**MIAMI UNIVERSITY**

**The Graduate School**

# Certificate for Approving the Dissertation

**We hereby approve the Dissertation**

**of**

Yingjia Shen

Candidate for the Degree:

**Doctor of Philosophy**

_____

Director
Dr. Qingshun Q. Li

_____

Reader
Dr. John Z. Kiss

_____

Reader
Dr. Chun Liang

_____

Reader
Dr. Nancy Smith-Huerta

_____

Graduate School Representative
Dr. Jack C. Vaughn

ABSTRACT

GENOME WIDE STUDIES OF MRNA 3'-END PROCESSING SIGNALS AND ALTERNATIVE POLYADENYLATION IN PLANTS

By Yingjia Shen

Chapter one is an overview of the entire dissertation. In this chapter, I provide background information about current understanding of polyadenylation [poly(A)], poly(A) signals and alternative polyadenylation (APA) in plants and other organisms.

Chapter two presents a survey of rice polyadenylation landscape using 55,742 authenticated poly(A) sites. A substantial similarity was found between rice and Arabidopsis in term of *cis*-elements, suggesting that the polyadenylation machinery is conserved in higher plants. We also found an extensive APA profile in rice where 50% of the genes analyzed have more than one unique poly(A) site and about 4% of the analyzed genes possess alternative poly(A) sites that could result in different protein products.

In Chapter three, we analyzed the nuclear mRNA polyadenylation mechanisms in the model alga *Chlamydomonas reinhardtii* with 16,952 *in silico* authenticated poly(A) sites. We found an unique and complex poly(A) signal profile that is different from higher plants and mammals. A high level of APA was also found in the *Chlamydomonas* genome.

In Chapter four, we used over 300 million Arabidopsis and rice sequence signatures by Massively Parallel Signature Sequencing (MPSS) and Illumina's GAII Sequence by Synthesis methods for the analysis of APA and its relationship with differential gene expression. We discovered a large number of genes undergo APA that have not been found previously by other methods. In both species, APA events upstream of stop codons

are evident from about 50% of the signatures, corresponding to about 10% of whole transcriptome abundance.

Chapter 5 concludes what was discovered in these studies and also gives some future perspectives for the research directions of mRNA polyadenylation in general, particularly in plants.

# GENOME WIDE STUDIES OF MRNA 3'-END PROCESSING SIGNALS AND ALTERNATIVE POLYADENYLATION IN PLANTS

A DISSERTATION

Submitted to the Faculty of

Miami University in partial

fulfillment of the requirements

for the degree of

Doctor of Philosophy

Department of Botany

by

Yingjia Shen

Miami University

Oxford, Ohio

2009

Dissertation Director: Dr. Qingshun Q. Li

## Copyright

## TABLE OF CONTENTS

## List of tables

## List of figures

**Acknowledgements**

I would like to express my gratitude to all those people who have made this dissertation possible and because of whom I will cherish my graduate experience forever.

My deepest gratitude is to my advisor, Dr. Qingshun Q. Li. I have been amazingly fortunate to have an advisor who gave me the freedom to explore on my own, and the guidance to cogitate when I was in doubt. His patience and support helped me overcome many crisis situations and finish this dissertation.

I would like to thank my dissertation committee members: Dr. John Z Kiss, Dr. Chun Liang, Dr. Nancy Smith-Huerta and Dr. Jack Vaughn for insightful comments, helpful suggestions and discussions.

This is also a great opportunity to express my respect to fellow members in the Li lab with whom I have interacted during the course of my graduate study. Particularly, I would like to acknowledge Dr. Denghui Xing and Dr. Man Liu, Jun Zheng and Zhangyang Liu for technical support and meaningful discussions. I appreciate the efforts of the undergraduate students; my sincere thanks to Dennis Keselman for his contributions to the lab.

None of this would have been possible without the love and patience of my family. My immediate family, to whom this dissertation is dedicated to, has been a constant source of love, concern, support and strength all these years. I would like to express my heart-felt gratitude to my wife Zhan Li and my wonderful son Charles. My parents and in-laws have also been wonderful to provide aid to take care of Charles and let me focus on my research.

Finally, I would like to express my gratitude to the Cell, Molecular and Structure Biology graduate program for their PAEA Research Assistantship Award for one year and the Department of Botany at Miami University for dissertation fellowship for the last semester of my graduate study. I also appreciate the financial support from Department of Botany's Academic challenge program, *Sigma Xi*'s grant-in-aid and Miami University's DUOS for funding supports of the research presented in this dissertation.

## Abbreviations:

APA: Alternative Polyadenylation

CE: Cleavage Element

CF: Cleavage Factors

CPSF: Cleavage and Polyadenylation Specificity Factor

CS: Cleavage Element

CstF: Cleavage Stimulating Factor

EST: Expressed Sequence Tag

FCA: Flowering Locus C

FUE: Far Upstream Elements

FY: a RNA 3′ processing protein

MPSS: Massively Parallel Signature Sequencing

nt: nucleotides

NUE: Near Upstream Elements

PAP: Poly(A) Polymerase

PASS: Poly(A) Site Sleuth

PCFS: an Arabidopsis polyadenylation factor

Poly(A): Polyadenylation or Poly-adenosine

RSAT: Regulatory Sequence Analysis Tools

RT-PCR: Reverse Transcription and Polymerase Chain Reaction

SBS: Sequencing By Synthesis

UTR: Untranslated Region

# CHAPTER 1: INTRODUCTION

## *Messenger RNA 3'-end formation and polyadenylation factors*

For a functional gene to be expressed, messenger RNA (mRNA) is first transcribed from DNA, and then directs polypeptide synthesis in the translation process. In eukaryotic cells, RNA polymerase II makes a copy of a gene from DNA to a pre-mRNA, which is further processed to a mature mRNA (Zhao et al., 1999). Processing steps such as 5'- capping, splicing and polyadenylation are crucial for the mRNA to be functional (Proudfoot et al., 2002). Polyadenylation is the addition of a poly-adenosine [poly (A)] tract to the 3'-end of an mRNA molecule and serves many important biological functions in eukaryotic cells (Zhao et al., 1999; Gilmartin, 2005).

In eukaryotic cells, the first function of mRNA polyadenylation is to protect mature mRNA from unregulated degradation (Coller et al., 1998). The mRNA moiety is subjected to many degradation pathways to ensure the desired level of mRNA. A poly(A) tail in the 3'-end protects mRNA from 3'- exonuclease digestion (Kuhn and Wahle, 2004). Secondly, a poly(A) tail also promotes export of mRNA to the cytoplasm. Mutations in polyadenylation related proteins result in nuclear accumulation of all mRNA, thereby linking mRNA exporting to the process of polyadenylation (Brodsky and Silver, 2000). Other studies have also shown that poly(A) binding proteins recognize the poly(A) tail and promotes export from the nucleus (Coller et al., 1998). Furthermore, poly(A) binding proteins interact with several translation-related proteins, such as initiation factor-4G which in turn recruits the 40S ribosomal subunit, thus facilitating translation initiation ( Jacobson and Peltz, 1996; Zhao et al., 1999; Gilmartin, 2005; Siddiqui et al., 2007).

During the formation of mRNA 3'-ends, multiple protein factors are required to recognize nucleotide (nt) sequences in the 3'-UTR [called poly(A) signals, discussed in

the next section] and initiate the cleavage. For example, CPSF (cleavage and polyadenylation specificity factor), CstF (cleavage stimulation factor) and CF (cleavage factor) are all multi-subunit complexes that either bind to nascent transcripts and induce cleavage reaction directly or play an accessory role in this process (Ryan et al., 2004; Danckwardt et al., 2008). In yeast, homologues of these polyadenylation factors have been identified and found to be generally conserved with mammalian ones (Zhao et al., 1999; Garber et al., 1999).

Unlike yeast and mammals, information about plant polyadenylation factors is less well understood. Genetic and biochemical studies identified most plant homologs of mammalian polyadenylation genes in rice and Arabidopsis (Xu *et al*., 2006; Hunt et al., 2008). However, many of these genes remain putative polyadenylation related protein, without knowing their real functions or to which part of residues they bind in 3'UTR.

## *Polyadenylation signals in mammals and yeast*

A key question in the studies of mRNA 3'-end processing is to understand how the poly(A) site is recognized by polyadenylation related proteins. The use of alternative sites could drastically change the coding capacity or alter the inclusion or exclusion of regulatory elements on the mRNA. Previous studies suggest that an interaction of signal residues in pre-mRNA and polyadenylation-related protein factors is crucial in this process (Zhao et al., 1999). In mammalian cells, classical core polyadenylation signals were defined by three elements: the highly conserved hexanucleotide AAUAAA or a close variant found between 10- and 30-nt upstream of the cleavage (so called polyadenylation signal) in half of mammalian genes, a less conserved U-rich or GU-rich element in the downstream of the cleavage site and the cleavage site itself ( Zarudnaya et al., 2003; Hu et al., 2005). To address whether there are additional *cis*-elements besides these well-characterized signals, Legendre and Gautheret (2003) analyzed the polyadenylation signals of 4,956 human EST sequences and their -300/+300 nt flanking regions. They visualized the upstream and downstream sequence elements (USE and

DSE, respectively), characterized by U-rich segments. The presence of a USE and DSE can distinguish true polyadenylation sites from randomly occurring A(A/U)UAAA hexamers. Hu et al (2005) developed a computer program named PROBE to identify *cis*-elements by comparing poly(A) regions of frequently used poly(A) sites and less frequently used ones (Hu et al., 2005). Fifteen *cis*-elements were identified in four regions surrounding cleavage sites and several *cis*-elements existing in yeast and plants are also found in the human poly (A) region, suggesting that many *cis*-elements are evolutionarily conserved among eukaryotes.

Signals which regulate mRNA 3'-end formation in the yeast are somewhat different from these in higher eukaryotes. Polyadenylation signals in yeast are less highly conserved and more complex than those in higher eukaryotes (Zhao et al., 1999). A yeast polyadenylation signal was previously proposed to consist of three elements: an UA-rich efficiency element, an A-rich positioning element and the cleavage site itself (Zhao et al., 1999). Graber et al. (1999) and his colleagues investigated 1352 unique pre-mRNA 3'-end-processing sites and two new polyadenylation signals, a predominance of U-rich sequences located on either side of the cleavage site, were discovered. Another research group analyzed the oligonucleotide composition on the sequences located downstream of the stop codon of all yeast genes. Several oligonucleotide families were found to play a role as an efficiency element and were mainly distributed around 35-bp after the stop codon (van Helden et al., 2000a).

With the increasing knowledge of polyadenylation signals, another promising aspect in polyadenylation research is the detection or prediction of polyadenylation sites. One of the earliest and most well-known software is Polyadq, a program designed for detection of human polyadenylation signals (Tabaska and Zhang, 1999). Polyadq can predict poly(A) signals with a correlation coefficient of 0.413 on whole genes and 0.512 in the last two exons of genes (Tabaska and Zhang, 1999). It is also the first program that is able to consistently detect the AAUAAA variant of the poly(A) signal in human. Based on an analysis of the polyadenylation signals of 4956 human EST sequences, the signal

profile was used to develop a software based on ERPIN program and achieved a prediction with 56% sensitivity and 69 to 85% specificity (Legendre and Gautheret, 2003). In addition to ERPIN, a new web-based software tool box, named DNAFSMiner, can also be used to predict polyadenylation signals in human DNA sequences (Liu et al., 2005). In yeast, Graber et al. (2002) designed a tool for the prediction of polyadenylation site using a discrete state-space mode or hidden Markov model (Graber et al., 2002). Based on putative element positioning and previous knowledge on yeast polyadenylation signals, the software has the ability to find probable 3'-processing sites as well as alternative polyadenylation sites. When an optimized threshold is selected, the sensitivity value of precise polyadenylation sites is over 33% and 87% of predicted polyadenylation sites are within 10-nt of the actual polyadenylation sites (Graber et al., 2002).

## *Current understanding of polyadenylation signals in plants*

In plants, the nucleotide profile of the 3'-UTR is different from that of animals and polyadenylation signals are more diverse (Zhao et al., 1999). Conventional genetic mutagenesis experiments revealed three major groups of poly(A) signals: far upstream elements (FUE), near upstream elements (NUE; an AAUAAA-like element) and cleavage site (CS) itself (Rothnie, 1996; Li and Hunt, 1997; Rothnie et al., 2001). Recent bioinformatics research in Arabidopsis confirmed the presence of NUE and FUE and found a new element, namely cleavage element (CE), an expansion of the original CS resides on both sides of the cleavage site at a genomic level. Figure 1-3 shows the length and position of each signal element in the 3-'UTR (Loke et al., 2005). Polyadenylation signals in Arabidopsis are also less conserved than those in mammals. The canonical hexamer AAUAAA in mammals was found only in 10% of Arabidopsis genes about 10 to 30 nt upstream of the cleavage site of over 8,000 expressed sequence tags (ESTs) examined (Loke et al., 2005). Another element, FUE, spans across an approximately 125-nt region just upstream of NUE and is a control or enhancing element with dominant UG-rich motifs (Li and Hunt, 1995; Loke et al., 2005). The new signal element, CE,

includes two U-rich regions located before and after the CS, both spanning about 5 to 10 nt (Loke et al., 2005). Genetic and computational analyses suggest that the total efficiency of polyadenylation is the function of all elements and no single signal sequence element is sufficient for the processing ( Rothnie, 1996; Li and Hunt, 1997). These complex patterns indicate that identification of plant 3'-end processing signals will require a full understanding of plant poly(A) signal elements.

In order to better understand the polyadenylation mechanism, the first challenge is to identify the position of poly(A) site along the transcripts. This challenge can first be met by using mRNA sequencing results that carries information of poly(A) tails. One widely used approach to identify the position of poly(A) is through ESTs or short sequences of complementary DNA (cDNA) that are reverse transcribed from mRNA (Graber et al., 1999b; MacDonald and Redondo, 2002; Loke et al., 2005; Shen et al., 2008a; Shen et al., 2008b). cDNAs are produced from polyadenylated RNA. Thus, oligo (dT) primer is normally used for reverse transcription, and the sequencing result should carry an oligo(A) tail in the cDNA sequence, which is then used to verify the poly(A) sites. However, most ESTs from the GenBank collection (at the National Center for Biological Information, or NCBI, www.ncbi.nlm.nih.gov) were processed before submission and oligo(A) or poly(A) signature sequences are typically trimmed off. This makes it difficult to identify the position of poly(A) sites without their signature sequences. To circumvent the problem, one solution is to obtain and process the original EST sequencing trace files (raw data from sequencers) that contain the signature and other valuable information, including linker sequences, restriction sites, and vector junction regions,, that can positively identify poly(A) sites. To extract poly(A) sites from these raw data, WebTraceMinner (http://www.conifergdb.org/software/wtm1.0/), a public web service for processing and mining EST trace files, was developed (Liang et al., 2007a). More than 800,000 trace files from six plant species are available at GenBank and are valuable sources for the studies of ploy(A) signals. Other trace files are also available from plant genome sequencing project resources [e.g. poplar; ( Sterky et al., 2004; Pavy et al., 2005)].

In addition to EST and cDNA sequencing, recent breakthroughs in sequencing technologies have made the large-scale studying of the polyadenylated transcription easier than before. Two related technologies called Massively Parallel Signature Sequencing (MPSS) and Illumina's sequencing by synthesis (SBS) both enable sequencing of over one million of sequences in one single reaction (Brenner et al., 2000b; Meyers et al., 2004a; Meyers et al., 2004b; Meyers et al., 2004d; Simon et al., 2009). MPSS normally yields millions of 17 or 20 nucleotides called signatures. In rice and Arabidopsis, each of these signatures is derived from a specific restriction enzyme (e.g. *DpnII*) in 3' most occurrence of a polyadenylated transcript (Meyers et al., 2004a; Meyers et al., 2004b; Meyers et al., 2004d). Therefore, MPSS can also be useful in identifying the locations of poly(A) sites besides tracking the expressional frequency of the transcripts as it was originally designed (Meyers et al., 2004d). The SBS method is simple similar method in sample preparation but the sequencing part is achieved by a polymerase-based SBS method. It provides over 75-bp per signature and >2,000 Mb of sequence data per run. This method has been widely used in DNA, mRNA and small RNA sequencing (Simon et al., 2009).

Traditionally, in order to obtain information from a poly(A) site, we need to clone and sequence EST. One sequencing reaction costs at least $5 to perform. Now with the availabilities of new technologies, millions of signatures carrying information of poly(A) sites can be sequenced within one week at the cost of about $0.001 per signature (Simon et al., 2009). The massive number of signatures produced from MPSS and SBS methods provides a unique advantage over EST-based analysis in which the number of poly(A) sites can be studied. Previously in Arabidopsis and rice, over 36 million and 46 million MPSS signatures from 14 and 22 different libraries have been generated, respectively (Meyers et al., 2004b; Nobuta et al., 2007a). However, the detailed analysis was mainly focused on RNA expression profile while the ability of using high-throughput methods in studying polyadenylation has not been fully appreciated. Chapter 4 of this dissertation

will be devoted for such analyses.

## *Alternative polyadenylation and its regulation*

Besides the *cis*-element detection and probabilistic prediction of the polyadenylation site, research on Alternative Polyadenylation (APA) is also the important aspect of polyadenylation research ( Lutz, 2008; Andreassi and Riccio, 2009). It is common for an individual gene to carry multiple sets of poly(A) signal or poly(A) sites. The phenomena of one gene could produce different 3'-end is called APA. It has been documented that APA plays an important role in gene expression regulation (Andreassi and Riccio, 2009; Lutz, 2008). Similar to alternative initiation and alternative splicing, APA is an important mechanism which generates the diversity of mature transcripts by producing mRNAs with different 3'-UTRs or coding regions. Cleavage and polyadenylation can also occur at alternate poly(A) site(s) and which are referred to as APA sites, or regulated polyadenylation, since it is often spatially and/or temporally regulated, resulting in a variety of transcripts and/or protein products from a single gene (Zhang et al., 2005b) . Importantly, APA site usage may result in truncated mRNA that produces non-functional or structurally altered proteins when poly(A) tails are added to the coding region. Such alteration may have serious consequences.

APA is believed to occur in a tissue or disease specific manner and the regulation mechanism of APA is not understood (Zhang et al., 2005). Several computational studies have surveyed APA in mammals. Gautheret et al.(1998) found 189 alternative polyadenylation sites in 1000 human ESTs  Follow-up studies found that over 29% of human genes have more than one polyadenylation site (Gautheret et al., 1998); while another study found that at least half of the human genes were alternatively polyadenylated (Iseli et al., 2002). Yan et al. (2005) analyzed the genome-wide patterns of APA in the human, mouse and rate using the 3'-end of ESTs. Four distinct classes of patterns categorized as polyadenylation-tandem poly (A) sites, composite exons, hidden exons, and truncated exons were observed (Yan and Marr, 2005). Tian et al. (2005)

surveyed 13,942 human and 11,155 mouse genes containing cleavage sites and found that 54% of human genes and 32% of mouse genes have alternative polyadenylation. They also found that the conservation of polyadenylation configuration between human and mouse orthologs is statistically significant, indicating that the both species employed alternative polyadenylation to produce alternative gene transcripts and this process may be evolutionarily conserved (Tian et al., 2005).

In plants, the best-known example of APA occurs when the FCA, a gene controlling flower time in Arabidopsis, pre-mRNA undergoes APA in an intron and yields a truncated mRNA encoding a smaller and presumably non-functional protein (Simpson et al., 2003). The ratio of this truncated mRNA and the full-length mRNA is crucial for the regulation of Arabidopsis flowering time (Quesada et al., 2005). Importantly, such an APA scheme has been implicated in a number of different plant species, both dicots and monocots (Simpson et al., 2003; Lee et al., 2005; Winichayakul et al., 2005), which suggests an evolutionarily conserved mechanism for gene expression regulation. Another case of APA regulation is Arabidopsis *OXT6* gene which encodes two proteins that may link both polyadenylation and splicing processing. Polyadenylation within intron-2 produces a 30 kDa proteins while splicing of intron-2 results in a 68 kDa protein. Interestingly, the smaller protein is a subunit of CPSF while the larger one contains a domain implicated in mRNA splicing. Therefore, the *OXT6* might regulate its own expression through APA-related pathway (Delaney et al., 2006a; Hunt et al., 2008).

Although APA plays an important role in gene expression regulation, little is known about the extent of APA or its overall impact on the transcriptome in plants. A study using MPSS technology estimated that about 25% of Arabidopsis genes undergo APA (Meyers et al., 2004c). This study, however, didn't provide detailed information about the location or abundance of APA events, thus requiring further fine scale analysis. We have further examed the poly(A) signal and extent of APA using both the EST-based method and large-scale sequencing methods. This dissertation provides information regarding: 1.

Polyadenylation sites and their location in the genomes if applicable; 2. Putative signals regulating the polyadenylation process; 3. Information of alternative polyadenylation sites. A collection of all the information will be an important resource for studying the role of polyadenylation in plant gene expression regulation.

## *Overall objectives*

The purpose of this dissertation research was to understand how polyadenylation sites are recognized and regulated in plant cells. The first question I attempted to answer was: What are polyadenylation signals in plant species other than Arabidopsis? To address these questions, we first retrieved ESTs from two model plant species, rice and the alga *Chlamydomonas*, to look for authentic poly(A) sites. These ESTs were further mapped to genome in order to identify the poly(A) sites. A series of computer programs were developed and used to study the signal profile near poly(A) sites. A better understanding of the poly(A) signal elements will not only enhance gene predictions by finding the position of the end of a gene but also extend our knowledge of how post-transcriptional processes regulate gene expression through differential processing of a gene transcript.

The second question I attempted to answer was: How many genes have utilized APA and where are these APA sites located? Sequences generated from newer sequencing technologies were used for such analysis. Based on these studies, I evaluated the extent and importance of APA in two important species, rice and Arabidopsis.

### *Outline of chapters*

Chapter one is an overview of the entire dissertation. In this chapter, I provide background information about current understanding of the polyadenylation signal and alternative polyadenylation in plants.

Chapter two presents a survey of the rice polyadenylation landscape using 55,742 authenticated poly(A) sites. A substantial similarity was found between rice and Arabidopsis in term of *cis*-element, suggesting that the polyadenylation machinery is conserved in higher plants. We also found an extensive alternative polyadenylation profile, where 50% of the genes analyzed had more than one unique poly(A) site and about 4% of the analyzed genes possess alternative poly(A) sites in their introns, 5'-UTRs, or protein coding regions.

In Chapter three, we analyzed the nuclear mRNA polyadenylation mechanisms in the model alga *Chlamydomonas reinhardtii* using 16,952 authenticated poly(A) sites. We found an unique and complex polyadenylation signal profile that is different from higher plants and mammals. We also found a high level of alternative polyadenylation in the *Chlamydomonas* genome, with a range of up to 33% of the 4,057 genes analyzed having at least two unique poly(A) sites and ~1% of these genes having poly(A) sites residing in predicted coding sequences, introns, or 5'-UTRs.

In Chapter four, we used over 300 million Arabidopsis and rice signatures sequenced from MPSS (Massively Parallel Signature Sequencing) and Illumina's GAII SBS (sequence by synthesis) methods . We discovered that a large amount of genes undergo APA that have not been found previously by other methods. In both species, APA events upstream of stop codons are evident from about half of the signatures, which corresponding to 10% of whole transcript abundance. This work as well as that in the previous two chapters suggests that APA is a common strategy used by cells to increase transcriptome complexity.

Chapter 5 concludes with what was discovered in this studies, and also gives some future perspectives of the research directions of mRNA polyadenylation in general, with plants in particular.

## *References*


Andreassi, C., and Riccio, A. (2009). To localize or not to localize: mRNA fate is in 3'UTR ends. Trends Cell Biol *19*, 465-474.

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D.H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M.*, et al.* (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nature Biotechnology *18*, 630-634.

Brodsky, A.S., and Silver, P.A. (2000). Pre-mRNA processing factors are required for nuclear export. RNA *6*, 1737-1749.

Coller, J.M., Gray, N.K., and Wickens, M.P. (1998). mRNA stabilization by poly(A) binding protein is independent of poly(A) and requires translation. Genes Dev *12*, 3226-3235.

Gautheret, D., Poirot, O., Lopez, F., Audic, S., and Claverie, J.M. (1998). Alternate polyadenylation in human mRNAs: a large-scale analysis by EST clustering. Genome Res *8*, 524-530.

Gilmartin, G.M. (2005). Eukaryotic mRNA 3' processing: a common means to different ends. Genes & Development *19*, 2517-2521.

Graber, J.H., Cantor, C.R., Mohr, S.C., and Smith, T.F. (1999). *In silico* detection of control signals: mRNA 3'-end-processing sequences in diverse species. PNAS *96*, 14055-14060.

Graber, J.H., McAllister, G.D., and Smith, T.F. (2002). Probabilistic prediction of *Saccharomyces cerevisiae* mRNA 3'-processing sites. Nucleic Acids Research *30*, 1851-1858.

Hu, J., Lutz, C.S., Wilusz, J., and Tian, B. (2005). Bioinformatic identification of candidate *cis*-regulatory elements involved in human mRNA polyadenylation. RNA *11*, 1485-1493.

Iseli, C., Stevenson, B.J., de Souza, S.J., Samaia, H.B., Camargo, A.A., Buetow, K.H., Strausberg, R.L., Simpson, A.J., Bucher, P., and Jongeneel, C.V. (2002). Long-range heterogeneity at the 3' ends of human mRNAs. Genome Res *12*, 1068-1074.

Jacobson, A., and Peltz, S.W. (1996). Interrelationships of the pathways of mRNA decay and translation in eukaryotic cells. Annu Rev Biochem *65*, 693-739.

Kuhn, U., and Wahle, E. (2004). Structure and function of poly(A) binding proteins. Biochim Biophys Acta *1678*, 67-84.

Lee, J.H., Cho, Y.S., Yoon, H.S., Suh, M.C., Moon, J., Lee, I., Weigel, D., Yun, C.H., and Kim, J.K. (2005). Conservation and divergence of FCA function between Arabidopsis and rice. Plant Mol Biol *58*, 823-838.

Legendre, M., and Gautheret, D. (2003). Sequence determinants in human polyadenylation site selection. BMC Genomics *4*, 7.

Li, QQ., and Hunt, A.G. (1995). A near-upstream element in a plant polyadenylation signal consists of more than six nucleotides. Plant Mol Biol *28*, 927-934.

Li, QQ., and Hunt, A.G. (1997). The polyadenylation of RNA in plants. Plant Physiol *115*, 321-325.

Liang, C., Wang, G., Liu, L., Ji, G., Liu, Y., Chen, J., Webb, J.S., Reese, G., and Dean, J.F. (2007). WebTraceMiner: a web service for processing and mining EST sequence trace files. Nucleic Acids Res *35*, W137-142.

Liu, H., Han, H., Li, J., and Wong, L. (2005). DNAFSMiner: a web-based software toolbox to recognize two types of functional sites in DNA sequences. Bioinformatics *21*, 671-673.

Loke, J.C., Stahlberg, E.A., Strenski, D.G., Haas, B.J., Wood, P.C., and Li, Q.Q. (2005). Compilation of mRNA polyadenylation signals in Arabidopsis revealed a new signal element and potential secondary structures. Plant Physiol *138*, 1457-1468.

Lutz, C.S. (2008). Alternative polyadenylation: a twist on mRNA 3' end formation. ACS Chem Biol *3*, 609-617.

MacDonald, C.C., and Redondo, J.L. (2002). Reexamining the polyadenylation signal: were we wrong about AAUAAA? Mol Cell Endocrinol *190*, 1-8.

Meyers, B.C., Lee, D.K., Vu, T.H., Tej, S.S., Edberg, S.B., Matvienko, M., and Tindell, L.D. (2004a). Arabidopsis MPSS. An online resource for quantitative expression analysis. Plant Physiology *135*, 801-813.

Meyers, B.C., Tej, S.S., Vu, T.H., Haudenschild, C.D., Agrawal, V., Edberg, S.B., Ghazal, H., and Decola, S. (2004b). The use of MPSS for whole-genome transcriptional analysis in Arabidopsis. Genome research *14*, 1641-1653.

Meyers, B.C., Vu, T.H., Tej, S.S., Ghazal, H., Matvienko, M., Agrawal, V., Ning, J., and Haudenschild, C.D. (2004c). Analysis of the transcriptional complexity of

Arabidopsis thaliana by massively parallel signature sequencing. Nat Biotechnol *22*, 1006-1011.

Meyers, B.C., Vu, T.H., Tej, S.S., Ghazal, H., Matvienko, M., Agrawal, V., Ning, J., and Haudenschild, C.D. (2004d). Analysis of the transcriptional complexity of *Arabidopsis thaliana* by massively parallel signature sequencing. Nature biotechnology *22*, 1006-1011.

Natalizio, B.J., Muniz, L.C., Arhin, G.K., Wilusz, J., and Lutz, C.S. (2002). Upstream elements present in the 3'-untranslated region of collagen genes influence the processing efficiency of overlapping polyadenylation signals. J Biol Chem *277*, 42733-42740.

Nobuta, K., Vemaraju, K., and Meyers, B.C. (2007). Methods for analysis of gene expression in plants using MPSS. Methods Mol Biol (Clifton, NJ) *406*, 387-408.

Pavy, N., Laroche, J., Bousquet, J., and Mackay, J. (2005). Large-scale statistical analysis of secondary xylem ESTs in pine. Plant Mol Biol *57*, 203-224.

Proudfoot, N.J., Furger, A., and Dye, M.J. (2002). Integrating mRNA processing with transcription. Cell *108*, 501-512.

Quesada, V., Dean, C., and Simpson, G.G. (2005). Regulated RNA processing in the control of Arabidopsis flowering. Int J Dev Biol *49*, 773-780.

Rothnie, H.M. (1996). Plant mRNA 3'-end formation. Plant Mol Biol *32*, 43-61.

Rothnie, H.M., Chen, G., Futterer, J., and Hohn, T. (2001). Polyadenylation in rice tungro bacilliform virus: *cis*-acting signals and regulation. J Virol *75*, 4184-4194.

Shen, Y., Ji, G., Haas, B.J., Wu, X., Zheng, J., Reese, G.J., and Li, Q.Q. (2008a). Genome level analysis of rice mRNA 3'-end processing signals and alternative polyadenylation. Nucleic Acids Res *36*, 3150-3161.

Shen, Y., Liu, Y., Liu, L., Liang, C., and Li, Q.Q. (2008b). Unique features of nuclear mRNA poly(A) signals and alternative polyadenylation in *Chlamydomonas reinhardtii*. Genetics *179*, 167-176.

Siddiqui, N., Mangus, D.A., Chang, T.C., Palermino, J.M., Shyu, A.B., and Gehring, K. (2007). Poly(A) nuclease interacts with the C-terminal domain of polyadenylate-binding protein domain from poly(A)-binding protein. J Biol Chem *282*, 25067-25075.

Simon, S.A., Zhai, J., Nandety, R.S., McCormick, K.P., Zeng, J., Mejia, D., and Meyers, B.C. (2009). Short-read sequencing technologies for transcriptional analyses. Annu Rev Plant Biol *60*, 305-333.

Simpson, G.G., Dijkwel, P.P., Quesada, V., Henderson, I., and 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.

Sterky, F., Bhalerao, R.R., Unneberg, P., Segerman, B., Nilsson, P., Brunner, A.M., Charbonnel-Campaa, L., Lindvall, J.J., Tandre, K., Strauss, S.H.*, et al.* (2004). A Populus EST resource for plant functional genomics. PNAS *101*, 13951-13956.

Tabaska, J.E., and Zhang, M.Q. (1999). Detection of polyadenylation signals in human DNA sequences. Gene *231*, 77-86.

Tian, B., Hu, J., Zhang, H., and Lutz, C.S. (2005). A large-scale analysis of mRNA polyadenylation of human and mouse genes. Nucleic Acids Res *33*, 201-212.

van Helden, J., Andre, B., and Collado-Vides, J. (2000). A web site for the computational analysis of yeast regulatory sequences. Yeast *16*, 177-187.

Winichayakul, S., Beswick, N.L., Dean, C., and Macknight, R.C. (2005). Components of the Arabidopsis autonomous floral promotion pathway, FCA and FY, are conserved in monocots. Functional Plant Biol *32*, 345-355.

Yan, J., and Marr, T.G. (2005). Computational analysis of 3'-ends of ESTs shows four classes of alternative polyadenylation in human, mouse, and rat. Genome Res *15*, 369-375.

Zarudnaya, M.I., Kolomiets, I.M., Potyahaylo, A.L., and Hovorun, D.M. (2003). Downstream elements of mammalian pre-mRNA polyadenylation signals: primary, secondary and higher-order structures. Nucleic Acids Res *31*, 1375-1386.

Zhang, H., Lee, J.Y., and Tian, B. (2005). Biased alternative polyadenylation in human tissues. Genome Biol *6*, R100.

Zhao, J., Hyman, L., and Moore, C. (1999). Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. Microbiol Mol Biol Rev *63*, 405-445.

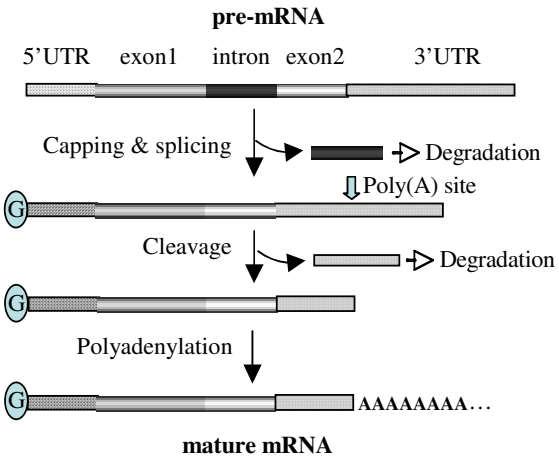



Figure 1-1: A representation of eukaryotic pre-mRNA processing (emphasis on 3'-end formation).

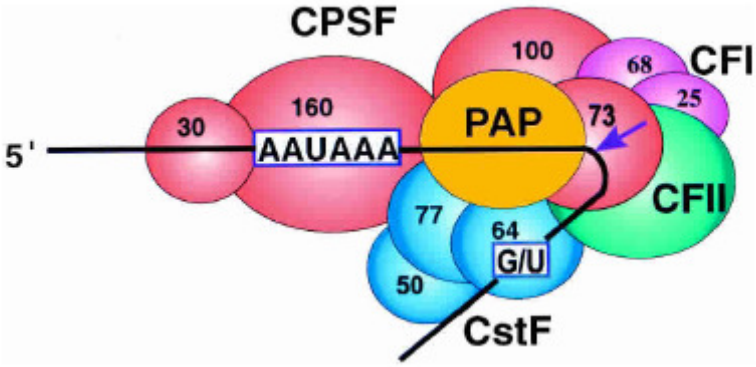



Figure 1-2: A simplified mammalian pre-mRNA 3'-end processing complex.

CPSF: cleavage and polyadenylation specificity factor; PAP: poly (A) polymerase; CstF: cleavage stimulating factor; CFI and CFII cleavage factors I and II. From Ryan et al., 2004.

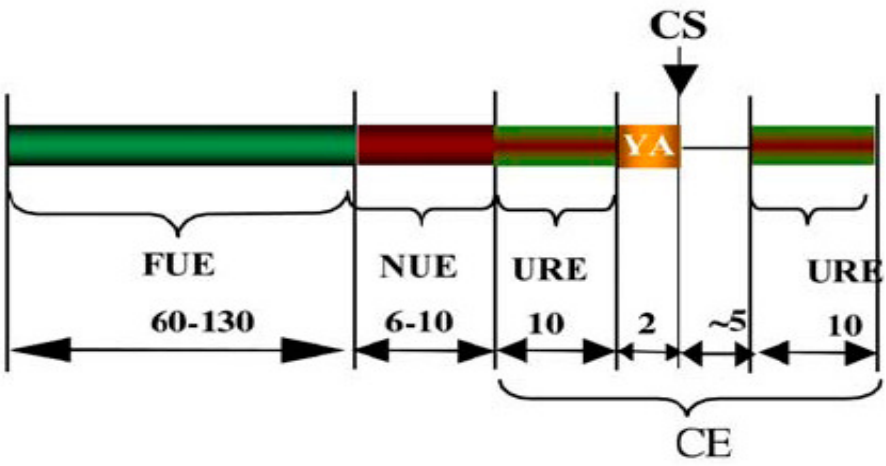



Figure 1-3: mRNA poly(A) signals in Arabidopsis.

FUE; Far Upstream Element; NUE; Near Upstream Element; CE; Cleavage element; CS; Cleavage site; URE; U-rich region; YA, CA or UA right before the cleavage site in the 3'UTR. (Loke et al., 2005)

# CHAPTER 2: GENOME LEVEL ANALYSIS OF RICE mRNA 3'-END PROCESSING SIGNALS AND ALTERNATIVE POLYADENYLATION

The material in this chapter has been published as: Yingjia Shen, Guoli Ji, Brian J. Haas, Xiaohui Wu, Jianti Zheng, Greg J. Reese and Qingshun Quinn Li. **Genome level analysis of rice mRNA 3'-end processing signals and alternative polyadenylation**. *Nucleic Acids Res*. 2008 36(9): 3150-3161. Contributions to this chapter: Guoli Ji, Xiaohui Wu and Jianti Zheng developed the PASS software. Brian Haas generated the initial rice poly(A) data set. Greg Reese analyzed the local cluster sequence patterns. All of the other data were collected and analyzed by Yingjia Shen.

Supplementary data from this chapter can be downloaded from http://nar.oxfordjournals.org/cgi/content/full/gkn158/DC1

## *Abstract*

The position of a poly(A) site of eukaryotic mRNA is determined by sequence signals in pre-mRNA and a group of polyadenylation factors. To reveal rice poly(A) signals at a genome level, we constructed a dataset of 55,742 authenticated poly(A) sites and characterized the poly(A) signals. This resulted in identifying the typical tripartite *cis*-elements, including FUE, NUE and CE, as previously observed in Arabidopsis. The average size of the 3'-UTR was 289 nucleotides. When mapped to the genome, however, 15% of these poly(A) sites were found to be located in the currently annotated intergenic regions. Moreover, an extensive alternative polyadenylation profile was evident where 50% of the genes analyzed had more than one unique poly(A) site (excluding microheterogeneity sites), and 13% had 4 or more poly(A) sites. About 4% of the analyzed genes possessed alternative poly(A) sites at their introns, 5'-UTRs, or protein coding regions. The authenticity of these alternative poly(A) sites was partially confirmed using MPSS data. Analysis of nucleotide profile and signal patterns indicated that there may be a different set of poly(A) signals for those poly(A) sites found in the coding regions. Based on the features of rice poly(A) signals, an updated algorithm termed PASS-Rice was designed to predict poly(A) sites.

## *Introduction*

During gene expression in eukaryotes, one of the mRNA processing steps is 3'-end formation, which includes cleavage and addition of a polyadenine tract [poly(A)] to the newly formed end. This polyadenylation process is tightly associated with transcription termination ( Zhao et al., 1999; Proudfoot et al., 2002a), and the poly(A) tail is crucial for the mRNA's functions because it serves multiple facets of common cellular functions. These functions include transport of mRNA from nucleus into cytoplasm, enhancement of mRNA stability, and regulation of mRNA translation (Zhao et al., 1999; Proudfoot et al., 2002). Previous studies show that sequence signals on pre-mRNA determine the specific position of the poly(A) site as well as the processing efficiency. In vertebrates cells, there are three elements defined as the core polyadenylation signal in the 3' untranslated region (UTR) of pre-mRNA: the highly conserved AAUAAA, about 10 to 30 nucleotides (nt) upstream of the cleavage site, and a downstream U-rich or GU-rich element ( Zhao et al., 1999;Gilmartin, 2005; Hu et al., 2005; Salisbury et al., 2006). A less conserved third element of the form UGUA at variable distances upstream of the cleavage sites has also been shown to potentially play a role, particularly in those genes that do not have AAUAAA (Venkataraman et al., 2005). In yeast, however, poly(A) signals are different from those observed in mammals in both signal sequence patterns and organization. Specifically, the signals are less conserved, with a lack of downstream elements (Graber et al., 1999a, b; Zhao et al., 1999). Further studies showed that there are also two U-rich elements flanking cleavage sites in yeast (Graber et al., 1999a).

Polyadenylation signals in plant mRNA are also less conserved than those found in mammals and therefore share some features in common with yeast (Li and Hunt, 1997; Graber et al., 1999a; Hunt, 2007;). Conventional genetic mutagenesis experiments have revealed three major groups of poly(A) signals in plants: the far upstream elements (FUE), the near upstream elements (NUE, an AAUAAA-like element), and the cleavage site (CS) itself (Rothnie, 1996; Li and Hunt, 1997; Rothnie et al., 2001). Recent

bioinformatics studies in Arabidopsis confirmed the presence of NUE and FUE. The canonical hexamer AAUAAA signal in mammals is only found in 10% of Arabidopsis transcripts (Loke et al., 2005). In addition, that study identified a new element termed the cleavage element (CE), which is an expansion of the original CS (as noted above) and resides on both sides of the cleavage site (Loke et al., 2005). The CE includes two U-rich regions, before and after the CS, both spanning about 10nt. The FUE, on the other hand, spans across an approximate 125-nt region upstream of the NUE and has dominant UG-rich motifs. Genetic analyses suggest that the efficiency of polyadenylation is the result of the cooperative efforts of all elements because no single signal sequence element is sufficient for the processing (Rothnie, 1996; Li and Hunt, 1997). These complex patterns indicate that understanding the plant 3'-end processing mechanism requires a full elucidation of plant poly(A) signal elements, which is one of the foci of this report.

It has been documented that Alternative PolyAdenylation (APA) plays an important role in gene expression regulation. Similar to alternative initiation and alternative splicing, APA is an important mechanism which generates the diversity of mature transcripts by producing mRNAs with different 3'-UTRs or coding regions. More than half of human genes (Zhang et al., 2005) and over 25% of Arabidopsis genes (Meyers et al., 2004) are estimated to have multiple poly(A) sites. Moreover, gene expression regulation through APA can result in altered 3'-UTRs which may affect mRNA stability, translatability, or ability to produce proteins (Chuvpilo et al., 1999; Peterson, 2007). The best-known example of APA in plants occurs when the pre-mRNA, encoded by the FCA gene, undergoes APA in an intron and yields a truncated mRNA that encodes a smaller and presumably non-functional protein (Simpson et al., 2003). The partition of this truncated mRNA and the full-length mRNA is crucial for the regulation of Arabidopsis flowering time (Quesada et al., 2005). Importantly, such an APA scheme has been implicated in a number of different plant species, both dicots and monocots (Simpson et al., 2003; Lee et al., 2005; Winichayakul et al., 2005), which suggests an evolutionarily conserved mechanism for gene expression regulation. Recently, we have also demonstrated the

involvement of other polyadenylation factors in the APA of FCA transcript (Xing et al., 2008). In another seemingly conserved case of APA, Tang et al. (Tang et al., 2002) described how the use of two intronic alternative poly(A) sites of a gene locus produced a shorter transcript encoding Lysine-ketoglutarate reductase leading to the fine-tuning of amino acid metabolism in plants. Interestingly, if these poly(A) sites are bypassed, the same gene produces a transcript encoding a bifunctional protein in the Lysine biosynthesis pathway. An Arabidopsis transcript encoding a polyadenylation factor can be alternatively processed to generate two different proteins, one being AtCPSF30, the other a potential splicing factor (Delaney et al., 2006). Recently, extensive APA has also been noted in the disease resistant gene transcripts in plants (Tan et al., 2007). However, the full extent of plant APA remains unclear.

Although rice is a dominant staple food crop, its mRNA polyadenylation machinery and *cis*-elements are largely unknown. We are therefore interested in analyzing the polyadenylation signals as the first step in understanding this important gene expression process in rice. With the rice genome sequences being made available, it is now feasible to perform large-scale analysis on rice poly(A) signals. Recently, two groups performed analyses on rice poly(A) signals based on 12,969 and 9,911 rice poly(A) sites ( Lu et al., 2006; Dong et al., 2007), respectively. However, these analyses failed to address some important issues. First, the number of genes tested only accounted for less than one third of all rice genes in both cases. Second, Lu et al. (Lu et al., 2006) only tested 40nt up- and downstream of the poly(A) sites, which was too narrow to include all poly(A) signals according to previous mutagenesis-based and bioinformatics studies in plants (Li and Hunt, 1997; Loke et al., 2005; Rothnie, 1996). Most importantly, none of the studies analyzed APA, which, as suggested above, may play a crucial role in the regulation of plant expression.

Here, we present an extensive analysis of the *cis*-elements around rice polyadenylation sites based on a new dataset containing 55,742 unique poly(A) sites. Using the features

of the rice poly(A) signals, we also build a model with which to effectively predict poly(A) sites. In the course of our work, we find that a significant number of rice genes have alternative poly(A) sites and that some of them are located in regions of the genes which could lead to production of altered transcripts and/or protein products.

## *Materials and methods*

### The rice 55K poly(A) site dataset and signal analysis

The sequences around rice poly(A) sites from ESTs were retrieved using the same criteria as previously described (Loke et al., 2005). Briefly, ESTs with oligo(A) stretches (8 to 15 nucleotides with at least 80% adenine content) were extracted and compared to the genomic DNA sequences to ensure that these oligo(A) stretches were not from the genome, which would indicate that they had been added post-transcriptionally. Internal priming contaminations were also eliminated this way. Thus, if the 10 genomic nucleotides past the cleavage site were at least 80% A, the poly(A) site candidate was excluded as a potential source of mispriming. When collecting poly(A) sites, the first adenine of the oligo(A) was generally saved as a poly(A) site nucleotide because previous biochemical and genetic evidence indicated that the first adenine is normally transcribed from DNA, and much less likely to be added post-transcriptionally (Chen et al., 1995; Moore et al., 1986; Sheets et al., 1990). A spike of adenine at the poly(A) sites of this dataset is also seen in yeast and mammal datasets (Graber et al., 1999b).

After alignment to the genome, a 300nt sequence upstream plus a 100nt sequence downstream were extracted for each authenticated poly(A) site. A total of 55,742 such sequences were found from about 1,156,000 rice ESTs (Campbell et al., 2006), and make up the dataset called 55K (available through our web site, www.polyA.org).

SignalSleuth, used in the studies on Arabidopsis poly(A) signals (Loke et al., 2005), was also used to perform an exhaustive search of varying size patterns within sub-regions. The output of SignalSleuth included a matrix file with the occurrence of each designated

length of poly(A) signals in the entire dataset of 55K 3'-UTR sequences. The signal patterns were sorted and ranked based on their frequency compared to the background and then used for further analysis.

**Predictive modeling of poly(A) sites**

A previously described algorithm, Poly(A) Site Sleuth, or PASS (Ji et al., 2007a; Ji et al., 2007b), was modified for use in our rice poly(A) site prediction (hereinafter termed PASS-Rice). Modification include the incorporation of the signal pattern features (NUE, FUE, and CE) and the single nucleotide profile from rice 3'-UTR. The topological structure of the algorithm was based on the Generalized Hidden Markov Model (GHMM) as previously described (Ji et al., 2007b). GHMM recognizes the signals from left to right and only allows the recognition of signals from the current state to the next state in one direction. A background state was added between every two signal states to represent the background sequences around the signals. In addition, a first order inhomogeneous Markov sub-model was built to characterize NUE and CS signals which possessed relatively better conservation. Since this sub-model could then represent the interactions of NUE and CS signals, feature information could be described more clearly.

The performance of PASS-Rice was evaluated by employing two common standards, sensitivity (Sn) and specificity (Sp), as defined previously (Ji et al., 2007b). The parameters of the forward-backward algorithm for the rice poly(A) site recognition system are listed in Supplemental Table S1. In the model, the size, or nucleotide length, of each signal (FUE, NUE, and CE, respectively) was fixed, as shown in Supplemental Table S2. Because there is little conservation in FUE, CE-L and CE-R, we calculated the nucleotide output probability B of these signals directly in their respective regions. The NUE and CS signals are slightly more conserved; therefore, we used a subset of the first order inhomogeneous Markov model to describe the feature information of these two signals. A matrix of transition probabilities was first generated by the best signals (for NUE, the top 50 patterns were used; for CS, CA and UA, dinucleotide frequencies were

used). Then, using the matrix, the nucleotide output probability of NUE and CS signals was calculated by the program automatically.

**Signal logos and the calculation of percentage hits**

We used the method described by Hu et al. (Hu et al., 2005) to generate sequence logos and calculate the percentage of hits. Using dynamic programming, we grouped the selected hexamers based on their distance, computed when gaps were not allowed. Then, Agnes, an agglomeration package in the R language (www.r-project.org), was used to cluster hexamers based on their dissimilarity distance. The suggested cutoff value of 2.6 was used to group them. Hexamers in the same group were further aligned by ClustalW. The length of each sequence logo was determined from the result of ClustalW, and spaces at both ends of the sequence (after alignment) were filled by nucleotides randomly selected from background sequence in the studied region. Finally, the weight of each hexamer in the group was also computed based on its frequency in each studied region, and the Web Logo Tool (Crooks et al., 2004) was used to generate the final images of sequence logos.

To detect if a sequence logo was represented in the studied region, we generated a position-specific scoring matrix for each logo (Hu et al., 2005). For each position, the score S was calculated as follows: $S=\sum_{p=1}^{L}\log_2\frac{f(n,p)}{f(n)}$, where L is the length of the sequence logo, f(n, p) is the frequency of nucleotide n at the position p of the sequence logo, and f(n) is the background frequency of occurrence of nucleotide n in a specific poly(A) region, e.g., NUE.

**Finding alternative polyadenylation sites**

The Build 3 rice genome sequences and corresponding annotation file were downloaded

from the Annotation Dataset of The First Rice Annotation Project Meeting (RAP1) (http://rapdownload.lab.nig.ac.jp/). The rice genes were defined using the full-cDNA sequences. BLAT (Kent, 2002) was used to align all of the 55K sequences to the rice genomic sequences. Each sequence was required to have at least 100nt surrounding its poly(A) site matched to the genome, and the ones that had multiple perfect matches to the genome were eliminated from the final analysis to avoid ambiguity. Finally, a Perl script was written to read the result of BLAT and mark the positions of poly(A) sites on the annotated genome map.

## *Results*

### Profile of rice 3'-UTR

The mRNA poly(A) site positions are determined by the interaction of *cis*-elements on the pre-mRNA and a set of polyadenylation factors. It follows that characterization of the *cis*-elements would lead to understanding poly(A) site selection, as well as finding the potential alternative poly(A) sites that could be used for differential gene expression. In order to study these *cis*-elements, we analyzed any given sequence 300nt upstream plus 100nt downstream for each authenticated poly(A) site in our 55K dataset using SignalSleuth for an exhaustive pattern search algorithm (Loke et al., 2005). First, we examined the single nucleotide profile around the poly(A) sites and the 3'-UTR of all sequences in the dataset. As shown in Figure 2-1A, the 3'-UTR is notably rich in A and U nucleotides and has distinct A and U profiles in which the -225 to -30 region has a high U content, while the -40 to -10 region has a high A content, with a clear transition between the two regions. Previously known YA dinucleotide (Y= C or U) at the cleavage site is indicated by a sharp spike of C (position -4, 18%; -3, 21%; -2, 33%, -1, 7%) which occurs right before the poly(A) site (Li and Hunt, 1997). Based on previous knowledge of poly(A) signals in plants, we further profiled hexamers and octamers near rice poly(A) sites in three distinct regions. Based on nucleotide composition and signal profiling, the locations (relative to the cleavage site, the -1 position) of rice signal elements are as

follows: -150 to -35 for FUE; -35 ~ -10 for NUE and -10 ~ +15 for CE, respectively.

In comparing rice with Arabidopsis as shown in Figure 2-1A and 1B (Loke et al., 2005), we find the general distribution pattern of nucleotides is similar, although the FUE region in rice is slightly expanded towards the coding region. The U-richness is also slightly reduced in rice as the gap between the U- and A- curves is smaller. This trend, however, is changed after the cleavage site, where the gap between U- and A-curves is wider in rice than in Arabidopsis. The U-rich sequences in the CE intersect with a region of high A and C at the cleavage site (termed CE-R and CE-L; 13). This is similar on both rice and Arabidopsis while a slightly higher U-rich peak is seen in the latter.

To examine the length of the 3'-UTR, we calculated the distance between the annotated stop codon and the poly(A) site for each gene. As reflected in longer U- and A-curves in the FUE, the size of the 3'-UTRs in rice is also larger than that in Arabidopsis. The average length of all 3'-UTRs in rice is 289nt and the majority of them are distributed in the range of 150 and 400nt, and the 3'-UTR length distributions among different subsets of poly(A) sites are not significant (Fig. 2-1C). In contrast, the average size of the Arabidopsis 3'-UTR is 223nt, as calculated based on the 3'-UTR dataset downloaded from The Arabidopsis Information Resources (www.Arabidopsis.org).

**Polyadenylation signals in rice**

Based on the scanning results of SignalSleuth, the three signal elements, including FUE, NUE, and CE, that are found in Arabidopsis are also identified in rice, as determined from top-ranked hexamer profiles in each section of the 3'-UTR (Fig. 2-2). This indicates conservation between two groups of plants, dicot and monocot.

To statistically analyze the significance of the signal patterns in these polyadenylation signal elements, we applied an oligo analyzer called Regulatory Sequence Analysis Tools, or RSAT (Van Helden et al., 2000). The full results are listed in the supplemental files (Table S3). Here, we present only the signals in the FUE with length 8nt, and length 6nt in the NUE and CE. These choices are based on our observation of the prevailing signal size in plants (Ji et al., 2007b). A standard score (the Z-score) was used to measure the standard deviation of each pattern from its expected occurrence based on Markov Chain models (Van Helden et al., 2000). Many experimentally characterized poly(A) signals, such as AAUAAA and AUAAAA in NUE, were found on the short list according to the order of Z-scores, indicating the efficacy of such a ranking. The top signal pattern is still AAUAAA, similar to that seen in Arabidopsis (Loke et al., 2005), and it accounts for about 7% of the total poly(A) signals in NUE.

To further study the individual signals of the three signal elements, we first used a word search program developed to compare the frequency of individual signal patterns in the 3'-UTR and coding sequence. Interestingly, a motif of 4 nucleotides, UGUA, the most over-represented tetramer, was found at least once in 76.9% of the FUEs, which range from -125 to -30. By comparison, randomized sequences preserving the nucleotide composition (AU-richness) of the same region only yield 46.9±3.1% (average calculated from testing randomized sequences 1000 times). Hence, UGUA appears 63.8% [100x(76.9-46.9)/46.9] more frequently in the FUE than in the randomized sequences. Moreover, when compared to the region of sequences with a similar nucleotide composition (downstream of the cleavage site, +1 to +96), UGUA was only found in 41.2% of the sequences, demonstrating a significant over-representation in the FUE, where it was 86.7% [100x(76.9-41.2)/41.2] more frequently found. This agrees with findings reported in yeast (Graber et al., 2002) where the UGUA motif was found to have high frequency in similar poly(A) signal regions. The same motif was also found in mammal genes, particularly those that lack AAUAAA in their NUEs (Venkataraman et al., 2005).

To compile and present these results in a concise format, we use a sequence logo program (Hu et al., 2005). The primary advantage of such sequence logos is that each logo represents multiple poly(A) signals corresponding to their occurrences. This reduces the number of signal patterns and, at the same time, ensures that potentially overlapping signals, such as AAUAAA and AAAUAA, are concisely presented. The top signals were those that have a Z-score higher than 8.53, a suggested cutoff for standard hit determination [$p<0.0001$; as described in (Seiler et al., 2007)]. These sequence patterns were clustered to generate sequence logos according to their similarity. Fig. 2-3 shows an example of how the sequence logos were generated in the NUE, where groups were identified according to similarities among the signal patterns. Using this method, we identified 12 major signal clusters for all three polyadenylation signal elements. Their sequence logos, the number of clustered hexamers, the top hexamers with the highest Z-scores and the frequency of occurrence in specific regions are listed in Table 1.

To compare polyadenylation signals of Arabidopsis and rice, we generated a similar set of logos (Supplemental Table S4) from the 8K dataset of Arabidopsis (13) using the same criteria as we did in rice. While comparison at such an abstract level of signal logos may be difficult, there are some obvious differences. One such difference is that a GC-rich *cis*-element was found in the rice FUE region (FUE.7) but not in Arabidopsis. There is only one NUE logo of Arabidopsis instead of 4 in rice. This may suggest that the similar NUE signals are more frequently utilized in Arabidopsis than in rice. In contrast, the CE is much less conserved in Arabidopsis, where a total of 9% of genes carry two *cis*-elements (compared to 67% in rice), indicating potentially less stringent CE to determine the position of poly(A) sites in Arabidopsis. The validity of these observations remains to be confirmed by other methods.

**Analysis of alternative polyadenylation of rice genes**

Alternative polyadenylation is an important mechanism in generating a diversity of mature transcripts. In order to study the extent of APA in rice, we first studied the overall distribution of authenticated poly(A) sites in 55K dataset. We aligned all the poly(A) sites to the full-length cDNA sequences and found that only about 50% of poly(A) sites in the 55K dataset are located within 30nt of annotated poly(A) sites of the rice genome Build 3 (Supplemental Table S5). We then examined the relative distance between neighboring poly(A) sites. In about 70% of these neighboring sites, at least one site was located within 30 nt of another in the same gene. The distribution of the distances among the poly(A) sites is shown in Supplemental Figure S1. This phenomenon, which we term "microheterogeneity", could result from the generally slack nature of the polyadenylation machinery, causing, in turn, the likelihood of overestimating the number of APA sites in the genome. Therefore, to minimize the impact of microheterogeneity in our analysis, we aggregated poly(A) sites that were within 30 nt of each other and considered this grouping to be one unique poly(A) site. Table 2 lists the number of unique poly(A) sites on each gene. Over 50% of the genes have more than one unique poly(A) sites with a maximum number of 19 unique poly(A) sites in a single gene. These poly(A) sites represent the extent of the APA in rice genome.

To further study the position of these alternative poly(A) sites on the genes, we aligned all the 3'-UTR sequences to the annotated rice genome. The results (Table 3) showed that 53.41% of authenticated poly(A) sites are located in the annotated genic regions and that the majority of them (51.45%) are in the 3'-UTR, as expected. Surprisingly, about half of the poly(A) sites were found in the annotated intergenic regions. To gain an understanding of this group of poly(A) sites, we next examined whether they were located close to the ends of the genes. Indeed, 31.26% (17,127) of the poly(A) sites were mapped to the region between the annotated poly(A) site and 100nt downstream of it. By comparison, only 34 poly(A) sites (0.06%) were found within 100nt upstream of

the annotated 5' end of the gene. If the region after poly(A) site is extended to include the region between 1 to 500 nt (for those genes that do not have an annotated 3'-end, 1 to 1000 nt range is used), there is only slight increase (from 31.26% to 35.40%; Table 3). These results suggest that the identification of many poly(A) sites located downstream of an annotated poly(A) site may simply be the result of inaccurate or incomplete annotation from an insufficient number of EST or full-length cDNA sequences. To our surprise, 11.12% (6,092; Table 3) of poly(A) sites are located in the intergenic region, which we define in this paper as being at least 500 nt (or 1000 nt for genes without an annotated 3'-UTR) away from 3'-ends and 100 nt away from 5'-ends of currently annotated full-length cDNA. These poly(A) sites might have originated from unannotated genes or from small, non-coding, or antisense RNAs. A similar observation has been made in human and mouse genomes where some poly(A) sites located in intergenic regions are thought to arise from novel transcripts (Lopez et al., 2006).

Interestingly, close to 2% of the poly(A) sites (1054 out of 55K) are located in the coding sequences (CDS), introns, or 5'-UTRs. Further analysis shows that about 4% of total genes (662 out of 17169 genes that were mapped by the 55K poly(A) sites) use these non-conventional poly(A) sites. To verify these results, we manually mapped some poly(A) sites to the rice MPSS plus database (http://mpss.udel.edu/rice/) (Nobuta et al., 2007b). Massively Parallel Signature Sequencing (MPSS) is a high throughput transcriptional profiling technology used for studying the comprehensive expression atlas (Brenner et al., 2000a). Although the exact locations of poly(A) sites cannot be deduced from this database, MPSS has been used to predict the extent of APA in Arabidopsis since the MPSS signatures are located on the closest DpnII restriction enzyme sites upstream of poly(A) sites (Meyers et al., 2004e). In the rice MPSS plus database, we manually searched over 100 of these non-conventional poly(A) sites and found over 50% to be supported by MPSS signatures, confirming that at least half of the cases use non-conventional poly(A) sites. Differences in annotations or incompleteness of MPSS data, among other possibilities, could account for the remaining unverified poly(A) sites.

To demonstrate how multiple poly(A) sites are located in the genes and the features of MPSS signatures, we use the poly(A) sites of a WRKY DNA binding domain containing protein LOC_Os01g47560 as an example (Fig. 2-4). It has 3 unique poly(A) sites found in the 55K dataset, which are located in the CDS, 3'-UTR, and downstream of annotated poly(A) sites. The poly(A) site in the CDS truncated 42% of the total coding sequence, making it unlikely to produce a functional protein product. The two poly(A) sites that are located downstream of the stop codon have a 255nt gap between them, thus increasing the likelihood that they carry different regulatory elements in their 3'-UTR. These results imply that APA could produce different proteins or non-functional proteins, or mRNA with different 3'-UTR properties, and could also serve as a regulatory mode in the gene expression regulation.

To further study if these APA sites use different cis-elements, we examined the polyadenylation signals around these APA sites. While the single nucleotide profiles of the 5'-UTR and intronic sites (Supplemental Fig. S2) are similar to the general profile (as seen in Fig. 2-1), that profile is very different around the APA sites that are located in the coding region. There, the transitions of A and U in the upstream of the poly(A) are no longer seen and the G and C contents are apparently higher (Fig. 2-5A). Such a difference is not due to smaller sample size because when a similar number (about 250) of sequences from intronic and 5'-UTR APA sites were used, the profiles were similar to the general one (compare Fig. 2-1 with Fig. S2). Next, we investigated if the signal patterns for the coding region APA are different from the regular ones. As shown in Fig. 5B, the NUE signal pattern logos of the coding region APA sites are highly G-rich elements when compared with the overall NUE logos in Table 1. This result, while reflecting the higher GC content in the coding region, implies that the poly(A) signals that direct the formation of APA in the coding sequences may be distinctly different from those signals of other poly(A) sites. It seems possible that these signals might be recognized by different polyadenylation factors, or assisted by other yet unknown proteins.

**Predictive modeling of rice polyadenylation sites**

The unique features of the polyadenylation signals and nucleotide profiles (Fig. 2-1) prompted us to devise an algorithm to predict rice poly(A) sites in an attempt to assist genome annotation and to scan transgenes to eliminate cryptic poly(A) sites that may hamper their expression in rice. We previously designed a program called Poly(A) Site Sleuth (or PASS) to predict poly(A) sites in Arabidopsis based on the Generalized Hidden Markov Model (GHMM; Ji et al., 2007a; Ji et al., 2007b). For the new model, we modified PASS using the features of rice polyadenylation signals (see Methods) and named it PASS-Rice.

To evaluate the performance of our model, we employed two common measures: sensitivity (Sn) and specificity (Sp) (Ji et al., 2007b). Sensitivity is defined as the fraction of true poly(A) sites correctly identified as positive, and specificity is the fraction of non-poly(A) sites correctly predicted as negative by PASS-Rice. Thus, high Sp and Sn values positively correlate to the increased validity of prediction model results. In the model, the rice sequences containing a single poly(A) site were used to calculate Sn. Because not all poly(A) sites have been identified in each sequence of the dataset, we cannot calculate the real Sp value. Therefore, we use several negative control datasets, and a dataset with randomly generated sequences that preserves the trinucleotide distributions in the 3'-UTR, to evaluate Sp. As shown in Fig. 2-6A, the descending line shows the variation of Sn, while the ascending lines show the variation of Sp in different datasets. Sp_Intron, Sp_5UTR, Sp_CDS and Sp_MC represent different Sp values calculated using rice introns, 5'-UTRs, coding sequences, and a randomly generated sequence dataset, respectively. Sn-3UTR represents the Sn calculated using the rice 3'-UTR sequences containing a single poly(A) site, and the prediction site is exactly the validated site. The PASS algorithm reached its best combination of specificity and sensitivity ( ~90% each) when the threshold (score) was set at 4 (Fig. 2-6A).

To test the validity of PASS-Rice, we examined many rice genes that have multiple poly(A) sites. The example given in Fig. 2-6B shows a gene Los_Os03g61890 tested by PASS-Rice and indicates that most of the experimentally validated sites are within the highly-scored (around 4) area of the 3'-UTR. However, PASS-Rice predicted peaks at around locations 350nt and 370nt, which were not validated by EST. This very likely results from the relatively small number of authenticated poly(A) sites corresponding to each gene (average ~2 ESTs for each gene) or some other components, such as protein factors or RNA secondary structures, which were not considered when modeling.

We also used PASS-Rice to scan a 50kb genomic sequence to see how it works in large-scale analysis. The results shown in Supplemental Fig. S3 indicate that PASS-Rice can clearly detect ends of genes, thus making it potentially useful in genome annotation by predicting the ends of transcripts. This predictive model can also be used to screen for potentially undesirable poly(A) sites and eventually eliminate them through targeted mutations in the transgenes. The PASS-Rice program is available through our web site (www.polyA.org).

## *Discussion*

Using SignalSleuth and RSAT, we performed a detailed analysis of rice poly(A) signals covering 55,742 authenticated poly(A) sites, and the results were used to build a predictive model for rice poly(A) sites. We also found that APA is extensive in rice, with about 50% of the genes having at least two poly(A) sites that are 30 nt apart. In addition, many poly(A) sites, including some confirmed by MPSS, were found in the exon or intron regions of the genes. This could be an alternative mechanism for regulating gene activities. More interestingly, we suggested that the APA sites in the coding sequences may use a different set of polyadenylation signals. A significant amount of polyadenylated transcripts (~11%) was found at least 500 nt outside 3'-end or 100 nt

outside 5'-end of the currently annotated genic regions, indicating the presence of some unannotated transcripts.

The distribution of the poly(A) signal regions in rice is generally similar to the previous working model of Arabidopsis (Loke et al., 2005). However, by comparing the Arabidopsis and rice models, differences are noticeable, both in pattern compositions and length of elements. First, the AAUAAA signal (known as a canonical signal in mammals), still ranked the first on the NUE signal list, was only found in approximately 7% of all tested rice 3'-UTRs in contrast to about 10% in Arabidopsis. This is also reflected in the signal logos (Table S4). Second, the FUE and CE occupy wider regions in rice than in Arabidopsis. Since the average length of the rice 3'-UTR is larger than that of Arabidopsis, this wider distribution of poly(A) signals possibly results from the less compact nature of the rice genome (Fig. 2-1). Using the datasets of authentic poly(A) sites from Arabidopsis and rice, we are able to compare the usage of poly(A) signals in two model plants of monocot and dicot, respectively. RSAT results showed that signals from rice are more over-representative (shorter list of good signals with higher Z-scores) than those in Arabidopsis, suggesting that monocot plants tend to require stronger signals to guide the cleavage reaction. Moreover, GC-rich signal elements are found in the rice FUE region (Table 1, Table S3 and S4). This might indicate that monocot plants can use more diverse FUE signals than dicot plants.

By making sequence logos, we identified 12 logos that concisely represent the three *cis*-elements. In the NUE, the logo with the largest percentage of hits is the one associated with AAUAAA. When using the logo to search the dataset, we found that this logo covers about 80% of sequences, whereas use of the single pattern count resulted in finding only 7% of sequences containing AAUAAA. These results suggest that many sequences contain signals similar to AAUAAA. Indeed, the AAUAAA signal can tolerate mutations so well that one- or two-nucleotide alterations may not even affect polyadenylation efficiency significantly (Li and Hunt, 1995; Rothnie et al., 1994). This

clearly contrasts to the polyadenylation signals in mammals where AAUAAA signals can be found in over 50% of the genes (Hu et al., 2005) and much less tolerance to mutations (Wilusz et al., 1989), while only about 7-10% of plant mRNA poly(A) signals possess AAUAAA signals (Fig. 2-2; (Loke et al., 2005). In the FUE region, FUE2, one of elements with the highest percentage of hits, contains a UGUA motif, which was also found to be highly distinctly present in FUE over coding by using another approach. Interestingly, the same UGUA motif has also been implicated in human and yeast poly(A) site recognition by both computational studies and biochemical experiments (Graber et al., 1999a; Venkataraman et al., 2005). In plants, a longer signal, UUUGUA, was previous known to be important for the FUE function (Rothnie et al., 1994). In addition, a GC-rich element in the FUE region of rice (Table 1, FUE7) can be found in human 3'-UTRs too (Hu et al., 2005). Taken together, our data support the notion that there is a commonality of some *cis*-elements among yeast, animal and plant poly(A) signals.

Microheterogeneity, as defined above, is used here to describe a number of poly(A) sites located in a short region of mRNA. Essentially resulting from the disorderly nature of the polyadenylation machinery, microheterogeneity can cause misinterpretation and/or overestimation of the prevalence of APA and, hence, the number of poly(A) sites. Poly(A) sites with a distance of around 30 nt are most likely to be determined by the same set of poly(A) signals since the NUE signals can function in this range. In this report, we therefore set the length of microheterogeneity to be 30 nt and aggregated poly(A) sites within 30 nt of each other as one unique poly(A) site. This step avoids repeat counts of similar poly(A) sites with the likelihood of no significant biological consequence. Excluding the effects of microheterogeneity, then, we found that 50% of rice genes have two or more poly(A) sites. This is very interesting because about 50% of human genes were also found to have alternative poly(A) sites (Ara et al., 2006). Previous studies in Arabidopsis using MPSS technology reported that APA was observed in 25% of genes and occurred in the exons, introns, or 3'-UTRs (Meyers et al., 2004). However, the MPSS approach was unable to detect the exact position of poly(A) sites in

the genes, indicating the potential for underestimating the true extent of APA. On the other hand, our EST-based analysis is able to distinguish the poly(A) sites with highest resolution at the level of individual nucleotide, thus providing a more accurate survey of APA in plants. Overall, the significance of extensive APA in plants is still to be elucidated.

Through poly(A) site mapping of the rice genome, we also found that about 2% of the 55K poly(A) sites are located in the region beyond 3'-UTRs. These account for 3.86% (662 out of 17169 analyzed here) of rice genes using this type of APA to produce transcripts encoding truncated or altered proteins. Moreover, about 50% of these non-conventional poly(A) sites are supported by MPSS evidence. The scope of such extensive APA suggests a widespread role of APA as an important mechanism for plant gene expression regulation. Further study of this mechanism should give rise meaningful insight into this phenomenon in plants.

In animal cells, the difference in 3'-UTR lengths is related, in a degree, to regulation of miRNAs (Bartel, 2004). In plant cells, since most miRNAs target sites located in the coding regions (Jones-Rhoades and Bartel, 2004), variation of 3'-UTR length could be implicated in the regulation of transportation, stability and translation, a hypothesis that remains to be tested. In contrast to the variants within 3'-UTR, the presence of alternative poly(A) sites in the other regions of a gene (e.g., those matching annotated introns or exons) may truncate the open reading frame, producing different types of transcripts and/or protein products. In addition, the question of if such altered transcripts can be targets of miRNA remains to be answered.

Based on our previous work involving poly(A) site prediction in Arabidopsis, we designed a new algorithm for the prediction of poly(A) in rice. This modified version is termed PASS-Rice. Using PASS-Rice, we can find regular and alternative poly(A) sites,

or the ends of genes, and predict unwanted poly(A) sites in transgenes, thus making PASS-Rice a potential useful tool in genome annotation and crop genetic engineering applications. Given the fact that there are some levels of species specificity of the poly(A) signals, as discussed above, each predictive model may need to be modified by using species-unique poly(A) signal features, as is the case when using the Arabidopsis model in rice. The quality of the prediction is similar to the original PASS (Ji et al., 2007b). As the field of bioinformatics advances, one would expect that other modeling techniques become available (e.g., (Cheng et al., 2006). At the same time, adaptation of advanced feature generation, selection, and classification methods to the prediction of poly(A) sites in plants remains a future task. Still, prediction accuracy is not likely to be dramatically improved without significant improvement of characterized poly(A) signals. Such improvement may possibly arise from the availability of data gained from analysis of polyadenylation signals pertinent to subsets of genes involved in different developmental stages, tissue and/or pathway specificities. Although such information has been made available for human genes (Zhang et al., 2005), it is still largely missing in plants due to a lack of large scale collection of poly(A) sites that are associated with these tissue and developmental stage specific samples. Further improvement of the prediction algorithm will doubtlessly enhance our ability to annotation poly(A) sites and currently unknown transcripts.

Figures:

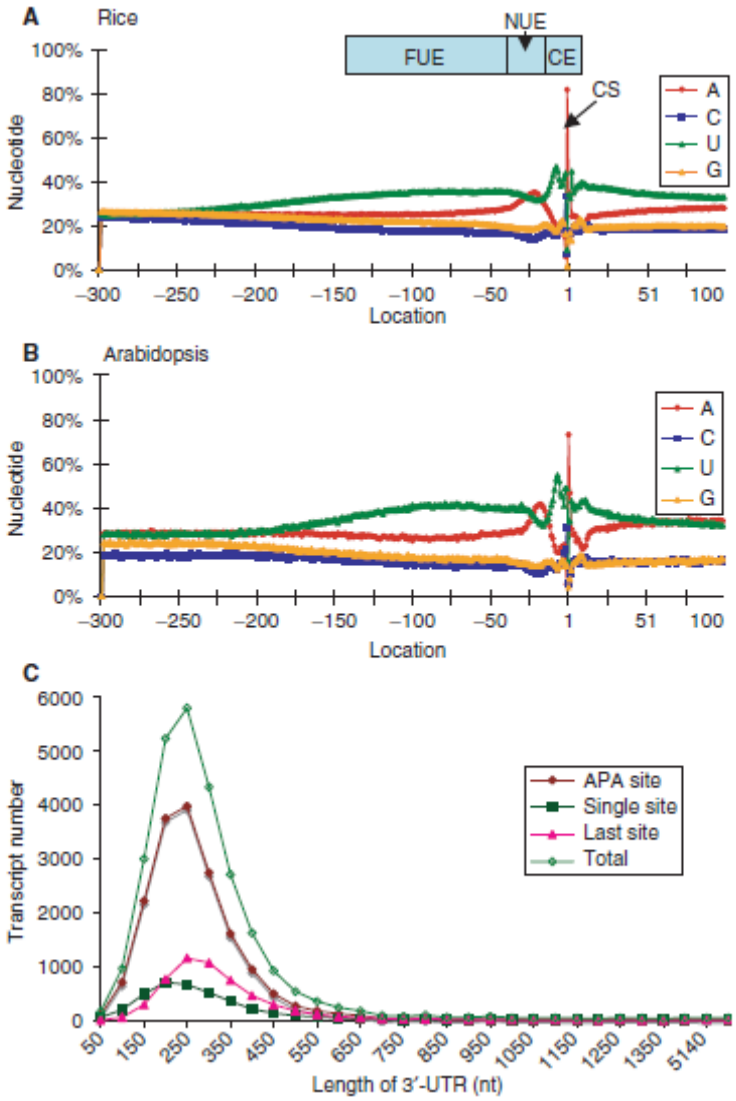



Figure 2-1: Single nucleotide profile comparison and the length of the 3'-UTRs of rice.

(A) One nucleotide profile of the rice 3'-UTR and 100-nt downstream of poly(A) sites. The regions of the poly(A) signals are shown. FUE, far upstream element; NUE, near upstream element; CE, cleavage element; CS, cleavage site or poly(A) site. The poly(A) site is at position -1. The upstream sequence (300-nt) of the poly(A) site is the minus designation, and downstream (100-nt) sequence is the plus designation. (B) One nucleotide profile of Arabidopsis 3'-UTR for comparison purposes. The arrangement is the same as in (A), and the dataset is as described (13). (C) Distribution of the 3'-UTR

lengths in rice. Single sites, transcripts with only one poly(A) site found. APA, sites found in the 3'-UTR with more than one poly(A) site. Last sites, the furthest sites of the APA sites from stop codon. Total, based on all the 3'-UTR lengths. The average length of 289-nt is calculated from the total.

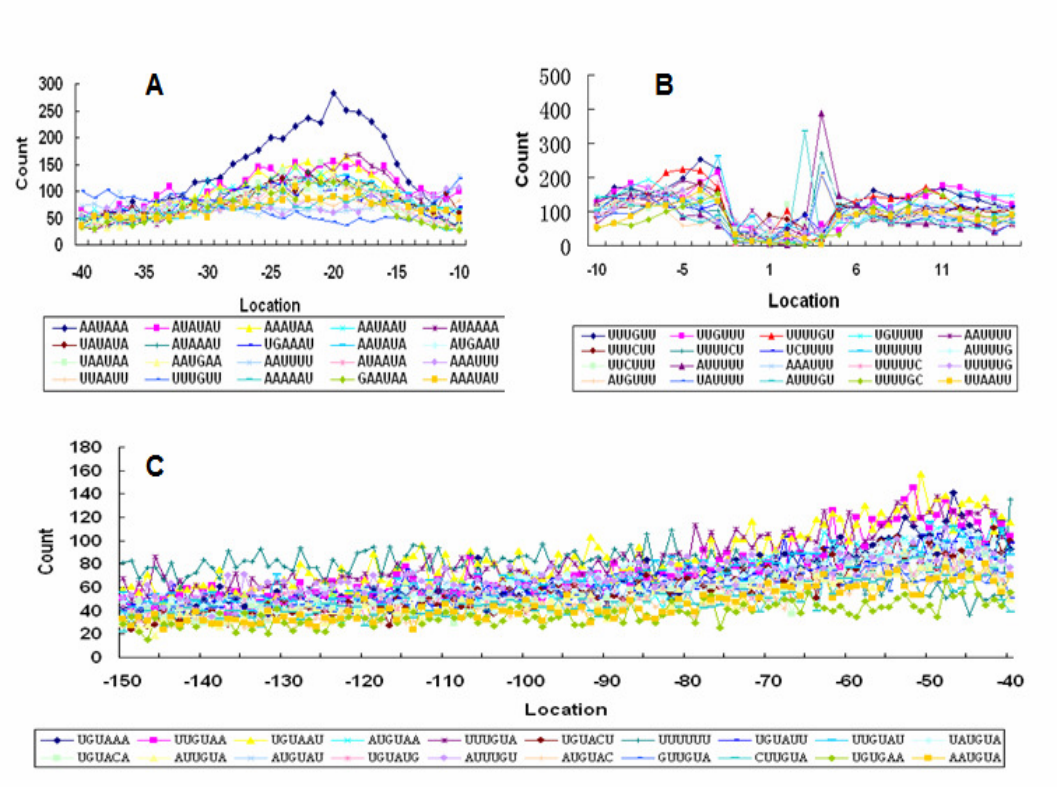



Figure 2-2: Top-ranked hexamers in the rice poly(A) signal elements.

(A) Hexamers from -35 to -10 in the NUE. (B) Hexamers from -10 to +15 in the CE. (C) Hexamers from -200 to -35 in the FUE. See Fig. 1 legend for position annotation.

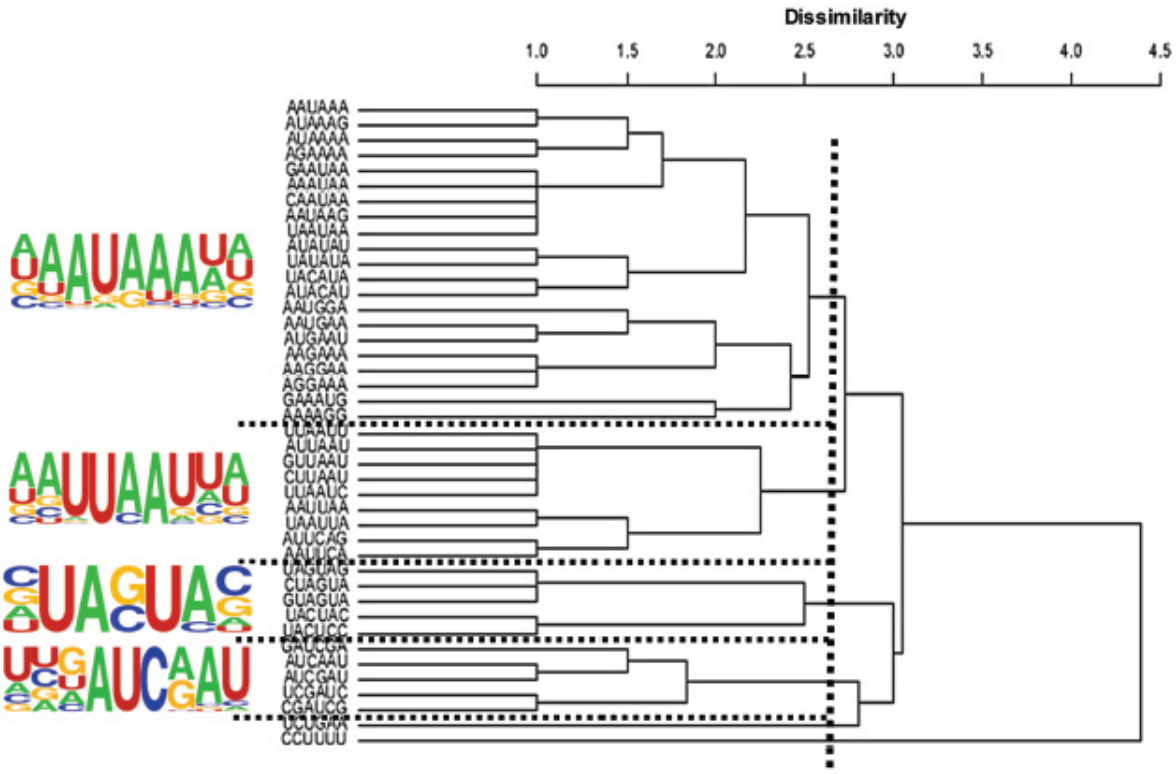



Figure 2-3: An example (NUE) of how sequence logos were constructed.

Dissimilarity distances between signals are calculated by using dynamic programming and then agglomerated by an R program. The suggested cutoff value of 2.6 was used. Hexamers in the same group were further aligned by ClustalW. The logos were generated using Web Logo tool based on ClustalW and their relative frequency in the derived region. The dotted lines indicate grouping regions.

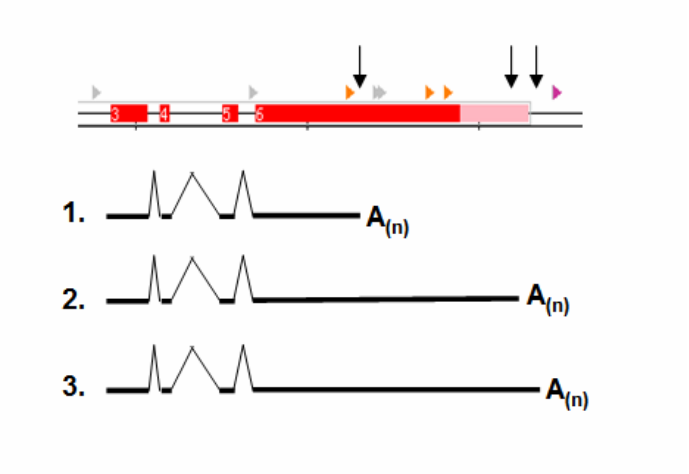



Figure 2-4: An example of APA of WRKY DNA binding domain-containing protein (LOC_Os01g47560) that is supported by rice MPSS data.

The red and pink boxes represent exons and the 3'-UTR, respectively. Vertical arrows show the positions of poly(A) sites. Triangles in orange indicate MPSS signatures inside annotated gene/feature, and the triangle in purple indicates MPSS signatures between genes. The grey triangles are potential MPSS signatures, but not confirmed. The top panel (except for the arrows) was an output from the MPSS-rice web site. The numbers indicate three different transcripts resulting from the use of different poly(A) sites. A(n) indicates a poly(A) tail. The vertical lines indicate splicing of the introns.

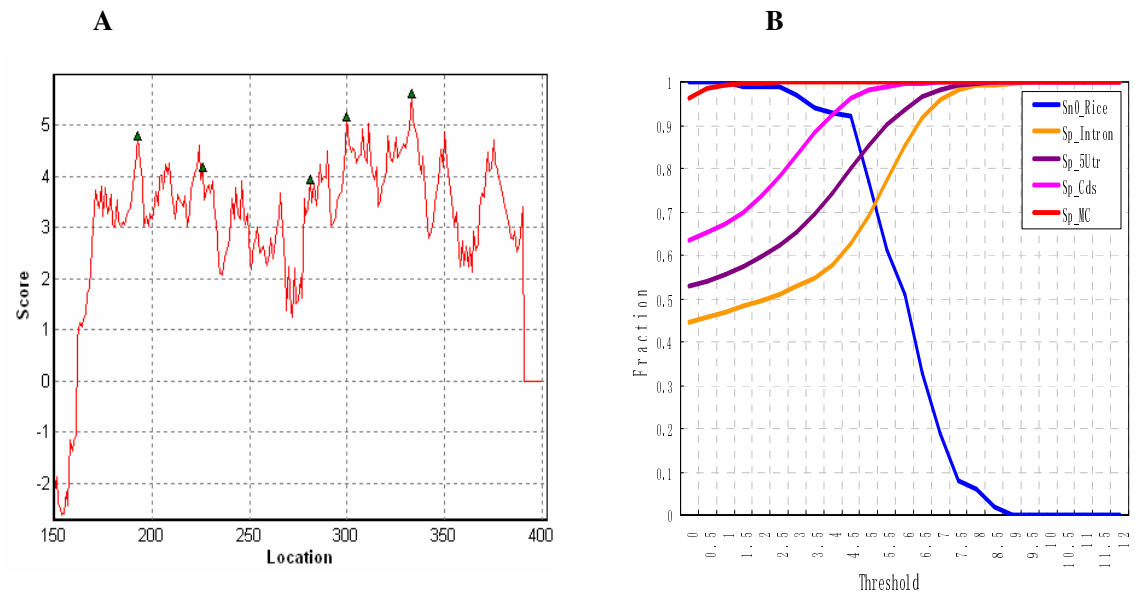



Figure 2-5: Representative outputs and evaluation parameters of PASS-Rice.

(A) The Sn and Sp based on PASS-Rice. The Sp values were calculated based on rice intron, 5'-UTR, coding sequences (CDS) and a random sequence set generated by Markov chain (MC) based on the 2nd order trinucleotide distribution of rice 3'-UTR. (B) An example output of PASS-Rice using Los_03g61890 with multiple poly(A) sites. Triangles indicate the poly(A) sites confirmed by ESTs.

Table 2-1: *Cis*-elements for mRNA polyadenylation in rice.

| Region | Sequence logo | Name | # of hexamer* | Top hexamer | % of hits** |
|---|---|---|---|---|---|
| -150/-36 |  | FUE.1 | 12 | UUAAUU | 93% |
| |  | FUE.2 | 16 | UGUAAA | 99% |
| |  | FUE.3 | 12 | AAUAAA | 80% |
| |  | FUE.4 | 8 | UUCAGA | 94% |
| |  | FUE.5 | 18 | UAGUAG | 99% |
| |  | FUE.6 | 13 | UUCUUU | 99% |
| |  | FUE.7 | 13 | GCGGCG | 98% |
| -35/-10 |  | NUE.1 | 21 | AAUAAA | 80% |
| |  | NUE.2 | 9 | UUAAUU | 53% |
| |  | NUE.3 | 5 | UAGUAG | 10% |
| |  | NUE.4 | 5 | GAUCGA | 23% |
| -10/+15 |  | CE.1 | 5 | UAAUUA | 67% |

Note: * The number of hexamers that were used to produce the logo. **Indicate the percentage of signal patterns the logo can represent in the defined region (FUE, NUE or CE).

Table 2-2: Number of genes with alternative poly(A) sites.

| Number of unique poly(A) site / gene | Number of genes | % |
|---|---|---|
| 1 | 8315 | 49.17 |
| 2 | 4062 | 24.02 |
| 3 | 2240 | 13.25 |
| 4 or more | 2294 | 13.57 |
| Total | 16911 | 100% |

Table 2-3: The locations of poly(A) sites in the rice genome.

| Category | Sub-category | Number of transcripts | % |
|---|---|---|---|
| Aligned to genome | - | 54,786* | 100 |
| Located in the full-length cDNA | Coding sequences | 244 | 0.45 |
| | Introns | 511 | 0.93 |
| | 5'-UTR | 299 | 0.54 |
| | 3'-UTR | 28,209 | 51.45 |
| | Subtotal | 29,263 | 53.41 |
| Located nearby annotated transcript ends | Within 500 nt ** downstream of 3'-end | 19,397 | 35.40 |
| | Within 100 nt upstream of5'-end | 34 | 0.06 |
| | Subtotal | 19,431 | 35.46 |
| Located in the intergenic region | At least 500 nt ** beyond currently annotated genes | 6092 | 11.12 |

* Only those that were mapped to unique genomic sequences are shown.

**For those genes that do not have an annotated 3'-UTR, 1000-nt (instead of 500-nt) downstream from their stop codons was used.

## *References*


Ara, T., F. Lopez, W. Ritchie, P. Benech, and D. Gautheret, 2006, Conservation of alternative polyadenylation patterns in mammalian genes, *BMC Genomics* 7, 189.

Bartel, D.P., 2004, Micrornas: Genomics, biogenesis, mechanism, and function, *Cell* 116, 281-297.

Brenner, S., M. Johnson, J. Bridgham, G. Golda, D.H. Lloyd, D. Johnson, S. Luo, S. McCurdy, M. Foy, M. Ewan, R. Roth, D. George, S. Eletr, G. Albrecht, E. Vermaas, S.R. Williams, K. Moon, T. Burcham, M. Pallas, R.B. DuBridge, J. Kirchner, K. Fearon, J. Mao, and K. Corcoran, 2000, Gene expression analysis by massively parallel signature sequencing (mpss) on microbead arrays, *Natture Biotechnol.* 18, 630-634.

Campbell, M. A. , B. J. Haas, J. P. Hamilton, S. M. Mount, and C. R. Buell, 2006, Comprehensive analysis of alternative splicing in rice and comparative analyses with arabidopsis, *BMC Genomics* 7.

Chen, F., C. C. Macdonald, and J. Wilusz, 1995, Cleavage site determinants in the mammalian polyadenylation signal, *Nucleic Acids Research* 23, 2614-2620.

Cheng, Y. , R. M. Miura, and B. Tian, 2006, Prediction of mRNA polyadenylation sites by support vector machine, *Bioinformatics* 22, 2320-2325.

Chuvpilo, S., M. Zimmer, A. Kerstan, J. Glockner, A. Avots, C. Escher, C. Fischer, I. Inashkina, E. Jankevics, F. Berberich-Siebelt, E. Schmitt, and E. Serfling, 1999, Alternative polyadenylation events contribute to the induction of nf-atc in effector t cells, *Immunity* 10, 261-9.

Crooks, G.E., G. Hon, J.M. Chandonia, and S.E. Brenner, 2004, Weblogo: A sequence logo generator, *Genome Res.* 14, 1188-1190.

Delaney, K. J., R. Q. Xu, J. X. Zhang, Q. Q. Li, K. Y. Yun, D. L. Falcone, and A. G. Hunt, 2006, Calmodulin interacts with and regulates the RNA-binding activity of an arabidopsis polyadenylation factor subunit, *Plant Physiology* 140, 1507-1521.

Dong, H. T., Y. Deng, J. Chen, S. Wang, S. H. Peng, C. Dai, Y. Q. Fang, J. Shao, Y. C. Lou, and D. B. Li, 2007, An exploration of 3 '-end processing signals and their tissue distribution in oryza sativa, *Gene* 389, 107-113.

Gilmartin, G. M., 2005, Eukaryotic mRNA 3' processing: A common means to different ends, *Genes Dev* 19, 2517-21.

Graber, J. H., C. R. Cantor, S. C. Mohr, and T. F. Smith, 1999, Genomic detection of new yeast pre-mRNA 3'-end-processing signals, *Nucleic Acids Res* 27, 888-94.

Graber, J. H., C. R. Cantor, S. C. Mohr, and T. F. Smith, 1999, In silico detection of control signals: mRNA 3'-end-processing sequences in diverse species, *Proc Natl Acad Sci U S A* 96, 14055-60.

Graber, J. H., G. D. McAllister, and T. F. Smith, 2002, Probabilistic prediction of saccharomyces cerevisiae mrna 3'-processing sites, *Nucleic Acids Res* 30, 1851-8.

Hu, J., C. S. Lutz, J. Wilusz, and B. Tian, 2005, Bioinformatic identification of candidate cis-regulatory elements involved in human mrna polyadenylation, *RNA* 11, 1485-93.

Hunt, A. G., 2007, Messenger RNA 3′ -end formation and the regulation of gene expression, in C.L. Bassett, ed.: *Regulation of gene expression in plants: The role of transcript structure and processing* (Springer).

Ji, G. , J. Zheng, Y. Shen, X. Wu, R. Jiang, Y. Lin, J. C. Loke, K. M. Davis, G. J. Reese, and Q. Q. Li, 2007, Predictive modeling of plant messenger RNA polyadenylation sites, *BMC Bioinformatics* 8, 43.

Ji, G., X. Wu, J. Zheng, Y. Shen, and Q. Q. Li, 2007, Modeling plant mrna poly(a) sites: Software design and implementation, *J. Computational and Theoretical Nanoscience* 4, 1365-1368.

Jones-Rhoades, M.W., and D.P. Bartel, 2004, Computational identification of plant micrornas and their targets, including a stress-induced mirna, *Mol Cell.* 14, 787-9.

Kent, W.J., 2002, Blat--the blast-like alignment tool, *Genome Res.* 12, 656-664.

Lee, J. H., Y. S. Cho, H. S. Yoon, M. C. Suh, J. Moon, I. Lee, D. Weigel, C. H. Yun, and J. K. Kim, 2005, Conservation and divergence of fca function between arabidopsis and rice, *Plant Mol Biol* 58, 823-38.

Li, Q. Q., and A. G. Hunt, 1995, A near upstream element in a plant polyadenylation signal consists of more than six bases, *Plant Molecular Biology* 28, 927-934.

Li, Q. S., and A. G. Hunt, 1997, The polyadenylation of rna in plants, *Plant Physiology* 115, 321-325.

Loke, J. C., E. A. Stahlberg, D. G. Strenski, B. J. Haas, P. C. Wood, and Q. Q. Li, 2005, Compilation of mRNA polyadenylation signals in arabidopsis revealed a new signal element and potential secondary structures, *Plant Physiol* 138, 1457-1468.

Lopez, F., S. Granjeaud, T. Ara, B. Ghattas, and D. Gautheret, 2006, The disparate nature of "Intergenic" Polyadenylation sites, *RNA* 12, 1794-1801.

Lu, Y., C-X. Gao, and B. Han, 2006, Sequence analysis of mRNA polyadenylation signals of rice genes, *Chinese Science Bulletin* 51, 1069-1077.

Meyers, B. C., T. H. Vu, S. S. Tej, H. Ghazal, M. Matvienko, V. Agrawal, J. C. Ning, and C. D. Haudenschild, 2004, Analysis of the transcriptional complexity of *Arabidopsis thaliana* by massively parallel signature sequencing, *Nature Biotechnology* 22, 1006-1011.

Moore, C. L., H. Skolnikdavid, and P. A. Sharp, 1986, Analysis of RNA cleavage at the adenovirus-2 l3 polyadenylation site, *EMBO Journal* 5, 1929-1938.

Nobuta, K., R.C. Venu, C. Lu, A. Beló, K. Vemaraju, K. Kulkarni, W. Wang, M. Pillay, P.J. Green, G.L. Wang, and B.C. Meyers, 2007, An expression atlas of rice mRNAs and small RNAs, *Nature Biotechnol.* 25, 473-477.

Peterson, M. L., 2007, Mechanisms controlling production of membrane and secreted immunoglobulin during B cell development, *Immunol Res* 37, 33-46.

Proudfoot, N. J., A. Furger, and M. J. Dye, 2002, Integrating mRNA processing with transcription, *Cell* 108, 501-12.

Proudfoot, N. J., A. Furger, and M. J. Dye, 2002, Integrating rnRNA processing with transcription, *Cell* 108, 501-512.

Quesada, V., C. Dean, and G. G. Simpson, 2005, Regulated RNA processing in the control of arabidopsis flowering, *Int J Dev Biol* 49, 773-80.

Rothnie, H. M., 1996, Plant mrna 3'-end formation, *Plant Mol Biol* 32, 43-61.

Rothnie, H. M., G. Chen, J. Futterer, and T. Hohn, 2001, Polyadenylation in rice tungro bacilliform virus: cis-acting signals and regulation, *J Virol* 75, 4184-94.

Rothnie, H. M., J. Reid, and T. Hohn, 1994, The contribution of AAUAAA and the upstream element UUUGUA to the efficiency of mRNA 3'-end formation in plants, *EMBO J* 13, 2200-10.

Salisbury, J., K. W. Hutchison, and J. H. Graber, 2006, A multispecies comparison of the metazoan 3 '-processing downstream elements and the cstf-64 RNA recognition motif, *BMC Genomics* 7, 55.

Seiler, K.P., G.A. George, M.P. Happ, N.E. Bodycombe, H.A. Carrinski, S. Norton, S. Brudz, J.P. Sullivan, J. Muhlich, M. Serrano, P. Ferraiolo, N.J. Tolliday, S.L. Schreiber, and P.A. Clemons, 2007, Chembank: A small-molecule screening and cheminformatics resource database., *Nucleic Acids Res.* 36,351-9 doi:10.1093/nar/gkm843.

Sheets, M. D., S. C. Ogg, and M. P. Wickens, 1990, Point mutations in AAUAAA and the poly(A) addition site - effects on the accuracy and efficiency of cleavage and polyadenylation invitro, *Nucleic Acids Research* 18, 5799-5805.

Simpson, G. G., P. P. Dijkwel, V. Quesada, I. Henderson, and C. Dean, 2003, Fy is an RNA 3' end-processing factor that interacts with FCA to control the arabidopsis floral transition, *Cell* 113, 777-87.

Tan, X., B.C. Meyers, A. Kozik, M.A. West, M. Morgante, D.A. St Clair, A.F. Bent, and R.W. Michelmore, 2007, Global expression analysis of nucleotide binding site-leucine rich repeat-encoding and related genes in Arabidopsis, *BMC Plant Biol.* 7, 56.

Tang, G. L., X. H. Zhu, B. Gakiere, H. Levanony, A. Kahana, and G. Galili, 2002, The bifunctional lkr/sdh locus of plants also encodes a highly active monofunctional lysine-ketoglutarate reductase using a polyadenylation signal located within an intron, *Plant Physiology* 130, 147-154.

Van Helden, J., M. Del Olmo, and J. E. Perez-Ortin, 2000, Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals, *Nucleic Acids Res* 28, 1000-10.

Venkataraman, K., K. M. Brown, and G. M. Gilmartin, 2005, Analysis of a noncanonical poly(A) site reveals a trinartite mechanism for vertebrate poly(A) site recognition, *Genes & Development* 19, 1315-1327.

Wilusz, J., S. M. Pettine, and T. Shenk, 1989, Functional analysis of point mutations in the AAUAAA motif of the SV40 late polyadenylation signal, *Nucleic Acids Res* 17, 3899-908.

Winichayakul, S., N. L. Beswick, C. Dean, and R. C. Macknight, 2005, Components of the Arabidopsis autonomous floral promotion pathway, fca and fy, are conserved in monocots, *Functional Plant Biology* 32, 345-355.

Xing, D., H. Zhao, R. Xu, and Q. Q. Li, 2008, Arabidopsis PCFS4 regulates flowering time and alternative polyadenylation of FCA., *Plant J* 54, 899-910

Zhang, H., J. Y. Lee, and B. Tian, 2005, Biased alternative polyadenylation in human tissues, *Genome Biol* 6, R100.

Zhao, J., L. Hyman, and C. Moore, 1999, Formation of mRNA 3' ends in eukaryotes: Mechanism, regulation, and interrelationships with other steps in mrna synthesis, *Microbiology and molecular biology reviews* 63, 405-445.

# CHAPTER 3: UNIQUE FEATURES OF NUCLEAR mRNA POLY(A) SIGNALS AND ALTERNATIVE POLYADENYLATION IN *CHLAMYDOMONAS REINHARDTII*

The material in this chapter has been published as: Yingjia Shen, Yuansheng Liu, Lin Liu, Chun Liang and Qingshun Q. Li. **Unique features of nuclear mRNA poly(A) signals and alternative polyadenylation in *Chlamydomonas reinhardtii***. *Genetics*. 2008 179:167-176. Contributions to this chapter: Yuansheng Liu, Lin Liu and Chun Liang generated the initial rice poly(A) data set of *Chlamydomonas*. All of the other data were collected and analyzed by Yingjia Shen. Supplementary data can be downloaded from:

http://www.genetics.org/cgi/content/full/179/1/167/DC1

## *Abstract*

To understand nuclear mRNA polyadenylation mechanisms in the model alga *Chlamydomonas reinhardtii*, we generated a dataset of 16,952 *in-silico* verified poly(A) sites from EST sequencing traces based on *Chlamydomonas* Genome Assembly v.3.1. Analysis of this dataset revealed a unique and complex polyadenylation signal profile that is setting *Chlamydomonas* apart from other organisms. In contrast to the high-AU content in 3'-UTR of other organisms, *Chlamydomonas* shows a high-guanylate content that transits to high-cytidylate around the poly(A) site. The average length of 3'-UTR is 595-nt, significantly longer than that of Arabidopsis and rice. The dominant poly(A) signal, UGUAA, was found in the Near-Upstream-Elements, and its occurrence may be positively correlated with higher gene expression levels. The UGUAA signal also exists in Arabidopsis and in some mammalian genes but mainly in the Far-Upstream-Elements, suggesting a shift in function. The C-rich region after poly(A) sites with unique signal elements is a characteristic downstream element that is lacking in higher plants. We also found high-level of alternative polyadenylation in the *Chlamydomonas* genome, with a range up to 33% genes having at least two unique poly(A) sites, and a range of 1% to 11% (depending upon the stringency of the criteria) of the genes having poly(A) sites residing in predicted coding sequences, introns, and 5'-UTR. These potentially contribute to transcriptome diversity and gene expression regulation.

As an essential post-transcriptional processing step in eukaryotic nuclear gene expression, messenger RNA (mRNA) 3'-end formation, which includes cleavage and polyadenylation, is tightly integrated with pre-mRNA capping, splicing, and transcription termination (Proudfoot, 2004). After being transcribed, pre-mRNA is cleaved at the poly(A) site to generate a new 3'-end where a poly(A) tail is then added. The functions of the poly(A) tail include protection of mature mRNA from unregulated degradation, recognition by mRNA cytoplasmic export machinery, and recognition by translational apparatus as a intact mRNA (Zhao et al., 1999). Since the mRNA becomes functional only with the correct configuration of 3'-ends, the process of 3'-end formation of mRNA is a crucial step in gene expression regulation. That is, correct configurations imply right location of the cleavage site on the pre-mRNA and an adequate length of poly(A) tail (Zhao et al., 1999). Since a poly(A) site marks the end of a transcript, alternative poly(A) locations may truncate or elongate the mRNA, possibly resulting in additional regulation by excluding or including *cis*-elements, or different protein products.

Appreciable understanding of the mRNA 3'-end formation process in animals, yeast and plants has been reached. Both cleavage and polyadenylation reactions require the pre-mRNA to have a set of *cis*-elements known as polyadenylation signals that are recognized by a group of polyadenylation factors (Gilmartin, 2005; Hunt, 2007; Zhao et al., 1999). The designated location of a poly(A) site on mature mRNA indicates that this process requires specific poly(A) signals on the mRNA to direct the process. While unique mRNA poly(A) signals exist in different domains of eukaryotes, a unifying theme has just emerged after confirmation of the existence of up-stream *cis*-elements in some mammalian genes (Gilmartin, 2005). In general, poly(A) signals can be divided into four different groups: the cleavage or poly(A) site and its surrounding sequences called cleavage element (or CE); a strong signal (e.g., AAUAAA found in mammals) about 20-30 nucleotides (nt) upstream from the poly(A) site, which is termed near upstream element (or NUE); a far upstream element (FUE) that is about 40 to150 nt upstream of the poly(A) site; and a downstream element located 20 to 40 nt beyond the cleavage site. In mammalian cells, these four *cis*-elements are all required with the highly conserved

NUE in the form of AAUAAA and seemingly weaker FUE (Venkataraman et al., 2005). The downstream elements are typically only found in mammals (Zhao et al., 1999). In contrast, yeast and plant pre-mRNA do not have downstream elements, and the other three elements are much less conserved in yeast and plants (Graber et al., 1999; Li and Hunt, 1997; Loke et al., 2005).

The polyadenylation signals of nuclear genes in algae are largely uncharacterized. In contrast to polyadenylation of chloroplast encoded genes, which use a very different system where poly(A) tails promote mRNA degradation, polyadenylation of algal nuclear genes are protecting mRNA and it is required for mRNA functions (Slomovic et al., 2006). Early research suggested that UGUAA could be the polyadenylation signal for *Chlamydomonas* (Silflow et al., 1985). While not confirmed by mutagenesis, the 3'-UTR containing this signal has been successfully used for expressing transgenes in *Chlamydomonas* (Berthold et al., 2002). It was also reported that the poly(A) signals in algae are different from those of higher plants and other eukaryotes (Wodniok et al., 2007), but the analysis was incomplete because genomic sequences were not available to extract information beyond the point of poly(A) sites. The recently finished *Chlamydomonas* genome offers an excellent system to examine poly(A) signals and their potential roles in gene expression regulation. It has been reported that the genome of *Chlamydomonas* has mixed features of both plants and animals in that the genome structure and gene families may have evolved in a pathway that is different from both plant and animal lineages (Merchant et al., 2007). Indeed, such intermediate genome structure may be the result of its peculiar cellular structure and "habitat" where a free-living mobile photosynthetic system can sustain unique challenges. These extraordinary features have prompted us to utilize this model to examine the extent of deviation between the mRNA processing event in animals and plants.

Alternative polyadenylation (APA) is a powerful pathway for gene expression regulation. There are many classical examples of APA where the use of an alternate poly(A) site results in the production of two or more different proteins (Cote et al., 1992; Delaney et al., 2006; Lou and Gagel, 1998; Peterson, 1994), or the production of a

non-functional variant of a functional one to regulate gene expression (Simpson et al., 2003). About half of human genes and an estimated 25% of Arabidopsis genes are subject to APA (Meyers et al., 2004; Zhang et al., 2005). Clearly, APA, in many cases with alternative splicing (Zhang et al., 2005), is an integral component of eukaryotic gene expression regulation. To gain initial understanding of such a gene expression regulation pathway in algae, it would be of interest to explore APA in algae like *Chlamydomonas*. Given the ample collection of the ESTs in *Chlamydomonas*, we have collected over 22,000 poly(A) sites in its draft genome. These data were obtained by processing raw EST sequencing trace files using a novel bioinformatics protocol that focuses on detecting in-silico verified cDNA termini or ends including poly(A) sites (Liang et al., 2007b; Liang et al., 2007c). Using this large dataset, we revealed that *Chlamydomonas* possesses unique features in polyadenylation signals, including nucleotide composition of 3'-UTR, signal arrangements and sequence patterns, and extensive APA. In terms of mRNA polyadenylation, then, it is this unique set of characteristics that distinguishes *Chlamydomonas* from other systems studied to date.

## *Materials and methods*

### The poly(A) site dataset

A total of 309,278 raw EST traces were obtained from the NCBI Trace Archive (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi), *Chlamydomonas* Resource Center (www.chlamy.org) and the Kazusa DNA Research Institute (http://est.kazusa.or.jp/en/plant/chlamy/EST/index.html). The raw trace files were processed as described (Liang et al., 2008; Liang et al., 2007b; Liang et al., 2007c), and the poly(A) sites were authenticated based on the features of each cDNA library construct. The *Chlamydomonas* Assembly v.3.1 genome sequences and corresponding annotation file were downloaded from the Joint Genome Institute of US Department of Energy (http://genome.jgi-psf.org/Chlre3/Chlre3.download.ftp.html). *In-silico* authenticated poly(A) tails in ESTs were defined as oligo(A) tracts that have a minimum length of 10-nt, allowing for a 2-nt error (i.e. mismatch, insertion or deletion) and were extensible. For every five adenine extension, we then allowed for one more nt error. In addition, the poly(A) tails found in the ends of ESTs must also be immediately followed by an XhoI restriction enzyme site (CTCGAG) and then by an extensible vector fragment (GGGGGGCCC…) that matched the expected structure in the relevant cDNA libraries. All raw sequences were mapped to the draft genome of *Chlamydomonas* Assembly v.3.1 DNA (http://genome.jgi-psf.org/Chlre3/Chlre3.download.ftp.html) by GMAP (Wu and Watanabe, 2005) to make sure that poly(A) tails were not from the genome sequences, thus eliminating internal priming contaminations. For ESTs with a valid genome mapping, the mapped length should be at least 70 nt with a minimum of 80% identity, and the matched coverage of the final clean portion of a raw EST sequence must be ≥ 80%. The poly(A) site is defined as the last nucleotide that matched to the genome sequence. In case an adenine was also found at a poly(A) site in the genome sequence, this adenine (right next to one of three other nucleotides, G, T, or C) was saved as a poly(A) site. This is because biochemical evidence indicate that the first A of a poly(A)

tail tends to be from transcription rather than added by poly(A) polymerase during polyadenylation (Chen et al., 1995; Moore et al., 1986; Sheets et al., 1990). Once a poly(A) site was identified, 300 nt sequence upstream plus 100 nt sequence downstream for each authenticated poly(A) site were extracted. This produced a dataset of 56,031 sequences, each with a poly(A) site, and this dataset became known as the 62K dataset. There were some redundant ESTs in the 62K dataset because this dataset reflects all the ESTs that were successfully mapped. When redundant ESTs with the same poly(A) sites in the genome were removed, a dataset totaling 16,952 unique sequences was generated and called the 17K dataset, which was used in most of the analyses presented here. These datasets are available from our web site (www.polyA.org).

To further study the locations of poly(A) sites in the genes, we mapped all ESTs to the annotated genome (v3.1) based upon GMAP results. 44,338 ESTs were found to be associated with currently annotated genes and corresponding to 11,730 non-redundant poly(A) sites. We further categorized these poly(A) sites based on their location on the genes into 4 groups [5' and 3' UTRs, coding sequences (CDS), and introns]. We also tested all poly(A) sites in the 5'UTR and CDS and some sites in intron by manually examining them through *Chlamydomonas* EST terminus database (http://www.conifergdb.org/chlamyest/; Liang et al. 2008)

**Analysis of poly(A) signals**

We previously used a program called SignalSleuth to find poly(A) signal patterns in connection with our studies on Arabidopsis (Loke et al., 2005). This program was also used to perform an exhaustive search of varying size patterns within a subregion of a large set of *Chlamydomonas* sequences. The program starts at the user-defined start nucleotide position of the first sequence and records the sequence pattern from this position onward before moving to the next nucleotide. This process continues until it reaches the end of the subregion defined by the user. The program then repeats this process for all the input sequences and generates a matrix file containing the occurrence of each designated length (3 to 12 nt) of sequence patterns with their location information

for the 17K dataset. The signal patterns were ranked on the basis of the frequency of their occurrence over the background, and such results were used for further analysis.

The other two poly(A) datasets from Arabidopsis (Loke et al., 2005) and human (provided by Dr. Bin Tian; (Tian et al., 2005) were used for comparison purposes.

To find the statistically significant signals in these polyadenylation *cis*-elements, we employed an oligo-analysis program called Regulatory Sequence Analysis Tools (or RSAT; http://rsat.ulb.ac.be/rsat/;(van Helden, 2003). Based on the Markov Chain model, this program uses the comparison of expected frequency of the particular sequence pattern on the region under study to the observed frequency. A standard score (so-called Z-score) is used to reveal the standard deviation of each pattern to its expected occurrence, also based on Markov Chain models (van Helden, 2003). The results of the calculation are presented as Z-scores and ranked according the statistical significance of each signal pattern.

**Construction of signal logos**

To generate sequence logos that could represent many signal variants, we adopted a method as described (Hu et al., 2005) to compile the signals and calculate the percentage of hits for each logo. With a dynamic programming method, we grouped the highly ranked sequence patterns (with z-score $\geq$8.53, except for FUE where z-score $\geq$5 is used) based on their mutual distance in which gap was not allowed. Then, an agglomeration package from program R (www.r-project.org) called Agnes was used to cluster patterns based on their dissimilarity distances. A cutoff value of 2.6 was used to group these patterns, as suggested (Hu et al., 2005), and they were aligned by using ClustalW. The size of a sequence logo was determined based on the ClustalW alignment results, and the openings at both ends in the aligned sequences were filled by nucleotides selected on the basis of the percentage of each nucleotide from the background sequence in the studied region. Each sequence pattern in the group was weighted based on its occurrence in each studied region, and the Web Logo Tool (Crooks et al., 2004) was used to generate the final images of sequence logos.

In order to evaluate if a sequence logo is represented in the studied region, we generated a position-specific scoring matrix for each logo (Hu et al., 2005). For each position, the score S was calculated as follows: $S=\sum_{p=1}^{L}\log_2\frac{f(n,p)}{f(n)}$, where L is the length of sequence logo, f(n, p) is the frequency of nucleotide n at the position p of the sequence logo, and f(n) is the background frequency of occurrence of nucleotide n in a specific poly(A) signal region, e.g., NUE.

Analysis of alternative polyadenylation sites

Positions of poly(A) sites detected by GMAP alignment were further marked on the annotated genome map using a Perl script. In order to avoid microheterogeneity, poly(A) sites in the same gene must have at least a 30 nt interval to be considered as unique alternative polyadenylation sites. This number was chosen based on the assumption that the same NUE could control more than one poly(A) site in the range of about 30 nt (Li and Hunt, 1997; Shen et al., 2008).

**Analysis of the size of *cis*-elements**

Since true signals should deviate from the background more than non-signals, the nucleotide sequence length of the *cis*-elements was justified by the degree of deviation of a particular signal size from the background. The degree of bias towards a certain size of *cis*-elements was calculated based on the difference between observed occurrence and expected value. The predicted values were calculated based on the fact that the increment of the pattern size for any given signal will be one half of the chance of its original occurrence if no bias occurs. For example, if a 3-mer (3 nt) signal appears 1000 times in a specific region of the sequences, then the 4-mer should be 500 based on the 1/4 chances of the nucleotides being incorporated onto either end of the 3-mer pattern. The difference between the predicted and observed is calculated using this formula: Score=([ΣObsn+1 [(A/T/C/G,A/T/C/G) – Obsn/2)]/ Obsn X 100%, where Obsn is the observed value of the first pattern size, Obsn+1 is the observed occurrence of the next successive size pattern and (A/T/C/G, A/T/C/G) is the sum of the occurrences of any

nucleotides incorporated to the ends of the pattern for Obsn+1. The results of this analysis are presented in the Supplemental Figure 1 (at http://www.genetics.org/supplemental/). The size with the highest score was used for SignalSleuth and RSAT analyses.

## *Results*

### The poly(A) site and 3'-UTR dataset of *Chlamydomonas*

Taking advantage of the recently published draft *Chlamydomonas* genome (Merchant et al., 2007), we mapped individual ESTs to the genome. Since a poly(A) tail is added post-transcriptionally, the nucleotide before a poly(A) stretch that also matches to the genome sequence can be defined as a poly(A) site. Because the libraries were constructed using oligo(dT) with a linker at its 5'-end, an authentic poly(A) tail should be found between this linker and a valid EST that can be mapped to the genome sequence. Moreover, the raw trace sequences should also include a part of the vector sequence right next to the linker, because primers for sequencing are generally match vector sequences. The presence of such a sequence and the linker were both used to verify the existence of the poly(A) site (Liang et al.,2008). Based on these in silico authenticated poly(A) sites, a dataset with 16,952 entries of 400 nt each was generated, where each sequence has a poly(A) site located at nucleotide 300 (from left right; the poly(A) site nucleotide is referred to as -1 position hereafter). This dataset, termed the 17K dataset, represents the largest poly(A) site collection in algae to date.

### The profile of 3'-UTR of transcripts in *Chlamydomonas*

Using SignalSleuth, an exhaustive pattern search algorithm (Loke et al., 2005), we first examined the single nucleotide profile around poly(A) sites of all sequences in the dataset. As shown in Fig. 3-1, the 3'-UTR of *Chlamydomonas* is notably rich in G nucleotide, except the -25 to -5 region where U and A are dominant, while the downstream +5 to +30

region has a high C content but the transition to high-C starts before the poly(A) site. This profile is distinctly different from the profiles of two land plant species, Arabidopsis (Figure 3-1B; (Loke et al., 2005) and rice (Dong et al., 2007; Shen et al., under revision), as well as yeast (Graber et al., 1999) and human (Figure 3-1C;(Tian et al., 2005), which are all AU rich in their 3'-UTR. It is known that the *Chlamydomonas* genome is uniquely GC-rich (64%; Merchant et al.2007), which would contribute, in part, to the G-richness in 3'-UTR. The previously known YA dinucleotide (Y= C or U) at the cleavage site (Loke et al. 2005; Graber et al.1999; Tian et al. 2005) is also missing in *Chlamydomonas* with only the A nucleotide showing at the cleavage site.

The average length of 3'-UTR in *Chlamydomonas* is 595 nt, which is calculated based on distance between annotated (JGI draft gene catalog) stop codon and authenticated poly(A) sites. This average length is more than double that of Arabidopsis and rice (223 and 289 nt, respectively; (Shen et al., under revision), as shown in Figure 3-2. The longer 3'-UTR may also reflect its less compact genome, the size of which (120 megabases) is similar to Arabidopsis, but predicted to encode only half the number (~15,000) of genes (Arabidopsis, 2000; Merchant et al., 2007).

**Nuclear polyadenylation signal regions in *Chlamydomonas***

Based on the scanning results of SignalSleuth, we plotted top-ranked signal profiles in each section of the poly(A) signal regions (Fig. 3-3). The full list of signals is in Supplemental Table 1 (at http://www.genetics.org/supplemental/). Considering the nucleotide composition and signal profiles, the locations (relative to the cleavage site, the -1 position) of *Chlamydomonas* poly(A) signal elements are defined as follows: -150 to -25 for FUE; -25 ~ -5 for NUE, and -5 ~ +5 for CE. One of the most noticeable features of *Chlamydomonas* polyadenylation signals is its NUE where UGUAA is over-represented (see below for more analysis). Also distinguishing *Chlamydomonas* from all other species studied are its FUE and CE. Located in the G-rich region, the predominant signals in FUE are apparently G-rich. As noted above, at the cleavage site, the YA dinucleotide is replaced by the A nucleotide. The distinct C-rich element,

located from +5 to +30, is termed the Downstream Element (DE). Such an element is unique to *Chlamydomonas* because it is different from the downstream element of animals, which is GU-rich (Gilmartin, 2005). In yeast and Arabidopsis, there is no clearly defined DE; rather, they have U-rich regions close to poly(A) sites (Graber et al., 1999; Loke et al., 2005).

In contrast to other species where hexamers are widely used as poly(A) signals, *Chlamydomonas* seems to use signals with more diverse lengths. We found that pentamers are dominant (deviating mostly from background signals) in FUE and NUE regions, while heptamers and hexamers are enriched in CE and DE, respectively (Supplemental Figure 1). For individual signal regions, pentamers have the highest score from the regions of -150 to -25 and -25 to -5, corresponding to FUE and NUE, respectively. Heptamers and hexamers have the biggest deviation in the region of -5 to +5 and +5 to +30 for CE and DE, respectively.

**Statistical analysis of *Chlamydomonas* poly(A) signal patterns**

To search for statistically significant poly(A) signal patterns from the general analysis above, we adopted an oligo analyzer called Regulatory Sequence Analysis Tools, or RSAT (van Helden, 2003). The full results of this analysis are listed in Supplemental Table 2 (at http://www.genetics.org/supplemental/). This online tool uses a standard score (the Z-score) to measure standard deviation of each pattern from its expected occurrence based on Markov Chain models (van Helden, 2003). Poly(A) signals with higher Z-scores are likely to be more significant in determining the position of poly(A) sites. The UGUAA signal has a very high Z-score and most occurrence, and this finding is supported by the SignalSleuth results in which UGUAA is also highly over-represented in the NUE (Figure 3-3). Compared to the predominant AAUAAA signal, which occurs in only about 10% of genes in Arabidopsis (Loke et al., 2005), *Chlamydomonas* genes use highly conserved UGUAA poly(A) signals with about 52% of transcripts in their NUE regions. UGUAA as a NUE signal was consistently ranked on the top of the list either by SignalSleuth or RAST. This is supported by previous

finding where UGUAA was regarded as the best poly(A) signal in *Chlamydomonas* (Wodniok et al. 2007). Two other signals were also found to have higher z-scores (AGUAC and UGCAA, Supplemental Table 2, NUE). However, due to extreme low occurrence (1/46 of UGUAA), AGUAC is very unlikely to be a realistic signal. The low occurrence of UGCAA (1/7 of UGUAA) could be the second best signal according to RSAT ranking. To assess the relationship of these two NUE signals, we examined the exclusiveness of their appearance in the dataset. Interestingly, for those NUEs that do not have UGUAA, 13.8% of them have UGCAA. This is in contrast to those sequences that use UGUAA, in which only 2.2% of sequences use UGCAA. In any case, this informatics analysis should be further confirmed by mutagenesis studies.

Since UGUAA is so conserved, we asked whether genes with higher expression levels tend to use UGUAA as their NUE signals. To test this, we classified different levels of EST redundancy (reflecting expression levels) and then examined the occurrence of UGUAA in each category using the dataset with 56,031 ESTs. As shown in Figure 3-4A, along with increase of EST copy number, the percentage of UGUAA found in these transcripts is also increased. This result suggests that the UGUAA, as a strong poly(A) signal, may be preferentially used by those genes with higher expression levels to facilitate RNA processing.

In order to find the relationship between different species in terms of poly(A) signal usage, we plotted the distributions of UGUAA and AAUAAA signals in *Chlamydomonas*, Arabidopsis, and human. Interestingly, while UGUAA and AAUAAA are predominant in the NUE (-30 to -10) in *Chlamydomonas* and human, respectively, these two signals are mutually exclusive (Figure 3-5). In contrast, AAUAAA is also dominant in NUE of Arabidopsis, while UGUAA occurs most frequently in the FUE region. This result suggests that AAUAAA may have evolved to be dominant in higher eukaryotes, while, conversely, the UGUAA signal might have shifted to upstream and thus assume a lesser role (Fig. 3-5).

To further compile and produce visual appreciation of the poly(A) signals in a concise format, we used a logo program (Hu et al., 2005) to make sequence logos of *Chlamydomonas* poly(A) signals. The primary advantage of such sequence logos is that

each logo represents multiple poly(A) signals corresponding to their occurrences. This reduces the number of signal patterns and, at the same time, ensures that potentially overlapping signals, such as UGUAAC and AUGUAA, are concisely presented. The top signals were those that have a Z-score higher than 8.53 (except 5.0 for the FUE because of its lower Z-score), a suggested cutoff for standard hit determination ($p<0.0001$; as described by (Seiler et al., 2007). These sequence patterns were clustered to generate sequence logos according to their similarity (Shen et al. 2008). Using this method, we identified 6 major signal clusters for all four polyadenylation signal elements. Their sequence logos, the number of clustered signals, the top signals with the high Z-scores and the frequency of occurrence in specific regions are all listed in Table 1. Two FUE signal elements are represented in most genes considered in this study, but they distribute in a board region of 3'-UTR (120-150 nt). For elements in the relative short region, one of the two NUE logos is the most conserved with over 78% of the genes containing UGUAA related element. Signals in CE and DE are less conserved, as each of their logos only cover a small percentage of sequences, but might play an auxiliary role in determining the position of poly(A) sites.

**Alternative polyadenylation in *Chlamydomonas***

APA is an important mechanism in generating diversities of transcriptome and proteome and contributes to gene expression regulation. In order to study the extent of APA in *Chlamydomonas*, we first studied the number of poly(A) sites for each gene based on the 17K dataset (Table 2). In order to avoid microheterogeneity, poly(A) sites in the same gene must have at least a 30 nt interval to be considered as unique APA sites. After excluding microheterogeneity, we found that over 33% (1341 of 4057) of the *Chlamydomonas* genes have 2 or more poly(A) sites. This estimation of APA is based on the currently annotated genes that have at least one poly(A) site at their 3'-UTRs authenticated by our 17K dataset (Table 2). A conservative estimate would be 9% (1341 divide by total predicted 15,000 genes, if no more APA is found in the rest of the

genes).

To further study the positions of these alternative poly(A) sites on the genes, we compared the locations of our authenticated poly(A) sites and annotated start and stop codons, coding sequences and intron boundaries of the draft *Chlamydomonas* Genome Assembly V.3.1. Because there are many genes that do not have annotated 3'-UTR sequences, which could result in possible inaccuracy, we extended the range of these genes by 1000 nt beyond their stop codons (meaning that if a poly(A) site is located within the range, it will be consider a site of this gene). Such a range was extended to 500 nt for those genes that have an annotated 3'-UTR. Our procedure makes sure that each PA site was on the same strand as the model to which it was attributed. After these operation, 44,338 ESTs corresponding to 11,730 non-redundant poly(A) sites were found to be associated with currently annotated genic regions (Table 3) and the majority of these sites (65%) are within 3'-UTRs, as expected. We attributed each PA site either to the gene model in which it lies, or to the immediate upstream model if it is less than 1000 nt away (500 nt if it had already a predicted 3'-UTR).

To our surprise, however, 719 (4.2%) of the poly(A) sites are located in the coding sequences (CDS), introns, or 5'-UTRs of 444 genes (10.9% of 4057 genes in Table 2) where no conventional poly(A) site should be located (Supplemental Table 3 at http://www.genetics.org/supplemental/). It is realized though that such a conclusion (extensive APA in the CDS, introns and 5'-UTRs) is drawn based on the current draft genome annotation information, from which discrepancy between those gene models and EST have been noted (Liang et al. 2008). To reach an accurate count, we used two more stringent conditions, one is that APAs in CDS, introns, and 5'-UTRs must also have another poly(A) site in the gene's 3'-UTR; the other is that the CDS, introns, and 5'-UTRs where APAs are found must be supported by at least another independent EST that validates the exon-intron junction (to prove that the APAs are within a gene's boundary, not in another gene's 3'-UTR). This exercise gave 140 poly(A) sites, and careful manual examination resulted in a total of 63 [0.53% of total poly(A) sites 11,730 mapped on transcripts, Table 3] highly confident and unique APA sites in the CDS, introns and 5'-UTRs that satisfy either conditions (listed in Supplemental Table 3). If

the number of the genes is considered, this represents 44 or about 1% of the 4057 genes with poly(A) data supports (Table 2).

Many of the poly(A) sites (30.8%) are found more than 500 nt from a currently annotated gene, indicating the extent of potential transcripts that could be produced in the currently unannotated region of the genome. Beyond transcripts of unknown genes, such transcripts could be from different sources, e.g., antisense transcripts or small RNA, among other possibilities. Our dataset offers a rich resource for such explorations.

## *Discussion*

In this paper, we were able to process and authenticate 16,952 poly(A) sites for the analysis of poly(A) signals and the extent of alternative polyadenylation in the *Chlamydomonas* model system. In doing so, we demonstrated polyadenylation features distinctive to this alga and the significance of those features in comparison to other species, both plants and animals. In addition to the unique characteristics of poly(A) signals in *Chlamydomonas*, our data clearly indicated that APA is extensive in *Chlamydomonas*, up to a third of the genes having at least two poly(A) sites. In addition, a significant amount of polyadenylated transcripts was found in the CDS, introns and 5'UTR, which could contribute to transcriptome, and hence, proteome diversity in this alga.

The characteristics of the poly(A) signal regions in *Chlamydomonas* are unique and differ from previous working models of mammals, yeast and plants (Graber et al., 1999; Loke et al., 2005; Tian et al., 2005). We found a unique poly(A) signal, UGUAA, which occurs in the NUE region of half the *Chlamydomonas* genes and which is equivalent to the AAUAAA signal found in the NUEs of half the mammalian genes (Tian et al., 2005). In stark contrast, there was barely a trace of the AAUAAA signal found in NUE of *Chlamydomonas*. However, in Arabidopsis, UGUAA is found in a different region, namely the FUE (Fig. 3-5), indicative a translocation of the signal. In mammalian poly(A) signals, it was demonstrated that UGUAA is an important poly(A)

signal, particularly for those transcripts that do not have AAUAAA (Venkataraman et al., 2005). For those human genes that do have the AAUAAA signal or its one-nucleotide variants (~90% of the genes; Zhang et al., 2005), the use of UGUAA signal seem to be minimal (Fig. 3-5). While the most conserved NUE signal for Arabidopsis and rice, AAUAAA, was found in approximately 10% and 7% of all tested genes, respectively (Loke et al., 2005; Shen et al., 2008), the UGUAA signal becomes dominant in FUE of these two species. All these data support our notion that UGUAA may be replaced by AAUAAA in higher eukaryotes, particularly in mammals. However, since UGUAA is still abundant in the FUE region of rice and Arabidopsis, it is possible that AAUAAA might have been a gain in higher plant species during evolution, while UGUAA signals in the FUE of plants might be a remnant from their algal ancestors. Interestingly, a recent study on the evolution of algal poly(A) signals suggested that UGUAA was invented in green algae but was not kept through evolution into land plants (Wodniok et al., 2007). Our data, on the other hand, offer an alternative explanation in which UGUAA, albeit a poly(A) signal, was relocated to other part of the 3'-UTR (FUE) to assist the polyadenylation process. Given the strength of the UGUAA signal (higher conservation level) in *Chlamydomonas*, it is puzzling why it is moved to the weaker position (FUE) in land plants. The primitive AAUAAA found in Streptophyta (Wodniok et al., 2007) never caught up the efficacy of UGUAA. This is because AAUAAA, while is the best signal (Li and Hunt 1995), still has not been adopted by more than 12% of the plants genes, at least in Arabidopsis and rice (Loke et al. 2005; Shen et al. 2008).

Several GC-rich signal elements were found in the *Chlamydomonas* FUE, CE and DE regions and might play an auxiliary role in determining the position of poly(A) sites. This could be an extension of what was seen in the *Chlamydomonas* genome where high GC-content was observed (64%; Merchant et al., 2007). The discrete G-rich FUE and C-rich DE are another set of signatures for *Chlamydomonas* poly(A) signals. Interestingly, however, the more broadly recognized YA, particularly CA signature, at the cleavage sites of plants, yeast and animals is no longer found at the cleavage site of

*Chlamydomonas*. Instead, only the A nucleotide remains predominant. This is directly contradictory to the G- and C-richness of the surrounding region, both before and after the cleavage site.

By making sequence logos, we identified 7 logos that concisely represent the four polyadenylation *cis*-elements in *Chlamydomonas*. In the NUE, the logo with the largest percentage of hits is the one associated with NUE-1 (UUGUAA; Table 1). When using the logo to search the dataset, we found that this logo covers about 78% of sequences, whereas the use of single pattern UGUAA resulted in covering only 52% of sequences. For FUE, both logos (Table 1) have a very high percentage of hits among genes. However, the SignalSleuth scan does not show a sharp spike for any signal (Figure 3-3A). One explanation is that FUE elements spread over a relative long region of 3'-UTR, whereas NUE signals are the determinant of the exact position of cleavage and polyadenylation so it carries stronger positional information. A few GC-rich elements are found in CE and DE regions. Although the presence of these elements is not likely to be the result of random chance based on their high Z-scores, they only account for a small fraction of genes, indicating that these are less conserved signals. These elements might play an auxiliary role in recruiting polyadenylation factors and determining the position of cleavage. In mammals, but not in yeast and plants, there is a downstream GU-rich polyadenylation signal that serves as another binding site for polyadenylation factors (Gilmartin, 2005). It seems that the C-rich downstream element in *Chlamydomonas* is somewhat different from that observed in animals in relation to both location (closer to the cleavage site) and sequence characteristics (less U and G in *Chlamydomonas*). The functionality of these signal elements during cleavage and polyadenylation reactions, however, remains to be tested.

Another important finding in this paper is the extensive usage of APA in *Chlamydomonas*. We estimate that a range of 9 to 33% of *Chlamydomonas* transcripts have two or more poly(A) sites. This number is less than human and rice (Shen et al., under revision; Zhang et al., 2005). Our EST-based analysis was able to distinguish the poly(A) sites with highest resolution (at a nucleotide level), thus providing a more

accurate survey of APA in algae. There are significant number of poly(A) sites, 1 to 11% of the genes depending on the criteria used, located in 5'UTRs, introns and CDS in *Chlamydomonas*. Compare to that of rice (~2%; Shen et al.,2008), the extent of APA in *Chlamydomonas* might be with the similar scope or even more. The presence of alternative poly(A) sites in other regions of a gene (e.g., those matching annotated introns or CDS) may truncate the open reading frame, producing different types of transcripts and/or protein products. Further study of this mechanism should result in meaningful insight into this phenomenon in algae. While such APAs and the extra length of 3'-UTR are revealed here, it is recognized that the accuracy of the data is relied on the current annotated draft genome. The latter, however, may only have limited accuracy due to its current annotation status. The confident level of our data would be improved when more concrete genome information becomes available for *Chlamydomonas*.

In addition, we revealed a high percentage of poly(A) sites found in the unannotated region of the genome (Table 3). These poly(A) sites could offer some clues about where to find additional genes, encoding proteins or not, in the *Chlamydomonas* genome. On the other hand, the unique features of the polyadenylation signals we revealed could also pave the way toward the design of predictive models to find other potential poly(A) sites that are not currently collected by the EST projects (Ji et al., 2007b). Achieving this, in turn, could improve *Chlamydomonas* genome annotation because poly(A) sites generally mark the ends of transcripts. The predictive results could, of course, also lead to further exploration of APA in *Chlamydomonas*.

Acknowledgements:

The authors wish to acknowledge Eric Stalhberg for porting the SignalSleuth program to a Linux platform; Anand Srinivasan and David Woods of the Miami University Research Computing Services group for support in running SignalSleuth on the Miami University Computing Cluster; David Martin for reviewing the manuscript; Bin Tian for the human

poly(A) dataset; and other lab members for helpful discussions. This project was funded in part by US NSF (MCB 0313472 to QQ Li), Ohio Plant Biotechnology Consortium, and Miami University Center for the Advancement of Computational Research (both to QQ Li and C Liang).

FIGURES

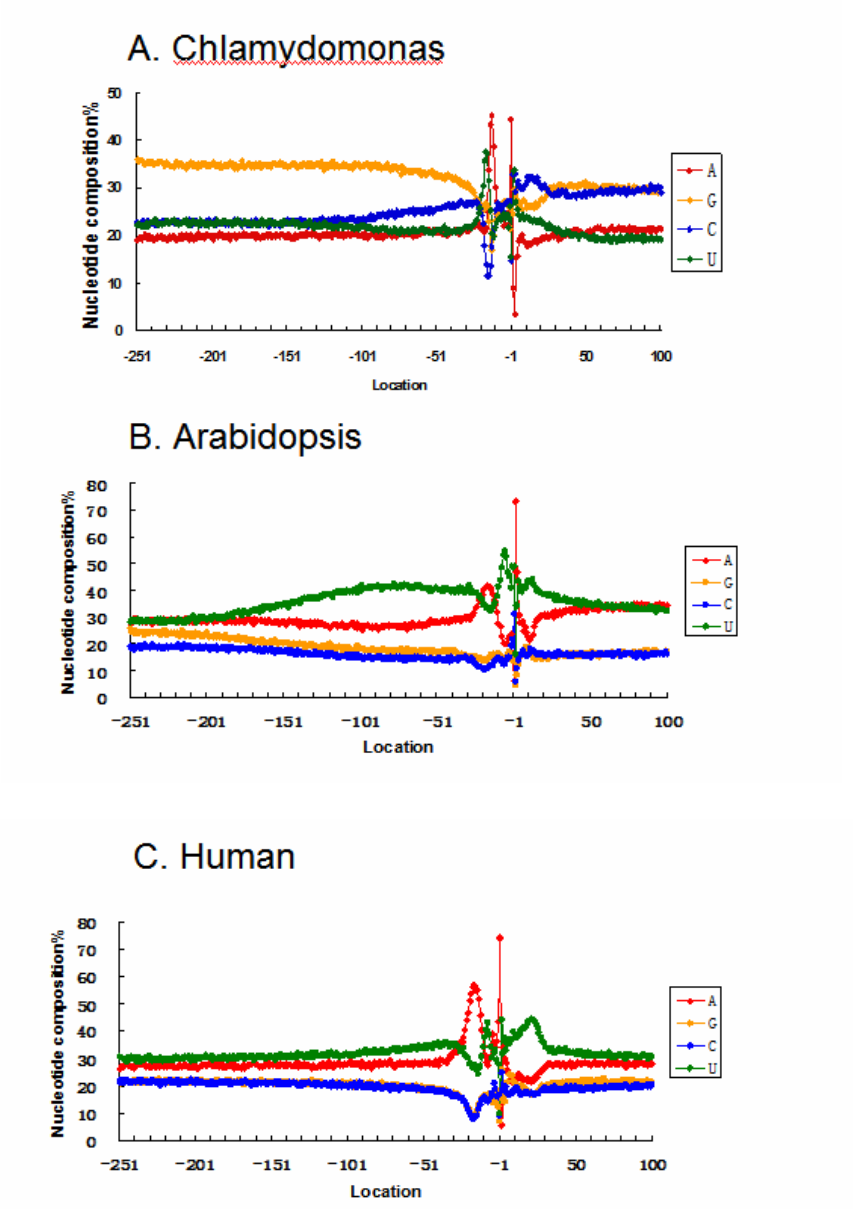



Figure 3-1: Single nucleotide profiles of the 3'-UTR for different species.

A. *Chlamydomonas*; B. Arabidopsis; C. Human. The poly(A) site is at position -1. The upstream sequence (300 *nt*) of the poly(A) site is in "-" designation, and downstream (100nt) sequence is in "+" designation.

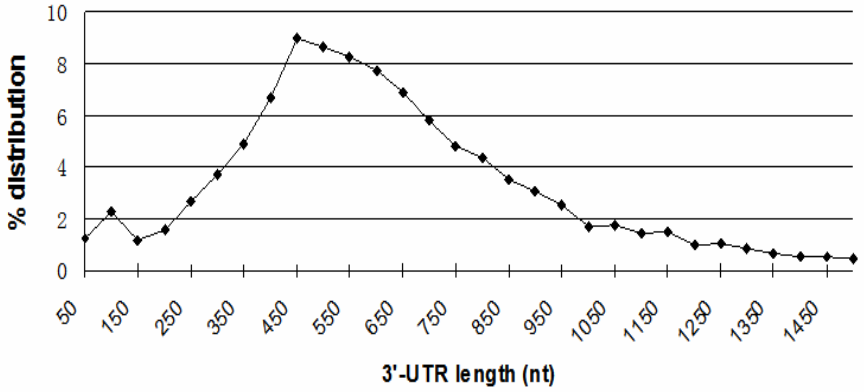



Figure 3-2: Distribution of the length of 3'-UTR in *Chlamydomonas*.

The average length is 595-nt.

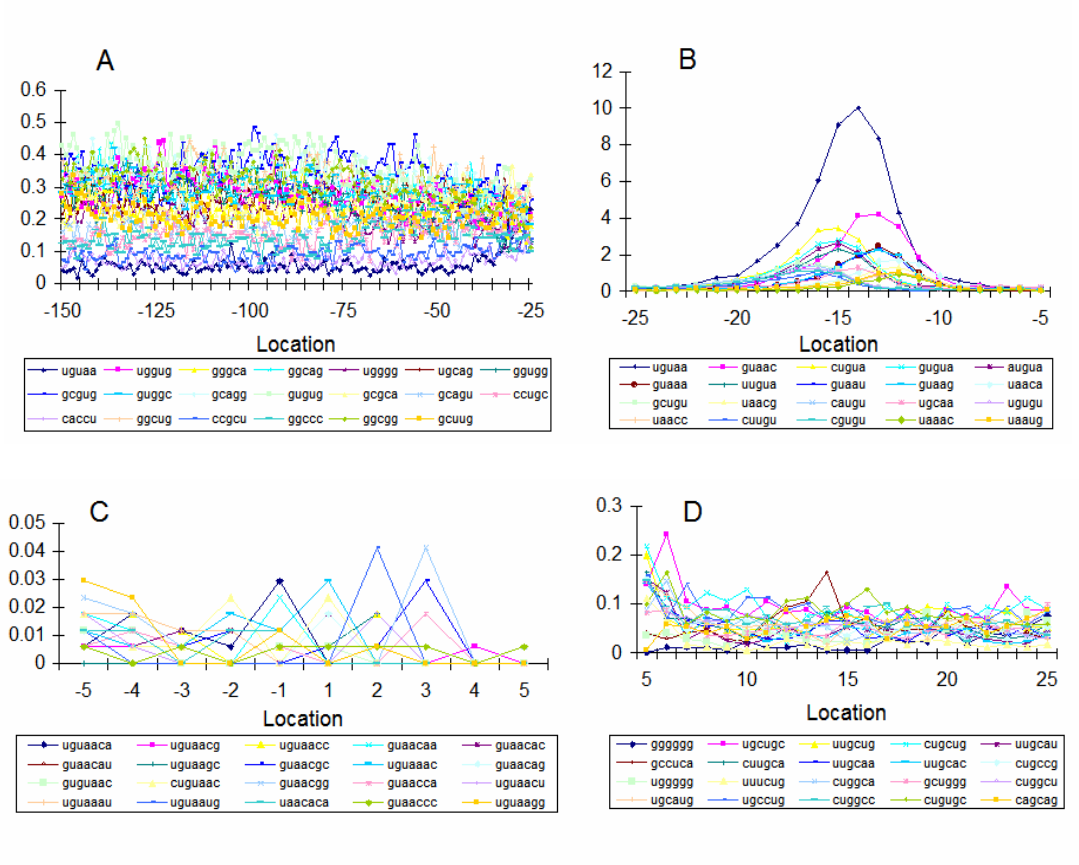



Figure 3-3: The 20 top-ranked signals in the designated poly(A) signal regions.

A. Pentamers from -150 to -25 in FUE. B. Pentamers from -25 to -5 in NUE. C. Heptamers from -5 to +5 in CE. D. Hexamers from +5 to +30 in DE. The poly(A) site is at position -1. The upstream sequence of the poly(A) site is in "-" designation, and the downstream sequence is in "+" designation.

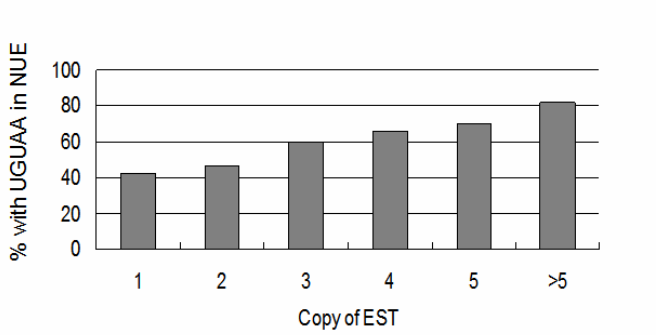



Figure 3-4:Correlation of UGUAA and gene expression level.

The higher the expression level (more EST copies), the better the chance of having UGUAA in their NUE as a poly(A) signal.

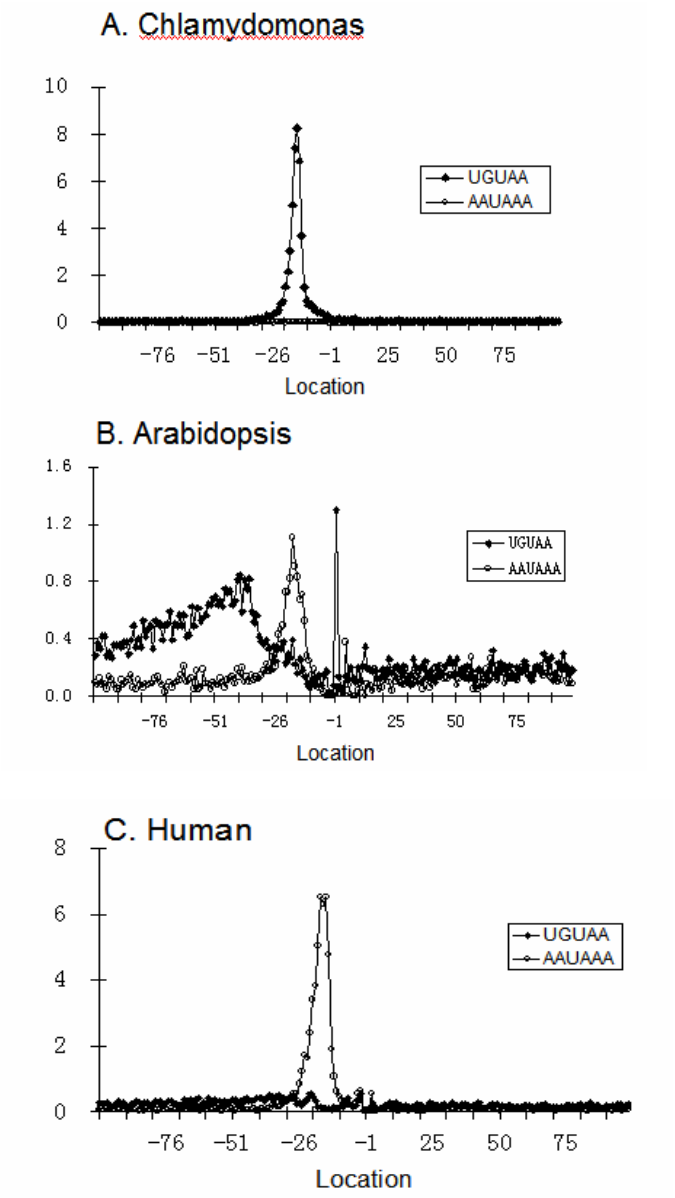



Figure 3-5: Distribution of UGUAA and AAUAAA signals in different species.

A. *Chlamydomonas*; B. Arabidopsis; C. Human. Filled-circle and open-circle lines show UGUAA and AAUAAA in the region of -100 to +100 surrounding the poly(A) site, respectively. The poly(A) site is at position -1.

## *References*


Arabidopsis, Genome Initiative, 2000, Analysis of the genome sequence of the flowering plant arabidopsis thaliana., *Nature* 408, 796-815.

Berthold, P., R. Schmitt, and W. Mages, 2002, An engineered streptomyces hygroscopicus aph 7" Gene mediates dominant resistance against hygromycin b in *Chlamydomonas reinhardtii*, *Protist* 153, 401-412.

Chen, F., C. C. Macdonald, and J. Wilusz, 1995, Cleavage site determinants in the mammalian polyadenylation signal, *Nucleic Acids Research* 23, 2614-2620.

Cote, G. J., D. T. Stolow, S. Peleg, S. M. Berget, and R. F. Gagel, 1992, Identification of exon sequences and an exon binding protein involved in alternative rna splicing of calcitonin/cgrp, *Nucleic Acids Res* 20, 2361-6.

Crooks, G. E., G. Hon, J. M. Chandonia, and S. E. Brenner, 2004, Weblogo: A sequence logo generator, *Genome Res* 14, 1188-90.

Delaney, K. J., R. Q. Xu, J. X. Zhang, Q. Q. Li, K. Y. Yun, D. L. Falcone, and A. G. Hunt, 2006, Calmodulin interacts with and regulates the rna-binding activity of an arabidopsis polyadenylation factor subunit, *Plant Physiology* 140, 1507-1521.

Dong, H. T., Y. Deng, J. Chen, S. Wang, S. H. Peng, C. Dai, Y. Q. Fang, J. Shao, Y. C. Lou, and D. B. Li, 2007, An exploration of 3 '-end processing signals and their tissue distribution in oryza sativa, *Gene* 389, 107-113.

Gilmartin, G. M., 2005, Eukaryotic mrna 3' processing: A common means to different ends, *Genes Dev* 19, 2517-21.

Graber, J. H., C. R. Cantor, S. C. Mohr, and T. F. Smith, 1999, Genomic detection of new yeast pre-mrna 3'-end-processing signals, *Nucleic Acids Res* 27, 888-94.

Hu, J., C. S. Lutz, J. Wilusz, and B. Tian, 2005, Bioinformatic identification of candidate cis-regulatory elements involved in human mrna polyadenylation, *RNA* 11, 1485-93.

Hunt, A. G., 2007, Messenger rna 3′ -end formation and the regulation of gene expression, in C.L. Bassett, ed.: *Regulation of gene expression in plants: The role of transcript structure and processing* (Springer).

Ji, G. , J. Zheng, Y. Shen, X. Wu, R. Jiang, Y. Lin, J. C. Loke, K. M. Davis, G. J. Reese, and Q. Q. Li, 2007, Predictive modeling of plant messenger rna polyadenylation sites, *BMC Bioinformatics* 8, 43.

Li, Q. Q., and A. G. Hunt, 1997, The polyadenylation of rna in plants, *Plant Physiology* 115, 321-325.

Liang, C., Y. Liu, L. Liu, A. C. Davis, Y. Shen, and Q.Q. Li, 2008, ESTs with cDNNNA termini – previously overlooked resources for gene annotation and transcriptome exploration in *Chlamydomonas reinhardtii*.179:83-93.

Liang, C., G. Wang, L. Liu, G. L. Ji, L. Fang, Y. S. Liu, K. Carter, J. S. Webb, and J. F. D. Dean, 2007, Coniferest: An integrated bioinformatics system for data reprocessing and mining of conifer expressed sequence tags (ests), BMC Genomics 8, 134.

Liang, C., G. Wang, L. Liu, G. Ji, Y. Liu, J. Chen, J. S. Webb, G. Reese, and J. F. D. Dean, 2007, Webtraceminer: A web service for processing and mining est sequence trace files, *Nucleic Acids Research* 35, 137-142

Loke, J. C., E. A. Stahlberg, D. G. Strenski, B. J. Haas, P. C. Wood, and Q. Q. Li, 2005, Compilation of mrna polyadenylation signals in arabidopsis revealed a new signal element and potential secondary structures, *Plant Physiol* 138, 1457-1468.

Lou, H., and R. F. Gagel, 1998, Alternative rna processing--its role in regulating expression of calcitonin/calcitonin gene-related peptide, *J Endocrinol* 156, 401-5.

Merchant, S. S., and S. E. Prochnik, and O. Vallon, and E. H. Harris, and S. J. Karpowicz, and G. B. Witman, and A. Terry, and A. Salamov, and L. K. Fritz-Laylin, and L. Marechal-Drouard, and W. F. Marshall, and L. H. Qu, and D. R. Nelson, and A. A. Sanderfoot, and M. H. Spalding, and V. V. Kapitonov, and Q. Ren, and P. Ferris, and E. Lindquist, and H. Shapiro, and S. M. Lucas, and J. Grimwood, and J. Schmutz, and P. Cardol, and H. Cerutti, and G. Chanfreau, and C. L. Chen, and V. Cognat, and M. T. Croft, and R. Dent, and S. Dutcher, and E. Fernandez, and H. Fukuzawa, and D. Gonzalez-Ballester, and D. Gonzalez-Halphen, and A. Hallmann, and M. Hanikenne, and M. Hippler, and W. Inwood, and K. Jabbari, and M. Kalanon, and R. Kuras, and P. A. Lefebvre, and S. D. Lemaire, and A. V. Lobanov, and M. Lohr, and A. Manuell, and I. Meier, and L. Mets, and M. Mittag, and T. Mittelmeier, and J. V. Moroney, and J. Moseley, and C. Napoli, and A. M. Nedelcu, and K. Niyogi, and S. V. Novoselov, and I. T. Paulsen, and G. Pazour, and S. Purton, and J. P. Ral, and D. M. Riano-Pachon, and W. Riekhof, and L. Rymarquis, and M. Schroda, and D. Stern, and J. Umen, and R. Willows, and N. Wilson, and S. L. Zimmer, and J. Allmer, and J. Balk, and K. Bisova, and C. J. Chen, and M. Elias, and K. Gendler, and C. Hauser, and M. R. Lamb, and H. Ledford, and J. C. Long, and J. Minagawa, and M. D. Page, and J. Pan, and W. Pootakham, and S. Roje, and A. Rose, and E. Stahlberg, and A. M. Terauchi, and P. Yang, and S. Ball, and C. Bowler, and C. L. Dieckmann, and V. N. Gladyshev, and P. Green, and R. Jorgensen, and S. Mayfield, and B. Mueller-Roeber, and S. Rajamani, and R. T. Sayre, and P. Brokstein, and I. Dubchak, and D. Goodstein, and L. Hornick, and Y. W. Huang, and J. Jhaveri, and Y. Luo, and D. Martinez, and W. C. Ngau, and B. Otillar, and A. Poliakov, and A. Porter, and L. Szajkowski, and G. Werner, and K. Zhou, and I. V. Grigoriev, and D. S. Rokhsar, and A. R. Grossman, 2007, The *Chlamydomonas* genome reveals the evolution of key animal and plant functions, *Science* 318, 245-50.

Meyers, B. C., T. H. Vu, S. S. Tej, H. Ghazal, M. Matvienko, V. Agrawal, J. Ning, and C. D. Haudenschild, 2004, Analysis of the transcriptional complexity of arabidopsis thaliana by massively parallel signature sequencing, *Nat Biotechnol* 22, 1006-11.

Moore, C. L., H. Skolnikdavid, and P. A. Sharp, 1986, Analysis of rna cleavage at the adenovirus-2 l3 polyadenylation site, *EMBO Journal* 5, 1929-1938.

Peterson, M. L., 1994, Regulated immunoglobulin (ig) rna processing does not require specific cis-acting sequences: Non-ig rna can be alternatively processed in b cells and plasma cells, *Mol Cell Biol* 14, 7891-8.

Proudfoot, N., 2004, New perspectives on connecting messenger rna 3' end formation to transcription, *Curr Opin Cell Biol* 16, 272-8.

Seiler, K.P., G.A. George, M.P. Happ, N.E. Bodycombe, H.A. Carrinski, S. Norton, S. Brudz, J.P. Sullivan, J. Muhlich, M. Serrano, P. Ferraiolo, N.J. Tolliday, S.L. Schreiber, and P.A. Clemons, 2007, Chembank: A small-molecule screening and cheminformatics resource database., *Nucleic Acids Res.* 36,351-359

Sheets, M. D., S. C. Ogg, and M. P. Wickens, 1990, Point mutations in aauaaa and the poly(a) addition site - effects on the accuracy and efficiency of cleavage and polyadenylation invitro, *Nucleic Acids Research* 18, 5799-5805.

Shen, Y., G. Ji, B. J. Haas, X. Wu, J. Zheng, G. J. Reese, and Q. Q. Li, 2008. Genome level analysis of rice mrna 3'-end processing signals and alternative polyadenylation *Nucleic Acids Res* 36: 3150-3161

Silflow, C. D., R. L. Chisholm, T. W. Conner, and L. P. Ranum, 1985, The two alpha-tubulin genes of *Chlamydomonas* reinhardi code for slightly different proteins, *Mol Cell Biol* 5, 2389-98.

Simpson, G. G., P. P. Dijkwel, V. Quesada, I. Henderson, and C. Dean, 2003, Fy is an rna 3' end-processing factor that interacts with fca to control the arabidopsis floral transition, *Cell* 113, 777-87.

Slomovic, S., V. Portnoy, V. Liveanu, and G. Schuster, 2006, Rna polyadenylation in prokaryotes and organelles; different tails tell different tales, *Critical Reviews in Plant Sciences* 25, 65-77.

Tian, B., J. Hu, H. B. Zhang, and C. S. Lutz, 2005, A large-scale analysis of mrna polyadenylation of human and mouse genes, *Nucleic Acids Research* 33, 201-212.

van Helden, J., 2003, Regulatory sequence analysis tools, *Nucleic Acids Res* 31, 3593-6.

Venkataraman, K., K. M. Brown, and G. M. Gilmartin, 2005, Analysis of a noncanonical poly(a) site reveals a trinartite mechanism for vertebrate poly(a) site recognition, *Genes & Development* 19, 1315-1327.

Wodniok, S., A. Simon, G. Glockner, and B. Becker, 2007, Gain and loss of polyadenylation signals during evolution of green algae, *BMC Evol Biol* 7, 65.

Wu, T. D., and C. K. Watanabe, 2005, Gmap: A genomic mapping and alignment program for mrna and est sequences, *Bioinformatics* 21, 1859-75.

Zhang, H., J. Y. Lee, and B. Tian, 2005, Biased alternative polyadenylation in human tissues, *Genome Biol* 6, R100.

Zhao, J., L. Hyman, and C. Moore, 1999, Formation of mrna 3' ends in eukaryotes: Mechanism, regulation, and interrelationships with other steps in mrna synthesis, *Microbiol Mol Biol Rev* 63, 405-45.

Table 3-1: Concise representation of polyadenylation *cis*-elements in *Chlamydomonas* as sequence logos.

| Region | Signal logo | Name | # of signals | Top signal | % of Hits |
|---|---|---|---|---|---|
| FUE<br>-150/-25 |  | FUE.1 | 5 | CGGUA | 98% |
| |  | FUE.2 | 2 | UUACA | 33% |
| NUE<br>-25/-5 |  | NUE.1 | 4 | UGUAA | 78% |
| |  | NUE.2 | 5 | AGUAU | 9% |
| CE<br>-5/+5 |  | CE.1 | 6 | UACCGUA | 6% |
| DE<br>+5/+30 |  | DE.1 | 2 | ACCGUA | 13% |

Table 3-2: Number of genes with unique alternative poly(A) sites.

| Number of poly(A) site(s) per mRNA | Number of genes* | % |
|---|---|---|
| 1 | 2716 | 67% |
| 2 | 913 | 23% |
| 3 | 296 | 7% |
| 4 or more | 132 | 3% |
| Sum | 4057 | 100% |

*These genes are selected from those that have at least one authenticated poly(A) site from the 17K dataset. Note that the genes are considered here is only a part of the total gene number from the gene catalog of about 15,000.

Table 3-3: The distribution of poly(A) sites on gene transcripts.

| Category | Sub-category | Number of transcripts | % |
|---|---|---|---|
| Total poly(A) sites | - | 16,952 | 100 |
| Located in the transcript (full-length cDNA) | In CDS | 45 (12)[#] | 0.3 |
| | In introns | 588 (39)[#] | 3.5 |
| | In 5'-UTR | 86 (12)[#] | 0.5 |
| | In 3'-UTR* | 11,011 | 65.0 |
| | Subtotal | 11,730 | 69.2 |
| Located in the intergenic region** | - | 5,222 | 30.8 |

* To avoid genome annotation error, the 3'-UTR defined here has been extended by 500-nt post the poly(A) site. For those genes that do not have an annotated 3'-UTR in the current version of the genome, the range were extended to 1000-nt after the annotated stop codons.

** The intergenic regions are the areas 500-nt or 1000-nt downstream of 3'-UTR (as defined above).

[#] The numbers in the parenthesis are the highly confident cases when using more stringent conditions and manual confirmation as described in the main text.

# **CHAPTER 4:** CONSERVED AND TISSUE-SPECIFIC ALTERNATIVE POLYADENYLATION IN ARABIDOPSIS AND RICE GENOMES

The material in this chapter is in preparation to be submitted: Yingjia Shen, Kan Nobuta, Varun Bala, Caghan Demirci, Blake C Meyers, RC Nenu, Guoliang Wang, and Qingshun Q Li. Contributions to this chapter: Kan Nobuta, Varun Bala, Caghan Demirci, Blake C Meyers, RC Nenu, Guoliang Wang generated the MPSS and SBS tags from rice and Arabidopsis. All of the other data were collected and analyzed by Yingjia Shen.

Supplymental data can be downloaded from www.ourshen.com/Supplemental_Data.zip

## *Abstract*

The location of the poly(A) site, as it marks the end of a transcript, is controlled by the interactions of poly(A) signals in the 3'-UTR and the *trans*-acting protein factors. However, it has been demonstrated that many transcripts contain poly(A) tails in alternative locations such as coding sequences, introns, or the 5'-UTR of their full-length isoforms. Such alternative polyadenylation (APA) may significantly alter the length of transcripts and/or coding sequences, a mechanism that can be effectively used to regulate gene expression and increase genome coding capacity. Using the rich collections of MPSS (Massively Parallel Signature Sequencing) and Illumina's GAII SBS (sequence by synthesis) tags, we discovered a large number of APAs that have not been found previously by other methods. In Arabidopsis, we found that 11,264 genes (42% of total protein coding genes) have multiple poly(A) sites based on 43,767 MPSS tags. In rice, we found that 12,091 (48%) and 24,877 (82%) of rice genes have more than one poly(A) site as evidenced by MPSS and SBS tags, respectively. In both species, 49% to 66% of the signatures were mapped as APA events upstream of stop codons. Although poly(A) sites in regions other than 3'UTR are common in both species, we found that only about 10% of the whole transcriptomes are made up of alternatively polyadenylated transcripts. Moreover, 3,034 genes were found to have APA that are conserved between rice and Arabidopsis. Our results suggest that plants tend to use APA at some tissue and developmental stages or when plants are under some stress,, and that enriched expression of some polyadenylation and/or splicing factors may be associated with determining abundance of APA events.

## *Introduction*

Messenger RNA (mRNA) polyadenylation is one of the essential post-transcriptional processing steps in eukaryotic gene expression. Immediately following being transcribed, pre-mRNAs are capped, spliced and cleaved at the 3'-untranslated region (3'-UTR) to generate new open ends allowing addition of poly(A) tails (Zhao et al. 1999). A poly(A) tail at the 3'-end protects the normal functionality of the mRNA from unregulated degradation, triggers export to the cytoplasm, and assists in recognition by the translational machinery (Zhao et al. 1999; Danckwardt et al. 2008). The choice of the location where pre-mRNA is cleaved (the poly(A) addition site) is heavily regulated by numerous of protein factors to ensure precision (Zhao et al. 1999; Hunt et al. 2008; Shi et al. 2009). Alternative polyadenylation (APA) is defined as the use of more than one poly(A) sites, excluding microheterogeneity which is defined as less than 30 nucleotides between two poly(A) sites of a same gene (Shen et al. 2008a). APA is a powerful pathway to increase the complexity of transcriptomes and proteomes because it leads to the production of two or more different proteins or non-functional variants from the same loci on the genome (Lutz 2008). Previous EST-based analyses showed that about half of human and rice genes and up to one third of algal protein-codinggenes undergo APA (Zhang et al. 2005; Shen et al. 2008a; Shen et al. 2008b), suggesting a global role of APA in gene expression regulation. The cost and efficiency of EST-based methodology in APA studies, however, limits the number of poly(A) sites that can be found and prevents studying APA in greater depth.

New technologies of large-scale DNA sequencing have made the study of polyadenylated transcriptome easier than before. Massively Parallel Signature Sequencing (MPSS) and Illumina's GAII Sequence by Synthesis (SBS) (Brenner et al. 2000; Meyers et al. 2004a; Meyers et al. 2004b; Meyers et al. 2004c; Simon et al. 2009) have yielded many millions of 17 or more nucleotides called signatures. Each of these signatures is derived from a specific restriction enzyme (e.g. *Dpn*II) in 3'-most occurrence of a polyadenylated

transcript. Thus, MPSS and the specially designed SBS can be useful in identifying the locations of poly(A) sites, in addition to besides tracking the frequency of the transcripts (Figure 4.1, Meyers et al. 2004c). In Arabidopsis and rice, over 36 million and 46 million MPSS signatures from 14 and 22 different libraries have been made available to the public, respectively (Meyers et al. 2004b; Nobuta et al. 2007). These signatures were further mapped to the annotated genomes resulting in over 67,000 and 81,000 unique genome-matched signatures, each representing one or a unique set of poly(A) sites (Meyers et al. 2004b; Nobuta et al. 2007). The massive number of signatures provides a unique advantage over EST-based analysis, in which the extent of APA can be studied. Based on MPSS signatures of five libraries from Arabidopsis, previous study estimates the extent of APA to be around 25% (Meyers et al. 2004c). In recent years, SBS based method features lower cost per tag and higher accuracy compared to MPSS, and has been used in genome resequencing, small RNA sequencing and epigenetic studies (Irizarry et al. 2008; Elling and Deng 2009; Fox et al. 2009; Li et al. 2009). However, the detailed analysis of plant APA using data generated by SBS has not been realized.

In this paper, we analyzed the extent and the conservation of APA in Arabidopsis and rice, two diverse (dicot and monocot) plant lineages, utilizing MPSS and SBS data. We also studied tissue specific usages of APA events and identified several candidate genes that account for differences between tissues. These analyses offer valuable information for studying alternative transcript processing and the potential role of these alternates in gene expression as well.

## *Results*

### Distributions of MPSS signatures among Arabidopsis and rice genes

Previously in Arabidopsis, a total of 67,735 signatures of 17-bp were sequenced and matched to 19,088 annotated genes (Meyers et al. 2004c). Analysis based on 5 different plant organs identified over 4,000 genes (account for 26.1% of total genes found in these 5 libraries) that have more than 1 signature in their sense strands, suggesting multiple 3'-ends. Since the work published in 2004, 12 additional libraries have been made available, including 4 inflorescences of different mutants, 2 stress treated leaves, germinating seedlings, and 4 additional organs (Meyers et al. 2004b). We have now analyzed all 17 libraries and found that 11,264 Arabidopsis genes have multiple signatures, which is about 60% of all 19,088 genes being detected by MPSS. In order to avoid unreliable signatures or counting the same signatures multiple times, all signatures we studied must match uniquely to the genome and also be significantly expressed (Tag Per Million or TPM >3) (Meyers et al. 2004a) we used 36,402 unique tags in the analysis. Among these 11,264 genes, only 530 of them have all their signatures localized in annotated 3'-UTR (based on Arabidopsis genome annotation version 8), suggesting a wide distribution of the signatures in other parts of the transcripts. On average, each gene on the list has about 3 unique signatures (36,403 divided by 11,264). Most of the genes have arelatively small number of signatures (76% of genes have less than 4 signatures), but some genes have numerous signatures, one of which has 27 unique signatures (At5g40450; Table 1).

In rice, we used 17-bp signatures from 22 MPSS libraries (Nobuta et al., 2007) and searched for genes with more than one signature. The number of genes with multiple signatures is 12,091 (48% of 25,500 genes, Table 1), slightly less than what was found in Arabidopsis. There are 1,681 genes with all their signatures in the 3'-UTR, which might be due to the fact that the length of 3'-UTR in rice is longer than that of Arabidopsis (289

to 223-nt on average; Shen et al., 2008a). The number of genes with APA is also slightly higher than previous calculation based on ESTs (8,596 genes, or 33% of 25,500 genes (Shen et al. 2008a)], suggesting MPSS is a more powerful tool in detecting rare APA events.

In addition to MPSS data, we also obtained information on 5,522,207 distinct 20bp tags from 48 libraries sequenced by the Illumina SBS method. After we applied the same filter (uniquely mapped to genome and TPM >3), 178,555 tags (from 30,288 genes) were used for further analysis. In these 30,288 genes, 24,877 (82%, Table 1) of them have multiple tags. The number of genes with APA doubled of that derived from MPSS study, suggesting that APA is a universal phenomenon in rice, and SBS is a more sensitive method of detecting APA.

**Locations of potential poly(A) sites on transcripts**

While the location of a MPSS or SBS signature does not directly coincide with a poly(A) site at the nucleotide level, it does indicate a poly(A) site in a defined region. Both MPSS and SBS signatures from Arabidopsis and rice were derived from restriction enzyme *Dpn*II sites immediately 5' to the poly(A) sites (Meyers et al. 2004a; Figure 4-1), The poly(A) site must be located between the sequenced signature and the immediate next *Dpn*II site downstream in the genomic sequence. Based on this notation, we developed a method to further categorize MPSS signatures according to both their locations and the poly(A) sites from which they are derived. To differentiate these signatures (derived from genomic sequences) from previous study of signatures, we call these "APA-class signatures" because the classification mainly deals with signatures associated with poly(A) sites. To reduce the complexity of the analysis, we only studied signatures found in genes with multiple signatures, and they were grouped into 8 different classes for easy discussion purposes (Figure 4-2).

In this APA-class system, signatures of APA-class 1 are located in the 3'-UTR of the gene (Figure 4-2). Signatures of APA-class 2 are located upstream of the stop codon while there are no other possible *Dpn*II site between the signature and the stop codon. Signatures in this class could be derived from two types of poly(A) sites: poly(A) sites located upstream of the stop codon (an APA site) or conventional poly(A) sites located in the 3'-UTR. The reason why APA-class 2 signatures also include poly(A) sites in the 3'-UTR is that there are no *Dpn*II sites between stop codon and 3'-UTR. We do not have a way to differentiate these APA sites from conventional sites in APA-class 2, but the higher expression levels and the similarity to APA-class 1 signatures based on the fact that they are more frequently used than signatures from other APA-class (see details below) made us believe that most signatures in this class are derived from conventional poly(A) sites. In Arabidopsis and rice, about 44% and 55% of MPSS signatures and 54% of SBS rice signatures belong to APA-classes 1 and 2 of conventional polyadenylation events. We further tested how often these poly(A) sites are utilized based on the their relative TPM value. In both species, about 90% of all sequenced signatures are from first two APA-classes, suggesting that mRNAs with their poly(A) sites in the 3'-UTR are still dominant over other alternative transcripts.

Signatures from all APA-classes 3 to 8 positively locate poly(A) sites upstream of the stop codons because there is at least one *Dpn*II site found to be upstream of the stop codon. Therefore, each signature from these 6 classes represents an APA event which leads to the production of truncated transcripts when compared to the full length ones. All signatures in APA-classes 3, 4 and 5 are located in exons. The difference among three classes is that the possible poly(A) sites derived from Class 3 extends over exons where the signatures are located, while the poly(A) sites of Classes 4 and 5 can be limited to one exon. Class 5 is more specifically designated to poly(A) sites located in the last exon because many of the truncated transcripts could still produce functional proteins due to only small deletions. Signatures located in the exons are one of most abundant APA groups in our analysis, mainly because MPSS signatures are collected from mature

mRNA where introns might have been spliced out. In Arabidopsis and rice, 48% and 33% of MPSS signatures and 35% of rice SBS tags belong to these three exon-located APA classes (Figure 4-2).

Signatures from APA-classes 6 and 7, however, are located in introns, where poly(A) sites of APA-class 6 signatures can also be positioned into one intron, and poly(A) sites of Class 7 might be located beyond that particular intron where the signature is located. Signatures from these two classes indicate the association of two tightly coupled mRNA processing events, alternative splicing and APA. Since MPSS signatures were collected from mostly mature mRNA where intron sequences were less expected, signatures from Classes 6 and 7 suggest that introns were retained in some mature mRNA. Such intron retentions were found to be common in plants (Ner-Gaon et al. 2004). The question remains as to whether they were the retention of unspliced introns that induced the APA events or cleavage in the introns [for generating poly(A) sites] that blocked the splicing process. About 2% and 5% of Arabidopsis and rice signatures, respectively, belong to these two classes. Tags sequenced by SBS technology, however, suggest a bigger role of introns in APA. Up to 15% of APA events could be associated with introns, correlating with our previous finding using ESTs (Shen et al., 2008a) and suggesting introns as major players in APA events. The differences between MPSS and SBS could be due to more in depth sequencing of the tags in SBS.

Signatures of APA-class 8 are located in the 5'-UTR of the genes, which can be found in 4% and 6% of Arabidopsis and rice signatures, respectively. There are evidences that some transcripts contain poly(A) sites in the 5'-UTR (Shen et al. 2008a; Shen et al. 2008b), although the significance of such events are not fully appreciated.

Although the last 6 APA-classes consist of about 50% of all unique signatures, analysis on the frequency of these signatures (representing the expression levels) shows they

consist of only 10% of all sequenced signature (Figure 4-2). In addition to lower abundance, signatures from these APA-classes are more library-specific (more detail below), suggesting their regulatory roles might be limit to certain tissues or development stages.

**Library/tissue specific APA events**

In order to determine whether or not APA events are associated with certain libraries (made from different tissues or developmental stages), we tested the number of signatures only expressed in one tissue and compared it with the number of all signatures in that particular APA-class. In both species, the percentages of library specific signatures in APA-classes 1 and 2 are significantly lower compared with the rest of 6 classes (10% compared with 37% in Arabidopsis, 26% compared with 41% in rice; see Supplementary Data S1). This suggests that conventional poly(A) sites represented by first two classes are more likely to be ubiquitously utilized and less tissue specific.

To further study the usage of alternative poly(A) sites in each library, we used a method similar to GAUGE (Zhang et al. 2005b; Lutz 2008) with slight modifications. In each library, the expression value of each signature (TPM value) was grouped based on their APA-classes. To normalize the differences between libraries, TPM values were divided by the total number of signatures sequenced in that library. The percentage of usage of a APA-class in a library can then be measured by the sum of all signatures in that APA-class. For each poly(A) site type in a library, its percentage of usage was compared with the average percentage of usage of all libraries. The difference was normalized to the mean and called as relative distance. Figure 4-4 shows libraries and their relative preference to usage of APA where positive value means that library use more APA, and vice versa. A complete list of values for all libraries are provided in Supplementary Data S2. The significance of difference to the mean usage of APA was further calculated by

Chi-Square test and libraries are ranked based on p-value with more significant group on the top (Figure 4-4).

Among all libraries of Arabidopsis, germinating seedling (GSE in Figure 4-4A) shows significant bias towards usage of APA sites (more positive values), particularly for APA-classes 4, 7 and 8. This suggests either APA might play an more important role in fast growing tissues. Another possibility is that the polyadenylation machinery is less tightly controlled when it comes to selecting a poly(A) site due to large amounts of mRNAs produced. Interestingly, some libraries just show preference to a certain APA group. There are two library samples with a single APA-class standing out, notably 21-day leaves with strong preference for APA-class 4, and *ap1-10* mutant with preference of APA-class 6. This suggests that some tissues or mutants respond differently in each library potentially under library-specific regulations. Beside dominant utility of different classes of APA sites, there may be general avoidance of APA (more negative values). In contrast, plants treated with salicylic acid tend to use less APA. While not as drastic as in the case of germinating seedlings in terms of preference on certain type of APA-classes, such a broad range reduction of APA would warrant further investigations.

In rice, 15 libraries were examined by the same method and the results are shown in Figure 4-4B. The most clearly demonstrated case is in the library of young leaves stressed in cold, which shows a very significantly increase in usage of poly(A) site in introns (APA-class 6), suggesting a specific role of APA in cold response in rice leaves through utilizing poly(A) sites in an intron. In contrast to germinating seeding in Arabidopsis, rice germinating seedlings grown in dark generally avoid the use of APA sites. This could probably due to the fact that many gene regulation pathways are not active in the dark condition. The rest of the 12 groups do not show significant difference compared to average usage of APA sites. Overall, the difference in usage of APA among different libraries as well as preference over certain APA class within a particular library suggests there is a complex mechanism regulating the degree of APA.

**Relationship of APA to the expression levels of *trans*-acting factors**

It is well known that some polyadenylation related protein factors play crucial roles in 3'-end formation (Simpson et al. 2003; Xing et al. 2008). To further determine the cause of different usage of APA sites in different tissue types, we examined the expression level of known polyadenylation factors in each MPSS library. The hypothesis was that different expression levels of these *trans*-acting factors might correspond to different APA levels among different libraries. Since Arabidopsis has more libraries which showed significant differences in APA site usages and the polyadenylation factors are better understood compared with those in rice, we only used the polyadenylation factors of Arabidopsis in this study. In 12 libraries (we did not use signatures sequenced by the classic MPSS method to make sure that comparisons between libraries are not affected by sequencing methods), the TPM values were used as the expression levels of genes involved in polyadenylation. In addition to polyadenylation factors, mRNA splicing is also known to play a role in polyadenylation (Millevoi et al. 2006; Tian et al. 2007). We thus included splicing factors as well in our analysis. The 24 known Arabidopsis polyadenylation factors (Hunt et al. 2008) and expression value of 19 splicing factors (Campbell et al. 2006) were examined in the same 12 libraries obtained from Arabidopsis MPSS databases.

To find genes that are critical to library-specific behavior of APA, we use Spearman correlation (Wissler 1905) to calculate the correlation of each gene's expression value with different usage of APA. We further categorized APA into two conditions, APA found in exons (classes 4, 5 and 6), and APA associated with alternative splicing (classes 6 and 7). In the first case, several polyadenylation genes *PAPS4, PABN3, PCFS1* and *U2AF65* were found to have a positive effect in increasing the usage of library-specific APA (Table 2; detail expression value of each genes in Supplemental Data S3). Meanwhile *CFIS25*, *SYM2* and *PABN2* have positive APA effect in introns. This result correlates some of our previous findings using traditional experimental methods. For example, *PCFS*, a

homologue of yeast polyadenylation factor Protein 1 of Cleavage Factor 1 (*Pcf11p*), is the only gene on the list significantly correlated with APA in exons and introns and is also known to regulate flowering time through regulation of APA (Xing et al. 2008). It is interesting that the splicing factor U2AF65 was recently connected to the splicing of the last exon and polyadenylation, through its interaction with polyadenylation factor CFI-m in mammals (Millevoi et al. 2006). Moreover, the homologue of CFI-m in Arabidopsis, CFIS25, is also implicated here for positive correlation with APA in introns.

It is surprising to see no gene on the list is significantly negatively correlated with APA in both cases (Supplemental Data S3). This result suggests a possibility that the expression of, rather than lack thereof, certain genes may promote APA in some tissues. Such correlations warrant further studies to reveal more specific relationships between polyadenylation and splicing factors and particular library preference towards APA.

**Conservation of APA between Arabidopsis and rice**

If the APA events identified herein really play a role in gene expression regulation or significantly increase proteomic diversity, they would be expected to be conserved in homologous genes in closely related plant species. To test this hypothesis, we compared orthologous genes with APA sites in exons or introns (APA-classes 3 through 7; Fig. 4-1) since they are likely to produce shorter than normal transcripts. Because protein sequences are more conserved, we used the percentage of proteins produced by these transcripts as a measurement. We consider that ±10% difference of final polypeptide length may be tolerated in the homolog group of APA. Using known orthologous groups downloaded from OrthoMCL (Fulton et al., 2006), a total of 3,034 genes (listed in Supplemental Data S4) were found to have conserved APA sites shared by orthologous genes. To further study whether certain functional gene groups are biased towards APA, we examined the relationships between gene functions based on Gene Ontology and their APA configurations. We included genes having conserved poly(A) sites between two

species and compared each functional group's proportion to all Arabidopsis genes with APA. The most significant changes were observed in category of cellular location (Figure 4-5A). Genes located in plastid and plasma membrane are enriched in conserved APA. In the category of biological processes, many conserved genes are stress related, suggesting the potential role of APA in stress responses (Figure 4-5B). In terms of molecular functions, genes having conserved APA events show slight enrichemnt in transferase and kinase activity (Figure 4-5C).

## *Discussion*

Advancements in DNA sequencing technologies provide us with a large number of DNA sequence information for targeted interrogation of the genome level transcriptome landscape (Siddiqui et al. 2007). In this paper, we utilized this opportunity to study the extent of APA in two plant species, which resulted in many novel sites that were not known before. Based on signatures from MPSS we found 11,264 and 12,091 genes potentially undergo APA which corresponding to 59% and 47% of rice and Arabidopsis genes, respectively. We then further refined the rice APA profile using tags from SBS and found that 24,877 (82%) rice genes undergo APA. This number is significantly higher than the previous estimated extent of APA using EST, which was believed to be around 50% (Shen et al. 2008a), demonstrating the power of current generation sequencing methods in finding new APA events. In addition, analyses based on MPSS and SBS provide us with more information in these APA events found in exons and introns. About 90% of APA genes found by ESTs have all their poly(A) sites located in the 3'-UTR, and most of them are only within dozens of nucleotides away from each other. MPSS and SBS signatures, however, tend to overlook these closely located poly(A) sites and cluster them together based on their immediate 5' *Dpn*II site. In rice, the EST based method found about 1,000 APA events located in CDS, introns or 5'UTR. In contrast, using the MPSS and SBS methods, over 20,000 and 97,000 events (APA-classes 3-8) were found, respectively. This suggests MPSS and SBS based analyses are more sensitive in finding

non-conventional APA events.

These sequencing methods, however, do have some disadvantages compared with classical EST collections. First, MPSS and SBS signatures do not provide information on the exact locations of poly(A) sites. This makes it hard to determine polyadenylation signals based on the location of poly(A) sites. Second, neighboring poly(A) sites are likely to be detected by the same signatures, thus overlooking the complexity of APA. The poly(A) site(s) could span a few nucleotides to a few hundred nucleotides over the genome and still be detected by the same signatures. Thus, other information is required to determine the possible range where poly(A) sites could located. Third, since tags from our MPSS and SBS datasets are usually short sequences so they are more vulnerable to sequence error or repeat sequences in the genome compared to long EST sequences. To overcome these limitations, we used strict filters (signatures must be uniquely mapped to the genome and have been sequenced at least three times) to eliminate unreliable signatures. Furthermore, we cross-checked the validity of MPSS signatures and SBS tags studied in this paper, since both methods use similar sample preparing procedures. About 86.5% (29,593 out of 34,195) of MPSS signatures can also be detected in the rice SBS data set, indicating that most MPSS tags were derived from real poly(A) sites.

Previous MPSS based analysis found that about 25% of Arabidopsis genes were alternatively polyadenylated (Meyers et al. 2004c). Our current analysis found more genes have more than one signature than previously calculated. This is mainly due to the factor that more libraries (17 compared with the original 5) were taken into consideration, so that more library-specific APA events are included in this analysis. In addition, in this paper we focused on the location of poly(A) sites instead of location of signatures. All signatures located in exons were grouped in one class, while we refined signatures in the class into several different APA-classes. Furthermore, we categorized signatures based on their locations and the regions of poly(A) sites located to find possible APA events that would produce truncated transcripts.

The determination of a poly(A) site is influenced by the interactions between *cis*-elements in the pre-mRNA sequence and *trans*-acting factors - the polyadenylation machinery - in the cell (Danckwardt et al. 2008). Many APA events usually are limited in certain libraries, thus behaving in a tissue-specific manner (Figure 4-4). The components in the polyadenylation apparatus thus may play an important role because sequences in pre-mRNA in different libraries were transcripts from the same genome. The current data provide us an opportunity to study how polyadenylation *trans*-acting factors affect the choices of poly(A) sites. First, expression levels of each polyadenylation factors were obtained simultaneously as APA genes, thus reflecting the real time conditions in the cells. Second, all libraries have been normalized so that the comparison between libraries would be robust. This is a novel observation on how plant polyadenylation factors affect APA in a large-scale, the results of which suggest that perhaps the expression levels of several factors affect APA. To our surprise, some splicing factors were also found to be important to the decision making of APA in the library. The presence of these factors increases the likelihood of cells undergoing APA. Of course, our data only open a new avenue to think of the relationship of the expression levels of these *trans*-acting factors and APA. Further studies are needed to support such a hypothesis.

Finally, we have compared APA gene between Arabidopsis and rice and found that over 3,000 gene pairs are orthologues and have APA sites in similar regions of the genes. This suggests that at least some of the APA events might play important regulatory roles, and that these regulation pathways are conserved across different species. Analysis based on GO indicates that in some subcellular compartments such as plasmid and plasma membrane and some stress related genes, genes with conserved APA are overrepresented. It is possible that a local concentration of *trans*-acting factors or activation of these factors to certain responses determines the choice of mRNA 3'-end. The utility of these different poly(A) sites may alter the inclusion or exclusion of some regulatory elements that may have an impact on transcript stability, among other possibilities. Examples have

been recently demonstrated in the APA of cancer cells (Mayr and Bartel 2009)

## *Materials and Methods*

### Sequencing data retrieving and processing

For MPSS analysis in Arabidopsis, we downloaded information of 17-base signatures (17bp_summary.txt) from MPSS plus database http://mpss.udel.edu/at/, which contains 297,313 rows of genome mapping and expression information. To ensure our data free from unreliable sequencing result, we only kept signatures uniquely mapped to the genome and having an expression level larger than TPM>3. Then the number of signatures in each gene was calculated and only signatures from genes with more than one signature were used in further analysis. A total of 36,403 signatures from 11,263 genes were used in this analysis. For rice, we used the same procedure to filter signatures and 34,195 distinct signatures from 12,091 rice genes were analyzed. For rice SBS data, we started from 5,522,207 signatures from 48 libraries and applied the same filter which resulted in 178,555 tags for further analysis. All data were saved in MySQL database and a series of Perl scripts were used for the following analyses.

In all signatures we analyzed, most of them were previously assigned as classes within exons, on the same DNA strand (Meyers et al., 2004a). In order to better categorize MPSS signatures according to both their locations and the poly(A) sites from which they were derived from, we developed a method to differentiate these MPSS signatures and grouped them based on APA-classes as detailed in Figure 4-2 and the text. In this APA-classification system, signatures of APA-class 1 are located in the 3'-UTR, thus inferring the poly(A) sites were also located in the 3'-UTR of the gene. This represents the conventional poly(A) sites and is the most abundant class in all 8 APA-classes. Signatures of APA-class 2 were located upstream of the stop codon while there are no other possible *Dpn*II sites between the signature and the stop codon. Signatures from

APA-classes 3 to 8 were all located upstream of the stop codon and there was at least one *Dpn*II site in between. All signatures in APA-classes 3, 4 and 5 were located in an exon. The difference between these three classes is the possible poly(A) sites derived from class 3 that extend over the exon where the signature was located, while the poly(A) sites of latter two classes could be positioned into a specific exon. Class 5 was more specifically designated to these poly(A) sites located in the last exon. Signatures from APA-classes 6 and 7, however, were located in introns where poly(A) sites of APA-class 6 signatures can also be positioned into that particular intron. Poly(A) sites of class 7 might be located beyond that particular intron where the signature was located. The last group of MPSS signatures is APA-class 8, which were located in the 5'UTR of the gene.

**Library specific data analysis**

To further study the usage of alternative poly(A) sites in each library, we used a method similar to GAUGE discussed in Zhang et al.(2005) with small modifications. In each library, the expression values of each signature (TPM value) were grouped based on their APA-classes. To normalize the differences between libraries, TPM values were first divided by the total number of signatures in that library. The percentages of usage of an APA-class were calculated by comparing to the sum of all signatures in that library. For each APA-class in a library, its percent of usage was compared with the average percent of usage of this particular APA-class in all libraries. The difference was normalized to the average and called the relative distance. A Chi-squared test was tested in each library against the null hypothesis that the usage of a given APA-class in this library is not different from the mean usage (complete lists of values for all libraries are provided in Supplemental Data S2).

To study which *trans*-acting factors play a role in determination of usage of APA in Arabidopsis, we looked for the known poly(A) and splicing factors (Hunt et al., 2008;

Campbell et al. 2006). For each library, expression levels of each factor were determined by the total number of TPMs (without considering the number of signatures each gene has) and expression values from all libraries were compared with relative distance calculated in the previous section. Spearman correlation values for each gene were calculated by comparing the expression value of the gene with the preference of APA discussed in previous paragraph. We grouped preference into two groups: Exon related group includes signatures from APA-classes 4, 5 and 6 while intron related group wasfrom APA-classes 6 and 7. Genes with p-value <0.1 were considered significant.

**Homologous analysis between rice and Arabidopsis**

The known orthologous groups were downloaded from orthoMCL (Chen et al. 2006) and examined for APA events leading to produce same percentage of truncated proteins in rice and Arabidopsis. A 10% difference in percentage of final protein length is tolerated, since MPSS signatures do not provide the exact location of the poly(A) sites. In total, 3,034 genes were found have conserved APA sites shared by orthologous genes. We further studied the relationship between gene functions (Gene Ontology data) and their APA configuration by comparing function of these 5,063 genes with all Arabidopsis genes using TAIR's GO web portal (http://www.arabidopsis.org/tools/bulk/go/index.jsp).

## *ACKNOWLEDGMENTS*

The authors wish to acknowledge Jianti Zheng and Michael Hughes for the help with statistical analyses. This project was funded in part by a grant from Committee for Faculty Research, Miami University (MU), and Academic Challenge grants from MU Botany Department. YS was funded in part by the MU Cell, Molecular and Structure Biology graduate program Research Assistantship, and MU Botany Department's dissertation fellowship.

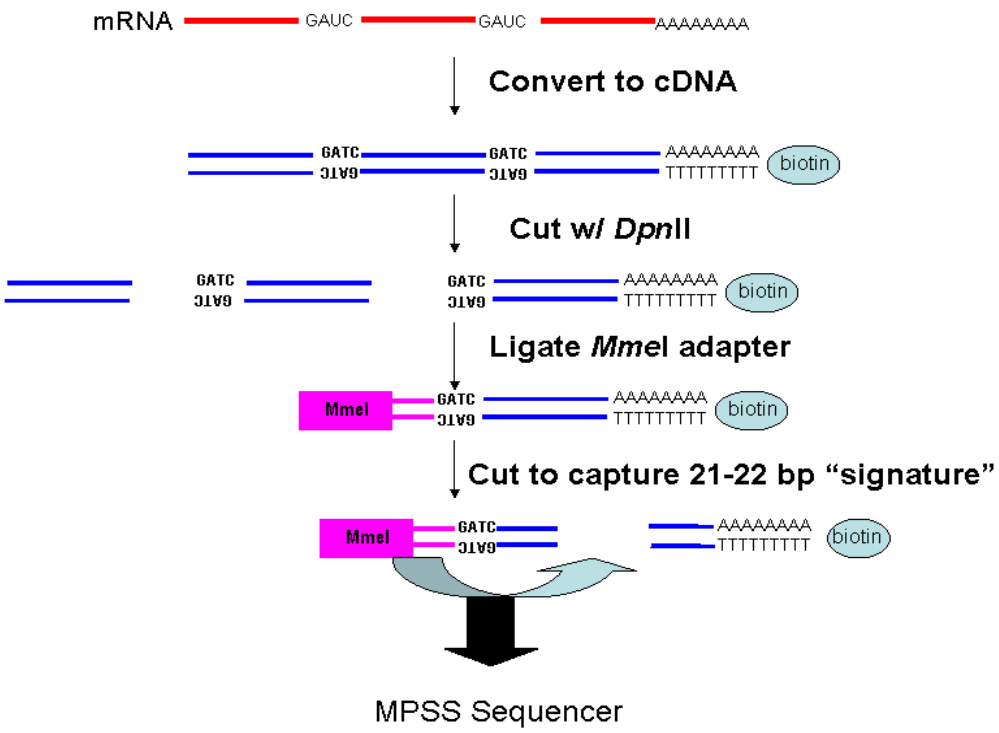



Figure4-1: Sample preparation for MPSS sequencing.

mRNAs were isolated and reverse-transcribed to cDNA with biotin tag attached to oligo-d(T) primer. *Dpn*II enzyme was then used and only cDNA fragment closest to poly(A) tail will be selected for the next step. A *Mme*I adapter were added to 5' of cDNA fragments and resulting fragments were digested with *Mme*I, which cuts 21-22 bp downstream of the recognition site, to generate 21-22bp short fragments called signatures. All signatures were then sequenced by Solexa MPSS sequencer directly.

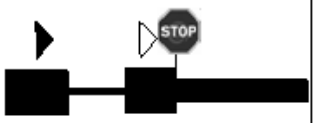
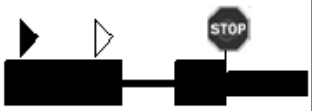
| APA-class | Alternative transcript event | Graphic representation | Arabidopsis MPSS | Rice MPSS | Rice SBS |
|---|---|---|---|---|---|
| 1 | 3'-UTR |  | 9,400 (21%; 40.1%) | 21,810 (41%; 53.1%) | 55,453 (32%; 83.6%) |
| 2 | Upstream of stop codon |  | 10,247 (23%; 50.3%) | 7,504 (14%; 29.8%) | 20,238 (12%; 10.3%) |
| 3 | Exclusively upstream of stop codon |  | 11,817 (27%; 6.9%) | 7,938 (15%; 5.7%) | 28,809 (17%; 1.7%) |
| 4 | Exclusively in an exon except the last one |  | 4,591 (10% 0.7%) | 6,400 (12%; 0.84%) | 11,961 (7%; 0.3%) |
| 5 | Last coding Exon |  | 5,002 (11%; 0.9%) | 3,707 (6% 0.6%) | 19,326 (11%; 1.5%) |
| 6 | Exclusively Intron |  | 102 (0.2%; 0.02%) | 957 (2%; 0.4%) | 14,517 (8%; 1.2%) |
| 7 | In intron |  | 1027 (2%; 0.8%) | 1,787 (3%; 0.3%) | 12,277 (7%; 0.6%) |
| 8 | 5'UTR |  | 1,578 (4%; 0.4%) | 3,069 (6%; 1.9%) | 10,590 (6%; 0.5%) |

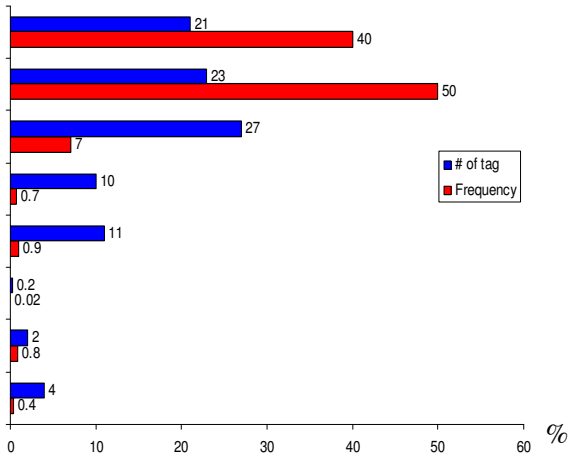

Figure 4-2: Distribution of the APA-classes and graphic illustrations of the locations of MPSS and SBS tags.

The graphic symbols are annotated on the bottom. Tag location marks where the tag is mapped on the gene. The next *Dpn*II site (open triangle) indicates a potential tag location but it is not found in the tag pool, meaning that a poly(A) site should be located between the tag and the next *Dpn*II site (Meyers et al. 2004b). The numbers on the right

three columns are the distribution of the MPSS or SBS tags in each APA-class. The percentages in the parenthesis are the fraction of each class as defined below: the first % is the tag abundance (the number of unique signatures in this class divided by the total number of unique signatures of all libraries); The second % is the transcript abundance (for MPSS, TPM value divided by total TPM value; in SBS, count of sequences divided by the sum of all TPM in that class).

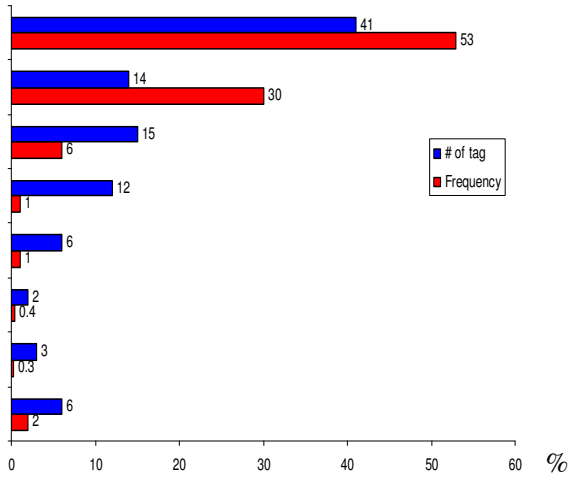
**A.**

| APA- | Alternative transcript event |
|---|---|
| 1 | 3'-UTR |
| 2 | Upstream of stop codon |
| 3 | Exclusively upstream of stop codon |
| 4 | Exclusively in an exon except the last one |
| 5 | Last coding Exon |
| 6 | Exclusively Intron |
| 7 | In intron |
| 8 | 5'UTR |




**B.**

| APA- | Alternative transcript event |
|---|---|
| 1 | 3'-UTR |
| 2 | Upstream of stop codon |
| 3 | Exclusively upstream of stop codon |
| 4 | Exclusively in an exon except the last one |
| 5 | Last coding Exon |
| 6 | Exclusively Intron |
| 7 | In intron |
| 8 | 5'UTR |




**C.**

| APA- | Alternative transcript event |
|---|---|
| 1 | 3'-UTR |
| 2 | Upstream of stop codon |
| 3 | Exclusively upstream of stop codon |
| 4 | Exclusively in an exon except the last one |
| 5 | Last coding Exon |
| 6 | Exclusively Intron |
| 7 | In intron |
| 8 | 5'UTR |

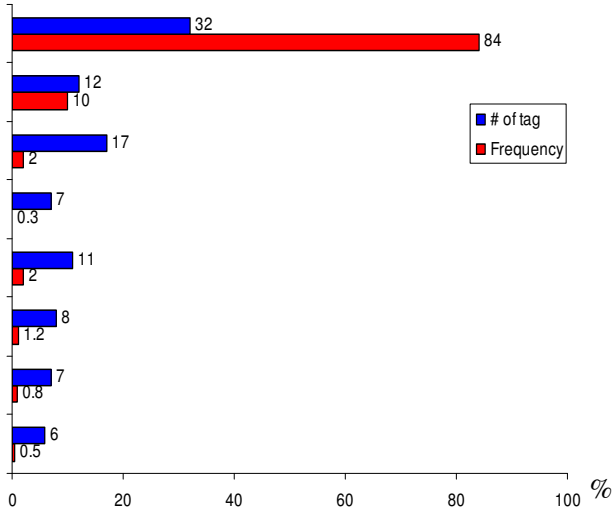


Figure 4-3: Number and frequency of different APA-classes.

The blue bars show percentage of signatures/tags in this class in term of unique signature/tag number. The red bars indicate percentage in term of frequency of signatures/tags in this class compared to frequency of all signatures. A. MPSS signatures from Arabidopsis. B. MPSS signatures from rice. C. SBS signatures from rice

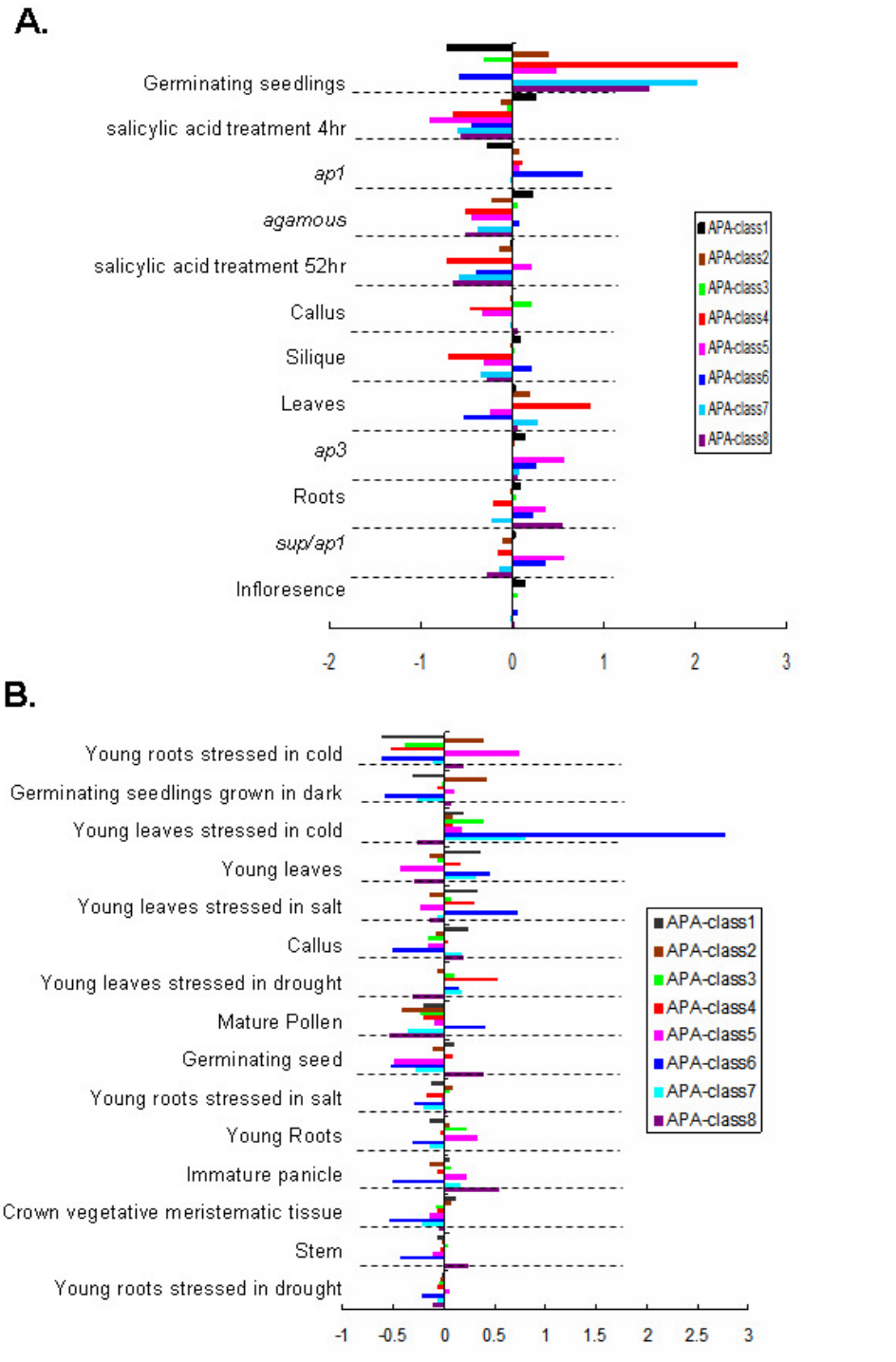



Figure 4-4: Tissue specific usage of different APA-classes.

X-axis shows the relative distance (or difference) to the average usage of all libraries available. The library designations are on the left. A. Libraries from Arabidopsis. B. Libraries from rice. Libraries most significantly different from average usage of APA are on top.

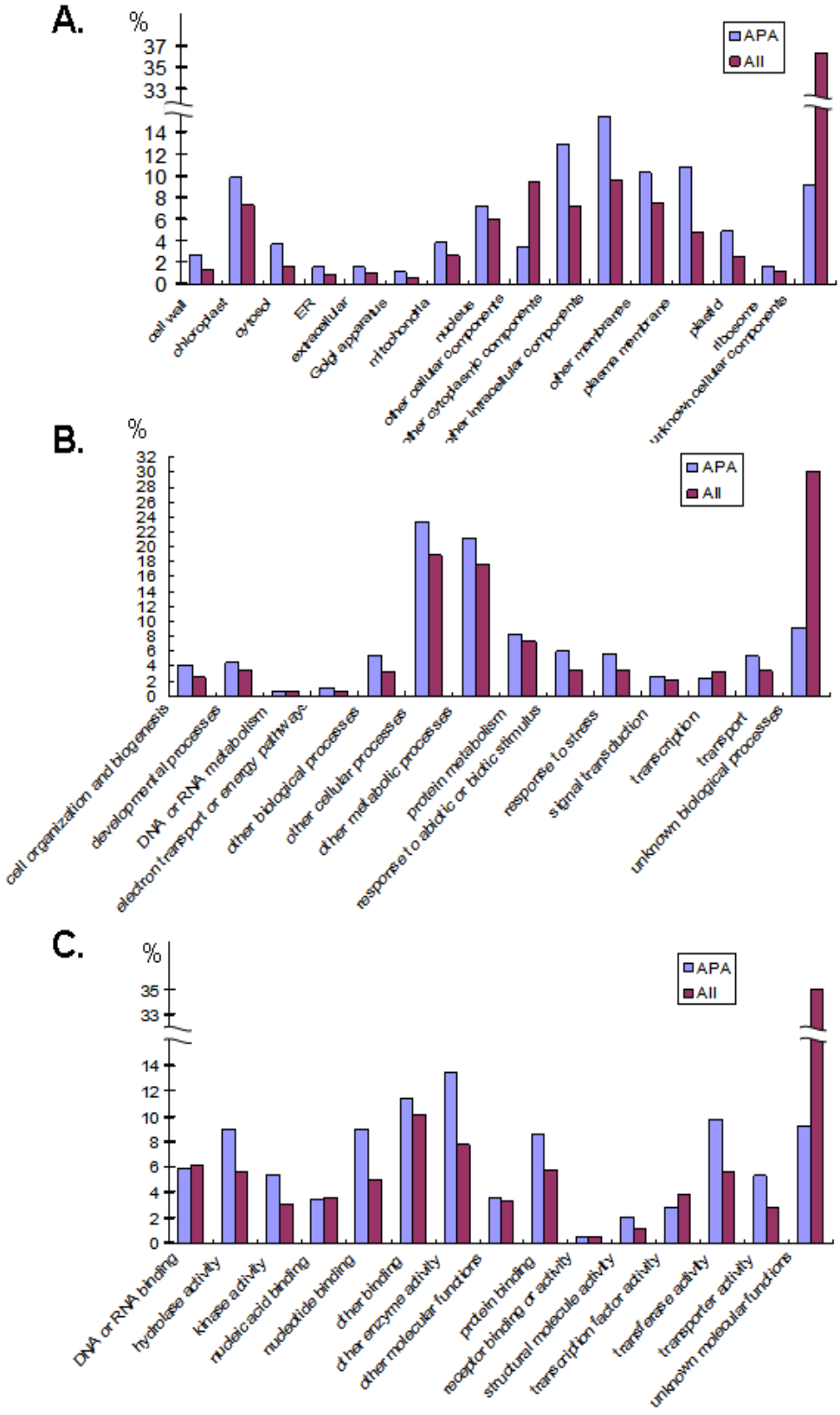

A.
%
APA
All
cell wall
chloroplast
cytosol
ER
extracellular
Golgi apparatus
mitochondria
nucleus
other cellular components
other cytoplasmic components
other intracellular components
other membranes
plasma membrane
plastid
ribosome
unknown cellular components
B.
%
APA
All
cell organization and biogenesis
developmental processes
DNA or RNA metabolism
electron transport or energy pathways
other biological processes
other cellular processes
other metabolic processes
protein metabolism
response to abiotic or biotic stimulus
response to stress
signal transduction
transcription
transport
unknown biological processes
C.
%
APA
All
DNA or RNA binding
hydrolase activity
kinase activity
nucleic acid binding
nucleotide binding
other binding
other enzyme activity
other molecular functions
protein binding
receptor binding or activity
structural molecule activity
transcription factor activity
transferase activity
transporter activity
unknown molecular functions

Figure 4-5: Functional distributions of 3,034 genes identified to have conserved APA between rice and Arabidopsis.

Comparisons were made using the functional distributions of these genes against the general distribution of all genes in Arabidopsis using GO categories. A. Cellular or subcellular localizations. B. Biological processes. C. Molecular functions.

Table 4-1: Numbers of genes with alternative polyadenylation sites as indicated by number of signatures in the gene

| # of signature(s) in a gene | # of genes in Arabidopsis (%) | # of genes in rice found by MPSS (%) | # of genes in rice found by SBS(%) |
|---|---|---|---|
| 1 | 7,824 (40) | 13,409 (52) | 5,411 (18) |
| 2 | 4,228 (22) | 61,96 (24) | 3,457 (11) |
| 3 | 2,371 (12) | 3,413 (13) | 2,872 (9) |
| >=4 | 4,665 (26) | 2,482 (11) | 18,548 (62) |
| Total | 19,088 [a] | 25,500 [a] | 30,288 [b] |

[a] The total numbers are from Meyers et al 2004. (Arabidopsis) and Nobuta et al. 2007 (rice).

[b] The total numbers are from number of genes have at least a unique tag mapped.

| Gene | Name | p-value-exon[a] | rho-value exon[b] | p-value-intron[a] | rho-value-intron[b] |
|---|---|---|---|---|---|
| At4g32850 | PAPS4 | 0.068633 | 0.545455 | 0.044262 | 0.594406 |
| At4g36690 | U2AF65 | 0.01709 | 0.67018 | - | - |
| At1g66500 | PCFS1 | 0.075394 | 0.53142 | - | - |
| At5g10350 | PABN3 | 0.07708 | 0.531469 | - | - |
| At5g51120 | PABN2 | - | - | 0.009865 | 0.708776 |
| At1g27595 | SYM2 | - | - | 0.056984 | 0.562398 |
| At4g25550 | CFIS2 | - | - | 0.072939 | 0.535225 |

Table 4-2: List of polyadenylation or splicing related proteins are significantly correlated with the usage of APA in exons or introns.

[a] p-value and [b]rho-value (Spearman ranking) were calculated based on expression values of the protein coding genes and APA usages in the 12 Arabidopsis libraries using Spearman statistic methods.

## *REFERENCES*


Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M et al. 2000. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature biotechnology* **18**: 630-634.

Campbell MA, Haas BJ, Hamilton JP, Mount SM, Buell CR. 2006. Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis. *BMC Genomics* **7**: 327.

Chen F, Mackey AJ, Stoeckert CJ, Jr., Roos DS. 2006. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res* **34**(Database issue): D363-368.

Danckwardt S, Hentze MW, Kulozik AE. 2008. 3' end mRNA processing: molecular mechanisms and implications for health and disease. *EMBO J* **27**: 482-498.

Elling AA, Deng XW. 2009. Next-generation sequencing reveals complex relationships between the epigenome and transcriptome in maize. *Plant Signal Behav* **4**: 760-762.

Fox S, Filichkin S, Mockler TC. 2009. Applications of Ultra-high-Throughput Sequencing. *Methods Mol Biol* **553**: 79-108.

Hunt AG, Xu R, Addepalli B, Rao S, Forbes KP, Meeks LR, Xing D, Mo M, Zhao H, Bandyopadhyay A et al. 2008. Arabidopsis mRNA polyadenylation machinery: comprehensive analysis of protein-protein interactions and gene expression profiling. *BMC Genomics* **9**: 220.

Irizarry RA, Ladd-Acosta C, Carvalho B, Wu H, Brandenburg SA, Jeddeloh JA, Wen B, Feinberg AP. 2008. Comprehensive high-throughput arrays for relative methylation (CHARM). *Genome Res* **18**: 780-790.

Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K. 2009. SNP detection for massively parallel whole-genome resequencing. *Genome Res* **19**: 1124-1132.

Lutz CS. 2008. Alternative polyadenylation: a twist on mRNA 3' end formation. *ACS Chem Biol* **3**: 609-617.

Meyers BC, Lee DK, Vu TH, Tej SS, Edberg SB, Matvienko M, Tindell LD. 2004a. Arabidopsis MPSS. An online resource for quantitative expression analysis. *Plant Physiology* **135**: 801-813.

Meyers BC, Tej SS, Vu TH, Haudenschild CD, Agrawal V, Edberg SB, Ghazal H, Decola S. 2004b. The use of MPSS for whole-genome transcriptional analysis in Arabidopsis. *Genome Research* **14**: 1641-1653.

Meyers BC, Vu TH, Tej SS, Ghazal H, Matvienko M, Agrawal V, Ning J, Haudenschild CD. 2004c. Analysis of the transcriptional complexity of Arabidopsis thaliana by massively parallel signature sequencing. *Nature Biotech* **22**: 1006-1011.

Millevoi S, Loulergue C, Dettwiler S, Karaa SZ, Keller W, Antoniou M, Vagner S. 2006. An interaction between U2AF 65 and CF I(m) links the splicing and 3' end processing machineries. *EMBO J* **25**: 4854-4864.

Ner-Gaon H, Halachmi R, Savaldi-Goldstein S, Rubin E, Ophir R, Fluhr R. 2004. Intron

retention is a major phenomenon in alternative splicing in Arabidopsis. *Plant J* **39**: 877-885.

Nobuta K, Vemaraju K, Meyers BC. 2007. Methods for analysis of gene expression in plants using MPSS. *Methods Mol Biol* **406**: 387-408.

Shen Y, Ji G, Haas BJ, Wu X, Zheng J, Reese GJ, Li QQ. 2008a. Genome level analysis of rice mRNA 3'-end processing signals and alternative polyadenylation. *Nucleic Acids Res* **36**: 3150-3161.

Shen Y, Liu Y, Liu L, Liang C, Li QQ. 2008b. Unique features of nuclear mRNA poly(A) signals and alternative polyadenylation in *Chlamydomonas reinhardtii*. *Genetics* **179**: 167-176.

Shi Y, Di Giammartino DC, Taylor D, Sarkeshik A, Rice WJ, Yates JR, 3rd, Frank J, Manley JL. 2009. Molecular architecture of the human pre-mRNA 3' processing complex. *Mol Cell* **33**: 365-376.

Siddiqui N, Mangus DA, Chang TC, Palermino JM, Shyu AB, Gehring K. 2007. Poly(A) nuclease interacts with the C-terminal domain of polyadenylate-binding protein domain from poly(A)-binding protein. *J Biol Chem* **282**: 25067-25075.

Simon SA, Zhai J, Nandety RS, McCormick KP, Zeng J, Mejia D, Meyers BC. 2009. Short-read sequencing technologies for transcriptional analyses. *Annu Rev Plant Biol* **60**: 305-333.

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.

Tian B, Pan Z, Lee JY. 2007. Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. *Genome Res* **17**: 156-165.

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. *Plant J* **54**: 899-910.

Zhang H, Lee JY, Tian B. 2005. Biased alternative polyadenylation in human tissues. *Genome biol* **6**: R100.

Zhao J, Hyman L, Moore C. 1999. Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. *Microbiol Mol Biol Rev* **63**: 405-445.

# CHAPTER 5: CONCLUSIONS

As one of the crucial steps in the biogenesis of mRNA, polyadenylation plays a significant role in mRNA stabilization, transportation and translation (Danckwardt et al. 2008). Emerging evidence suggests it is also associated with many other aspects of development processes such as transcription termination, splicing, etc. (Proudfoot 2004; Andreassi and Riccio 2009). Studying the mechanism of polyadenylation and its regulation is important to understand how cells maintain an accurate control of the mRNA status and respond differently to developmental or environmental signals.

In mammalian and yeast cells, polyadenylation signals have been studied for years. Key signal elements have been defined by experiments and computational studies. In plants, polyadenylation signals had been carefully studied only in a few genes genetically and one species bioinformatically (Rothnie 1996; Li and Hunt 1997; Rothnie et al. 2001; Loke et al. 2005) when this dissertation work was initiated. In Arabidopsis, three major groups of poly(A) signals are known: far upstream elements (FUE), near upstream elements (NUE; an AAUAAA-like element) and cleavage element (CE) (Loke et al. 2005). Substantial differences in nucleotide profile near poly(A) sites are observed between plant and mammalian cells, and plant polyadenylation signals are more diverse than those in animals (Zhao et al. 1999; Loke et al. 2005). The canonical hexamer AAUAAA in mammals was found in only about 10% of Arabidopsis genes about 10 to 30-nt upstream of the cleavage site of all 8,000 ESTs examined (Loke et al. 2005). The findings of plant-specific polyadenylation signal profiles warrant further study in other plant species.

The first species I chose to study was rice. This is not only an important model crop plant, but also has a rich resource of genome and transcriptome information. To reveal rice poly(A) signals at a genome level, I first characterized the poly(A) signals using a data set of 55,742 authenticated poly(A) sites constructed by a collaborator from publicly available EST sequences. A high similarity is observed between rice and Arabidopsis,

including similar tripartite *cis*-elements FUE, NUE and CE. I further compared the nucleotide distribution profile near poly(A) sites and found that the general distribution pattern of the four nucleotides is similar to that seen in Arabidopsis. However, FUE region in rice is slightly expanded towards the coding region, possibly caused by a longer 3'-UTR in rice. A statistic tool called RSAT is further used to compare the usage of poly(A) signals between rice and Arabidopsis. RSAT results showed that signals from rice are more over-representative (shorter list of good signals with higher Z-scores – a statistical evaluation score) than those in Arabidopsis, suggesting that monocot plants tend to require stronger signals to guide the cleavage reaction during 3'-end formation.

The second species I studied was the green alga *Chlamydomonas reinhardtii*. The complete genomic sequence is known and a large number of ESTs. More importantly, studies in green algae might shed light on how polyadenylation machinery evolves from single cell photosynthetic organisms to higher land plants. To understand nuclear mRNA polyadenylation mechanisms in *Chlamydomonas* we generated a data set of 16,952 *in silico*-verified poly(A) sites from ESTs. Analysis of this data set revealed a very unique and complex polyadenylation signal profile that is setting *Chlamydomonas* apart from other organisms. In contrast to the high-AU content in the 3'-UTRs of other organisms, *Chlamydomonas* shows a high-guanylate content that transits to high-cytidylate around the poly(A) sites. The average length of the 3'-UTR is 595-nt, significantly longer than that of Arabidopsis and rice (223 and 289, respectively). The dominant poly(A) signal, UGUAA, was found in 52% of the NUE, and its occurrence may be positively correlated with higher gene expression levels. The UGUAA signal also exists in Arabidopsis and in some mammalian genes but mainly in the FUE, suggesting a shift in function during evolution. The C-rich region after poly(A) sites with unique signal elements is a characteristic downstream element that is lacking in higher plants.

Another significant finding of my dissertation is the phenomenlon of APA, which is defined as the capacity for an individual gene to carry multiple sets of poly(A) signals or poly(A) sites. When I mapped rice ESTs to the genome, however, 50% of the genes

analyzed had more than one unique poly(A) site and 13% had four or more poly(A) sites. In *Chlamydomonas*, a range of up to 33% of the 4,057 genes analyzed have at least two unique poly(A) sites. These results suggest that APA plays a universal role across different plant species. To further localize these poly(A) sites, I then fine-mapped ESTs to genomes with coordinates of annotated genes. In rice, about 4% of the analyzed genes were found to possesse alternative poly(A) sites at their introns, 5'-UTRs, or protein coding regions. In *Chlamydomonas*, about 1% of these genes have poly(A) sites residing in predicted coding sequences, introns, and 5'-UTRs. This aspect of my research opened up a new avenue of studying the role of gene expression through APA.

Although my work on rice and *Chlamydomonas* was among the very first to examine the extent of APA in plants, the depth of study was limited by the number of ESTs available in public databases. In each species, I have found about 1,000 poly(A) sites are located in positions other than 3'-UTR. Increasing the number of authenticated poly(A) sites can undoubtedly improve our understanding in plants by discovering rarer but functionally determinant APA events. Using the rich collection of MPSS (Massively Parallel Signature Sequencing) and Illumina's GAII SBS (sequence by synthesis) tags in rice and Arabidopsis, we discovered a large number of APAs that have not been found previously by other methods. In Arabidopsis, we examined 43 million MPSS signatures from 17 libraries and mapped these signatures unambiguously to the annotated genome. We found that 11,264 genes (42% of the protein coding genes) have multiple poly(A) sites based on 43,767 unique MPSS signatures. In rice, we examined 47 millions MPSS signatures from 22 and 48 libraries and found 12,091 (48%) and 24,877 (82%) of rice gene have more than one poly(A) sites from MPSS and SBS tags, respectively. More importantly, in both species, APA events upstream of stop codons (e.g. introns, exons or 5'-UTR) are evident from about 50% of unique signatures, which corresponds to 10% of whole transcriptome abundance. This number is significantly higher compared with percentage obtained from methods such as ESTs (about 2%), suggesting MPSS and Illumina sequencing are very powerful methods for detecting rare APA events. In addition, we studied the possible library specificity of APA events and found that most of the APA sites in regions other

than the 3'-UTR were significantly more specific than poly(A) sites located in the 3'-UTR. These results imply that APA events are likely to regulate gene expression only in certain developmental stages, tissues, or mutation conditions (from which the libraries were generated).

Overall, my dissertation answered several fundamental questions about mRNA 3'-end formation processes: 1) What are the polyadenylation signals in species other than Arabidopsis? 2) What's the degree of APA in the different plant species? Our results in this dissertation further our understanding of regulation of the polyadenylation mechanisms and their potential to affect gene expression by choosing different 3'-ends of mRNA. The availability of information about genes that are regulated by APA should stimulate the research on gene structures and its functions in plant development and environmental responses through a case by case study.

The results presented in this dissertation answered a number of fundamental questions about polyadenylation. However, many additional questions/issues still remain:

1. What is the mRNA 3'-end processing signal in other plant species? Our results suggested that flowering plants might employ similar polyadenylation mechanisms. It will be interesting to see how polyadenylation is regulated in non-vascular plants, seedless plants as well as gymnosperms. Do they use a mechanism in 3'-end formation similar to flowering plants or green algae? What are the evolutionary trends of both polyadenylation signals and mechanisms?

2. The fact that *Chlamydomonas* uses a very different set of poly(A) signals suggests that this algal species might have acquired a unique set of poly(A) machinery since polyadenylation mechanisms are relatively conserved in higher eukaryotic species (Zhao et al., 1999). Whether this acquisition is unique in this species or more universally adapted in other species remains an open question. An examination of other closely related species could answer the question of the origin for this poly(A) machinery.

3. Lastly, we found that APA is universally adapted in plant species. Learning the destinations of these short transcripts with alternate 3'-ends will be an interesting direction to follow. Will these short mRNAs be found to be translated into proteins, or if they are translated, will these proteins be functional? Further biochemical and genetic studies should address the question of destinations for these mRNA and ultimately explain the importance of APA in plants and beyond.

## Reference:


Andreassi C, Riccio A. 2009. To localize or not to localize: mRNA fate is in 3'UTR ends. *Trends Cell Biol* **19**: 465-474.

Danckwardt S, Hentze MW, Kulozik AE. 2008. 3' end mRNA processing: molecular mechanisms and implications for health and disease. *EMBO J* **27**: 482-498.

Li Q, Hunt AG. 1997. The polyadenylation of RNA in plants. *Plant Physiol***115**: 321-325.

Loke JC, Stahlberg EA, Strenski DG, Haas BJ, Wood PC, Li QQ. 2005. Compilation of mRNA polyadenylation signals in Arabidopsis revealed a new signal element and potential secondary structures. *Plant Physiol* **138**: 1457-1468.

Proudfoot N. 2004. New perspectives on connecting messenger RNA 3' end formation to transcription. *Curr Opin Cell Biol* **16**: 272-278.

Rothnie HM. 1996. Plant mRNA 3'-end formation. *Plant Molecular Biology* **32**: 43-61.

Rothnie HM, Chen G, Futterer J, Hohn T. 2001. Polyadenylation in rice tungro bacilliform virus: cis-acting signals and regulation. *J Virol* **75**: 4184-4194.

Zhao J, Hyman L, Moore C. 1999. Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. *Microbiol Mol Biol Rev* **63**: 405-445.