of molecular oxygen, 2-oxoglutarate as one donor, and incorporation of one atom each of oxygen into both donors
AT1G18100	E12A11; phosphatidylethanolamine binding
AT1G20760	Calcium-binding EF hand family protein
AT1G21050	Similar to unknown protein [Arabidopsis thaliana] (TAIR: AT1G76610.1); similar to hypothetical protein [Vitis vinifera] (GB: CAN67637.1); contains InterPro domain Protein of unknown function DUF617, plant (InterPro:IPR006460)
AT1G22500	Zinc finger (C3HC4-type RING finger) family protein
AT1G22700	Tetratricopeptide repeat (TPR)-containing protein
AT1G26290	Unknown protein
AT1G28600	Lipase, putative
AT1G29120	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT4G25770.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO48063.1); contains InterPro domain Protein of unknown function DUF676, hydrolase-like (InterPro:IPR007751)
AT1G31780	Similar to unknown [Populus trichocarpa] (GB: ABK94646.1); similar to unnamed protein product [Vitis vinifera] (GB: CAO63439.1); similar to Hypothetical protein [Oryza sativa (japonica cultivar-group)] (GB: AAN52749.1); contains InterPro domain Conserved oligomeric complex COG6 (InterPro:IPR010490)
AT1G33050	Similar to unknown protein [Arabidopsis thaliana] (TAIR: AT4G10470.1)
AT1G33110	MATE efflux family protein
AT1G35670	ATCDPK2 (CALCIUM-DEPENDENT PROTEIN KINASE 2); calmodulin-dependent protein kinase/ kinase
AT1G37130	NIA2 (NITRATE REDUCTASE 2)
AT1G48170	Similar to expressed protein [Oryza sativa (japonica cultivar-group)] (GB:ABA95965.1); similar to hypothetical protein OsI_036449 [Oryza sativa (indica cultivar-group)] (GB:EAY82490.1); similar to Os12g0182800 [Oryza sativa (japonica cultivar-group)] (GB:NP_001066323.1)
AT1G50450	Binding / catalytic
AT1G51170	Protein kinase family protein
AT1G54100	ALDH7B4 (ALDEHYDE DEHYDROGENASE 7B4); 3-chloroallyl aldehyde dehydrogenase
AT1G59700	ATGSTU16 (Arabidopsis thaliana Glutathione S-transferase (class tau) 16); glutathione transferase
AT1G60080	3' exoribonuclease family domain 1-containing protein
AT1G60960	IRT3 (Iron regulated transporter 3); cation transmembrane transporter/ metal ion transmembrane transporter

Appendix 4.7 (continued)

Gene name	Annotation
AT1G64510	Ribosomal protein S6 family protein
AT1G65970	TPX2 (THIOREDOXIN-DEPENDENT PEROXIDASE 2); antioxidant
AT1G66180	Aspartyl protease family protein
AT1G66330	Senescence-associated family protein
AT1G66390	PAP2 (PRODUCTION OF ANTHOCYANIN PIGMENT 2); DNA binding / transcription factor
AT1G66760	MATE efflux family protein
AT1G69870	Proton-dependent oligopeptide transport (POT) family protein
AT1G70090	GATL9/LGT8 (Galacturonosyltransferase-like 9); polygalacturonate 4-alpha-galacturonosyltransferase/ transferase, transferring glycosyl groups / transferase, transferring hexosyl groups
AT1G72550	tRNA synthetase beta subunit family protein
AT1G77280	Protein kinase family protein
AT1G77450	ANAC032 (Arabidopsis NAC domain containing protein 32); transcription factor
AT1G77760	NIA1 (NITRATE REDUCTASE 1)
AT1G78800	Glycosyl transferase family 1 protein
AT1G78850	Curculin-like (mannose-binding) lectin family protein
AT1G80440	Kelch repeat-containing F-box family protein
AT1G80640	Protein kinase family protein
AT1G80670	Transducin family protein / WD-40 repeat family protein
AT2G01850	EXGT-A3 (endo-xyloglucan transferase A3); hydrolase, acting on glycosyl bonds / xyloglucan: xyloglucosyl transferase
AT2G02850	ARPN (PLANTACYANIN); copper ion binding
AT2G02930	ATGSTF3 (GLUTATHIONE S-TRANSFERASE 16); glutathione transferase
AT2G03420	Similar to expressed protein [Oryza sativa (japonica cultivar-group)] (GB: ABF94594.1); similar to hypothetical protein OsI_010237 [Oryza sativa (indica cultivar-group)] (GB: EAY89004.1)
AT2G05710	Aconitate hydratase, cytoplasmic, putative / citrate hydro-lyase/aconitase, putative
AT2G13440	Glucose-inhibited division family A protein
AT2G17150	RWP-RK domain-containing protein
AT2G18980	Peroxidase, putative
AT2G20120	COV1 (CONTINUOUS VASCULAR RING)
AT2G23150	NRAMP3 (NRAMP metal ion transporter 3); manganese ion transmembrane transporter/ metal ion transmembrane transporter
AT2G24280	Serine carboxypeptidase S28 family protein
AT2G26980	CIPK3 (CBL-INTERACTING PROTEIN KINASE 3); kinase/ protein kinase
AT2G27775	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT2G27800.1); similar to unknown [Populus trichocarpa] (GB:ABK95440.1)
AT2G29420	ATGSTU7 (GLUTATHIONE S-TRANSFERASE 25); glutathione transferase
AT2G29490	ATGSTU1 (GLUTATHIONE S-TRANSFERASE 19); glutathione transferase
AT2G30760	Unknown protein
AT2G32460	AtM1/AtMYB101/MYB101 (myb domain protein 101); DNA binding / transcription factor
AT2G32520	Dienelactone hydrolase family protein
AT2G34080	Cysteine proteinase, putative
AT2G36950	Heavy-metal-associated domain-containing protein
AT2G37460	Nodulin MtN21 family protein
AT2G40000	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT3G55840.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO41329.1); contains InterPro domain Hs1pro-1, C-terminal

Appendix 4.7 (continued)

Gene name	Annotation
	(InterPro:IPR009743); contains InterPro domain Hs1pro-1, N-terminal (InterPro:IPR009869)
AT2G42130	Identical to Probable plastid-lipid-associated protein 13, chloroplast precursor (PAP13) [Arabidopsis Thaliana] (GB: Q8S9M1; GB:O48521;GB: Q84X37; GB: Q84X38; GB: Q84X39; GB: Q8GY49); similar to unknown protein [Arabidopsis thaliana] (TAIR: AT3G58010.1); similar to unknown [Populus trichocarpa] (GB: ABK96151.1)
AT2G43510	ATTI1 (ARABIDOPSIS THALIANA TRYPSIN INHIBITOR PROTEIN 1)
AT2G43590	Chitinase, putative
AT2G43630	Similar to glycine-rich protein [Arabidopsis thaliana] (TAIR: AT3G59640.2); similar to glycine-rich protein [Arabidopsis thaliana] (TAIR: AT3G59640.1); similar to unnamed protein product [Vitis vinifera] (GB: CAO46269.1)
AT2G44040	Dihydrodipicolinate reductase family protein
AT2G46650	B5 #1 (cytochrome b5 family protein #1); heme binding / transition metal ion binding
AT2G46740	FAD-binding domain-containing protein
AT3G01200	Similar to phosphoprotein phosphatase/ protein kinase [Arabidopsis thaliana] (TAIR: AT4G21210.1); similar to unnamed protein product [Vitis vinifera] (GB: CAO69694.1); contains InterPro domain Protein of unknown function DUF299 (InterPro:IPR005177); contains InterPro domain Pyruvate Pi dikinase regulator (InterPro:IPR017409)
AT3G01970	WRKY45 (WRKY DNA-binding protein 45); transcription factor
AT3G02420	Similar to hypothetical protein [Cleome spinosa] (GB:ABD96906.1)
AT3G03710	RIF10 (RESISTANT TO INHIBITION WITH FSM 10); 3'-5'-exoribonuclease/ RNA binding / nucleic acid binding
AT3G04630	WDL1 (WVD2-LIKE 1)
AT3G08740	Elongation factor P (EF-P) family protein
AT3G10140	RECA3 (RECA HOMOLOG 3); DNA binding / DNA-dependent ATPase
AT3G14990	4-methyl-5(b-hydroxyethyl)-thiazole monophosphate biosynthesis protein, putative
AT3G15450	Similar to unknown protein [Arabidopsis thaliana] (TAIR: AT4G27450.1); similar to unknown [Populus trichocarpa] (GB: ABK93866.1); contains domain PTHR11772 (PTHR11772); contains domain G3DSA: 3.60.20.10 (G3DSA: 3.60.20.10); contains domain SSF56235 (SSF56235)
AT3G15770	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G25360.1); similar to unknown [Populus trichocarpa] (GB:ABK94402.1)
AT3G16050	A37 (PYRIDOXINE BIOSYNTHESIS 1.2); protein heterodimerization
AT3G16150	L-asparaginase, putative / L-asparagine amidohydrolase, putative
AT3G17970	ATTOC64-III (ARABIDOPSIS THALIANA TRANSLOCON AT THE OUTER MEMBRANE OF CHLOROPLASTS 64-III); binding / carbon-nitrogen ligase, with glutamine as amido-N-donor
AT3G17970	ATTOC64-III (ARABIDOPSIS THALIANA TRANSLOCON AT THE OUTER MEMBRANE OF CHLOROPLASTS 64-III); binding / carbon-nitrogen ligase, with glutamine as amido-N-donor
AT3G21300	RNA methyltransferase family protein
AT3G21690	MATE efflux family protein
AT3G22200	POP2 (POLLEN-PISTIL INCOMPATIBILITY 2); 4-aminobutyrate transaminase
AT3G23560	ALF5 (ABERRANT LATERAL ROOT FORMATION 5); antiporter/ transporter
AT3G23730	Xyloglucan: xyloglucosyl transferase, putative / xyloglucan endotransglycosylase, putative / endo-xyloglucan transferase, putative
AT3G23910	Similar to unknown protein
AT3G25410	Bile acid:sodium symporter family protein
AT3G26510	Octicosapeptide/Phox/Bem1p (PB1) domain-containing protein
AT3G27770	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G62960.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO14740.1)
AT3G27850	RPL12-C (RIBOSOMAL PROTEIN L12-C); structural constituent of ribosome
AT3G28740	Cytochrome P450 family protein
AT3G28900	60S ribosomal protein L34 (RPL34C)

Appendix 4.7 (continued)

Gene name	Annotation
AT3G29970	Germination protein-related
AT3G44830	Lecithin: cholesterol acyltransferase family protein / LACT family protein
AT3G48580	Xyloglucan: xyloglucosyl transferase, putative / xyloglucan endotransglycosylase, putative / endo-xyloglucan transferase, putative
AT3G49790	ATP binding
AT3G51140	Heat shock protein binding
AT3G53460	CP29 (chloroplast 29 kDa ribonucleoprotein); RNA binding / poly (U) binding
AT3G56140	Similar to unknown protein [Arabidopsis thaliana] (TAIR: AT2G40400.2); similar to unknown protein [Arabidopsis thaliana] (TAIR: AT2G40400.1); similar to hypothetical protein [Vitis vinifera] (GB: CAN64033.1); similar to unnamed protein product [Vitis vinifera] (GB: CAO41449.1); similar to hypothetical protein OsI_004191 [Oryza sativa (indica cultivar-group)] (GB: EAY76344.1); contains InterPro domain Protein of unknown function DUF399 (InterPro:IPR007314)
AT3G59140	ATMRP14 (Arabidopsis thaliana multidrug resistance-associated protein 14)
AT3G62910	APG3 (ALBINO AND PALE GREEN); translation release factor
AT4G00335	RHB1A (RING-H2 finger B1A); protein binding / zinc ion binding
AT4G02380	SAG21 (SENESCENCE-ASSOCIATED GENE 21)
AT4G05150	Octicosapeptide/Phox/Bem1p (PB1) domain-containing protein
AT4G08950	Phosphate-responsive protein, putative (EXO)
AT4G11175	Translation initiation factor IF-1, chloroplast, putative
AT4G12910	SCPL20 (serine carboxypeptidase-like 20); serine carboxypeptidase
AT4G14130	XTR7 (XYLOGLUCAN ENDOTRANSGLYCOSYLASE 7); hydrolase, acting on glycosyl bonds
AT4G14660	RNA polymerase Rpb7 N-terminal domain-containing protein
AT4G14930	Acid phosphatase survival protein SurE, putative
AT4G16563	Aspartyl protease family protein
AT4G17600	LIL3: 1; transcription factor
AT4G19880	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G45020.1); similar to Intracellular chloride channel [Medicago truncatula] (GB:ABC75353.2); contains InterPro domain Thioredoxin-like fold (InterPro:IPR012336); contains InterPro domain Glutathione S-transferase, C-terminal-like (InterPro:IPR010987); contains InterPro domain Glutathione S-transferase, C-terminal (InterPro:IPR004046); contains InterPro domain Glutathione S-transferase, predicted (InterPro:IPR016639)
AT4G23250	EMB1290 (EMBRYO DEFECTIVE 1290); kinase
AT4G25050	ACP4 (ACYL CARRIER PROTEIN 4)
AT4G26500	ATSUFE/CPSUFE/EMB1374 (EMBRYO DEFECTIVE 1374); enzyme activator/ transcription regulator
AT4G27250	Dihydroflavonol 4-reductase family / dihydrokaempferol 4-reductase family
AT4G27450	Similar to unknown protein [Arabidopsis thaliana] (TAIR: AT3G15450.1); similar to unnamed protein product [Vitis vinifera] (GB: CAO39242.1); contains domain G3DSA: 3.60.20.10 (G3DSA: 3.60.20.10); contains domain SSF56235 (SSF56235)
AT4G30270	MERI5B (MERISTEM-5); hydrolase, acting on glycosyl bonds / xyloglucan: xyloglucosyl transferase
AT4G30490	AFG1-like ATPase family protein
AT4G30620	Similar to unknown protein [Arabidopsis thaliana] (TAIR: AT2G24020.1); similar to unknown [Picea sitchensis] (GB: ABK26000.1); similar to Os02g0180200 [Oryza sativa (japonica cultivar-group)] (GB: NP_001046090.1); contains InterPro domain Conserved hypothetical protein CHP00103 (InterPro:IPR004401)
AT4G33666	Unknown protein
AT4G33980	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G42900.2); similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G42900.3); similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G42900.1); similar to hypothetical protein [Vitis vinifera] (GB:CAN64989.1)
AT4G34290	SWIB complex BAF60b domain-containing protein

Appendix 4.7 (continued)

Gene name	Annotation
AT4G35750	Rho-GTPase-activating protein-related
AT4G35760	Electron carrier/ protein disulfide oxidoreductase
AT4G36040	DNAJ heat shock N-terminal domain-containing protein (J11)
AT4G36760	ATAPP1 (aminopeptidase P1)
AT4G37580	HLS1 (HOOKLESS 1); N-acetyltransferase
AT4G37870	PCK1/PEPCK (PHOSPHOENOLPYRUVATE CARBOXYKINASE 1); ATP binding / phosphoenolpyruvate carboxykinase (ATP)
AT4G39675	Unknown protein
AT5G03500	Transcription coactivator
AT5G03610	GDSL-motif lipase/hydrolase family protein
AT5G04910	Similar to hypothetical protein at5g04910 [Brassica rapa] (GB: ABV89667.1)
AT5G05290	ATEXPA2 (ARABIDOPSIS THALIANA EXPANSIN A2)
AT5G07440	GDH2 (GLUTAMATE DEHYDROGENASE 2); oxidoreductase
AT5G07870	Transferase family protein
AT5G11090	Serine-rich protein-related
AT5G14320	30S ribosomal protein S13, chloroplast (CS13)
AT5G14530	Transducin family protein / WD-40 repeat family protein
AT5G15880	Similar to hypothetical protein [Cleome spinosa] (GB: ABD96950.1); contains InterPro domain Polyadenylate-binding protein/Hyperplastic disc protein; (InterPro:IPR002004)
AT5G18170	GDH1 (GLUTAMATE DEHYDROGENASE 1); oxidoreductase
AT5G20790	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT3G43110.1)
AT5G23900	60S ribosomal protein L13 (RPL13D)
AT5G25460	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G11420.1); similar to unknown [Ricinus communis] (GB:CAB02653.1); contains InterPro domain Protein of unknown function DUF642 (InterPro:IPR006946)
AT5G26340	MSS1 (SUGAR TRANSPORT PROTEIN 13); carbohydrate transmembrane transporter/ hexose:hydrogen symporter/ high-affinity hydrogen:glucose symporter/ sugar:hydrogen ion symporter
AT5G27380	GSH2/GSHB (GLUTATHIONE SYNTHETASE 2); glutathione synthase
AT5G37740	C2 domain-containing protein
AT5G37890	Seven in absentia (SINA) protein, putative
AT5G39040	ATTAP2 (Arabidopsis thaliana transporter associated with antigen processing protein 2); ATPase, coupled to transmembrane movement of substances
AT5G39050	Transferase family protein
AT5G39580	Peroxidase, putative
AT5G39800	60S ribosomal protein-related
AT5G40730	AGP24 (ARABINOGALACTAN PROTEIN 24)
AT5G42060	Similar to unknown protein [Arabidopsis thaliana] (TAIR: AT1G64490.1); contains InterPro domain DEK C terminal (InterPro:IPR014876)
AT5G45380	Sodium: solute symporter family protein
AT5G46640	DNA-binding family protein
AT5G46960	Invertase/pectin methylesterase inhibitor family protein
AT5G47550	Cysteine protease inhibitor, putative / cystatin, putative
AT5G48840	PANC (Arabidopsis homolog of bacterial panC); pantoate-beta-alanine ligase
AT5G51830	PfkB-type carbohydrate kinase family protein
AT5G52960	Similar to unnamed protein product [Vitis vinifera] (GB:CAO69341.1)

Appendix 4.7 (continued)

Gene name	Annotation
AT5G55050	GDSL-motif lipase/hydrolase family protein
AT5G55930	ATOPT1 (oligopeptide transporter 1); oligopeptide transporter
AT5G57220	CYP81F2 (cytochrome P450, family 81, subfamily F, polypeptide 2); oxygen binding
AT5G57350	AHA3 (Arabidopsis H (+)-ATPase 3); ATPase
AT5G59030	COPT1 (COPPER TRANSPORTER 1); copper ion transmembrane transporter
AT5G61820	Similar to MtN19-like protein [Pisum sativum] (GB:AAU14999.2); contains InterPro domain Stress up-regulated Nod 19 (InterPro:IPR011692)
AT5G64100	Peroxidase, putative
AT5G64250	2-nitropropane dioxygenase family / NPD family
AT5G64260	Phosphate-responsive protein, putative
AT5G65380	Ripening-responsive protein, putative
AT5G65510	AIL7 (AINTEGUMENTA-LIKE 7); DNA binding / transcription factor
AT5G65660	Hydroxyproline-rich glycoprotein family protein
No hits: AT1G14690; AT1G29030; AT1G52342; AT1G65032; AT2G14247; AT2G41905; AT2G44798; AT3G05937; AT3G14362; AT3G21820; AT3G45890; AT3G52561; AT3G58490; AT4G08555; AT4G15396; AT4G20690; AT4G32208; AT4G36850; AT5G21326	

Appendix 4.8: Genes whose products function in metal ion, ion, drug and transmembrane transport that are overrepresented in the *de novo* RNAs candidates gene list

Gene ID	Functions
AT1G16460	Encodes a cytoplasmic thiosulfate:cyanide sulfurtransferase; ATMST2, ATRDH2,
AT1G33110	MATE efflux family protein
AT1G60960	ATIRT3, IRON REGULATED TRANSPORTER 3, IRT3
AT1G66760	MATE efflux family protein
AT1G69870	NITRATE TRANSPORTER 1.7, NRT1.7, Responsible for Source-to-Sink Remobilization of Nitrate
AT2G23150	ATURAL RESISTANCE-ASSOCIATED MACROPHAGE PROTEIN 3,Encodes a member of the Nramp2 metal transporter family; like its homolog Atnramp4, localized in vacuolar membrane. Seedlings of double mutant, atnramp3-1 atnramp4-1, were arrested at early germination
AT2G36950	Heavy metal transport/detoxification superfamily protein
AT3G21690	MATE efflux family protein
AT3G23560	ABERRANT LATERAL ROOT FORMATION 5, ALF5
AT3G25410	Member of the multidrug and toxic compound extrusion (MATE) family protects roots from inhibitory compounds.
AT5G26340	SUGAR TRANSPORT PROTEIN 13
AT5G45380	DEGRADATION OF UREA 3 (DUR3)
AT5G57350	AHA3, ARABIDOPSIS THALIANA ARABIDOPSIS H(+)-ATPASE, ATAHA3,
AT5G59030	COPPER TRANSPORTER 1, COPT1
AT5G65380	MATE efflux family protein

Appendix 4.9: Genes that respond to inorganic substance such as toxin, metal ion, cadmium ion and nitric oxide, herbicide and oxidative stress that are overrepresented in *de novo* RNA candidate gene list

Gene ID	Functions
AT1G02930	Glutathione S-transferases (GSTs) 6,AtGSTF6
AT1G14870	ATPCR2, PCR2, PLANT CADMIUM RESISTANCE 2
AT1G37130	ARABIDOPSIS NITRATE REDUCTASE 2, ATNR2,NIA2
AT1G59700	GLUTATHIONE S-TRANSFERASE TAU 16,GSTU16
AT1G77760	NIA1, NITRATE REDUCTASE 1
AT2G02930	GLUTATHIONE S-TRANSFERASE 16,GST16
AT2G05710	ACO3, ACONITASE 3
AT2G18980	Peroxidase superfamily protein
AT2G23150	NATURAL RESISTANCE-ASSOCIATED MACROPHAGE PROTEIN 3,ATNRAMP3,
AT2G40000	ARABIDOPSIS ORTHOLOG OF SUGAR BEET HS1 PRO-1 2, ATHSPRO2
AT2G29420	ATGSTU7, GLUTATHIONE S-TRANSFERASE 25, GLUTATHIONE S-TRANSFERASE TAU 7, GST25, GSTU7
AT2G29490	GLUTATHIONE S-TRANSFERASE 19, GLUTATHIONE S-TRANSFERASE TAU 1, GST19, GSTU
AT3G14990	ATDJ1A, DJ-1 HOMOLOG A, DJ1A
AT3G22200	GABA-T, GAMMA-AMINOBUTYRATE TRANSAMINASE, HER1, HEXENAL RESPONSE1, POLLEN-PISTIL INCOMPATIBILITY 2, POP2
AT3G23560	ABERRANT LATERAL ROOT FORMATION 5,ALF5
AT4G02380	ARABIDOPSIS THALIANA LATE EMBRYOGENENSIS ABUNDANT LIKE 5, ATLEA5, SAG21, SENESCENCE-ASSOCIATED GENE 21
AT4G02520	Encodes glutathione transferase belonging to the phi class of GST
AT4G19880	Glutathione S-transferase family protein
AT4G37870	PCK1, PEPCK, PHOSPHOENOLPYRUVATE CARBOXYKINASE, PHOSPHOENOLPYRUVATE CARBOXYKINASE 1
AT5G07440	GDH2, GLUTAMATE DEHYDROGENASE 2
AT5G18170	GDH1, GLUTAMATE DEHYDROGENASE
AT5G39580	Peroxidase superfamily protein
AT5G51830	pfkB-like carbohydrate kinase family protein
AT5G64100	Peroxidase superfamily protein
AT5G64250	Aldolase-type TIM barrel family protein

Appendix 4.10: Genes whose products function in cell wall related events such as cell wall organization, cellular glucan metabolic process, glucan metabolic process; external encapsulating structure organization that are overrepresented in the *de novo* RNA candidate gene lists

Gene ID	Functions
AT2G01850	ATXTH27, ENDOXYLOGLUCAN TRANSFERASE A3, EXGT-A3, XTH27, XYLOGLUCAN ENDOTRANSGLUCOSYLASE/HYDROLASE 27
AT5G05290	ATEXP2, ATEXPA2, ATHEXP ALPHA 1.12, EXP2, EXPA2, EXPANSIN 2, EXPANSIN A2
AT4G14130	XTH15, XTR7, XYLOGLUCAN ENDOTRANSGLUCOSYLASE/HYDROLASE 15, XYLOGLUCAN ENDOTRANSGLYCOSYLASE 7
AT3G23730	XTH16, XYLOGLUCAN ENDOTRANSGLUCOSYLASE/HYDROLASE 16
AT3G48580	XTH11, XYLOGLUCAN ENDOTRANSGLUCOSYLASE/HYDROLASE 11
AT4G30270	MERI-5, MERI5B, MERISTEM 5, MERISTEM-5, SEN4, SENESCENCE 4, XTH24, XYLOGLUCAN ENDOTRANSGLUCOSYLASE/HYDROLASE 24

Appendix 4.11: High confidence gene list for stored, degraded mRNAs

Gene ID	Annotation
AT1G01240	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT2G46550.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO48635.1)
AT1G01470	LEA14 (LATE EMBRYOGENESIS ABUNDANT 14)
AT1G01650	Peptidase
AT1G01720	ATAF1 (Arabidopsis NAC domain containing protein 2); transcription factor
AT1G02310	Glycosyl hydrolase family protein 5 / cellulase family protein / (1-4)-beta-mannan endohydrolase, putative
AT1G02660	Lipase class 3 family protein
AT1G02700	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT4G02140.1); similar to hypothetical protein [Vitis vinifera] (GB:CAN70483.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO40074.1)
AT1G02890	AAA-type ATPase family protein
AT1G03470	Kinase interacting family protein
AT1G03770	Protein binding / zinc ion binding
AT1G03790	Zinc finger (CCCH-type) family protein
AT1G03880	CRU2 (CRUCIFERIN 2); nutrient reservoir
AT1G03890	Cupin family protein
AT1G03990	Alcohol oxidase-related
AT1G04560	AWPM-19-like membrane family protein
AT1G04660	Glycine-rich protein
AT1G05060	Similar to hypothetical protein [Vitis vinifera] (GB:CAN75913.1)
AT1G05340	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT2G32210.1)
AT1G05510	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT2G31985.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO45247.1); contains InterPro domain Protein of unknown function DUF1264 (InterPro:IPR010686)
AT1G06110	SKIP16 (SKP1/ASK-INTERACTING PROTEIN 16); protein binding
AT1G07310	C2 domain-containing protein
AT1G07400	17.8 kDa class I heat shock protein (HSP17.8-CI)
AT1G07430	Protein phosphatase 2C, putative / PP2C, putative
AT1G07500	Unknown protein
AT1G08050	Zinc finger (C3HC4-type RING finger) family protein
AT1G08170	Histone H2B family protein
AT1G08570	Thioredoxin family protein
AT1G09500	Cinnamyl-alcohol dehydrogenase family / CAD family
AT1G10070	ATBCAT-2; branched-chain-amino-acid transaminase/ catalytic
AT1G12130	Flavin-containing monooxygenase family protein / FMO family protein
AT1G13090	CYP71B28 (cytochrome P450, family 71, subfamily B, polypeptide 28); oxygen binding
AT1G13340	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G34220.2); similar to unknown [Carica papaya] (GB:ABS01355.1); contains InterPro domain Protein of unknown function DUF292, eukaryotic (InterPro:IPR005061)
AT1G13960	WRKY4 (WRKY DNA-binding protein 4); transcription factor
AT1G13990	Similar to unnamed protein product [Vitis vinifera] (GB:CAO68469.1)
AT1G14200	Zinc finger (C3HC4-type RING finger) family protein
AT1G14530	(TOM THREE HOMOLOG); virion binding
AT1G14930	Major latex protein-related / MLP-related

Appendix 4.11 (continued)

Gene ID	Annotation
AT1G14940	Major latex protein-related / MLP-related
AT1G14950	Major latex protein-related / MLP-related
AT1G15330	CBS domain-containing protein
AT1G16030	HSP70B (heat shock protein 70B); ATP binding
AT1G16730	Similar to unknown [Picea sitchensis] (GB:ABK21208.1)
AT1G16770	Similar to unnamed protein product [Vitis vinifera] (GB:CAO41707.1)
AT1G16850	Unknown protein
AT1G17010	Oxidoreductase, 2OG-Fe(II) oxygenase family protein
AT1G17640	RNA recognition motif (RRM)-containing protein
AT1G17810	BETA-TIP (BETA-TONOPLAST INTRINSIC PROTEIN); water channel
AT1G19540	Isoflavone reductase, putative
AT1G19660	Wound-responsive family protein
AT1G20070	Unknown protein
AT1G20870	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G54850.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO66281.1); contains InterPro domain HSP20-like chaperone (InterPro:IPR008978)
AT1G21400	2-oxoisovalerate dehydrogenase, putative / 3-methyl-2-oxobutanoate dehydrogenase, putative / branched-chain alpha-keto acid dehydrogenase E1 alpha subunit, putative
AT1G21410	SKP2A; protein binding
AT1G21680	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G21670.1); similar to hypothetical protein [Vitis vinifera] (GB:CAN73514.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO61906.1); similar to hypothetical protein OsJ_012725 [Oryza sativa (japonica cultivar-group)] (GB:EAZ29242.1); contains InterPro domain WD40-like Beta Propeller (InterPro:IPR011659); contains InterPro domain Six-bladed beta-propeller, TolB-like (InterPro:IPR011042)
AT1G22370	ATUGT85A5 (UDP-GLUCOSYL TRANSFERASE 85A5); transferase, transferring glycosyl groups
AT1G22380	ATUGT85A3 (UDP-GLUCOSYL TRANSFERASE 85A3); glucuronosyltransferase/ transcription factor/ transferase, transferring glycosyl groups
AT1G22600	Similar to late embryogenesis abundant domain-containing protein / LEA domain-containing protein [Arabidopsis thaliana] (TAIR:AT1G72100.1); similar to seed maturation protein PM27 [Glycine max] (GB:AAD30426.1); contains domain PTHR23241:SF1 (PTHR23241:SF1); contains domain PTHR23241 (PTHR23241)
AT1G22985	AP2 domain-containing transcription factor, putative
AT1G23050	Hydroxyproline-rich glycoprotein family protein
AT1G23070	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT4G38360.2); similar to unnamed protein product [Vitis vinifera] (GB:CAO65220.1); contains InterPro domain Protein of unknown function DUF300 (InterPro:IPR005178)
AT1G24600	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G67920.1)
AT1G24735	Caffeoyl-CoA 3-O-methyltransferase, putative
AT1G26400	FAD-binding domain-containing protein
AT1G27461	Similar to unnamed protein product [Vitis vinifera] (GB:CAO61483.1)
AT1G27990	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G52420.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO41629.1)
AT1G28360	ATERF12/ERF12 (ERF domain protein 12); DNA binding / transcription factor/ transcription repressor
AT1G29680	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G45690.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO47983.1); contains InterPro domain Protein of unknown function DUF1264 (InterPro:IPR010686)
AT1G30860	Protein binding / zinc ion binding
AT1G31170	ATSRX/SRX (SULFIREDOXIN); DNA binding / oxidoreductase, acting on sulfur group of donors
AT1G31750	Proline-rich family protein

Appendix 4.11 (continued)

Gene ID	Annotation
AT1G32380	Ribose-phosphate pyrophosphokinase 2 / phosphoribosyl diphosphate synthetase 2 (PRS2)
AT1G32560	Late embryogenesis abundant group 1 domain-containing protein / LEA group 1 domain-containing protein
AT1G34370	STOP1 (SENSITIVE TO PROTON RHIZOTOXICITY 1); nucleic acid binding / transcription factor/ zinc ion binding
AT1G46768	RAP2.1 (related to AP2 1); DNA binding / transcription factor
AT1G47540	Trypsin inhibitor, putative
AT1G47980	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT3G62730.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO44946.1)
AT1G48130	ATPER1 (Arabidopsis thaliana 1-cysteine peroxiredoxin 1); antioxidant
AT1G48470	GLN1;5 (GLUTAMINE SYNTHETASE 1;5); glutamate-ammonia ligase
AT1G50020	Similar to unnamed protein product [Vitis vinifera] (GB:CAO49863.1)
AT1G51090	Heavy-metal-associated domain-containing protein
AT1G52560	26.5 kDa class I small heat shock protein-like (HSP26.5-P)
AT1G52690	Late embryogenesis abundant protein, putative / LEA protein, putative
AT1G53540	17.6 kDa class I small heat shock protein (HSP17.6C-CI) (AA 1-156)
AT1G54050	17.4 kDa class III heat shock protein (HSP17.4-CIII)
AT1G54130	RSH3 (RELA/SPOT HOMOLOG 3); catalytic
AT1G54860	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G19250.1); similar to GPI-anchored protein-like protein II [Cucumis melo] (GB:ABR67421.1)
AT1G54870	Oxidoreductase
AT1G56600	ATGOLS2 (ARABIDOPSIS THALIANA GALACTINOL SYNTHASE 2); transferase, transferring glycosyl groups / transferase, transferring hexosyl groups
AT1G60680	AGD2 (ARF-GAP DOMAIN 2); aldo-keto reductase
AT1G62710	BETA-VPE (vacuolar processing enzyme beta); cysteine-type endopeptidase
AT1G64810	APO1 (ACCUMULATION OF PHOTOSYSTEM ONE 1)
AT1G64900	CYP89A2 (CYTOCHROME P450 89A2); oxygen binding
AT1G66770	Nodulin MtN3 family protein
AT1G67600	Similar to catalytic [Arabidopsis thaliana] (TAIR:AT1G24350.1); similar to unknown [Picea sitchensis] (GB:ABK26930.1); contains InterPro domain Acid phosphatase/vanadium-dependent haloperoxidase related (InterPro:IPR003832)
AT1G67920	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G24600.1)
AT1G68340	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G25370.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO68084.1); contains InterPro domain Protein of unknown function DUF1639 (InterPro:IPR012438)
AT1G68570	Proton-dependent oligopeptide transport (POT) family protein
AT1G69260	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G13740.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO41856.1); contains InterPro domain Protein of unknown function DUF1675 (InterPro:IPR012463)
AT1G69800	CBS domain-containing protein
AT1G70580	AOAT2 (GLUTAMATE: GLYOXYLATE AMINOTRANSFERASE 2); alanine transaminase
AT1G70810	C2 domain-containing protein
AT1G70840	MLP31 (MLP-LIKE PROTEIN 31)
AT1G71000	DNAJ heat shock N-terminal domain-containing protein
AT1G71140	MATE efflux family protein
AT1G72100	Late embryogenesis abundant domain-containing protein / LEA domain-containing protein
AT1G72660	Developmentally regulated GTP-binding protein, putative
AT1G73190	ALPHA-TIP/TIP3;1 (ALPHA-TONOPLAST INTRINSIC PROTEIN); water channel

Appendix 4.11 (continued)

Gene ID	Annotation
AT1G74310	ATHSP101 (HEAT SHOCK PROTEIN 101); ATP binding / ATPase
AT1G74370	Zinc finger (C3HC4-type RING finger) family protein
AT1G74410	Zinc finger (C3HC4-type RING finger) family protein
AT1G75490	DNA binding / transcription factor
AT1G75810	Similar to unnamed protein product [Vitis vinifera] (GB:CAO60969.1)
AT1G76590	Zinc-binding family protein
AT1G77000	ATSKP2;2/SKP2B (ARABIDOPSIS HOMOLOG OF HOMOLOG OF HUMAN SKP2 2); ubiquitin-protein ligase
AT1G77930	DNAJ heat shock N-terminal domain-containing protein
AT1G77950	AGL67; transcription factor
AT1G78070	WD-40 repeat family protein
AT1G80160	Lactoylglutathione lyase family protein / glyoxalase I family protein
AT1G80380	Phosphoribulokinase/uridine kinase-related
AT1G80570	F-box family protein (FBL14)
AT1G80920	J8; heat shock protein binding / unfolded protein binding
AT2G02120	LCR70/PDF2.1 (Low-molecular-weight cysteine-rich 70); protease inhibitor
AT2G02710	PAC motif-containing protein
AT2G02930	ATGSTF3 (GLUTATHIONE S-TRANSFERASE 16); glutathione transferase
AT2G03340	WRKY3 (WRKY DNA-binding protein 3); transcription factor
AT2G03520	ATUPS4 (ARABIDOPSIS THALIANA UREIDE PERMEASE 4)
AT2G04690	Cellular repressor of E1A-stimulated genes (CREG) family
AT2G04890	SCL21 (SCARECROW-LIKE 21); transcription factor
AT2G05580	Pseudogene, glycine-rich protein
AT2G06005	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G20580.1); similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G20580.2); similar to unknown [Populus trichocarpa] (GB:ABK93352.1)
AT2G06010	ORG4 (OBP3-RESPONSIVE GENE 4)
AT2G14520	CBS domain-containing protein
AT2G15010	Thionin, putative
AT2G16070	PDV2 (PLASTID DIVISION2)
AT2G16890	UDP-glucoronosyl/UDP-glucosyl transferase family protein
AT2G18250	ATCOAD (4-PHOSPHOPANTETHEINE ADENYLYLTRANSFERASE); nucleotidyltransferase/ pantetheine-phosphate adenylyltransferase
AT2G18340	Late embryogenesis abundant domain-containing protein / LEA domain-containing protein
AT2G18540	Cupin family protein
AT2G18570	UDP-glucoronosyl/UDP-glucosyl transferase family protein
AT2G18915	LKP2 (LOV KELCH PROTEIN 2); ubiquitin-protein ligase
AT2G19320	Unknown protein
AT2G19900	ATNADP-ME1 (NADP-MALIC ENZYME 1); malate dehydrogenase (oxaloacetate-decarboxylating) (NADP+)/ malic enzyme/ oxidoreductase, acting on NADH or NADPH, NAD or NADP as acceptor
AT2G19930	RNA-dependent RNA polymerase family protein
AT2G20560	DNAJ heat shock family protein
AT2G20770	GCL2 (GCR2-LIKE 2); catalytic
AT2G20920	Similar to unnamed protein product [Vitis vinifera] (GB:CAO47410.1)
AT2G21490	LEA (DEHYDRIN LEA)

Appendix 4.11 (continued)

Gene ID	Annotation
AT2G21780	Unknown protein
AT2G21820	Similar to hypothetical protein MtrDRAFT_AC155884g16v2 [Medicago truncatula] (GB:ABN08202.1)
AT2G22080	Unknown protein
AT2G22240	Inositol-3-phosphate synthase isozyme 2 / myo-inositol-1-phosphate synthase 2 / MI-1-P synthase 2 / IPS 2
AT2G23110	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT2G23120.1); similar to seed maturation protein PM35 [Glycine max] (GB:AAD51623.1)
AT2G23240	Plant EC metallothionein-like family 15 protein
AT2G23640	Reticulon family protein (RTNLB13)
AT2G25340	ATVAMP712 (Arabidopsis thaliana vesicle-associated membrane protein 712)
AT2G25625	Similar to unnamed protein product [Vitis vinifera] (GB:CAO44157.1)
AT2G25890	Glycine-rich protein / oleosin
AT2G26000	Zinc finger (C3HC4-type RING finger) family protein
AT2G26740	ATSEH (Arabidopsis thaliana soluble epoxide hydrolase); epoxide hydrolase
AT2G27310	F-box family protein
AT2G27380	ATEPR1 (Arabidopsis thaliana extensin proline-rich 1)
AT2G27940	Zinc finger (C3HC4-type RING finger) family protein
AT2G28400	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G60680.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO21845.1); contains InterPro domain Protein of unknown function DUF584 (InterPro:IPR007608)
AT2G28420	Lactoylglutathione lyase family protein / glyoxalase I family protein
AT2G28490	Cupin family protein
AT2G29300	Tropinone reductase, putative / tropine dehydrogenase, putative
AT2G29340	Short-chain dehydrogenase/reductase (SDR) family protein
AT2G29460	ATGSTU4 (GLUTATHIONE S-TRANSFERASE 22); glutathione transferase
AT2G30100	Ubiquitin family protein
AT2G30760	Unknown protein
AT2G32120	HSP70T-2; ATP binding
AT2G33070	Jacalin lectin family protein
AT2G33520	Similar to proline-rich family protein [Arabidopsis thaliana] (TAIR:AT1G12810.1)
AT2G33590	Cinnamoyl-CoA reductase family
AT2G34315	Disease resistance protein-related
AT2G34740	Protein phosphatase 2C, putative / PP2C, putative
AT2G35550	ATBPC7/BBR/BPC7/BPC7 (BASIC PENTACYSTEINE 7); DNA binding / transcription factor
AT2G36270	ABI5 (ABA INSENSITIVE 5); DNA binding / transcription activator/ transcription factor
AT2G36490	DML1/ROS1 (REPRESSOR OF SILENCING1); DNA N-glycosylase/ DNA-(apurinic or apyrimidinic site) lyase/ protein binding
AT2G36640	ATECP63 (EMBRYONIC CELL PROTEIN 63)
AT2G36780	UDP-glucoronosyl/UDP-glucosyl transferase family protein
AT2G37970	SOUL-1; binding
AT2G38820	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT3G22970.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO22614.1); contains InterPro domain Protein of unknown function DUF506, plant (InterPro:IPR006502)
AT2G38900	Serine protease inhibitor, potato inhibitor I-type family protein
AT2G38905	Hydrophobic protein, putative / low temperature and salt responsive protein, putative

Appendix 4.11 (continued)

Gene ID	Annotation
AT2G40170	ATEM6/GEA6 (ARABIDOPSIS EARLY METHIONINE-LABELLED 6)
AT2G41260	M17
AT2G41280	M10
AT2G42000	Plant EC metallothionein-like family 15 protein
AT2G42400	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G28520.2); similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G28520.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO69825.1)
AT2G42560	Late embryogenesis abundant domain-containing protein / LEA domain-containing protein
AT2G42750	DNAJ heat shock N-terminal domain-containing protein
AT2G42950	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G29820.1); similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G29820.2); similar to unnamed protein product [Vitis vinifera] (GB:CAO66903.1)
AT2G44990	CCD7 (more axillary growth 3)
AT2G45040	Matrix metalloproteinase
AT2G45210	Auxin-responsive protein-related
AT2G45290	Transketolase, putative
AT2G45510	CYP704A2 (cytochrome P450, family 704, subfamily A, polypeptide 2); oxygen binding
AT2G45560	CYP76C1 (cytochrome P450, family 76, subfamily C, polypeptide 1); heme binding / iron ion binding / monooxygenase
AT2G45570	CYP76C2 (cytochrome P450, family 76, subfamily C, polypeptide 2); oxygen binding
AT2G46240	BAG6 (ARABIDOPSIS THALIANA BCL-2-ASSOCIATED ATHANOGENE 6); calmodulin binding / protein binding
AT2G47180	ATGOLS1 (ARABIDOPSIS THALIANA GALACTINOL SYNTHASE 1); transferase, transferring hexosyl groups
AT2G47770	Benzodiazepine receptor-related
AT2G47820	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G09050.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO17561.1); contains InterPro domain Homeodomain-like (InterPro:IPR009057)
AT2G47890	Zinc finger (B-box type) family protein
AT3G01570	Glycine-rich protein / oleosin
AT3G01650	RGLG1 (RING DOMAIN LIGASE1); protein binding / zinc ion binding
AT3G01990	ACR6 (ACT Domain Repeat 6)
AT3G02875	ILR1 (IAA-LEUCINE RESISTANT 1); metallopeptidase
AT3G03310	Lecithin: cholesterol acyltransferase family protein / LACT family protein
AT3G03520	Phosphoesterase family protein
AT3G03620	MATE efflux family protein
AT3G04640	Glycine-rich protein
AT3G05120	ATGID1A/GID1A (GA INSENSITIVE DWARF1A); hydrolase
AT3G05200	ATL6 (Arabidopsis T?xicos en Levadura 6); protein binding / zinc ion binding
AT3G05260	Short-chain dehydrogenase/reductase (SDR) family protein
AT3G05510	Phospholipid/glycerol acyltransferase family protein
AT3G06380	ATTLP9 (TUBBY-LIKE PROTEIN 9); phosphoric diester hydrolase/ protein binding / transcription factor
AT3G06420	ATG8H (AUTOPHAGY 8H); microtubule binding
AT3G07250	Nuclear transport factor 2 (NTF2) family protein / RNA recognition motif (RRM)-containing protein
AT3G07370	ATCHIP/CHIP (CARBOXYL TERMINUS OF HSC70-INTERACTING PROTEIN); ubiquitin-protein ligase
AT3G07565	DNA binding

Appendix 4.11 (continued)

Gene ID	Annotation
AT3G07700	ABC1 family protein
AT3G10020	Similar to Os12g0147200 [Oryza sativa (japonica cultivar-group)] (GB:NP_001066153.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO15981.1); similar to Os11g0149200 [Oryza sativa (japonica cultivar-group)] (GB:NP_001065754.1)
AT3G10450	SCPL7; serine carboxypeptidase
AT3G11050	ATFER2 (FERRITIN 2); ferric iron binding
AT3G12580	HSP70 (heat shock protein 70); ATP binding
AT3G12955	Auxin-responsive protein-related
AT3G12960	Similar to seed maturation protein [Glycine tomentella] (GB:ABB72392.1)
AT3G13350	High mobility group (HMG1/2) family protein / ARID/BRIGHT DNA-binding domain-containing protein
AT3G14050	RSH2 (RELA-SPOT HOMOLOG); catalytic
AT3G14130	(S)-2-hydroxy-acid oxidase, peroxisomal, putative / glycolate oxidase, putative / short chain alpha-hydroxy acid oxidase, putative
AT3G14330	Pentatricopeptide (PPR) repeat-containing protein
AT3G14595	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G17080.1); similar to unknown [Populus trichocarpa] (GB:ABK95219.1); contains domain PTHR10052 (PTHR10052); contains domain PTHR10052:SF2 (PTHR10052:SF2)
AT3G15260	Protein phosphatase 2C, putative / PP2C, putative
AT3G15280	Similar to unnamed protein product [Vitis vinifera] (GB:CAO66421.1)
AT3G15670	Late embryogenesis abundant protein, putative / LEA protein, putative
AT3G15780	Unknown protein
AT3G16120	Dynein light chain, putative
AT3G16990	TENA/THI-4 family protein
AT3G17520	Late embryogenesis abundant domain-containing protein / LEA domain-containing protein
AT3G18570	Glycine-rich protein / oleosin
AT3G20250	APUM5 (ARABIDOPSIS PUMILIO 5); RNA binding
AT3G21370	Glycosyl hydrolase family 1 protein
AT3G21380	Similar to MBP1 (MYROSINASE-BINDING PROTEIN 1) [Arabidopsis thaliana] (TAIR:AT1G52040.1); similar to jasmonate inducible protein [Brassica napus] (GB:CAA72270.1); contains InterPro domain Mannose-binding lectin (InterPro:IPR001229)
AT3G22490	Late embryogenesis abundant protein, putative / LEA protein, putative
AT3G22500	ATECP31 (late embryogenesis abundant protein ECP31)
AT3G22640	Cupin family protein
AT3G22740	HMT3 (Homocysteine S-methyltransferase 3); homocysteine S-methyltransferase
AT3G23340	CKL10 (Casein Kinase I-like 10); casein kinase I/ kinase
AT3G23920	BAM1/BMY7/TR-BAMY (BETA-AMYLASE 1); beta-amylase
AT3G25870	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G13360.1); similar to unknown [Populus trichocarpa] (GB:ABK92948.1)
AT3G26580	Binding
AT3G26770	Short-chain dehydrogenase/reductase (SDR) family protein
AT3G27330	Zinc finger (C3HC4-type RING finger) family protein
AT3G27660	OLEO4 (OLEOSIN4)
AT3G27870	Haloacid dehalogenase-like hydrolase family protein
AT3G29090	Pectinesterase family protein
AT3G30460	Zinc finger (C3HC4-type RING finger) family protein

Appendix 4.11 (continued)

Gene ID	Annotation
AT3G44830	Lecithin: cholesterol acyltransferase family protein / LACT family protein
AT3G45900	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G53020.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO66703.1); contains domain PTHR21596:SF1 (PTHR21596:SF1); contains domain PTHR21596 (PTHR21596)
AT3G46230	ATHSP17.4 (Arabidopsis thaliana heat shock protein 17.4)
AT3G46660	UDP-glucoronosyl/UDP-glucosyl transferase family protein
AT3G46670	UDP-glucoronosyl/UDP-glucosyl transferase family protein
AT3G47080	Binding
AT3G47340	ASN1 (DARK INDUCIBLE 6)
AT3G47950	AHA4 (Arabidopsis H(+)-ATPase 4); ATPase
AT3G50970	LTI30 (LOW TEMPERATURE-INDUCED 30)
AT3G50980	XERO1 (DEHYDRIN XERO 1)
AT3G51810	ATEM1 (Early methionine labelled)
AT3G51860	CAX3 (cation exchanger 3); cation:cation antiporter
AT3G51880	HMGB1 (HIGH MOBILITY GROUP B 1); transcription factor
AT3G52230	Similar to unknown [Populus trichocarpa] (GB:ABK93315.1)
AT3G53040	Late embryogenesis abundant protein, putative / LEA protein, putative
AT3G53960	Proton-dependent oligopeptide transport (POT) family protein
AT3G54940	Cysteine proteinase, putative
AT3G56350	Superoxide dismutase (Mn), putative / manganese superoxide dismutase, putative
AT3G57020	Strictosidine synthase family protein
AT3G58450	Universal stress protein (USP) family protein
AT3G59940	Kelch repeat-containing F-box family protein
AT3G60190	ADL4/ADLP2/DRP1E/EDR3 (DYNAMIN-LIKE PROTEIN 4); GTP binding / GTPase
AT3G61040	CYP76C7 (cytochrome P450, family 76, subfamily C, polypeptide 7); oxygen binding
AT3G62090	PIL2 (PHYTOCHROME INTERACTING FACTOR 3-LIKE 2); transcription factor
AT3G62590	Lipase class 3 family protein
AT3G62700	ATMRP10 (Arabidopsis thaliana multidrug resistance-associated protein 10)
AT3G63040	Unknown protein
AT3G63340	Protein phosphatase 2C-related / PP2C-related
AT4G02280	SUS3; UDP-glycosyltransferase/ sucrose synthase/ transferase, transferring glycosyl groups
AT4G02410	Lectin protein kinase family protein
AT4G02690	Glutamate binding /
AT4G03200	Catalytic
AT4G04870	CLS (CARDIOLIPIN SYNTHASE); cardiolipin synthase/ phosphatidyltransferase
AT4G05070	Unknown protein
AT4G09600	GASA3 (GAST1 PROTEIN HOMOLOG 3)
AT4G09610	GASA2 (GAST1 PROTEIN HOMOLOG 2)
AT4G10020	ATHSD5 (HYDROXYSTEROID DEHYDROGENASE 5); oxidoreductase
AT4G11310	Cysteine proteinase, putative
AT4G11570	Haloacid dehalogenase-like hydrolase family protein

Appendix 4.11 (continued)

Gene ID	Annotation
AT4G12130	Aminomethyltransferase
AT4G12290	Copper amine oxidase, putative
AT4G12400	Stress-inducible protein, putative
AT4G12750	Sequence-specific DNA binding / transcription factor
AT4G13010	Oxidoreductase, zinc-binding dehydrogenase family protein
AT4G13160	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT4G13630.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO43544.1); contains InterPro domain Protein of unknown function DUF593 (InterPro:IPR007656)
AT4G13250	Short-chain dehydrogenase/reductase (SDR) family protein
AT4G13530	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT4G10080.1); similar to unknown protein [Oryza sativa (japonica cultivar-group)] (GB:BAD82228.1)
AT4G13830	J20 (DNAJ-LIKE 20); heat shock protein binding
AT4G15620	Integral membrane family protein
AT4G16160	ATOEP16-2/ATOEP16-S; P-P-bond-hydrolysis-driven protein transmembrane transporter
AT4G16620	Integral membrane family protein / nodulin MtN21-related
AT4G17840	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT2G35260.1); similar to hypothetical protein 40.t00061 [Brassica oleracea] (GB:ABD65174.1)
AT4G18130	PHYE (PHYTOCHROME DEFECTIVE E); G-protein coupled photoreceptor/ signal transducer
AT4G18530	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G11170.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO46884.1); contains InterPro domain Protein of unknown function DUF707 (InterPro:IPR007877)
AT4G18650	Transcription factor-related
AT4G18920	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G45690.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO47983.1); contains InterPro domain Protein of unknown function DUF1264 (InterPro:IPR010686)
AT4G19170	NCED4 (NINE-CIS-EPOXYCAROTENOID DIOXYGENASE 4)
AT4G19390	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G13720.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO63752.1); contains InterPro domain Uncharacterised conserved protein UCP022348 (InterPro:IPR016804); contains InterPro domain Protein of unknown function UPF0114 (InterPro:IPR005134)
AT4G21020	Late embryogenesis abundant domain-containing protein / LEA domain-containing protein
AT4G21680	Proton-dependent oligopeptide transport (POT) family protein
AT4G22753	SMO1-3 (STEROL 4-ALPHA METHYL OXIDASE); catalytic
AT4G22920	ATNYE1/NYE1 (NON-YELLOWING 1)
AT4G23990	ATCSLG3 (Cellulose synthase-like G3); transferase/ transferase, transferring glycosyl groups
AT4G25140	OLEO1 (OLEOSIN1)
AT4G25170	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G61490.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO60860.1); contains InterPro domain Uncharacterised conserved protein UCP012943 (InterPro:IPR016606)
AT4G25200	ATHSP23.6-MITO (MITOCHONDRION-LOCALIZED SMALL HEAT SHOCK PROTEIN 23.6)
AT4G25230	RIN2 (RPM1 INTERACTING PROTEIN 2); protein binding / zinc ion binding
AT4G25580	Stress-responsive protein-related
AT4G26050	Leucine-rich repeat family protein
AT4G26700	ATFIM1 (Arabidopsis thaliana fimbrin 1); actin binding
AT4G26740	ATS1 (ARABIDOPSIS THALIANA SEED GENE 1); calcium ion binding
AT4G27070	TSB2 (TRYPTOPHAN SYNTHASE BETA-SUBUNIT); tryptophan synthase
AT4G27140	2S seed storage protein 1 / 2S albumin storage protein / NWMU1-2S albumin 1
AT4G27150	2S seed storage protein 2 / 2S albumin storage protein / NWMU2-2S albumin 2

Appendix 4.11 (continued)

Gene ID	Annotation
AT4G27160	AT2S3; lipid binding / nutrient reservoir
AT4G27170	2S seed storage protein 4 / 2S albumin storage protein / NWMU2-2S albumin 4
AT4G27410	RD26 (RESPONSIVE TO DESSICATION 26); transcription factor
AT4G27460	CBS domain-containing protein
AT4G27530	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G53895.1)
AT4G27780	ACBP2 (ACYL-COA BINDING PROTEIN ACBP 2)
AT4G27990	YGGT family protein
AT4G28020	Similar to unnamed protein product [Vitis vinifera] (GB:CAO43327.1); contains InterPro domain Protein of unknown function UPF0066 (InterPro:IPR001378)
AT4G28390	AAC3 (ADP/ATP CARRIER 3); ATP:ADP antiporter/ binding
AT4G28520	CRU3 (CRUCIFERIN 3); nutrient reservoir
AT4G29820	ATCFIM-25/CFIM-25 (ARABIDOPSIS HOMOLOG OF CFIM-25)
AT4G30935	WRKY32 (WRKY DNA-binding protein 32); transcription factor
AT4G31270	Transcription factor
AT4G31540	ATEXO70G1 (exocyst subunit EXO70 family protein G1); protein binding
AT4G31830	Similar to unnamed protein product [Vitis vinifera] (GB:CAO44350.1)
AT4G31860	Protein phosphatase 2C, putative / PP2C, putative
AT4G32040	KNAT5 (KNOTTED1-LIKE HOMEOBOX GENE 5); transcription factor
AT4G32300	Lectin protein kinase family protein
AT4G32770	VTE1 (VITAMIN E DEFICIENT 1)
AT4G33980	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G42900.2); similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G42900.3); similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G42900.1); similar to hypothetical protein [Vitis vinifera] (GB:CAN64989.1)
AT4G34890	ATXDH1 (XANTHINE DEHYDROGENASE 1); xanthine dehydrogenase
AT4G35160	O-methyltransferase family 2 protein
AT4G36530	Hydrolase, alpha/beta fold family protein
AT4G36600	Late embryogenesis abundant domain-containing protein / LEA domain-containing protein
AT4G36620	Zinc finger (GATA type) family protein
AT4G36700	Cupin family protein
AT4G36900	RAP2.10 (related to AP2 10); DNA binding / transcription factor
AT4G37370	CYP81D8 (cytochrome P450, family 81, subfamily D, polypeptide 8); oxygen binding
AT4G38380	MATE efflux protein-related
AT4G38740	ROC1 (rotamase CyP 1); peptidyl-prolyl cis-trans isomerase
AT4G38810	Calcium-binding EF hand family protein
AT4G39800	MI-1-P SYNTHASE (Myo-inositol-1-phosphate synthase); inositol-3-phosphate synthase
AT4G39890	AtRABH1c (Arabidopsis Rab GTPase homolog H1c); GTP binding
AT5G01300	Phosphatidylethanolamine-binding family protein
AT5G01520	Zinc finger (C3HC4-type RING finger) family protein
AT5G01670	Aldose reductase, putative
AT5G01880	Zinc finger (C3HC4-type RING finger) family protein
AT5G02020	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT3G55646.1); similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G59080.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO15731.1)

Appendix 4.11 (continued)

Gene ID	Annotation
AT5G03180	Zinc finger (C3HC4-type RING finger) family protein
AT5G03210	Unknown protein
AT5G03795	Oxidoreductase
AT5G04010	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G03920.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO23344.1); contains domain SSF81383 (SSF81383)
AT5G04250	OTU-like cysteine protease family protein
AT5G04500	Glycosyltransferase family protein 47
AT5G05220	Similar to hypothetical protein [Vitis vinifera] (GB:CAN82940.1)
AT5G05230	Ubiquitin-protein ligase
AT5G05250	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT3G56360.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO41488.1)
AT5G05320	Monooxygenase, putative (MO3)
AT5G05410	DREB2A (DRE-BINDING PROTEIN 2A); DNA binding / transcription activator/ transcription factor
AT5G06750	Protein phosphatase 2C family protein / PP2C family protein
AT5G06760	Late embryogenesis abundant group 1 domain-containing protein / LEA group 1 domain-containing protein
AT5G07330	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G63060.1); similar to unknown [Xerophyta humilis] (GB:AAT45004.1)
AT5G07360	Amidase family protein
AT5G09990	PROPEP5 (Elicitor peptide 5 precursor)
AT5G10650	Zinc finger (C3HC4-type RING finger) family protein
AT5G10695	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G57123.1); similar to unknown [Picea sitchensis] (GB:ABK22689.1)
AT5G11840	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G67370.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO43828.1); contains InterPro domain Protein of unknown function DUF1230 (InterPro:IPR009631)
AT5G12030	AT-HSP17.6A (Arabidopsis thaliana heat shock protein 17.6A)
AT5G13800	Hydrolase, alpha/beta fold family protein
AT5G14120	Nodulin family protein
AT5G15330	SPX (SYG1/Pho81/XPR1) domain-containing protein
AT5G16460	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G29760.1); similar to hypothetical protein [Vitis vinifera] (GB:CAN83562.1); contains domain PTHR21212 (PTHR21212)
AT5G16650	DNAJ heat shock N-terminal domain-containing protein
AT5G16970	AT-AER (ALKENAL REDUCTASE); 2-alkenal reductase
AT5G18130	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT3G03870.2); similar to unknown [Medicago truncatula] (GB:ABK28852.1)
AT5G18250	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT3G04040.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO23501.1)
AT5G18450	AP2 domain-containing transcription factor, putative
AT5G19430	Zinc finger (C3HC4-type RING finger) family protein
AT5G20960	AAO1 (ALDEHYDE OXIDASE 1)
AT5G22290	ANAC089 (Arabidopsis NAC domain containing protein 89); transcription factor
AT5G22470	NAD+ ADP-ribosyltransferase
AT5G22690	Disease resistance protein (TIR-NBS-LRR class), putative
AT5G23230	NIC2 (NICOTINAMIDASE 2); catalytic/ nicotinamidase
AT5G23340	Protein binding
AT5G24160	Squalene monooxygenase 1,2 / squalene epoxidase 1,2 (SQP1,2)

Appendix 4.11 (continued)

Gene ID	Annotation
AT5G24970	ABC1 family protein
AT5G25180	CYP71B14 (cytochrome P450, family 71, subfamily B, polypeptide 14); oxygen binding
AT5G35660	Pseudogene similar to protease inhibitor/seed storage/lipid transfer protein (LTP) family protein
AT5G37670	15.7 kDa class I-related small heat shock protein-like (HSP15.7-CI)
AT5G37680	ATARLA1A (ADP-ribosylation factor-like A1A); GTP binding
AT5G39520	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G39530.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO15021.1)
AT5G39660	CDF2 (CYCLING DOF FACTOR 2); DNA binding / protein binding / transcription factor
AT5G39720	AIG2L (AVIRULENCE INDUCED GENE 2 LIKE PROTEIN)
AT5G40420	OLEO2 (OLEOSIN 2)
AT5G40840	SYN2 (Sister chromatid cohesion 1 (SCC1) protein homolog 2)
AT5G41610	ATCHX18 (cation/hydrogen exchanger 18); monovalent cation:proton antiporter
AT5G42290	Transcription activator-related
AT5G42690	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT4G37080.2); similar to unnamed protein product [Vitis vinifera] (GB:CAO61290.1); contains InterPro domain Protein of unknown function DUF547 (InterPro:IPR006869)
AT5G43770	Proline-rich family protein
AT5G44000	Glutathione S-transferase C-terminal domain-containing protein
AT5G44120	CRA1 (CRUCIFERINA); nutrient reservoir
AT5G44280	Protein binding / zinc ion binding
AT5G44310	Late embryogenesis abundant domain-containing protein / LEA domain-containing protein
AT5G44670	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT4G20170.1); similar to Os06g0328800 [Oryza sativa (japonica cultivar-group)] (GB:NP_001057533.1); similar to Os02g0712500 [Oryza sativa (japonica cultivar-group)] (GB:NP_001047907.1); similar to unknown protein [Oryza sativa (japonica cultivar-group)] (GB:BAD72474.1); contains InterPro domain Protein of unknown function DUF23 (InterPro:IPR008166)
AT5G45160	Root hair defective 3 GTP-binding (RHD3) family protein
AT5G45310	Similar to unnamed protein product [Vitis vinifera] (GB:CAO63757.1)
AT5G45630	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT4G18980.1); similar to unknown [Cucumis sativus] (GB:ABY56081.1); similar to unknown [Cucumis melo] (GB:ABQ53634.1); contains InterPro domain Protein of unknown function DUF584 (InterPro:IPR007608)
AT5G45690	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT4G18920.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO47983.1); contains InterPro domain Protein of unknown function DUF1264 (InterPro:IPR010686)
AT5G45830	DOG1 (DELAY OF GERMINATION 1)
AT5G47810	Phosphofructokinase family protein
AT5G49990	Xanthine/uracil permease family protein
AT5G50170	C2 domain-containing protein / GRAM domain-containing protein
AT5G50360	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G63350.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO68256.1)
AT5G51070	ERD1 (EARLY RESPONSIVE TO DEHYDRATION 1); ATP binding / ATPase
AT5G51210	OLEO3 (OLEOSIN3)
AT5G51760	AHG1 (ABA-HYPERSENSITIVE GERMINATION 1); protein serine/threonine phosphatase
AT5G51830	PfkB-type carbohydrate kinase family protein
AT5G52300	LTI65/RD29B (RESPONSIVE TO DESSICATION 29B)
AT5G52420	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G23920.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO49441.1)
AT5G52570	BETA-OHASE 2 (BETA-CAROTENE HYDROXYLASE 2); beta-carotene hydroxylase

Appendix 4.11 (continued)

Gene ID	Annotation
AT5G52580	RAB GTPase activator
AT5G52580	RAB GTPase activator
AT5G52640	HSP81-1 (HEAT SHOCK PROTEIN 81-1); ATP binding / unfolded protein binding
AT5G53220	Similar to unnamed protein product [Vitis vinifera] (GB:CAO69343.1)
AT5G54000	Oxidoreductase, 2OG-Fe(II) oxygenase family protein
AT5G54070	AT-HSFA9 (ARABIDOPSIS THALIANA HEAT SHOCK TRANSCRIPTION FACTOR A9); DNA binding / transcription factor
AT5G54290	Cytochrome c biogenesis protein family
AT5G54730	AtATG18f (Arabidopsis thaliana homolog of yeast autophagy 18 (ATG18) f)
AT5G54740	Protease inhibitor/seed storage/lipid transfer protein (LTP) family protein
AT5G55240	Caleosin-related family protein / embryo-specific protein, putative
AT5G55750	Hydroxyproline-rich glycoprotein family protein
AT5G56100	Glycine-rich protein / oleosin
AT5G57050	ABI2 (ABA INSENSITIVE 2); protein serine/threonine phosphatase
AT5G57260	CYP71B10 (cytochrome P450, family 71, subfamily B, polypeptide 10); oxygen binding
AT5G57550	XTR3 (XYLOGLUCAN ENDOTRANSGLYCOSYLASE 3); hydrolase, acting on glycosyl bonds
AT5G57790	Unknown protein
AT5G57900	SKIP1 (SKP1 INTERACTING PARTNER 1)
AT5G58160	Actin binding
AT5G59170	Proline-rich family protein
AT5G60220	TET4 (TETRASPANIN4)
AT5G60760	2-phosphoglycerate kinase-related
AT5G61590	AP2 domain-containing transcription factor family protein
AT5G62220	Exostosin family protein
AT5G62490	ATHVA22B (Arabidopsis thaliana HVA22 homologue B)
AT5G63190	MA3 domain-containing protein
AT5G64080	Protease inhibitor/seed storage/lipid transfer protein (LTP) family protein
AT5G64210	AOX2 (alternative oxidase 2); alternative oxidase
AT5G65070	MAF4 (MADS AFFECTING FLOWERING 4)
AT5G65100	Ethylene insensitive 3 family protein
AT5G65550	UDP-glucoronosyl/UDP-glucosyl transferase family protein
AT5G65890	ACR1 (ACT DOMAIN REPEAT 1)
AT5G66110	Metal ion binding
AT5G66400	RAB18 (RESPONSIVE TO ABA 18)
AT5G66430	S-adenosyl-L-methionine:carboxyl methyltransferase family protein
AT5G66730	Zinc finger (C2H2 type) family protein
AT5G66780	Similar to unknown [Ammopiptanthus mongolicus] (GB:AAW33981.1)
AT5G67030	ABA1 (ABA DEFICIENT 1); zeaxanthin epoxidase
AT5G67480	BT4 (BTB AND TAZ DOMAIN PROTEIN 4); protein binding / transcription regulator
No hits: AT1G03120; AT1G07473; AT1G07645; AT1G07985; AT1G11175; AT1G12060; AT1G25422; AT1G32710; AT1G46554; AT1G47056; AT1G48990; AT1G55152; AT1G62420; AT1G64065; AT1G67365; AT1G69540; AT1G70800; AT2G01275; AT2G01340; AT2G13665; AT2G21720; AT2G22821; AT2G31985; AT2G33770; AT2G34355; AT2G45360; AT3G03341; AT3G07500; AT3G10130; AT3G14590; AT3G44290; AT3G48270;	

Appendix 4.11 (continued)

Gene ID	Annotation
AT3G48660; AT4G06746; AT4G11910; AT4G11911; AT4G12680; AT4G13395; AT4G15396; AT4G15563; AT4G18680; AT4G21320; AT4G25707; AT4G33467; AT4G36720; AT5G02750; AT5G10000; AT5G15820; AT5G20510; AT5G36100; AT5G40382; AT5G45870; AT5G50240; AT5G53260; AT5G53270; AT5G53895; AT5G54165; AT5G55135; AT5G56550; AT5G57123; AT5G61490; AT5G63350; AT5G64750; AT5G65165; AT5G65495; AT5G66580	

Appendix 4.12: The genes encoding lipid storage or localization proteins that are overrepresented in the stored, degraded mRNA gene list

Gene ID	Function
AT1G48990	Oleosin family protein, involve in lipid storage
AT2G18250	4-PHOSPHOPANTETHEINE ADENYLYLTRANSFERASE, ATCOAD, COAD,response to osmotic stress, lipid metabolic process
AT2G25890	Oleosin family protein, involve in lipid storage and seed dormancy process
AT3G01570	Oleosin family protein, involve in seed germination
AT3G18570	Oleosin family protein, involve in lipid storage
AT3G27660	OLEOSIN 4, involve in seed germination, seed oilbody biogenesis, embryo development ending in seed dormancy,lipid storage
AT3G27870	ATPase E1-E2 type family protein / haloacid dehalogenase-like hydrolase family protein
AT4G25140	OLEOSIN 1, involve in seed germination, seed oilbody biogenesis, protein import into nucleus
AT4G27140	SEED STORAGE ALBUMIN 1, SESA1, lipid transport
AT4G27150	SEED STORAGE ALBUMIN 2, SESA2, GA signaling pathway, GA biosynthetic process, lipid transport
AT4G27160	SEED STORAGE ALBUMIN 3, SESA3, lipid storage, lipid transport, seed germination, response to freezing
AT4G27170	SEED STORAGE ALBUMIN 4, SESA4,lipid transport
AT4G27780	ACBP2, ACYL-COA BINDING PROTEIN 2
AT5G40420	OLEOSIN 1, involve in seed germination, seed oilbody biogenesis, lipid storage, embryo development ending in seed dormancy
AT5G51210	OLEOSIN3, lipid storage
AT5G54740	STORAGE ALBUMIN 5, SESA5, lipid transport
AT5G56100	glycine-rich protein / oleosin,lipid storage
AT5G64080	ATXYP1, XYLOGEN PROTEIN 1, XYP1,Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein, lipid transport

Appendix 4.13: The relative expression level of genes whose products function in lipid storage or localization in the globular-, heart, torpedo, or cotyledon-stage embryo, in dry seed or 24hr imbibed seed

Lipid storage or localization	Globular	Heart	Torpedo	Cotyledon	Dry seed	24hr imbibed seed
AT2G18250	77	63	77	149	340	75
AT2G25890	6	26	360	4593	2460	314
AT3G01570	10	1266	5108	10174	2541	600
AT3G18570	10	71	1156	3810	1408	41
AT3G27660	53	903	4576	10779	2230	396
AT3G27870	70	58	84	433	873	86
AT4G25140	3	956	4930	11583	1199	977
AT4G27140	5	470	5007	10839	2019	486
AT4G27150	0	732	3799	7835	2104	491
AT4G27160	3	995	4299	8912	2641	1051
AT4G27170	2	489	5575	12101	2821	594
AT4G27780	66	79	173	886	742	69
AT5G40420	3	1264	4770	9777	1950	992
AT5G51210	20	2110	4551	1370	768	203
AT5G54740	6	168	3205	12132	1670	1685
AT5G56100	71	79	94	1513	2393	384
AT5G64080	696	889	3021	4554	1567	726

Appendix 4.14: The normalized expression level of genes whose products function in lipid storage or localization in the globular-, heart, torpedo, or cotyledon-stage embryo, in dry seed or 24hr imbibed seed

Gene name	Globular	Heart	Torpedo	Cotyledon	Dry seed	24hr imbibed seed
AT2G18250	0.591549296	0.483994878	0.591549296	1.1446863	2.612035851	0.576184379
AT2G25890	0.004639773	0.020105684	0.27838639	3.551746359	1.902306998	0.242814796
AT3G01570	0.00304584	0.38560333	1.555815016	3.098837504	0.773947916	0.182750393
AT3G18570	0.009236453	0.065578818	1.06773399	3.51908867	1.300492611	0.037869458
AT3G27660	0.016792523	0.286106564	1.449860062	3.415218884	0.706553308	0.125468659
AT3G27870	0.261845387	0.216957606	0.314214464	1.619700748	3.265586035	0.321695761
AT4G25140	0.000916124	0.291938111	1.505496743	3.537153909	0.366144137	0.298350977
AT4G27140	0.001593541	0.14979284	1.595771805	3.45447785	0.643471794	0.15489217
AT4G27150	0	0.293563264	1.523561259	3.142169641	0.843793864	0.196911971
AT4G27160	0.00100553	0.333500922	1.440925088	2.987095693	0.885201944	0.352270823
AT4G27170	0.000556019	0.135946622	1.549902697	3.364192383	0.784264665	0.165137615
AT4G27780	0.196526055	0.235235732	0.515136476	2.6382134	2.20942928	0.205459057
AT5G40420	0.000959693	0.404350608	1.525911708	3.127639155	0.623800384	0.317338452
AT5G51210	0.01330082	1.403236533	3.02660164	0.911106185	0.510751496	0.135003325
AT5G54740	0.001908195	0.05342945	1.019293968	3.858369554	0.531114174	0.53588466
AT5G56100	0.093956771	0.104543449	0.124393472	2.002205558	3.166740185	0.508160565
AT5G64080	0.364620623	0.465729503	1.582642103	2.385750458	0.820920283	0.38033703

Appendix 4.15: The genes that respond to heat, temperature stimulus, oxidative stress and light intensity, that are overrepresented in the stored, degraded mRNA gene list

Gene ID	Functions
AT1G01470	LATE EMBRYOGENESIS ABUNDANT 14, LEA14, LIGHT STRESS-REGULATED 3, LSR3
AT1G07400	HSP20-like chaperones superfamily protein,responsive to heat and oxidative stress
AT1G13340	Regulator of Vps4 activity in the MVB pathway protein
AT1G16030	HEAT SHOCK PROTEIN 70B, HSP70B
AT1G31170	Encodes a cysteine-sulfinic acid reductase (sulfiredoxin - EC 1.8.98.2) capable of reducing overoxidized plastidic 2-Cys-Prx involved in peroxide detoxification and response to oxidative stress
AT1G52560	HSP20-like chaperones superfamily protein
AT1G53540	HSP20-like chaperones superfamily protein
AT1G54050	HSP20-like chaperones superfamily protein
AT1G74310	THSP101, HEAT SHOCK PROTEIN 101, HOT1, HSP101
AT1G77000	ARABIDOPSIS HOMOLOG OF HOMOLOG OF HUMAN SKP2 2, ATSKP2;2, SKP2B
AT2G22080	Unknown protein
AT2G22240	ATIPS2, ATMIPS2, INOSITOL 3-PHOSPHATE SYNTHASE 2, MIPS2, MYO-INOSITOL-1-PHOSPHATE SYNTHASE 2, MYO-INOSITOL-1-PHOSTPATE SYNTHASE 2
AT2G32120	HEAT-SHOCK PROTEIN 70T-2, HSP70T-2
AT2G38905	Low temperature and salt responsive protein family
AT2G46240	ARABIDOPSIS THALIANA BCL-2-ASSOCIATED ATHANOGENE 6, ATBAG6, BAG6, BCL-2-ASSOCIATED ATHANOGENE 6
AT2G47180	ATGOLS1, GALACTINOL SYNTHASE 1, GOLS1
AT3G07370	ATCHIP, CARBOXYL TERMINUS OF HSC70-INTERACTING PROTEIN, CHIP
AT3G10020	Unknown protein
AT3G11050	Ferritin 2 (FER2)
AT3G12580	ARABIDOPSIS HEAT SHOCK PROTEIN 70, ATHSP70, HEAT SHOCK PROTEIN 70, HSP70
AT3G46230	ARABIDOPSIS THALIANA HEAT SHOCK PROTEIN 17.4, ATHSP17.4, HEAT SHOCK PROTEIN 17.4, HSP17.4
AT3G50970	LOW TEMPERATURE-INDUCED 30, LTI30, XERO2
AT3G56350	Iron/manganese superoxide dismutase family protein
AT4G12400	HOP3
AT4G25200	ATHSP23.6-MITO, HSP23.6-MITO, MITOCHONDRION-LOCALIZED SMALL HEAT SHOCK PROTEIN 23.6
AT4G32770	ATSDX1, SUCROSE EXPORT DEFECTIVE 1, VITAMIN E DEFICIENT 1, VTE1
AT5G05410	DEHYDRATION-RESPONSIVE ELEMENT BINDING PROTEIN 2, DRE-BINDING PROTEIN 2A, DREB2, DREB2A
AT5G12030	AT-HSP17.6A, HEAT SHOCK PROTEIN 17.6, HEAT SHOCK PROTEIN 17.6A, HSP17.6, HSP17.6A
AT5G16970	AER, ALKENAL REDUCTASE, AT-AER
AT5G37670	HSP20-like chaperones superfamily protein
AT5G52300	LOW-TEMPERATURE-INDUCED 65, LTI65, RD29B, RESPONSIVE TO DESSICATION 29B
AT5G52640	a cytosolic heat shock protein AtHSP90.1
AT5G57050	ABA INSENSITIVE 2, ABI2, ATABI2
AT5G67030	Encodes a single copy zeaxanthin epoxidase gene that functions in first step of the biosynthesis of the abiotic stress hormone abscisic acid (ABA).

Appendix 4.16: The relative expression level of genes that respond to heat, temperature stimulus, oxidative stress or light intensity, in globular-, heart, torpedo, or cotyledon-stage embryos, in dry seed or 24hr imbibed seed

Gene ID	Globular	Heart	Torpedo	Cotyledon	Dry seed	24hr imbibed seed
AT1G01470	676	1037	892	3901	2879	195
AT1G07400	11	37	26	102	188	32
AT1G13340	14	79	110	139	346	15
AT1G16030	37	34	36	1792	2608	66
AT1G31170	154	247	109	223	528	71
AT1G52560	4	8	13	401	1635	17
AT1G53540	6	6	6	113	992	23
AT1G54050	85	32	45	743	906	19
AT1G74310	7	42	68	1472	1908	44
AT1G77000	112	104	109	557	1494	71
AT2G22080	121	79	47	321	160	15
AT2G22240	728	284	445	2340	2476	19
AT2G32120	30	27	27	1184	2076	44
AT2G38905	30	120	1909	5430	2385	365
AT2G46240	14	28	24	105	281	12
AT2G47180	65	68	8	1801	2194	51
AT3G07370	31	32	22	58	134	35
AT3G10020	289	656	146	1501	1795	130
AT3G11050	20	37	399	6157	2800	218
AT3G12580	3	260	446	4269	2358	794
AT3G46230	4	7	6	4784	3044	148
AT3G50970	147	361	139	3830	2892	71
AT3G56350	5	9	215	6915	2550	288
AT4G12400	6	13	8	83	317	29
AT4G25200	1	1	1	2	1115	7
AT4G32770	199	360	760	922	525	66
AT5G05410	44	495	62	308	1790	466
AT5G12030	3	11	12	4268	2623	130
AT5G16970	75	128	183	833	1043	185
AT5G37670	14	8	12	260	673	21
AT5G52300	2	6	17	4779	2898	216
AT5G52640	9	80	89	948	2079	263
AT5G57050	40	89	42	283	451	22
AT5G67030	417	791	154	835	1672	125

Appendix 4.17: The normalized expression level of genes that respond to heat, temperature stimulus, oxidative stress and light intensity, in globular-, heart, torpedo, cotyledon-stage embryos, in dry seed or 24hr imbibed seed

Gene name	Globular	Heart	Torpedo	Cotyledon	Dry seed	24hr imbibed seed
AT1G01470	0.423382046	0.649478079	0.558663883	2.443215031	1.803131524	0.122129436
AT1G07400	0.166666667	0.560606061	0.393939394	1.545454545	2.848484848	0.484848485
AT1G13340	0.119487909	0.674253201	0.93883357	1.186344239	2.953058321	0.12802276
AT1G16030	0.048545812	0.044609665	0.047233763	2.351191778	3.421823748	0.086595233
AT1G31170	0.693693694	1.112612613	0.490990991	1.004504505	2.378378378	0.31981982
AT1G52560	0.011549567	0.023099134	0.037536092	1.157844081	4.720885467	0.049085659
AT1G53540	0.031413613	0.031413613	0.031413613	0.591623037	5.193717277	0.120418848
AT1G54050	0.278688525	0.104918033	0.147540984	2.436065574	2.970491803	0.062295082
AT1G74310	0.011861056	0.071166337	0.115221689	2.494210675	3.232985032	0.07455521
AT1G77000	0.274621986	0.25500613	0.26726604	1.365753984	3.663261136	0.174090723
AT2G22080	0.977119785	0.63795424	0.379542396	2.592193809	1.292059219	0.121130552
AT2G22240	0.694214876	0.270820089	0.424348379	2.231404959	2.361093452	0.018118245
AT2G32120	0.053128689	0.047815821	0.047815821	2.096812279	3.676505313	0.077922078
AT2G38905	0.017579842	0.070319367	1.118663932	3.181951362	1.397597422	0.213888075
AT2G46240	0.181034483	0.362068966	0.310344828	1.357758621	3.63362069	0.155172414
AT2G47180	0.09314545	0.097444471	0.011464055	2.580845474	3.144017196	0.073083353
AT3G07370	0.596153846	0.615384615	0.423076923	1.115384615	2.576923077	0.673076923
AT3G10020	0.383883108	0.871374806	0.193934027	1.993801195	2.38432588	0.172680983
AT3G11050	0.012459765	0.023050566	0.248572319	3.83573876	1.744367148	0.135811442
AT3G12580	0.002214022	0.191881919	0.329151292	3.150553506	1.740221402	0.58597786
AT3G46230	0.003002627	0.005254598	0.004503941	3.591142249	2.284999374	0.11109721
AT3G50970	0.118548387	0.291129032	0.112096774	3.088709677	2.332258065	0.057258065
AT3G56350	0.00300541	0.005409738	0.129232619	4.156481667	1.532758966	0.173111601
AT4G12400	0.078947368	0.171052632	0.105263158	1.092105263	4.171052632	0.381578947
AT4G25200	0.005323869	0.005323869	0.005323869	0.010647737	5.936113576	0.037267081
AT4G32770	0.421610169	0.762711864	1.610169492	1.953389831	1.112288136	0.139830508
AT5G05410	0.083412322	0.938388626	0.117535545	0.583886256	3.393364929	0.883412322
AT5G12030	0.002554278	0.009365688	0.010217114	3.63388676	2.233290762	0.110685398
AT5G16970	0.183898651	0.313853698	0.448712709	2.042501022	2.557417246	0.453616673
AT5G37670	0.085020243	0.048582996	0.072874494	1.578947368	4.087044534	0.127530364
AT5G52300	0.001515534	0.004546603	0.012882041	3.621369033	2.196009093	0.163677696
AT5G52640	0.015570934	0.138408304	0.153979239	1.640138408	3.596885813	0.455017301
AT5G57050	0.258899676	0.57605178	0.27184466	1.83171521	2.919093851	0.142394822
AT5G67030	0.626439659	1.188282424	0.231347021	1.254381572	2.511767651	0.187781673

LITERATURE CITED

1. Proudfoot NJ, Furger A, & Dye MJ (2002) Integrating mRNA processing with transcription. *Cell* 108:501-512.
2. Xing D & Li QQ (2011) Alternative polyadenylation and gene expression regulation in plants. *Wiley interdisciplinary reviews. RNA* 2:445-458.
3. Mayr C & Bartel DP (2009) Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* 138(4):673-684.
4. Tollervey D (2004) Molecular biology: termination by torpedo. *Nature* 432(7016):456-457.
5. Proudfoot N (2004) New perspectives on connecting messenger RNA 3' end formation to transcription. *Current opinion in cell biology* 16:272-278.
6. Tian B, Pan Z, & Lee JY (2007) Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. *Genome research* 17(2):156-165.
7. Macknight R*, et al.* (2002) Functional Significance of the Alternative Transcript Processing of the Arabidopsis Floral Promoter FCA. 14:877-888.
8. Delaney KJ*, et al.* (2006) Calmodulin interacts with and regulates the RNA-binding activity of an Arabidopsis polyadenylation factor subunit. *Plant physiology* 140(4):1507-1521.
9. Millevoi S & Vagner S (2010) Molecular mechanisms of eukaryotic pre-mRNA 3' end processing regulation. *Nucleic acids research* 38:2757-2774.
10. Chan S, Choi E-A, & Shi Y (2011) Pre-mRNA 3'-end processing complex assembly and function. *Wiley interdisciplinary reviews. RNA* 2:321-335.
11. Tian B & Graber JH (2012) Signals for pre-mRNA cleavage and polyadenylation. *Wiley interdisciplinary reviews. RNA* 3:385-396.
12. Hunt A, Chu, NathanM., Odell, JoanT., Nagy, Ferenc., Chua, Nam-Hai (1987) Plant cells do not properly recognize animal gene polyadenylation signals. *Plant Molecular Biology* 8(1):23-35.
13. Wu L, Ueda T, & Messing J (1995) The formation of mRNA 3'-ends in plants. *The Plant journal : for cell and molecular biology* 8(3):323-329.
14. Bentley DR*, et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456(7218):53-59.
15. Rothnie HM, Reid J, & Hohn T (1994) The contribution of AAUAAA and the upstream element UUUGUA to the efficiency of mRNA 3'-end formation in plants. *The EMBO journal* 13(9):2200-2210.
16. Kühn U*, et al.* (2009) Poly(A) tail length is controlled by the nuclear poly(A)-binding protein regulating the interaction between poly(A) polymerase and the cleavage and polyadenylation specificity factor. *The Journal of biological chemistry* 284:22803-22814.
17. Hunt AG*, et al.* (2008) Arabidopsis mRNA polyadenylation machinery: comprehensive analysis of protein-protein interactions and gene expression profiling. *BMC genomics* 9:220.
18. Chen JIE & Moore C (1992) polyadenylation of yeast pre-mRNA . Separation of Factors Required for Cleavage and Polyadenylation of Yeast Pre-mRNA. 12.

19. Mandel CR, Bai Y, & Tong L (2008) Protein factors in pre-mRNA 3'-end processing. *Cellular and molecular life sciences : CMLS* 65(7-8):1099-1122.
20. Hunt AG (2008) Messenger RNA 3' end formation in plants. *Current topics in microbiology and immunology* 326:151-177.
21. Zhao H, Xing D, & Li QQ (2009) Unique features of plant cleavage and polyadenylation specificity factor revealed by proteomic studies. *Plant physiology* 151:1546-1556.
22. Rao S, Dinkins RD, & Hunt AG (2009) Distinctive interactions of the Arabidopsis homolog of the 30 kD subunit of the cleavage and polyadenylation specificity factor (AtCPSF30) with other polyadenylation factor subunits. *BMC cell biology* 10:51.
23. Xu R, Ye X, & Quinn Li Q (2004) AtCPSF73-II gene encoding an Arabidopsis homolog of CPSF 73 kDa subunit is critical for early embryo development. *Gene* 324:35-45.
24. Zhang J*, et al.* (2008) A polyadenylation factor subunit implicated in regulating oxidative signaling in Arabidopsis thaliana. *PloS one* 3:e2410.
25. Bienroth S, Wahle E, Suter-Crazzolara C, & Keller W (1991) Purification of the cleavage and polyadenylation factor involved in the 3'-processing of messenger RNA precursors. *The Journal of biological chemistry* 266(29):19768-19776.
26. Murthy KG & Manley JL (1992) Characterization of the multisubunit cleavage-polyadenylation specificity factor from calf thymus. *The Journal of biological chemistry* 267:14804-14811.
27. Shi Y*, et al.* (2009) Molecular architecture of the human pre-mRNA 3' processing complex. *Molecular cell* 33(3):365-376.
28. Takagaki Y & Manley JL (2000) Complex protein interactions within the human polyadenylation machinery identify a novel component. *Molecular and cellular biology* 20:1515-1525.
29. Kaufmann I, Martin G, Friedlein A, Langen H, & Keller W (2004) Human Fip1 is a subunit of CPSF that binds to U-rich RNA elements and stimulates poly(A) polymerase. *The EMBO journal* 23:616-626.
30. Keller W, Bienroth S, Lang KM, & Christofori G (1991) Cleavage and polyadenylation factor CPF specifically interacts with the pre-mRNA 3' processing signal AAUAAA. *The EMBO journal* 10(13):4241-4249.
31. Sheets MD, Ogg SC, & Wickens MP (1990) Point mutations in AAUAAA and the poly (A) addition site: effects on the accuracy and efficiency of cleavage and polyadenylation in vitro. *Nucleic acids research* 18(19):5799-5805.
32. Moore CL, Chen J, & Whoriskey J (1988) Two proteins crosslinked to RNA containing the adenovirus L3 poly(A) site require the AAUAAA sequence for binding. *The EMBO journal* 7(10):3159-3169.
33. Murthy KG & Manley JL (1995) The 160-kD subunit of human cleavage-polyadenylation specificity factor coordinates pre-mRNA 3'-end formation. *Genes & development* 9:2672-2683.
34. Dichtl B*, et al.* (2002) Yhh1p/Cft1p directly links poly(A) site recognition and RNA polymerase II transcription termination. *The EMBO journal* 21:4125-4135.

35. Barabino SM, Hubner W, Jenny a, Minvielle-Sebastia L, & Keller W (1997) The 30-kD subunit of mammalian cleavage and polyadenylation specificity factor and its yeast homolog are RNA-binding zinc finger proteins. *Genes & development* 11:1703-1716.
36. Ohnacker M, Barabino SM, Preker PJ, & Keller W (2000) The WD-repeat protein pfs2p bridges two essential factors within the yeast pre-mRNA 3'-end-processing complex. *The EMBO journal* 19:37-47.
37. Ryan K, Calvo O, & Manley JL (2004) Evidence that polyadenylation factor CPSF-73 is the mRNA 3 processing endonuclease.565-573.
38. Mandel CR, *et al.* (2006) Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease. *Nature* 444:953-956.
39. Chanfreau G, Noble SM, & Guthrie C (1996) Essential Yeast Protein with Unexpected Similarity to Subunits of Mammalian Cleavage and Polyadenylation Specificity Factor (CPSF). *Science (New York, N.Y.)* 274(5292):1511-1514.
40. Ryan K (2007) Pre-mRNA 3 ' Cleavage is Reversibly Inhibited In Vitro by Cleavage Factor Dephosphorylation ND ES SC Key woRds RIB. 4:26-33.
41. Kyburz a (2003) The role of the yeast cleavage and polyadenylation factor subunit Ydh1p/Cft2p in pre-mRNA 3'-end formation. *Nucleic acids research* 31:3936-3945.
42. de Vries H, *et al.* (2000) Human pre-mRNA cleavage factor II(m) contains homologs of yeast proteins and bridges two other cleavage factors. *The EMBO journal* 19:5895-5904.
43. Preker PJ, Ohnacker M, Minvielle-Sebastia L, & Keller W (1997) A multisubunit 3' end processing factor from yeast containing poly(A) polymerase and homologues of the subunits of mammalian cleavage and polyadenylation specificity factor. *The EMBO journal* 16:4727-4737.
44. Barabino SM, Ohnacker M, & Keller W (2000) Distinct roles of two Yth1p domains in 3'-end cleavage and polyadenylation of yeast pre-mRNAs. *The EMBO journal* 19:3778-3787.
45. Helmling S & Zhelkovsky A (2001) Fip1 Regulates the Activity of Poly (A) Polymerase through Multiple Interactions. 21:2026-2037.
46. Tzafrir I, *et al.* (2004) Identification of Genes Required for Embryo Development in Arabidopsis 1 [w]. 135:1206-1220.
47. Herr AJ, Molnàr A, Jones A, & Baulcombe DC (2006) Defective RNA processing enhances RNA silencing and influences flowering of Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America* 103:14994-15001.
48. Xu R, *et al.* (2006) The 73 kD subunit of the cleavage and polyadenylation specificity factor (CPSF) complex affects reproductive development in Arabidopsis. *Plant molecular biology* 61:799-815.
49. Manzano D, *et al.* (2009) Altered interactions within FY/AtCPSF complexes required for Arabidopsis FCA-mediated chromatin silencing. *Proceedings of the National Academy of Sciences of the United States of America* 106:8772-8777.

50. Addepalli B & Hunt aG (2007) A novel endonuclease activity associated with the Arabidopsis ortholog of the 30-kDa subunit of cleavage and polyadenylation specificity factor. *Nucleic acids research* 35:4453-4463.
51. Addepalli B & Hunt AG (2008) Redox and heavy metal effects on the biochemical activities of an Arabidopsis polyadenylation factor subunit. *Archives of biochemistry and biophysics* 473:88-95.
52. Addepalli B, Limbach Pa, & Hunt AG (2010) A disulfide linkage in a CCCH zinc finger motif of an Arabidopsis CPSF30 ortholog. *FEBS letters* 584:4408-4412.
53. Thomas PE*, et al.* (2012) Genome-Wide Control of Polyadenylation Site Choice by CPSF30 in Arabidopsis. *The Plant cell* 24(11):4376-4388.
54. Forbes KP, Addepalli B, & Hunt AG (2006) An Arabidopsis Fip1 homolog interacts with RNA and provides conceptual links with a number of other polyadenylation factor subunits. *The Journal of biological chemistry* 281:176-186.
55. Simpson GG, Dijkwel PP, Quesada V, Henderson I, & Dean C (2003) FY is an RNA 3' end-processing factor that interacts with FCA to control the Arabidopsis floral transition. *Cell* 113:777-787.
56. Colgan DF & Manley JL (1997) Mechanism and regulation of mRNA polyadenylation. *Genes & development* 11:2755-2766.
57. Takagaki Y, Manley JL, MacDonald CC, Wilusz J, & Shenk T (1990) A multisubunit factor, CstF, is required for polyadenylation of mammalian pre-mRNAs. *Genes & development* 4:2112-2120.
58. Wallace aM*, et al.* (1999) Two distinct forms of the 64,000 Mr protein of the cleavage stimulation factor are expressed in mouse male germ cells. *Proceedings of the National Academy of Sciences of the United States of America* 96:6763-6768.
59. Takagaki Y & Manley JL (1997) RNA recognition by the human polyadenylation factor CstF. *Molecular and cellular biology* 17:3907-3914.
60. Monarez RR, MacDonald CC, & Dass B (2007) Polyadenylation proteins CstF-64 and tauCstF-64 exhibit differential binding affinities for RNA polymers. *The Biochemical journal* 401:651-658.
61. Dass B*, et al.* (2007) Loss of polyadenylation protein tauCstF-64 causes spermatogenic defects and male infertility. *Proceedings of the National Academy of Sciences of the United States of America* 104(51):20374-20379.
62. Bai Y*, et al.* (2007) Crystal structure of murine CstF-77: dimeric association and implications for polyadenylation of mRNA precursors. *Molecular cell* 25:863-875.
63. Legrand P, Pinaud N, Minvielle-Sébastia L, & Fribourg S (2007) The structure of the CstF-77 homodimer provides insights into CstF assembly. *Nucleic acids research* 35:4515-4522.
64. McCracken S*, et al.* (1997) The C-terminal domain of RNA polymerase II couples mRNA processing to transcription. in *Nature*, pp 357-361.
65. Kessler MM, Zhao J, & Moore CL (1996) Purification of the Saccharomyces cerevisiae Cleavage / Polyadenylation Factor I. 271:27167-27175.

66. Noble CG, Walker Pa, Calder LJ, & Taylor Ia (2004) Rna14-Rna15 assembly mediates the RNA-binding capability of Saccharomyces cerevisiae cleavage factor IA. *Nucleic acids research* 32:3364-3375.
67. Gross S & Moore CL (2001) Rna15 Interaction with the A-Rich Yeast Polyadenylation Signal Is an Essential Step in mRNA 3J-End Formation. 21:8045-8055.
68. Liu F, Marquardt S, Lister C, Swiezewski S, & Dean C (2010) Targeted 3' processing of antisense transcripts triggers Arabidopsis FLC chromatin silencing. *Science (New York, N.Y.)* 327:94-97.
69. Yao Y, Song L, Katz Y, & Galili G (2002) Cloning and characterization of Arabidopsis homologues of the animal CstF complex that regulates 3' mRNA cleavage and polyadenylation. *Journal of experimental botany* 53(378):2277-2278.
70. Colgan DF, Murthy KG, Prives C, & Manley JL (1996) Cell-cycle related regulation of poly(A) polymerase by phosphorylation. *Nature* 384(6606):282-285.
71. Kim H, Lee JH, & Lee Y (2003) Regulation of poly(A) polymerase by 14-3-3epsilon. *The EMBO journal* 22(19):5208-5219.
72. Shimazu T, Horinouchi S, & Yoshida M (2007) Multiple histone deacetylases and the CREB-binding protein regulate pre-mRNA 3'-end processing. *The Journal of biological chemistry* 282:4470-4478.
73. Vethantham V, Rao N, & Manley JL (2008) Sumoylation regulates multiple aspects of mammalian poly (A) polymerase function.499-511.
74. Lee YJ, Lee Y, & Chung JH (2000) An intronless gene encoding a poly(A) polymerase is specifically expressed in testis. *FEBS letters* 487:287-292.
75. Shi Y, Chan S, & Martinez-Santibañez G (2009) An up-close look at the pre-mRNA 3'-end processing complex. *RNA biology* 6:522-525.
76. Mellman DL*, et al.* (2008) A PtdIns4,5P2-regulated nuclear poly(A) polymerase controls expression of select mRNAs. *Nature* 451:1013-1017.
77. Mangus DA, Evans MC, & Jacobson A (2003) Protein family review Poly (A) -binding proteins : multifunctional scaffolds for the post- transcriptional control of gene expression.1-14.
78. Kühn U & Wahle E (2004) Structure and function of poly(A) binding proteins. *Biochimica et biophysica acta* 1678:67-84.
79. Bienroth S (1993) Assembly of a processive polyadenylation complex. 12:585-594.
80. Jenal M*, et al.* (2012) The Poly(A)-Binding Protein Nuclear 1 Suppresses Alternative Cleavage and Polyadenylation Sites. *Cell* 149(3):538-553.
81. Kessler MM*, et al.* (1997) Hrp1, a sequence-specific RNA-binding protein that shuttles between the nucleus and the cytoplasm, is required for mRNA 3'-end formation in yeast. *Genes & development* 11(19):2545-2556.
82. Minvielle-Sebastia L, Preker PJ, & Keller W (1994) RNA14 and RNA15 proteins as components of a yeast pre-mRNA 3'-end processing factor. *Science (New York, N.Y.)* 266(5191):1702-1705.

83. Preker PJ, Lingner J, Minvielle-Sebastia L, & Keller W (1995) The FIP1 gene encodes a component of a yeast pre-mRNA polyadenylation factor that directly interacts with poly(A) polymerase. *Cell* 81(3):379-389.
84. Sachs AB, Bond MW, & Kornberg RD (1986) A single gene from yeast for both nuclear and cytoplasmic polyadenylate-binding proteins: domain structure and expression. *Cell* 45(6):827-835.
85. Brune C, Munchel SE, Fischer N, Podtelejnikov AV, & Weis K (2005) Yeast poly(A)-binding protein Pab1 shuttles between the nucleus and the cytoplasm and functions in mRNA export. *RNA (New York, N.Y.)* 11(4):517-531.
86. Meeks LR, Addepalli B, & Hunt AG (2009) Characterization of genes encoding poly(A) polymerases in plants: evidence for duplication and functional specialization. *PloS one* 4:e8082.
87. Mangus DA, Evans MC, & Jacobson A (2003) Poly(A)-binding proteins: multifunctional scaffolds for the post-transcriptional control of gene expression. *Genome biology* 4(7):223.
88. Dufresne PJ, Ubalijoro E, Fortin MG, & Laliberte JF (2008) Arabidopsis thaliana class II poly(A)-binding proteins are required for efficient multiplication of turnip mosaic virus. *The Journal of general virology* 89(Pt 9):2339-2348.
89. Hunt AG, Xing D, & Li QQ (2012) Plant polyadenylation factors: conservation and variety in the polyadenylation complex in plants. *BMC genomics* 13:641.
90. Coseno M*, et al.* (2008) Crystal structure of the 25 kDa subunit of human cleavage factor Im. *Nucleic acids research* 36:3474-3483.
91. Yang Q, Gilmartin GM, & Doublié S (2010) Structural basis of UGUA recognition by the Nudix protein CFI(m)25 and implications for a regulatory role in mRNA 3' processing. *Proceedings of the National Academy of Sciences of the United States of America* 107:10062-10067.
92. Rüegsegger U, Beyer K, & Keller W (1996) Purification and characterization of human cleavage factor Im involved in the 3' end processing of messenger RNA precursors. *The Journal of biological chemistry* 271:6107-6113.
93. Dettwiler S, Aringhieri C, Cardinale S, Keller W, & Barabino SML (2004) Distinct sequence motifs within the 68-kDa subunit of cleavage factor Im mediate RNA binding, protein-protein interactions, and subcellular localization. *The Journal of biological chemistry* 279:35788-35797.
94. Venkataraman K, Brown KM, & Gilmartin GM (2005) Analysis of a noncanonical poly (A) site reveals a tripartite mechanism for vertebrate poly (A) site recognition.1315-1327.
95. Rüegsegger U, Blank D, & Keller W (1998) Human pre-mRNA cleavage factor Im is related to spliceosomal SR proteins and can be reconstituted in vitro from recombinant subunits. *Molecular cell* 1:243-253.
96. Millevoi S*, et al.* (2006) An interaction between U2AF 65 and CF I(m) links the splicing and 3' end processing machineries. *The EMBO journal* 25:4854-4864.
97. Zhou Z, Licklider LJ, Gygi SP, & Reed R (2002) Comprehensive proteomic analysis of the human spliceosome. *Nature* 419:182-185.

98. Weitzer S & Martinez J (2007) The human RNA kinase hClp1 is active on 3' transfer RNA exons and short interfering RNAs. *Nature* 447:222-226.
99. Paushkin SV, Patel M, Furia BS, Peltz SW, & Trotta CR (2004) Identification of a human endonuclease complex reveals a link between tRNA splicing and pre-mRNA 3' end formation. *Cell* 117:311-321.
100. Ramirez A, Shuman S, & Schwer B (2008) Human RNA 59-kinase (hClp1) can function as a tRNA splicing enzyme in vivo.1737-1745.
101. Xing D, Zhao H, & Li QQ (2008) Arabidopsis CLP1-SIMILAR PROTEIN3, an ortholog of human polyadenylation factor CLP1, functions in gametophyte, embryo, and postembryonic development. *Plant physiology* 148(4):2059-2069.
102. Xing D, Zhao H, Xu R, & Li QQ (2008) Arabidopsis PCFS4, a homologue of yeast polyadenylation factor Pcf11p, regulates FCA alternative processing and promotes flowering time. *The Plant journal : for cell and molecular biology* 54:899-910.
103. Shi Y (2012) Alternative polyadenylation: New insights from global analyses. *RNA (New York, N.Y.)*:2105-2117.
104. Lutz CS & Moreira A (2011) Alternative mRNA polyadenylation in eukaryotes: an effective regulator of gene expression. *Wiley interdisciplinary reviews. RNA* 2(1):22-31.
105. Alt FW*, et al.* (1980) Synthesis of secreted and membrane-bound immunoglobulin mu heavy chains is directed by mRNAs that differ at their 3' ends. *Cell* 20(2):293-301.
106. Early P*, et al.* (1980) Two mRNAs can be produced from a single immunoglobulin mu gene by alternative RNA processing pathways. *Cell* 20(2):313-319.
107. Rogers J*, et al.* (1980) Two mRNAs with different 3' ends encode membrane-bound and secreted forms of immunoglobulin mu chain. *Cell* 20(2):303-312.
108. Setzer DR, McGrogan M, Nunberg JH, & Schimke RT (1980) Size heterogeneity in the 3' end of dihydrofolate reductase messenger RNAs in mouse cells. *Cell* 22(2 Pt 2):361-370.
109. Edwalds-Gilbert G, Veraldi KL, & Milcarek C (1997) Alternative poly(A) site selection in complex transcription units: means to an end? *Nucleic acids research* 25(13):2547-2561.
110. Tian B, Hu J, Zhang H, & Lutz CS (2005) A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic acids research* 33(1):201-212.
111. Yan J & Marr TG (2005) Computational analysis of 3'-ends of ESTs shows four classes of alternative polyadenylation in human, mouse, and rat. *Genome research* 15(3):369-375.
112. Sandberg R, Neilson JR, Sarma A, Sharp PA, & Burge CB (2008) Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. *Science (New York, N.Y.)* 320(5883):1643-1647.
113. Ji Z & Tian B (2009) Reprogramming of 3' untranslated regions of mRNAs by alternative polyadenylation in generation of pluripotent stem cells from different cell types. *PloS one* 4(12):e8419.

114. Ozsolak F, *et al.* (2010) Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell* 143(6):1018-1029.
115. Jan CH, Friedman RC, Ruby JG, & Bartel DP (2011) Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs. *Nature* 469(7328):97-101.
116. Mangone M, *et al.* (2010) The landscape of C. elegans 3'UTRs. *Science (New York, N.Y.)* 329(5990):432-435.
117. Haenni S, *et al.* (2012) Analysis of C. elegans intestinal gene expression and polyadenylation by fluorescence-activated nuclei sorting and 3'-end-seq. *Nucleic acids research* 40(13):6304-6318.
118. Smibert P, *et al.* (2012) Global patterns of tissue-specific alternative polyadenylation in Drosophila. *Cell reports* 1(3):277-289.
119. Wu X, *et al.* (2011) Genome-wide landscape of polyadenylation in Arabidopsis provides evidence for extensive alternative polyadenylation. *Proc Natl Acad Sci U S A* 108(30):12533-12538.
120. Sherstnev A, *et al.* (2012) Direct sequencing of Arabidopsis thaliana RNA reveals patterns of cleavage and polyadenylation. *Nature structural & molecular biology* 19(8):845-852.
121. Fox-Walsh K, Davis-Turak J, Zhou Y, Li H, & Fu XD (2011) A multiplex RNA-seq strategy to profile poly(A+) RNA: application to analysis of transcription response and 3' end formation. *Genomics* 98(4):266-271.
122. Fu Y, *et al.* (2011) Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and normal cells by high-throughput sequencing. *Genome research* 21(5):741-747.
123. Shepard PJ, *et al.* (2011) Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA (New York, N.Y.)* 17(4):761-772.
124. Derti A, *et al.* (2012) A quantitative atlas of polyadenylation in five mammals. *Genome research* 22(6):1173-1183.
125. Lin Y, *et al.* (2012) An in-depth map of polyadenylation sites in cancer. *Nucleic acids research* 40(17):8460-8471.
126. Ulitsky I, *et al.* (2012) Extensive alternative polyadenylation during zebrafish development. *Genome research* 22(10):2054-2066.
127. Liu F, *et al.* (2007) The Arabidopsis RNA-binding protein FCA requires a lysine-specific demethylase 1 homolog to downregulate FLC. *Molecular cell* 28:398-407.
128. Evsikov AV, *et al.* (2006) Cracking the egg: molecular dynamics and evolutionary aspects of the transition from the fully grown oocyte to embryo. *Genes & development* 20(19):2713-2727.
129. Ji Z, Lee JY, Pan Z, Jiang B, & Tian B (2009) Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proceedings of the National Academy of Sciences of the United States of America* 106(17):7028-7033.
130. Zhang H, Lee JY, & Tian B (2005) Biased alternative polyadenylation in human tissues. *Genome biology* 6(12):R100.
131. Lee YS & Dutta A (2007) The tumor suppressor microRNA let-7 represses the HMGA2 oncogene. *Genes & development* 21(9):1025-1030.

132. Mayr C, Hemann MT, & Bartel DP (2007) Disrupting the pairing between let-7 and Hmga2 enhances oncogenic transformation. *Science (New York, N.Y.)* 315(5818):1576-1579.
133. Wiestner A*, et al.* (2007) Point mutations and genomic deletions in CCND1 create stable truncated cyclin D1 mRNAs that are associated with increased proliferation rate and shorter survival. *Blood* 109(11):4599-4606.
134. Keller T*, et al.* (1998) A plant homolog of the neutrophil NADPH oxidase gp91phox subunit gene encodes a plasma membrane protein with Ca2+ binding motifs. *The Plant cell* 10(2):255-266.
135. Cheng Y, Kato N, Wang W, Li J, & Chen X (2003) Two RNA binding proteins, HEN4 and HUA1, act in the processing of AGAMOUS pre-mRNA in Arabidopsis thaliana. *Developmental cell* 4(1):53-66.
136. Michaels SD & Amasino RM (2001) Loss of FLOWERING LOCUS C activity eliminates the late-flowering phenotype of FRIGIDA and autonomous pathway mutations but not responsiveness to vernalization. *The Plant cell* 13(4):935-941.
137. Simpson GG (2004) The autonomous pathway: epigenetic and post-transcriptional gene regulation in the control of Arabidopsis flowering time. *Current opinion in plant biology* 7:570-574.
138. Shen Y*, et al.* (2011) Transcriptome dynamics through alternative polyadenylation in developmental and environmental responses in plants revealed by deep sequencing.1478-1486.
139. Monroy A, Maggio R, & Rinaldi AM (1965) Experimentally induced activation of the ribosomes of the unfertilized sea urchin egg. *Proceedings of the National Academy of Sciences of the United States of America* 54(1):107-111.
140. Hultin T (1961) Activation of ribosomes in sea urchin eggs in response to fertilization. *Experimental cell research* 25:405-417.
141. Maggio R, Vittorelli ML, Rinaldi AM, & Monroy A (1964) In vitro incorporation of amino acids into proteins stimulated by RNA from unfertilized sea urchin eggs. *Biochemical and biophysical research communications* 15(5):436-441.
142. Baltus E, Quertier J, Ficq A, & Brachet J (1965) BIOCHEMICAL STUDIES OF NUCLEATE AND ANUCLEATED FRAGMENTS ISOLATED FROM SEA-URCHIN EGGS. A COMPARISON BETWEEN FERTILIZATION AND PARTHENOGENETIC ACTIVATION. *Biochimica et biophysica acta* 95:408-417.
143. Gross PR & Cousineau GH (1963) Effects of actinomycin D on macromolecule synthesis and early development in sea urchin eggs. *Biochemical and biophysical research communications* 10:321-326.
144. Colot HV & Rosbash M (1982) Behavior of individual maternal pA+ RNAs during embryogenesis of Xenopus laevis. *Developmental biology* 94(1):79-86.
145. Rosenthal ET, Tansey TR, & Ruderman JV (1983) Sequence-specific adenylations and deadenylations accompany changes in the translation of maternal messenger RNA after fertilization of Spisula oocytes. *Journal of molecular biology* 166(3):309-327.

146. Dworkin MB & Dworkin-Rastl E (1985) Changes in RNA titers and polyadenylation during oogenesis and oocyte maturation in Xenopus laevis. *Developmental biology* 112(2):451-457.
147. Huarte J, Belin D, Vassalli A, Strickland S, & Vassalli JD (1987) Meiotic maturation of mouse oocytes triggers the translation and polyadenylation of dormant tissue-type plasminogen activator mRNA. *Genes & development* 1(10):1201-1211.
148. Vassalli JD, *et al.* (1989) Regulated polyadenylation controls mRNA translation during meiotic maturation of mouse oocytes. *Genes & development* 3(12B):2163-2171.
149. Rosenthal ET & Ruderman JV (1987) Widespread changes in the translation and adenylation of maternal messenger RNAs following fertilization of Spisula oocytes. *Developmental biology* 121(1):237-246.
150. Goldman DS, Kiessling AA, & Cooper GM (1988) Post-transcriptional processing suggests that c-mos functions as a maternal message in mouse eggs. *Oncogene* 3(2):159-162.
151. Huez G, *et al.* (1974) Role of the polyadenylate segment in the translation of globin messenger RNA in Xenopus oocytes. *Proceedings of the National Academy of Sciences of the United States of America* 71(8):3143-3146.
152. Yisraeli JK & Melton DA (1988) The material mRNA Vg1 is correctly localized following injection into Xenopus oocytes. *Nature* 336(6199):592-595.
153. Fox CA, Sheets MD, & Wickens MP (1989) Poly(A) addition during maturation of frog oocytes: distinct nuclear and cytoplasmic activities and regulation by the sequence UUUUUAU. *Genes & development* 3(12B):2151-2162.
154. Simon R, Tassan JP, & Richter JD (1992) Translational control by poly(A) elongation during Xenopus development: differential repression and enhancement by a novel cytoplasmic polyadenylation element. *Genes & development* 6(12B):2580-2591.
155. Simon R & Richter JD (1994) Further analysis of cytoplasmic polyadenylation in Xenopus embryos and identification of embryonic cytoplasmic polyadenylation element-binding proteins. *Molecular and cellular biology* 14(12):7867-7875.
156. Radford HE, Meijer Ha, & de Moor CH (2008) Translational control by cytoplasmic polyadenylation in Xenopus oocytes. *Biochimica et biophysica acta* 1779:217-229.
157. Kashiwabara S, Nakanishi T, Kimura M, & Baba T (2008) Non-canonical poly(A) polymerase in mammalian gametogenesis. *Biochimica et biophysica acta* 1779(4):230-238.
158. Lalancette C, Miller D, Li Y, & Krawetz SA (2008) Paternal contributions: new functional insights for spermatozoal RNA. *Journal of cellular biochemistry* 104(5):1570-1579.
159. Miller D, Tang P-Z, Skinner C, & Lilford R (1994) Differential RNA fingerprinting as a tool in the analysis of spermatozoal gene expression. *Human Reproduction* 9(5):864-869.

160. Wykes SM, Visscher DW, & Krawetz SA (1997) Haploid transcripts persist in mature human spermatozoa. *Molecular human reproduction* 3(1):15-19.
161. Cheng YS*, et al.* (2006) Association of spermatogenic failure with decreased CDC25A expression in infertile men. *Human reproduction (Oxford, England)* 21(9):2346-2352.
162. Mao XM, Ma WL, Feng CQ, Zou YG, & Zheng WL (2004) [An initial examination of the spermatozoal gene expression profile]. *Di 1 jun yi da xue xue bao = Academic journal of the first medical college of PLA* 24(9):1033-1036.
163. Dadoune JP, Pawlak A, Alfonsi MF, & Siffroi JP (2005) Identification of transcripts by macroarrays, RT-PCR and in situ hybridization in human ejaculate spermatozoa. *Molecular human reproduction* 11(2):133-140.
164. De Ambrogi M, Spinaci M, Galeati G, & Tamanini C (2007) Leptin receptor in boar spermatozoa. *International journal of andrology* 30(5):458-461.
165. Fiore C*, et al.* (2006) Identification of the mineralocorticoid receptor in human spermatozoa. *International journal of molecular medicine* 18(4):649-652.
166. Teng YN*, et al.* (2007) Expression of various CDC25B isoforms in human spermatozoa. *Fertility and sterility* 88(2):379-382.
167. Yeung CH & Cooper TG (2008) Potassium channels involved in human sperm volume regulation--quantitative studies at the protein and mRNA levels. *Molecular reproduction and development* 75(4):659-668.
168. Zhang JS*, et al.* (2006) Genome-wide profiling of segmental-regulated transcriptomes in human epididymis using oligo microarray. *Molecular and cellular endocrinology* 250(1-2):169-177.
169. Zhao Y*, et al.* (2006) Characterization and quantification of mRNA transcripts in ejaculated spermatozoa of fertile men by serial analysis of gene expression. *Human reproduction (Oxford, England)* 21(6):1583-1590.
170. Ostermeier GC, Miller D, Huntriss JD, Diamond MP, & Krawetz SA (2004) Reproductive biology: delivering spermatozoan RNA to the oocyte. *Nature* 429(6988):154.
171. Villalba A, Coll O, & Gebauer F (2011) Cytoplasmic polyadenylation and translational control. *Current opinion in genetics & development* 21:452-457.
172. Liang CG, Su YQ, Fan HY, Schatten H, & Sun QY (2007) Mechanisms regulating oocyte meiotic resumption: roles of mitogen-activated protein kinase. *Molecular endocrinology (Baltimore, Md.)* 21(9):2037-2055.
173. Belloc E, Pique M, & Mendez R (2008) Sequential waves of polyadenylation and deadenylation define a translation circuit that drives meiotic progression. *Biochemical Society transactions* 36(Pt 4):665-670.
174. Paillard L, Maniey D, Lachaume P, Legagneux V, & Osborne HB (2000) Identification of a C-rich element as a novel cytoplasmic polyadenylation element in Xenopus embryos. *Mechanisms of development* 93(1-2):117-125.
175. Charlesworth A, Cox LL, & MacNicol AM (2004) Cytoplasmic polyadenylation element (CPE)- and CPE-binding protein (CPEB)-independent mechanisms regulate early class maternal mRNA translational activation in Xenopus oocytes. *The Journal of biological chemistry* 279(17):17650-17659.

176. Wu L, Good PJ, & Richter JD (1997) The 36-kilodalton embryonic-type cytoplasmic polyadenylation element-binding protein in Xenopus laevis is ElrA, a member of the ELAV family of RNA-binding proteins. *Molecular and cellular biology* 17(11):6402-6409.
177. Slevin MK, Gourronc F, & Hartley RS (2007) ElrA binding to the 3'UTR of cyclin E1 mRNA requires polyadenylation elements. *Nucleic acids research* 35(7):2167-2176.
178. Charlesworth A, Wilczynska A, Thampi P, Cox LL, & MacNicol AM (2006) Musashi regulates the temporal order of mRNA translation during Xenopus oocyte maturation. *The EMBO journal* 25(12):2792-2801.
179. Bardwell VJ*, et al.* (1991) Site-directed ribose methylation identifies 2'-OH groups in polyadenylation substrates critical for AAUAAA recognition and poly(A) addition. *Cell* 65(1):125-133.
180. Paris J & Philippe M (1990) Poly(A) metabolism and polysomal recruitment of maternal mRNAs during early Xenopus development. *Developmental biology* 140(1):221-224.
181. McGrew LL, Dworkin-Rastl E, Dworkin MB, & Richter JD (1989) Poly(A) elongation during Xenopus oocyte maturation is required for translational recruitment and is mediated by a short sequence element. *Genes & development* 3(6):803-815.
182. Stebbins-Boaz B, Hake LE, & Richter JD (1996) CPEB controls the cytoplasmic polyadenylation of cyclin, Cdk2 and c-mos mRNAs and is necessary for oocyte maturation in Xenopus. *The EMBO journal* 15(10):2582-2592.
183. Fox CA, Sheets MD, Wahle E, & Wickens M (1992) Polyadenylation of maternal mRNA during oocyte maturation: poly(A) addition in vitro requires a regulated RNA binding activity and a poly(A) polymerase. *The EMBO journal* 11(13):5021-5032.
184. Gebauer F, Xu W, Cooper GM, & Richter JD (1994) Translational control by cytoplasmic polyadenylation of c-mos mRNA is necessary for oocyte maturation in the mouse. *The EMBO journal* 13(23):5712-5720.
185. Zhao J, Kessler M, Helmling S, O'Connor JP, & Moore C (1999) Pta1, a component of yeast CF II, is required for both cleavage and poly(A) addition of mRNA precursor. *Molecular and cellular biology* 19(11):7733-7740.
186. Hofmann I, Schnolzer M, Kaufmann I, & Franke WW (2002) Symplekin, a constitutive protein of karyo- and cytoplasmic particles involved in mRNA biogenesis in Xenopus laevis oocytes. *Molecular biology of the cell* 13(5):1665-1676.
187. Barnard DC, Ryan K, Manley JL, & Richter JD (2004) Symplekin and xGLD-2 are required for CPEB-mediated cytoplasmic polyadenylation. *Cell* 119(5):641-651.
188. Dickson KS, Bilger A, Ballantyne S, & Wickens MP (1999) The cleavage and polyadenylation specificity factor in Xenopus laevis oocytes is a cytoplasmic factor involved in regulated polyadenylation. *Molecular and cellular biology* 19(8):5707-5717.

189. Wang L, Eckmann CR, Kadyk LC, Wickens M, & Kimble J (2002) A regulatory cytoplasmic poly(A) polymerase in Caenorhabditis elegans. *Nature* 419(6904):312-316.
190. Kwak JE, Wang L, Ballantyne S, Kimble J, & Wickens M (2004) Mammalian GLD-2 homologs are poly(A) polymerases. *Proceedings of the National Academy of Sciences of the United States of America* 101(13):4407-4412.
191. Dure L & Waters L (1965) LONG-LIVED MESSENGER RNA: EVIDENCE FROM COTTON SEED GERMINATION. *Science (New York, N.Y.)* 147(3656):410-412.
192. Barker GR, Bray CM, & Detlefsen MA (1971) An examination of the evidence for stable messenger ribonucleic acid in seed. *The Biochemical journal* 124(2):5P-6P.
193. Walbot V, Capdevila A, & Dure LS, 3rd (1974) Action of 3'd adenosine (cordycepin) and 3'd cytidine on the translation of the stored mRNA of cotton cotyledons. *Biochemical and biophysical research communications* 60(1):103-110.
194. Delseny MA, L. Guitton, Y. (1977) Disappearance of stored polyadenylic acid and mRNA during early germination of radish (Raphanus sativus L.) embryo axes. . *Planta* 135:125-128.
195. Chen SS & Park WM (1973) Early Actions of Gibberellic Acid on the Embryo and on the Endosperm of Avena fatua Seeds. *Plant physiology* 52(2):174-176.
196. Abdul-Baki AA (1969) Metabolism of barley seed during early hours of germination. *Plant physiology* 44(5):733-738.
197. Aspart L, Meyer Y, Laroche M, & Penon P (1984) Developmental regulation of the synthesis of proteins encoded by stored mRNA in radish embryos. *Plant physiology* 76(3):664-673.
198. Nakabayashi K, Okamoto M, Koshiba T, Kamiya Y, & Nambara E (2005) Genome-wide profiling of stored mRNA in Arabidopsis thaliana seed germination: epigenetic and genetic regulation of transcription in seed. *The Plant journal : for cell and molecular biology* 41(5):697-709.
199. Kimura M & Nambara E (2010) Stored and neosynthesized mRNA in Arabidopsis seeds: effects of cycloheximide and controlled deterioration treatment on the resumption of transcription during imbibition. *Plant molecular biology* 73(1-2):119-129.
200. Marcus A & Feeley J (1964) ACTIVATION OF PROTEIN SYNTHESIS IN THE IMBIBITION PHASE OF SEED GERMINATION. *Proceedings of the National Academy of Sciences* 51(6):1075-1079.
201. Marcus A, Feeley J, & Volcani T (1966) Protein Synthesis in Imbibed Seeds III. Kinetics of Amino Acid Incorporation Ribosome Activation, and Polysome Formation. *Plant physiology* 41(7):1167-1172.
202. Hollstein U (1974) Actinomycin. Chemistry and mechanism of action. *Chemical Reviews* 74(6):625-652.
203. Sobell HM (1985) Actinomycin and DNA transcription. *Proceedings of the National Academy of Sciences of the United States of America* 82(16):5328-5331.

204. Chen D, Sarid S, & Katchalski E (1968) STUDIES ON THE NATURE OF MESSENGER RNA IN GERMINATING WHEAT EMBRYOS. *Proceedings of the National Academy of Sciences* 60(3):902-909.
205. Siev M, Weinberg R, & Penman S (1969) The selective interruption of nucleolar RNA synthesis in HeLa cells by cordycepin. *The Journal of cell biology* 41(2):510-520.
206. Kondrashov A*, et al.* (2012) Inhibition of polyadenylation reduces inflammatory gene induction. *RNA (New York, N.Y.)* 18(12):2236-2250.
207. Harris B & Dure L, 3rd (1978) Developmental regulation in cotton seed germination: polyadenylation of stored messenger RNA. *Biochemistry* 17(16):3250-3256.
208. Tao KL & Khan AA (1976) Differential effects of actinomycin d and cordycepin in lettuce seed germination and RNA synthesis. *Plant physiology* 58(6):769-772.
209. Suzuki Y & Minamikawa T (1985) On the Role of Stored mRNA in Protein Synthesis in Embryonic Axes of Germinating Vigna unguiculata Seeds. *Plant physiology* 79(2):327-331.
210. Ishibashi N, Yamauchi D, & Minamikawa T (1990) Stored mRNA in cotyledons of Vigna unguiculata seeds: nucleotide sequence of cloned cDNA for a stored mRNA and induction of its synthesis by precocious germination. *Plant molecular biology* 15(1):59-64.
211. Kuligowski J, Ferrand M, & Chenou E (1991) Stored mRNA in early embryos of a fern Marsilea vestita: a paternal and maternal origin. *Molecular reproduction and development* 30(1):27-33.
212. Beltran-Pena E, Ortiz-Lopez A, & Sanchez de Jimenez E (1995) Synthesis of ribosomal proteins from stored mRNAs early in seed germination. *Plant molecular biology* 28(2):327-336.
213. Rajjou L*, et al.* (2004) The effect of alpha-amanitin on the Arabidopsis seed proteome highlights the distinct roles of stored and neosynthesized mRNAs during germination. *Plant physiology* 134(4):1598-1613.
214. Obrig TG, Culp WJ, McKeehan WL, & Hardesty B (1971) The mechanism by which cycloheximide and related glutarimide antibiotics inhibit peptide synthesis on reticulocyte ribosomes. *The Journal of biological chemistry* 246(1):174-181.
215. Jiménez-López S*, et al.* (2011) Expression profile of maize (Zea mays L.) embryonic axes during germination: translational regulation of ribosomal protein mRNAs. *Plant & cell physiology* 52:1719-1733.
216. Shen Y*, et al.* (2008) Genome level analysis of rice mRNA 3'-end processing signals and alternative polyadenylation. *Nucleic acids research* 36:3150-3161.
217. Loke JC*, et al.* (2005) Compilation of mRNA Polyadenylation Signals in Arabidopsis Revealed a New Signal Element and Potential Secondary Structures 1 [w]. 138:1457-1468.
218. Chang S, Puryear, Jeff., Cairney, John. (1993) A simple and efficient method for isolating RNA from pine trees. *Plant Molecular Biology Reporter* 11(2):113-116.

219. Browne KA (2002) Metal ion-catalyzed nucleic acid alkylation and fragmentation. *Journal of the American Chemical Society* 124(27):7950-7962.
220. Lodish H, Berk, A., Zipursky, L.S., Matsudaira, P., Baltimore, D.& Darnell, J. (2000) Processing of rRNA and tRNA. *Molecular Cell Biology, 4th edition, W. H. Freeman and Company, New York.*
221. Zhu YY, Machleder EM, Chenchik a, Li R, & Siebert PD (2001) Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *BioTechniques* 30:892-897.
222. Kanagawa T (2003) Bias and artifacts in multitemplate polymerase chain reactions (PCR). *Journal of bioscience and bioengineering* 96(4):317-323.
223. Kozarewa I*, et al.* (2009) Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nature methods* 6(4):291-295.
224. Ni T*, et al.* (2010) A paired-end sequencing strategy to map the complex landscape of transcription initiation. *Nature methods* 7(7):521-527.
225. Slomovic S, Laufer D, Geiger D, & Schuster G (2006) Polyadenylation of ribosomal RNA in human cells. *Nucleic acids research* 34(10):2966-2975.
226. Rajjou L*, et al.* (2012) Seed germination and vigor. *Annual review of plant biology* 63:507-533.
227. Hunt AG (2012) RNA Regulatory Elements and Polyadenylation in Plants. *Frontiers in Plant Science* 2:1-5.
228. Wang ET*, et al.* (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* 456(7221):470-476.
229. Dean C*, et al.* (1986) mRNA transcripts of several plant genes are polyadenylated at multiple sites in vivo. *Nucleic acids research* 14(5):2229-2240.
230. Klahre U, Hemmings-Mieszczak M, & Filipowicz W (1995) Extreme heterogeneity of polyadenylation sites in mRNAs encoding chloroplast RNA-binding proteins in Nicotiana plumbaginifolia. *Plant molecular biology* 28(3):569-574.
231. Shen Y*, et al.* (2011) Transcriptome dynamics through alternative polyadenylation in developmental and environmental responses in plants revealed by deep sequencing. *Genome research* 21(9):1478-1486.
232. Ziemienowicz A, Haasen D, Staiger D, & Merkle T (2003) Arabidopsis transportin1 is the nuclear import receptor for the circadian clock-regulated RNA-binding protein AtGRP7. *Plant molecular biology* 53(1-2):201-212.
233. Gille S*, et al.* (2011) O-acetylation of Arabidopsis hemicellulose xyloglucan requires AXY4 or AXY4L, proteins with a TBL and DUF231 domain. *The Plant cell* 23(11):4041-4053.
234. Gu Y*, et al.* (2010) Identification of a cellulose synthase-associated protein required for cellulose biosynthesis. *Proceedings of the National Academy of Sciences of the United States of America* 107(29):12866-12871.
235. Wu X*, et al.* (2011) Genome-wide landscape of polyadenylation in Arabidopsis provides evidence for extensive alternative polyadenylation. *Proceedings of the National Academy of Sciences of the United States of America* 108:12533-12538.

236. Preker P*, et al.* (2011) PROMoter uPstream Transcripts share characteristics with mRNAs and are produced upstream of all three major types of mammalian promoters. *Nucleic acids research* 39(16):7179-7193.
237. Kwak H, Fuda NJ, Core LJ, & Lis JT (2013) Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* 339(6122):950-953.
238. Klauer AA & van Hoof A (2012) Degradation of mRNAs that lack a stop codon: a decade of nonstop progress. *Wiley interdisciplinary reviews. RNA* 3(5):649-660.
239. Zhu Y, Rowley MJ, Bohmdorfer G, & Wierzbicki AT (2013) A SWI/SNF chromatin-remodeling complex acts in noncoding RNA-mediated transcriptional silencing. *Molecular cell* 49(2):298-309.
240. Ausin I*, et al.* (2012) INVOLVED IN DE NOVO 2-containing complex involved in RNA-directed DNA methylation in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America* 109(22):8374-8381.
241. Qian W*, et al.* (2012) A histone acetyltransferase regulates active DNA demethylation in Arabidopsis. *Science (New York, N.Y.)* 336(6087):1445-1448.
242. Axtell MJ, Westholm JO, & Lai EC (2011) Vive la difference: biogenesis and evolution of microRNAs in plants and animals. *Genome biology* 12(4):221.
243. Bewley JD, and M. Black (1994) Seed: Physiology of development and germination. *Plenum Press, New York.*
244. Floros J.D. NR, Fisher W., Barbosa-Cánovas G.V., Chen H., Dunne C.P., German J.B., Hall R.L., Heldman D.R., Karwe M.V., et al. (2010) Feeding the world today and tomorrow: The importance of food science and technology. *Compr. Rev. Food Sci. F.* 9:572–599.
245. Fargione J, Hill J, Tilman D, Polasky S, & Hawthorne P (2008) Land clearing and the biofuel carbon debt. *Science (New York, N.Y.)* 319:1235-1238.
246. Fernandez-Cornejo J (2004) The seed industry in U.S. agriculture: an exploration of data and information on crop seed markets, regulation, industry structure, and research and development. *p.81. In U.S.D.o.A (USDA) (ed.).*
247. Gallardo K*, et al.* (2001) Proteomic analysis of arabidopsis seed germination and priming. *Plant physiology* 126:835-848.
248. F. Han SEU, J.A. Clancy, V. Jitkov, A. Kilian, I. Romagosa (1996) Verification of barley seed dormancy loci via linked molecular markers. *Theor Appl Genet,* :87-91.
249. Ajtkhozhin MA DK, Akhanov AU (1976) Informosomes as a stored form of mRNA in wheat embryos. *FEBS Lett.* 66:124-126.
250. Datta K, Marsh L, & Marcus a (1983) Early growth of wheat embryonic axes and the synthesis of RNA and DNA. *Plant physiology* 72:394-397.
251. Jendrisak J (1980) The use of alpha-amanitin to inhibit in vivo RNA synthesis and germination in wheat embryos. *The Journal of biological chemistry* 255(18):8529-8533.

252. Guilfoyle TJ & Jendrisak JJ (1978) Plant DNA-dependent RNA polymerases: subunit structures and enzymatic properties of the class II enzymes from quiescent and proliferating tissues. *Biochemistry* 17(10):1860-1866.
253. de Mercoyrol L, Job C, & Job D (1989) Studies on the inhibition by alpha-amanitin of single-step addition reactions and productive RNA synthesis catalysed by wheat-germ RNA polymerase II. *The Biochemical journal* 258(1):165-169.
254. Bushnell DA, Cramer P, & Kornberg RD (2002) Structural basis of transcription: alpha-amanitin-RNA polymerase II cocrystal at 2.8 A resolution. *Proceedings of the National Academy of Sciences of the United States of America* 99(3):1218-1222.
255. Gong XQ, Nedialkov YA, & Burton ZF (2004) Alpha-amanitin blocks translocation by human RNA polymerase II. *The Journal of biological chemistry* 279(26):27422-27427.
256. Debeaujon I, Leon-Kloosterziel KM, & Koornneef M (2000) Influence of the testa on seed dormancy, germination, and longevity in Arabidopsis. *Plant physiology* 122(2):403-414.
257. Debeaujon I & Koornneef M (2000) Gibberellin requirement for Arabidopsis seed germination is determined both by testa characteristics and embryonic abscisic acid. *Plant physiology* 122(2):415-424.
258. Nesi N, Jond C, Debeaujon I, Caboche M, & Lepiniec L (2001) The Arabidopsis TT2 gene encodes an R2R3 MYB domain protein that acts as a key determinant for proanthocyanidin accumulation in developing seed. *The Plant cell* 13(9):2099-2114.
259. Barakat A*, et al.* (2001) The organization of cytoplasmic ribosomal protein genes in the Arabidopsis genome. *Plant physiology* 127(2):398-415.
260. Winter D*, et al.* (2007) An "Electronic Fluorescent Pictograph" browser for exploring and analyzing large-scale biological data sets. *PloS one* 2(8):e718.
261. Lellis AD*, et al.* (2010) Deletion of the eIFiso4G subunit of the Arabidopsis eIFiso4F translation initiation complex impairs health and viability. *Plant molecular biology* 74(3):249-263.
262. Chiba Y*, et al.* (2004) AtPARN is an essential poly(A) ribonuclease in Arabidopsis. *Gene* 328:95-102.
263. Roberts NJ, Scott RW, & J.T.C. T (2008) Recent Biotechnological Applications Using Oleosins. *The Open Biotechnology Journal* 2:13-21.
264. Siloto RM*, et al.* (2006) The accumulation of oleosins determines the size of seed oilbodies in Arabidopsis. *The Plant cell* 18(8):1961-1974.
265. Murphy DJ & Vance J (1999) Mechanisms of lipid-body formation. *Trends in biochemical sciences* 24(3):109-115.
266. Qin F*, et al.* (2008) Arabidopsis DREB2A-interacting proteins function as RING E3 ligases and negatively regulate plant drought stress-responsive gene expression. *The Plant cell* 20(6):1693-1707.
267. Manfre AJ, LaHatte GA, Climer CR, & Marcotte WR, Jr. (2009) Seed dehydration and the establishment of desiccation tolerance during seed maturation is altered in the Arabidopsis thaliana mutant atem6-1. *Plant & cell physiology* 50(2):243-253.

268. Khan AA & Karssen CM (1981) Changes during light and dark osmotic treatment independently modulating germination and ribonucleic acid synthesis in Chenopodium bonus-henricus seeds. *Physiologia Plantarum* 51(3):269-276.
269. Demarsy E, Courtois F, Azevedo J, Buhot L, & Lerbs-Mache S (2006) Building up of the plastid transcriptional machinery during germination and early plant development. *Plant physiology* 142(3):993-1003.
270. Sano N, *et al.* (2012) Proteomic analysis of embryonic proteins synthesized from long-lived mRNAs during germination of rice seeds. *Plant & cell physiology* 53(4):687-698.
271. He D, Han C, Yao J, Shen S, & Yang P (2011) Constructing the metabolic and regulatory pathways in germinating rice seeds through proteomic approach. *Proteomics* 11(13):2693-2713.
272. Bassel GW, *et al.* (2008) Elucidating the germination transcriptional program using small molecules. *Plant physiology* 147(1):143-155.
273. Lespinay Ad, Lequeux Hln, Lambillotte Ba, & Lutts S (2010) Protein synthesis is differentially required for germination in Poa pratensis and Trifolium repens in the absence or in the presence of cadmium. *Plant Growth Regul* 61:205–214.
274. Frankland B, Jarvis BC, & Cherry JH (1971) RNA synthesis and the germination of light-sensitive lettuce seeds. *Planta* 97:39-49.
275. Morris K, *et al.* (2011) Regulation of seed germination in the close Arabidopsis relative Lepidium sativum: a global tissue-specific transcript analysis. *Plant physiology* 155(4):1851-1870.
276. Walton DC (1966) Germination of phaseolus vulgaris I. Resumption of axis growth. *Plant physiology* 41(2):298-302.

VITA

Name: Liuyin Ma

Education:

2004-2008: Bachelor of Agriculture
Major: Plant Science & Technology
Northwest A & F University,
Yangling, Shaanxi, China.

Research papers:

Liuyin Ma, Pratap Kumar Pati, Qingshun Q. Li, and Arthur G. Hunt. (2013) High throughput characterizations of poly(A) site choice in plants. Manuscript submitted to *Methods*.

Liuyin Ma, Allan Bruce Downie, and Arthur G. Hunt. (2013) Identify genes producing alternative polyadenylated mRNA in seed germination. Manuscript in preparation.

Liuyin Ma, Pratap Pati, Allan Bruce Downie, and Arthur G. Hunt. (2014) The role of polyadenylation in seed germination: Defining the trans(crypto)me. Manuscript in preparation.