DOI: 10.1534/genetics.108.088971

Unique Features of Nuclear mRNA Poly(A) Signals and Alternative Polyadenylation in *Chlamydomonas reinhardtii*

Yingjia Shen, Yuansheng Liu, Lin Liu, Chun Liang and Qingshun Q. Li[1]

Department of Botany, Miami University, Oxford, Ohio 45056

Manuscript received March 5, 2008
Accepted for publication March 21, 2008

ABSTRACT

To understand nuclear mRNA polyadenylation mechanisms in the model alga *Chlamydomonas reinhardtii*, we generated a data set of 16,952 *in silico*-verified poly(A) sites from EST sequencing traces based on Chlamydomonas Genome Assembly v.3.1. Analysis of this data set revealed a 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) site. The average length of the 3′-UTR is 595 nucleotides (nt), significantly longer than that of Arabidopsis and rice. The dominant poly(A) signal, UGUAA, was found in 52% of 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 a high level of alternative polyadenylation in the Chlamydomonas genome, with a range of up to 33% of the 4057 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, and 5′-UTRs. 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 (Zhao *et al.* 1999; Gilmartin 2005; Hunt 2007). 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 general theme has just emerged after confirmation of the existence of upstream *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 the cleavage element (CE); a strong signal (*e.g.*, AAUAAA found in mammals) ~20–30 nucleotides (nt) upstream from the poly(A) site, which is termed the near upstream element (NUE); a far upstream element (FUE) that is ~40–150 nt upstream of the poly(A) site; and a downstream element located 20–40 nt beyond the cleavage site. In mammalian cells, these four *cis*-elements are all required: the highly conserved NUE in the form of AAUAAA and the seemingly weaker FUE (Venkataraman *et al.* 2005). The downstream elements are typically found only 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 (Li and Hunt 1997; Graber *et al.* 1999; Loke *et al.* 2005).

[1]*Corresponding author:* Botany Department, 316 Pearson Hall, Miami University, Oxford, OH 45056. E-mail: liq@muohio.edu

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 protects 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 an 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 events 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; Peterson 1994; Lou and Gagel 1998; Delaney *et al.* 2006) or the production of a nonfunctional 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 >16,952 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.* 2007a,b, 2008, this issue). Using this large data set, we revealed that Chlamydomonas possesses unique features in polyadenylation signals, including nucleotide composition of 3′-UTRs, signal arrangements and sequence patterns, and substantial 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 data set: A total of 309,278 raw EST traces were obtained from the NCBI Trace Archive (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi), the Chlamydomonas Center (http://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), and the poly(A) sites were authenticated on the basis of the features of each cDNA library construct. The *Chlamydomonas reinhardtii* Assembly v.3.1 genome sequences and corresponding annotation file were downloaded (September 2007) from the Joint Genome Institute of the U.S. 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 error. In addition, the poly(A) tails found in the ends of ESTs must also be immediately followed by an *Xho*I restriction enzyme site and then by an extensible vector fragment 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 the case that 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 indicates that the first A of a poly(A) tail tends to be from transcription rather than added by poly(A) polymerase during polyadenylation (Moore *et al.* 1986; Sheets *et al.* 1990; Chen *et al.* 1995). Once a poly(A) site was identified, 300 nt of sequence upstream plus 100 nt of sequence downstream for each authenticated poly(A) site were extracted. This produced a data set of 56,031 sequences, each with a poly(A) site, and this data set became known as the 56K data set. There were some redundant ESTs in the 56K data set because this data set reflects all the ESTs that were successfully mapped. When redundant ESTs with the same poly(A) sites in the genome were removed, a data set totaling 16,952 unique sequences was generated and called the 17K data set, which was used in most of the analyses presented here. These data sets 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) on the basis of GMAP results. A total of 44,338 ESTs were found to be associated with annotated genes [Joint Genome Institute (JGI) Gene Catalog 3.1] and corresponding to 11,730 nonredundant poly(A) sites. We further categorized these poly(A) sites on the basis of their location on the genes into four 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 introns by manually examining them through the ChlamyEST-

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 (3–12 nt) of sequence patterns with their location information for the 17K data set. 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) data sets from Arabidopsis (Loke *et al.* 2005) and human (provided by 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 (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 from its expected occurrence, also based on Markov chain models (van Helden *et al.* 2000). The results of the calculation are presented as *Z*-scores and ranked according to 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 in 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 a *Z*-score ≥8.53, except for the FUE, where a *Z*-score ≥5 is used) on the basis of their mutual distance in which a gap was not allowed. Then, an agglomeration package from the R program (http://www.r-project.org) called *Agnes* was used to cluster patterns on the basis of their dissimilarity distances. A cutoff value of 2.6 was used to group these patterns, as suggested in Hu *et al.* (2005), and they were aligned by using ClustalW. The size of a sequence logo was determined on the basis of 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 on the basis of 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.

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: $S = \sum_{p=1}^{L} \log_2(f(n, p)/f(n))$, 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) signal region, *e.g.*, the 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. 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 on the basis of the assumption that the same NUE could control more than one poly(A) site in the range of ~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 nonsignals, 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 toward a certain size of *cis*-elements was calculated on the basis of the difference between observed occurrence and expected value. The predicted values were calculated on the basis of 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 the observed is calculated using this formula: Score = $([\Sigma\ \mathrm{Obs}_{n+1}[(\mathrm{A/T/C/G,A/T/C/G}) - \mathrm{Obs}_n/2)]/\mathrm{Obs}_n) \times 100\%$, where Obs_n is the observed value of the first pattern size, Obs_{n+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 Obs_{n+1}. The results of this analysis are presented in supplemental Figure 1. The size with the highest score was used for SignalSleuth and RSAT analyses.

RESULTS

The poly(A) site and 3′-UTR data set 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 an adapter 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). On the basis of these *in silico*-authenticated poly(A) sites, a data set with 16,952 genomic sequences of 400 nt each was generated, where each sequence has a poly(A) site located at nucleotide 300 (from left to right; the poly(A) site nucleotide is referred to as −1 position hereafter). This data set, termed the 17K data set, represents the largest poly(A) site collection in algae to date.

The profile of the 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 data set. As shown in Figure 1, the 3′-UTR of Chlamydomonas is notably rich in G nucleotides, except the −25 to −5 region where U and A are

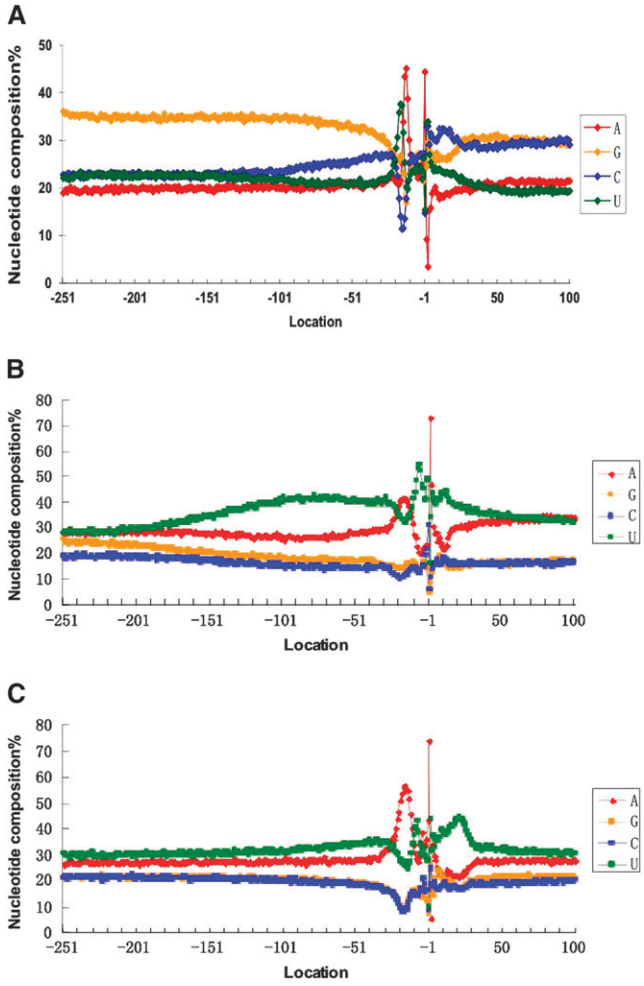

Figure 1.—Single-nucleotide profiles of the 3′-UTR for different species. (A) Chlamydomonas; (B) Arabidopsis; (C) human. The poly(A) site is at position −1. The upstream sequence (300 nt) of the poly(A) site is in "−" designation, and the downstream (100 nt) sequence is in "+" designation.

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 1B; Loke *et al.* 2005) and rice (Dong *et al.* 2007; Shen *et al.* 2008), as well as yeast (Graber *et al.* 1999) and human (Figure 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 the 3′-UTR. The previously known YA dinucleotide (Y equals C or U) at the cleavage site (Graber *et al.*1999; Loke *et al.* 2005; Tian *et al.* 2005) is also missing in Chlamydomonas with only the A nucleotide showing at the cleavage site.

The average length of the 3′-UTR in Chlamydomonas is 595 nt, which is calculated on the basis of the distance between the annotated (JGI draft gene catalog 3.1) 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.* 2008), as shown in Figure 2. The longer 3′-UTR may also reflect its less compact genome, the size of which (120 Mb) is similar to Arabidopsis, but predicted to encode only half the number (~15,000) of genes (Arabidopsis Genome Initiatives 2000; Merchant *et al.* 2007).

Nuclear polyadenylation signal regions in Chlamydomonas: On the basis of the scanning results of SignalSleuth, we plotted top-ranked signal profiles in each section of the poly(A) signal regions (Figure 3). The full list of signals is in supplemental Table 1. 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 the FUE; −25 to ~ −5 for the NUE, and −5 to ~ +5 for the CE. One of the most notable features of Chlamydomonas polyadenylation signals is its NUE, where UGUAA is overrepresented (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 the FUE are apparently G rich. As noted above, at the cleavage site, the YA dinucleotide is replaced by the A nucleotide, which is unique to Chlamydomonas. 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 regions, 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 the FUE and the NUE, respectively. Heptamers and hexamers have the biggest deviation in the region of −5 to +5 and +5 to +30 for the CE and the 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 RSAT (Van Helden 2003). The full results of this analysis are listed in supplemental Table 2. 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 *et al.* 2000). Poly(A) signals with higher *Z*-scores are likely to be more significant in determining the position of poly(A) sites.

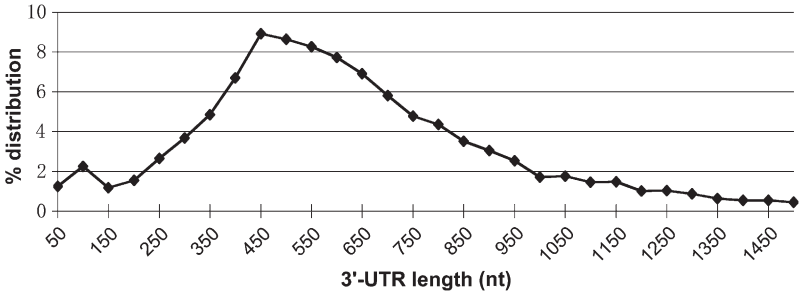

FIGURE 2.—Distribution of the length of the 3′-UTR in Chlamydomonas. The average length is 595 nt.

The UGUAA signal has a very high *Z*-score and the most occurrence, and this finding is supported by the SignalSleuth results in which UGUAA is also highly overrepresented in the NUE (Figure 3). Compared to the predominant AAUAAA signal, which occurs in only ~10% of genes in Arabidopsis (LOKE *et al.* 2005), Chlamydomonas genes use highly conserved UGUAA poly(A) signals with ~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 RSAT. This is supported by previous findings 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 extremely 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 data set. Interestingly, for those NUEs that do not have UGUAA,

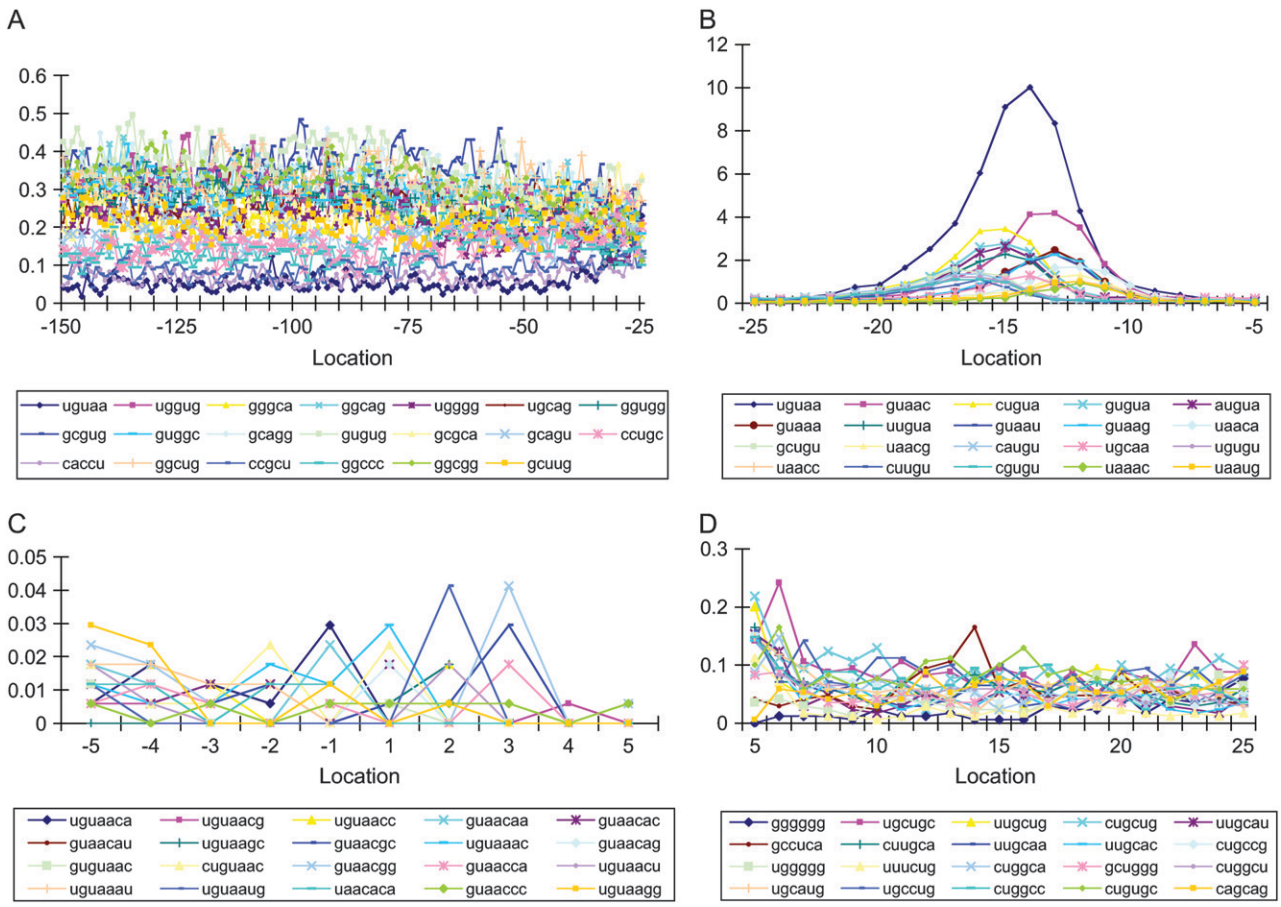

FIGURE 3.—The 20 top-ranked signals in the designated poly(A) signal regions. (A) Pentamers from −150 to −25 in FUEs. (B) Pentamers from −25 to −5 in NUEs. (C) Heptamers from −5 to +5 in CEs. (D) Hexamers from +5 to +30 in DEs. The poly(A) site is at position −1. The upstream sequence of the poly(A) site is in "−" designation, and the downstream sequence is in "+" designation.

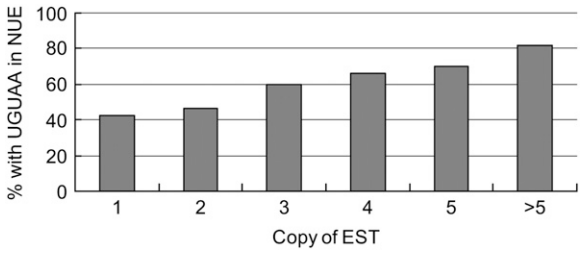

Figure 4.—Correlation of UGUAA and gene expression level. The higher the expression level is (more EST copies), the better the chance is of having UGUAA in their NUE as a poly(A) signal.

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 data set with 56,031 ESTs. As shown in Figure 4, 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.

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 5). In contrast, AAUAAA is also dominant in the 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 may assume a lesser role (Figure 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 >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).

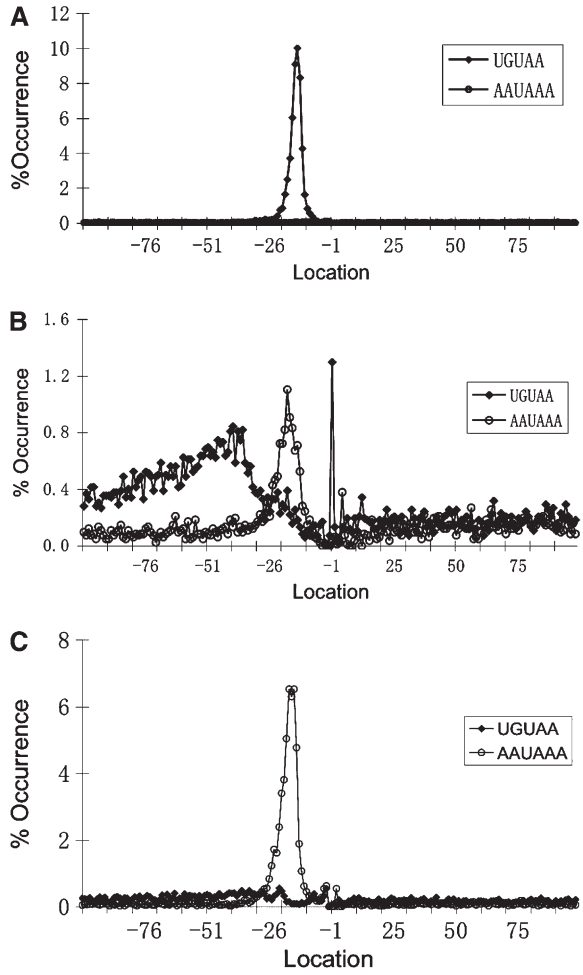

Figure 5.—Distribution of UGUAA and AAUAAA signals in different species. (A) Chlamydomonas; (B) Arabidopsis; (C) human. The region of −100 to +100 surrounding the poly(A) site is shown. The poly(A) site is at position −1.

These sequence patterns were clustered to generate sequence logos according to their similarity (Shen *et al.* 2008). Using this method, we identified six 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 wide region of 3′-UTRs (120–150 nt). For elements in the relative short region, one of the two NUE logos is the most conserved with >78% of the genes containing a UGUAA-related element. Signals in the CE and the DE are less conserved, as each of their logos covers only 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 transcriptomes and proteomes and contributes to gene expression regulation. To study the extent of APA in Chlamydomonas, we first studied the number of

TABLE 1

Concise representation of polyadenylation *cis*-elements in Chlamydomonas as sequence logos

Region	Signal logo	Name	No. of signals	Top signal	% of hits
FUE −150/−25		FUE.1	5	CGGUA	98
		FUE.2	2	UUACA	33
NUE −25/−5		NUE.1	4	UGUAA	78
		NUE.2	5	AGUAU	9
CE −5/+5		CE.1	6	UACCGUA	6
DE +5/+30		DE.1	2	ACCGUA	13

poly(A) sites for each gene on the basis of the 17K data set (Table 2). 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 33% (1341 of 4057) of the Chlamydomonas genes have two 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 data set (Table 2). A conservative estimate would be 9% (1341 divided by the total predicted 15,000 genes, if no more APA is found in the rest of the genes).

TABLE 2

Number of genes with unique alternative poly(A) sites

No. of poly(A) site(s) per mRNA	No. of genes[a]	%
1	2716	67
2	913	23
3	296	7
4 or more	132	3
Sum	4057	100

[a] These genes are selected from those that have at least one authenticated poly(A) site from the 17K data set. Note that the genes considered here are only a part of the total gene number from the gene catalog of ~15,000.

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 polyadenylation site was on the same strand as the model to which it was attributed. After these steps, 44,338 ESTs corresponding to 11,730 nonredundant 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 poly(A) site either to the gene model in which it lies or to the immediate upstream model if it is <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). We realize, though, that such a conclusion (extensive APA in the CDS,

TABLE 3

The distribution of poly(A) sites on gene transcipts

Category	Subcategory	No. of transcripts	%
Total poly(A) sites	—	16,952	100
Located in the transcript (full-length cDNA)	In CDS	45 (12)[c]	0.3
	In introns	588 (39)[c]	3.5
	In 5′-UTR	86 (12)[c]	0.5
	In 3′-UTR[b]	11,011	65.0
	Subtotal	11,730	69.2
Located in the intergenic region[a]	—	5,222	30.8

[a] The intergenic regions are the areas 500 or 1000 nt downstream of the 3′-UTR (as defined above).

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

[c] The numbers in parentheses are the cases with high confidence when using more stringent conditions and manual confirmation as described in the main text.

introns, and 5′-UTRs) is drawn on the basis of the current draft genome annotation information, from which discrepancies between those gene models and ESTs 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; and 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 a total of 11,730 poly(A) sites mapped on transcripts, Table 3] unique APA sites with high confidence in the CDS, introns, and 5′-UTRs that satisfy either condition (listed in supplemental Table 3). If the number of genes is considered, this represents 44 or ~1% of the 4057 genes with 3′-UTR EST supports (Table 2).

Many of the poly(A) sites (30.8%) are found >500 nt from a currently annotated gene, indicating 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 data set offers a rich resource for such explorations.

DISCUSSION

In this article, we were able to process and authenticate 16,952 poly(A) sites for the analysis of poly(A) signals and to estimate 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 substantial in Chlamydomonas, with up to one-third of the 4057 genes analyzed 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. We realize that our analysis would gain most by more accurate annotation of the Chlamydomonas genome. Thus, the data presented here will offer guidelines for future in-depth analysis and some clues for wet laboratory confirmations.

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 of the Chlamydomonas genes and may play a similar role as AAUAAA signal found in the NUEs of half of the mammalian genes (Tian *et al.* 2005). In stark contrast, there was barely a trace of the AAUAAA signal found in the NUE of Chlamydomonas. However, in Arabidopsis, UGUAA is found in a different region, namely the FUE (Figure 5), indicating a translocation of the signal. In mammals, 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; Tian *et al.* 2005), the use of UGUAA signal seems to be minimal (Figure 5). While the most conserved NUE signal for Arabidopsis and rice, AAUAAA, was found in ~10 and 7% of all tested genes, respectively (Loke *et al.* 2005; Shen *et al.* 2008), the UGUAA signal becomes dominant in FUEs 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 another 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 has moved to the weaker position (FUE) in land plants. The primitive AAUAAA found in Streptophyta (Wodniok *et al.* 2007) never caught up to the efficacy of UGUAA. This is because AAUAAA, while it is the best signal (Li and Hunt 1995), still has not been adopted by >12% of the plant 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 seven 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 data set, we found that this logo covers ~78% of sequences, whereas the use of single-pattern UGUAA resulted in covering only 52% of sequences. For the 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 3A). One explanation is that FUEs spread over a relative long region of the 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 account for only 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 article is the extensive usage of APA in Chlamydomonas. We estimate that a range of 9–33% of Chlamydomonas transcripts have two or more poly(A) sites. This number is less than that in human and rice (~50%; Zhang *et al.* 2005; Shen *et al.* 2008). A significant number of poly(A) sites, 1–11% of the genes depending on the criteria used, are located in 5′-UTRs, introns, and CDS in Chlamydomonas. Compared to that in rice (~2%; Shen *et al.* 2008), the extent of APA in Chlamydomonas might be within a similar scope or even greater. 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′-UTRs are revealed here, we recognize that the accuracy of the data relies on the current annotated draft genome. The latter, however, may have only limited accuracy due to its current annotation status. The confidence level of our data will be improved when more concrete genome information becomes available for Chlamydomonas. As there is no proven APA case for gene expression regulation in Chlamydomonas, we hope our data will offer some clues for such explorations.

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 EST projects (Ji *et al.* 2007a,b). 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.

The authors acknowledge Eric Stalhberg for exporting 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) data set, and other lab members for helpful discussions. This project was funded in part by the National Science Foundation (MCB 0313472 to Q.Q.L.) and by the Ohio Plant Biotechnology Consortium and the Miami University Center for the Advancement of Computational Research (to both Q.Q.L. and C.L.).

LITERATURE CITED

Arabidopsis Genome Initiatives, 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 Res. **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:** 2361–2366.

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 *et al.*, 2006 Calmodulin interacts with and regulates the RNA-binding activity of an Arabidopsis polyadenylation factor subunit. Plant Physiol. **140:** 1507–1521.

Dong, H. T., Y. Deng, J. Chen, S. Wang, S. H. Peng *et al.*, 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:** 2517–2521.

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:** 888–894.

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:** 1485–1493.

Hunt, A. G., 2007 Messenger RNA 3′-end formation and the regulation of gene expression, pp. 101–122 in *Regulation of Gene Expression in Plants: The Role of Transcript Structure and Processing*, edited by C. L. Bassett. Springer, New York.

Ji, G., J. Zheng, Y. Shen, X. Wu, R. Jiang *et al.*, 2007a Predictive modeling of plant messenger RNA polyadenylation sites. BMC Bioinformatics **8:** 43.

Ji, G., X. Wu, J. Zheng, Y. Shen and Q.Q. Li., 2007b Modeling plant mRNA poly(A) sites: software design and implementation. J. Comput. Theor. Nanosci. **4:** 1365–1368.

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

Li, Q. Q., and A. G. Hunt, 1997 The polyadenylation of RNA in plants. Plant Physiol. **115:** 321–325.

Liang, C., G. Wang, L. Liu, G. Ji, Y. Liu *et al.*, 2007a WebTraceMiner: a web service for processing and mining EST sequence trace files. Nucleic Acids Res. **35:** W137–W142.

Liang, C., G. Wang, L. Liu, G. L. Ji, L. Fang *et al.*, 2007b ConiferEST: an integrated bioinformatics system for data reprocessing and mining of conifer expressed sequence tags (ESTs). BMC Genomics **8:** 134.

Liang, C., Y. Liu, L. Liu, A. C. Davis, Y. Shen *et al.*, 2008 Expressed sequence tags with cDNA termini: previously overlooked resources for gene annotation and transcriptome exploration in *Chlamydomonas reinhardtii*. Genetics **179:** 83–93.

Loke, J. C., E. A. Stahlberg, D. G. Strenski, B. J. Haas, P. C. Wood *et al.*, 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:** 401–405.

Merchant, S. S., S. E. Prochnik, O. Vallon, E. H. Harris, S. J. Karpowicz *et al.*, 2007 The Chlamydomonas genome reveals the evolution of key animal and plant functions. Science **318:** 245–250.

Meyers, B. C., T. H. Vu, S. S. Tej, H. Ghazal, M. Matvienko *et al.*, 2004 Analysis of the transcriptional complexity of *Arabidopsis thaliana* by massively parallel signature sequencing. Nat. Biotechnol. **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 J. **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:** 7891–7898.

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

Seiler, K. P., G. A. George, M. P. Happ, N. E. Bodycombe, H. A. Carrinski *et al.*, 2007 ChemBank: a small-molecule screening and cheminformatics resource database. Nucleic Acids Res. **36:** D351–D359.

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 in vitro. Nucleic Acids Res. **18:** 5799–5805.

Shen, Y., G. Ji, B. J. Haas, X. Wu, J. Zheng *et al.*, 2008 Genome level analysis of rice mRNA 3′-end processing signals and alternative polyadenylation. Nucleic Acids Res. (http://nar.oxfordjournals.org/cgi/content/abstract/gkn158) (in press).

Silflow, C. D., R. L. Chisholm, T. W. Conner and L. P. Ranum, 1985 The two alpha-tubulin genes of *Chlamydomonas reinhardtii* code for slightly different proteins. Mol. Cell. Biol. **5:** 2389–2398.

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:** 777–787.

Slomovic, S., V. Portnoy, V. Liveanu and G. Schuster, 2006 RNA polyadenylation in prokaryotes and organelles; different tails tell different tales. Crit. Rev. Plant Sci. **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 Res. **33:** 201–212.

van Helden, J., 2003 Regulatory sequence analysis tools. Nucleic Acids Res. **31:** 3593–3596.

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. **15:** 1000–1010.

Venkataraman, K., K. M. Brown and G. M. Gilmartin, 2005 Analysis of a noncanonical poly(A) site reveals a tripartite mechanism for vertebrate poly(A) site recognition. Genes Dev. **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:** 1859–1875.

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:** 405–445.

Communicating editor: O. Vallon