PlantAPAdb: A Comprehensive Database for Alternative Polyadenylation Sites in Plants[^CC-BY]

Sheng Zhu,^a,b,2 Wenbin Ye,^a,b,2 Lishan Ye,^b,c,3 Hongjuan Fu,^a,b Congting Ye,^d Xuesong Xiao,^a Yuankaowei Ji,^e Weixu Lin,^a Guoli Ji,^a,b,3 and Xiaohui Wu^a,b,3,4

^aDepartment of Automation, Xiamen University, Xiamen, Fujian 361005, China
^bNational Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, Fujian 3611002, China
^cXiamen Health and Medical Big Data Center, Xiamen, Fujian 361008, China
^dKey Laboratory of the Ministry of Education for Coastal and Wetland Ecosystems, College of the Environment and Ecology, Xiamen University, Xiamen, Fujian 361102, China
^eSchool of Mathematics, Northwest University, Xi’an, Shanxi 710127, China

ORCID IDs: 0000-0001-5916-5625 (S.Z.); 0000-0003-4803-2098 (C.Y.); 0000-0002-7973-5434 (W.L.); 0000-0003-0356-7785 (X.W.).

Alternative cleavage and polyadenylation (APA) is increasingly recognized as an important regulatory mechanism in eukaryotic gene expression and is dynamically modulated in a developmental, tissue-specific, or environmentally responsive manner. Given the functional importance of APA and the rapid accumulation of APA sites in plants, a comprehensive and easily accessible APA site database is necessary for improved understanding of APA-mediated gene expression regulation. We present a database called PlantAPAdb that catalogs the most comprehensive APA site data derived from sequences from diverse sequencing protocols and biological samples in plants. Currently, PlantAPAdb contains APA sites in six species, *Oryza sativa* (japonica and indica), *Arabidopsis* (*Arabidopsis thaliana*), *Medicago truncatula*, *Trifolium pratense*, *Phyllostachys edulis*, and *Chlamydomonas reinhardtii*. APA sites in PlantAPAdb are available for bulk download and can be queried in a Google-like manner. PlantAPAdb provides rich information of the whole-genome APA sites, including genomic locations, heterogeneous cleavage sites, expression levels, and sample information. It also provides comprehensive poly(A) signals for APA sites in different genomic regions according to distinct profiles of cis-elements in plants. In addition, PlantAPAdb contains events of 3' untranslated region shortening/lengthening resulting from APA, which helps to understand the mechanisms underlying systematic changes in 3' untranslated region lengths. Additional information about conservation of APA sites in plants is also available, providing insights into the evolutionary polyadenylation configuration across species. As a user-friendly database, PlantAPAdb is a large and extendable resource for elucidating APA mechanisms, APA conservation, and gene expression regulation.

Cleavage and polyadenylation of precursor mRNA is a critical posttranscriptional process for mRNA maturation, which involves cleavage at the 3' end of a nascent transcript followed by the addition of a tract of adenosines [poly(A) tail]. Accumulating genomic studies, particularly those based on protocols capturing transcript 3' end (hereinafter referred to as 3' seq), have documented that most eukaryotic genes have multiple poly(A) sites (Ji et al., 2015; Tian and Manley, 2017; Grüber and Zavolan, 2019), offering the possibility of utilizing alternative cleavage and polyadenylation (APA) sites for gene expression regulation. APA has been indicated as a key layer of gene expression regulation by affecting mRNA stability, localization, exportation, and translational efficiency (Tian and Manley, 2017). APA also contributes to increased transcriptome complexity by generating diverse transcript isoforms with variable 3' untranslated regions (UTRs) and/or distinct coding potentials (Mayr, 2017; Tian and Manley, 2017). In animals, more than 70% of mammalian genes (Derti et al., 2012; Hoque et al., 2013) and ~50% of genes in fruit fly (*Drosophila melanogaster*) and zebrafish (*Danio rerio*) (Li et al., 2012; Smibert et al., 2012; Ulitsky et al., 2012) possess more than one poly(A) site. APA is also prevalent in plants. For example, 55% of genes in moso bamboo (*Phyllostachys edulis*) (Wang et al., 2017), more

[^CC-BY]: Article free via Creative Commons CC-BY 4.0 license. www.plantphysiol.org/cgi/doi/10.1104/pp.19.00943
involved in processes such as cancer, and neurodegenerative disorders (Tian and Bartel, 2009; Mayr, 2017). In plants, APA has been demonstrated in the regulation of genes that are responsive to oxidative stress responses (Thomas et al., 2012; Lin et al., 2017; Hong et al., 2018; Fu et al., 2019; Zhou et al., 2019). An increasing body of evidence has indicated that APA is dynamically modulated in a developmentally, tissue-specific, or environmentally responsive manner. Shortening or lengthening of 3′ UTRs mediated through APA, also called APA site switching, has been found in a number of biological processes and diseases, such as cell activation, growth, proliferation, cancer, and neurodegenerative disorders (Tian and Manley, 2013; Gruber et al., 2014; Mayr, 2017). Alternative 3′ UTRs can also differentially modulate gene expression through the binding of RNA-binding proteins (Mayr, 2017). Shorter 3′ UTRs that escape regulation by 3′ UTR elements, such as microRNA-binding sites, AU-rich elements, and GU-rich elements, may be more stable than their corresponding longer 3′ UTRs and result in the production of more protein (Mayr and Bartel, 2009; Mayr, 2017). In plants, APA has been demonstrated in the regulation of genes that are involved in processes such as flowering time control, amino acid biosynthesis, plant incompatibility, and oxidative stress responses (Thomas et al., 2012; Lin et al., 2017; Hong et al., 2018; Fu et al., 2019; Zhou et al., 2019). Given the importance of APA for gene expression and cellular function as well as the rapid accumulation of APA sites in diverse species, it is necessary that we have a comprehensive and easily accessible catalog of poly(A) sites with quantified expression levels on a genome-wide scale.

Currently, several poly(A) site databases are available, the majority of which are for animal species. Early databases, such as polyA_DB (Zhang et al., 2005), PACdb (Brockman et al., 2005), and polyA_DB2 (Lee et al., 2007), were built upon cDNAs and ESTs, which are limited in data scale. With the development of various 3′ seq technologies (for review, see Ji et al., 2015), a few databases recording poly(A) sites from 3′ seq are emerging, such as APADB (Müller et al., 2014), APASdb (You et al., 2015), PolyAStie (Gruber et al., 2016), and PolyA_DB 3 (Wang et al., 2018b). However, these databases mainly target nonplant species and typically provide little or limited integration with poly(A) sites in plants. Previously, we presented the PlantAPA database (Wu et al., 2016), which stores poly(A) sites derived from RNA sequencing (RNA-seq), ESTs, 454, and 3′ seq for four plant species. Whereas PlantAPA helps improve the initial understanding of the prevalence of APA in plants and contributes to the analysis of APA dynamics, it has several limitations. It is restrictive in data scale, as it fails to cover the entire set of 3′ seq data available to date. For example, poly(A) sites from rice and Chlamydomonas (Chlamydomonas reinhardtii) were collected from ESTs and RNA-seq, which are of limited quantities and cover only a few samples. Moreover, there is no unified data-processing and identification process for poly(A) sites from 3′ seq, hindering the integration of emerging 3′ seq data. Although users are allowed to analyze APA site switching and extract relevant sequences by uploading their own data, results of APA switching, poly(A) signals, and sequences for public data are not available for download in PlantAPA. Furthermore, the annotation of poly(A) sites is not comprehensive and lacks information of poly(A) signal, conservation, or sample-specific expression levels.

Here, we present a database named PlantAPAdb (accessible through http://bmi.xmu.edu.cn/plantAPAdb or http://www.bmibig.cn/plantAPAdb), which provides a comprehensive and manually curated catalog of APA sites in plants based on a large volume of data derived from diverse biological samples generated by 3′ seq. Currently, PlantAPAdb contains APA sites in six organisms, rice (japonica and indica), Arabidopsis, M. truncatula, Trifolium pratense (forage legume red clover), moso bamboo, and Chlamydomonas. It details APA sites in different tissues from diverse conditions, with rich information including their genomic locations, evidence of supported 3′ seq reads, poly(A) signals, and conservation of APA between species. Data in PlantAPAdb are available for bulk download and can be queried in a Google-like manner. PlantAPAdb also stores detailed 3′ UTR lengthening or shortening events between conditions, providing a rich resource for the study of mechanisms affecting the 3′ UTR length and APA-mediated gene regulation. Moreover, PlantAPAdb provides conservation information of APA sites across plants, which helps elucidate the evolutionary significance of APA. In addition, poly(A) signals and related sequences for APA sites located in different genomic regions are available in PlantAPAdb, which contributes to in-depth analysis of cis-elements related to APA, especially those in non-3′ UTRs such as introns and intergenic regions. Particularly, PlantAPAdb incorporates a rich repertory of back-end programs and standard pipelines, which makes it highly expandable.

| Species                        | 3′ Read No. | Sample No. | Experiment No. | High-Confidence PAC No. |
|-------------------------------|-------------|------------|----------------|------------------------|
| Arabidopsis thaliana          | 308,380,174 | 37         | 86             | 91,946                 |
| Oryza sativa indica           | 112,508,754 | 14         | 42             | 64,219                 |
| Oryza sativa japonica         | 135,802,234 | 14         | 42             | 63,544                 |
| Chlamydomonas reinhardtii     | 3,688,190   | 4          | 12             | 34,995                 |
| Trifolium pratense            | 4,281,782   | 3          | 9              | 49,832                 |
| Medicago truncatula           | 6,950,530   | 2          | 4              | 37,645                 |
| Phyllostachys edulis          | 50,801,147  | 3          | 9              | 44,066                 |

Table 1. Poly(A) site data sets in PlantAPAdb
to incorporate emerging 3’ seq data sets in the future. As a user-friendly database, PlantAPAdb is a large and extendable resource for improving genome annotation and for elucidating APA mechanisms, APA conservation, and APA-mediated gene expression regulation.

Figure 1. Data overview in PlantAPAdb. A, A web page showing an example of the data summary for individual species. B, Shown are the statistical results of poly(A) sites in Arabidopsis revealed by clicking the Stats link outlined in (A), including the distribution of poly(A) sites in different genomic regions, number of genes with different number of poly(A) sites, and single-nucleotide profiles of poly(A) sites in different regions. This example is available at http://bmi.xmu.edu.cn/plantAPAdb/index.php.
RESULTS

Poly(A) Site Data Sets in PlantAPAdb

Poly(A) sites from PlantAPAdb were all collected from 3’ seq, which are of much higher confidence and larger scale than those from ESTs or RNA-seq. The homepage of PlantAPAdb (http://bmi.xmu.edu.cn/ plantAPAdb/index.php) summarizes curated poly(A) site data involved in diverse cell and tissue types, biotic and/or abiotic stress responses, and physiological conditions in six organisms (Table 1). Full information of the samples used in PlantAPAdb is presented in Supplemental Data Set S1. As an example, for Arabidopsis, poly(A) sites were obtained from a total of 37 samples from 86 experiments (Fig. 1A). Mapped 3’ end reads within 24 nucleotides of each other were grouped into poly(A) site clusters (PACs; see “Materials and Methods”). After stringent filtering, the pooled sample contains 91,946 high-confidence PACs supported by more than 300 million reads. Detailed statistical results, such as the distribution of PACs in different genomic regions, number of genes with different number of PACs, and single-nucleotide profiles of PACs in different regions, are also available by clicking the Stats button on the summary table of a given species from the homepage (Fig. 1B).

The APA catalog page (http://bmi.xmu.edu.cn/ plantAPAdb/APACatalog.php) catalogs all PAC data sets, which are well organized by different categories such as sequencing protocols, tissues, environmental conditions, plant types, and relevant studies (Fig. 2A). For instance, if users are interested in rice PACs related to the study of Fu et al. (2016), they can click the Fu H, 2016 button in the Studies category to view all relevant data sets. Data sets are displayed by data set cards organized by tissue types. By clicking the Stats button at the bottom of a data set card [e.g. 20 Days Leaf (SRP073467)], users can view detailed descriptive and statistical information of the data set in a pop-up page named Statistics (Fig. 2B). Detailed information is given, including the sequencing protocol, tissue type, relevant study, literature, number of PACs, mapping information of the raw data, and external links to the National Center for Biotechnology Information (NCBI). Alternatively, by clicking the Download button in the data set card or the Statistics web page, users can download all potential PACs and high-confidence PACs in bed format or with full genome annotation (Fig. 2C). The list of heterogeneous cleavage sites before grouping into PACs is also available for download, which is especially useful for analyzing the microheterogeneity phenomenon of polyadenylation or testing the impact of a clustering procedure. Moreover, users can browse online high-confidence PACs in the genome browser by clicking the JBrowse button in the Statistics web page without downloading the full data (Fig. 2C).

Users can conduct a full-text search by choosing All databases from the drop-down list next to the search box without specifying any field. When queried by the full-text strategy, PlantAPAdb returns search results relevant to the provided keyword by automatically scanning commonly used information such as gene identifier, gene symbol/alias, gene description, gene function, and Gene Ontology (GO) identifier. Users can also conduct a limited-keyword search by limiting a field of gene, GO identifier, or species. All queries except for the sample-centric search will lead to a media page that shows the number of PACs in each species. Take the Yellow-Green Leaf8 (YGL8) gene, for example, which has been reported to have a strong impact on chlorophyll content and photosynthesis in rice (Zhou et al., 2019). This gene has multiple poly(A) sites and exhibits 3’ UTR APA site switching between two rice subspecies (Zhou et al., 2019). Users can directly search for YGL8 and then select Oryza sativa Japonica Group from the search result page (Fig. 3A). Then a media page opens, which shows a detailed gene list and the number of PACs in each gene associated with the input keyword in the selected species. By clicking a gene of interest (here Os01g0279100), a new page opens to show detailed information of the gene and its PACs, including the description of the gene, relevant GO identifiers, external links to Ensembl Plants and AmiGO (http://amigo.geneontology.org), and the PAC list (Fig. 3B). There are four PACs in YGL8; all are located in 3’ UTRs or extended 3’ UTRs. Among the four PACs, the proximal one is the most dominant PAC, which has a much higher expression level than other PACs. This result is consistent with that in the previous study (Zhou et al., 2019). In addition, by clicking the JBrowse link of a PAC item in the PAC list (Fig. 3B), users can browse the PAC and the related genome sequence and gene model in the JBrowse genome browser. Alternatively, users can also access the browser by clicking the APA browse tab in the main menu. Selected samples (tracks) from each species can be quickly loaded and graphically browsed. Users can zoom in on particular genomic regions by searching a gene or chromosome fragment via the search box at the top of the browser. Data tracks of interest can be downloaded onto a local computer.

In addition to the full-text and limited-keyword strategies, users can also conduct a sample-centric search using a keyword of tissue/sample (e.g. root), which will lead to a media page that lists numbers of matched samples in all species (Supplemental Fig. S1). Users can further click a species name to view detailed information of the selected sample and download PACs as in the APA catalog module. Collectively, the one-click search provided in PlantAPAdb is convenient and powerful for data query, which returns informative and well-organized search results that can be easily browsed, downloaded, and visualized.

Search and Browse

PlantAPAdb provides three strategies to query the database: full-text, limited-keyword, and sample-centric.
Figure 2. The APA catalog module. A, Catalogs of poly(A) sites by different categories such as sequencing protocols, tissues, environmental conditions, plant types, and relevant studies. Each data set is displayed by a data set card. B, Screenshot of the page detailing information of an example poly(A) site data set (20 Days Leaf of *japonica* rice). Shown are the summary of the relevant sequencing project, reference literature, and statistics of the read-mapping results. C, The download page and genome browser reached by clicking the Download and JBrowse link outlined in (B). This example is available at http://bmi.xmu.edu.cn/plantAPAdb/APAcatalog.php.
**Figure 3.** The one-click search in PlantAPAdb. A, An example search. The top part shows the search interface designed to query poly(A) sites by a gene symbol. The middle part is a media page of an example search using the keyword YGL8, which shows the number of resulting poly(A) sites in each species. The bottom part shows a returned gene list that meets the search criteria. B, A web page with detailed information and poly(A) sites in YGL8 (Os01g0279100) by clicking the link outlined in (A). This example is available at http://bmi.xmu.edu.cn/plantAPAdb/index.php.
Poly(A) Signals of Poly(A) Sites

Over the last decade, numerous genomic studies have revealed various cis-elements that control polyadenylation in a wide range of mechanisms, suggesting both conserved and divergent poly(A) signals across species (Xing and Li, 2011; Tian and Graber, 2012; Ji et al., 2015). However, the majority of studies focused on poly(A) signals of canonical poly(A) sites that are located in 3' UTRs, whereas little emphasis has been placed on poly(A) signals of noncanonical sites. In PlantAPAdb, poly(A) signals for both 3' UTR and non-3' UTR APA sites for different plant species were identified, jointly considering various cis-element regions surrounding poly(A) sites that have been discovered in plants. Full lists of poly(A) signals for poly(A) sites in different genomic regions and sequences surrounding poly(A) sites are available for download in PlantAPAdb.

The APA signal module (http://bmi.xmu.edu.cn/plantAPAdb/APASignal.php) provides comprehensive poly(A) signals for different species. According to previous studies on plant poly(A) signals (for review, see Xing and Li, 2011), two schemes of poly(A) signals, one for Chlamydomonas and the other for other plant species, were present in PlantAPAdb. For each species, poly(A) signal patterns for APA sites in different genomic regions, including 3' UTR, 5' UTR, coding sequence (CDS), and intron, were provided. For example, following the information displayed on the web site, users can easily view poly(A) signals in near-upstream element (NUE) of 3' UTR poly(A) sites in Arabidopsis by choosing 3' UTR and NUE on the web page (Fig. 4, top). Single-nucleotide compositions and the 20 most frequently occurring patterns in the selected NUE region are intuitively visualized (Fig. 4, middle). Based on the poly(A) signal model in Arabidopsis (Loke et al., 2005), hexamers (e.g. AAUAAA) were scanned in NUE. Accordingly, statistical information of each hexamer is given, such as the frequency of occurrence, occurrence probability, and number of overlapping occurrences (Fig. 4, bottom). Similarly, users can also choose Chlamydomonas from the drop-down list on the web page to view poly(A) signals in the NUE of Chlamydomonas (Supplemental Fig. S2). Different from Arabidopsis, the most dominant poly(A) signal pattern in Chlamydomonas is UGUAA instead of AAUAAA (Supplemental Fig. S2, bottom). Moreover, users can easily view poly(A) signals in different cis-element regions of noncanonical APA sites that are located in non-3’ UTRs, such as 5’ UTR, CDS, and intron.

In addition to the poly(A) signals provided in PlantAPAdb, users can also retrieve respective sequences for custom poly(A) signal analysis. Through the APA sequence page (http://bmi.xmu.edu.cn/plantAPAdb/APAsequence.php), users can download genomic sequences surrounding PACs onto their local computers. Sequences of PACs from different species in different genomic regions, including 3’ UTR, 5’ UTR, CDS, and intron, were exportable. Particularly, sequences of PACs located in extended 3’ UTRs were also provided, which may be useful for the study of cis-regulatory elements involved in 3’ UTR extension. Moreover, if users would like to identify poly(A) signal patterns from a custom region, they could download poly(A) sequences and then use scripts provided in PlantAPAdb and/or other existing tools for poly(A) signal identification.

Collectively, PlantAPAdb provides a flexible and intuitive way to show comprehensive poly(A) signals in plants. The categories of these diverse poly(A) signals in different plant species would provide new insights into the conservation of APA and the effect of poly(A) signals on polyadenylation machineries. Particularly, poly(A) signals for APA sites in non-3’ UTRs, such as introns, CDS, and 5’ UTRs, would provide valuable resources for in-depth analysis of noncanonical APA sites and their functions in gene regulation.

Poly(A) Site Usage

APA has been shown to be modulated in a tissue- and/or developmental stage-specific manner (Tian and Manley, 2013; Gruber et al., 2014; Mayr, 2017). The APA metric module (http://bmi.xmu.edu.cn/plantAPAdb/APAmetric.php) provides four PAC-level metrics and two gene-level metrics to quantify the relative poly(A) site usage, tissue specificity of a PAC, or 3’ UTR length change of a gene across samples (see “Materials and Methods”). The PAC-level metrics include number of samples expressed, percentage of samples expressed, ratio, and sample specificity. The two gene-level metrics are relative usage of distal PAC and weighted 3’ UTR length. Take the PAC-level metric of tissue specificity, for example. Users can view the tissue specificity of each PAC across all samples in a species (e.g. rice japonica) by choosing PAC index from the Type drop-down list and clicking the Tissue specificity radio button (Fig. 5, top). Different filtering conditions, such as total read number in all samples and percentage of expressed samples, can be set to filter the results. Top PACs ranked by the metric score (here the Shannon entropy score) can be visualized by a heatmap that shows the variation of tissue specificity across samples (Fig. 5, bottom). The full list of PACs and the data corresponding to the heatmap can also be downloaded. By clicking the JBrowse link of a PAC in the list, users are linked to the genome browser to view distributions of PACs and their expression levels across samples on the respective gene. Users can also filter PACs by their genomic locations (Fig. 5, top), which can further view the variability of tissue specificity across samples for both canonical 3’ UTR PACs and noncanonical PACs located in non-3’ UTRs. Therefore, through the APA metric module, users can easily retrieve PACs or genes specific to given samples by different methods and filtration conditions.

3’ UTR Shortening/Lengthening Events

APA has been indicated to regulate the expression of genes containing APA sites through impacting mRNA
metabolism and protein localization (Tian and Manley, 2017). Modifications in the length of 3’ UTRs, mediated through APA, have been found to play a fundamental role in posttranscriptional regulation in diverse tissues and developmental stages. Analysis of 3’ UTR switching has become a routine process in many APA studies (Fu et al., 2011, 2016; Gruber et al., 2014; Lin et al., 2017; Zhou et al., 2019); however, there seems to be no universal strategy to identify 3’ UTR lengthening/shortening events. In PlantAPAdb, we implemented two methods that have been widely applied in previous studies to identify 3’ UTR switching by leveraging the comprehensive APA sites from various biological samples available in our database.

The APA switching module (http://bmi.xmu.edu.cn/plantAPAdb/APAswitch.php) provides comprehensive...
lists of genes and PACs involving 3' UTR shortening/lengthening. Users are free to choose one species and a pair of samples to obtain the list of APA site switching genes. Different filtering conditions (e.g., adjusted P value and log fold change) can be set to filter the results. 3' UTR switching events between the selected two conditions are

Figure 5. The APA metric module. A list of poly(A) sites or genes can be obtained by choosing a metric (top). Top poly(A) sites ranked by the metric score (here the Shannon entropy score) can be visualized by a heatmap that shows the variation of tissue specificity across samples (middle). By clicking the buttons outlined in red, the full data and the data corresponding to the heatmap can be downloaded. This example is available at http://bmi.xmu.edu.cn/plantAPAdb/APAmetric.php.
visualized by a scatterplot that shows both shortening and lengthening events. Moreover, the list of genes with significant 3′ UTR switching is present in a table and can be downloaded, including information such as the respective gene identifier, 3′ UTR lengths, expression levels of involved PACs, and relevant statistical information. Through the APA switching module, users can easily retrieve 3′ UTR switching results by different methods and filtration conditions. Therefore, consensus or combined results from multiple methods can be obtained to mitigate potential bias caused by different methods.

Next, we utilized an example to show the retrieval of APA site switching events. A previous study (Fu et al., 2016) investigated APA profiles in 14 different rice tissues and developmental stages and found a distinct pattern of APA usage in mature pollen. Here, we attempted to identify APA site switching genes between mature pollen and anther using the linear trend method. By choosing the two sample groups from the drop-down lists and selecting the Linear Trend method on the web page (Fig. 6, top), 22 genes with significant 3′ UTR shortening/lengthening were obtained. These genes are visualized by a scatterplot to show the extent of 3′ UTR shortening/lengthening (Fig. 6, middle). The full list of genes is provided in a table, which clearly shows expression levels of each gene and related PACs in both samples (Fig. 6, bottom). By clicking the link of a gene in the list, users can view distributions of PACs and their expression levels between the two investigated samples of this gene in a genome browser.

APA Conservation

The conservation information of APA sites across species is crucial for understanding the evolutionary importance of APA and the divergence in gene expression. However, the majority of previous studies focused on conservation of gene expression profiles across species, and there is scarce work done at the level of APA, especially in plants. The collection of high-confidence APA sites in PlantAPAdb would provide advantages to study APA conservation in a wider range of plant species. The conservation information of APA would provide insights into the importance of APA in transcriptome diversification and help elucidate the extensive morphological and functional differences among plant species.

The APA conservation module (http://bmi.xmu.edu.cn/plantAPAdb/APAconservation.php) provides the conservation information of APA sites across species. Users can browse conserved PACs by species and genomic regions (Fig. 7A). By clicking the Stats link in a data set card, conserved PACs and statistical information can be shown in a pop-up window (Fig. 7B). For each PAC, the coordinates of both the original species and the reference species (here Arabidopsis) were recorded. Moreover, the conservation status (conserved or nonconserved) and the PAC counterparts in other species of the given PAC are also given. Users can also download the list of conserved PACs to their local computers.

Bulk Download

In addition to cataloging PACs from individual samples, PlantAPAdb also provides the PAC list of the pooled sample for bulk download (http://bmi.xmu.edu.cn/plantAPAdb/Bulkdownload.php; Supplemental Fig. S3). For the pooled data, different files are available for download to meet the user’s needs. The file in bed format records simple information such as chromosome, strand, coordinate, and total number of reads for each PAC. The file in text format tabulates full information of each PAC, including the total read count, the raw and normalized read count in each individual experiment, the respective gene, the genomic location (CDS, intron, 3′ UTR, 5′ UTR, or intergenic), and the distance to neighboring genes (if the PAC is located in the intergenic region). Moreover, the file of heterogeneous cleavage sites in bed format is also provided, which allows users to inspect the polyadenylation in higher resolution. In addition to PACs, a conserved PAC list for each species is also available for download.

DISCUSSION

We present a comprehensive database of APA sites in plants called PlantAPAdb, which catalogs the most comprehensive plant poly(A) site data sequenced from diverse 3′ seq protocols and biological samples. Poly(A) sites are sorted by categories such as experimental studies, sequencing protocols, and biological samples, facilitating the retrieval of data for specific usage. PlantAPAdb also provides rich annotation information of genome-wide poly(A) sites, including genomic locations, heterogeneous cleavage sites, expression levels, related poly(A) signals, sample information, and conservation information. APA sites and cleavage sites from the pooled sample and individual samples are available for bulk download as flat files.

PlantAPAdb not only provides various kinds of data for download but also provides biologists with an easy-to-use web service for data query, visualization, and browsing. A convenient and powerful one-click search based on the full-text search technique is integrated in PlantAPAdb for data query. Search results can then be visualized in their genomic context via the JBrowse genome browser. In addition, implemented with dynamic interaction technologies, each kind of data [e.g., poly(A) signals] are arranged in different categories and can be browsed in a card-like manner. In the card-like view, different categories are shown in individual cards with respective statistical information and download option. Moreover, PlantAPAdb provides an overview of the data by presenting rich statistics and intuitive charts. A wealth of help information and references are available throughout the web site for the user’s reference. These user-friendly features greatly improve the ease of use and data retrieval of our database.

In addition to various function modules, PlantAPAdb also incorporates a rich repertory of back-end
programs that facilitate processing, annotation, and analyses of APA sites. A uniform and flexible processing pipeline was designed for accurately identifying and quantifying high-confidence poly(A) sites from 3’ seq. PlantAPAdb also integrates a standard procedure for parsing a genome annotation file obtained from the unified portal (Ensembl Plants), facilitating the automatic annotation of poly(A) sites in a universal way.

Figure 6. The APA switching module. The top part shows parameters for retrieving an APA switching list between two sample groups. The middle part shows a scatterplot visualizing both 3’ UTR shortening and lengthening events between the selected two conditions. The bottom part tabulates the list of genes with significant 3’ UTR switching. By clicking the link of a gene in the list, users can view this gene in the genome browser. This example is available at http://bmi.xmu.edu.cn/plantAPAdb/APAswitch.php.
Figure 7. The APA conservation module. A, Conserved poly(A) sites are organized by different species and different genomic regions. Each data set is displayed by a data set card. B, Screenshot of the page showing detailed conservation information of 3’ UTR poly(A) sites in Arabidopsis. The top part shows the summary of conserved poly(A) sites in other species. The bottom part tabulates the list of conserved poly(A) sites for a selected species. This example is available at http://bmi.xmu.edu.cn/plantAPAdb/APAconservation.php.
These resources make PlantAPAdb highly expandable for the incorporation of emerging 3′ seq data sets in the future. Relevant scripts are available for download in PlantAPAdb, which allows users to perform similar analyses conducted in PlantAPAdb for their own data. However, it should be noted that there is no single best pipeline for poly(A) site identification from 3′ seq data. In different studies on 3′ seq, researchers may use different alignment tools for read mapping, different strategies for removing internal priming artifacts, different algorithms for grouping cleavage sites, and different criteria for filtering high-confidence poly(A) sites (Ji et al., 2015). Currently, there is no criterion or benchmark study to evaluate existing pipelines for identifying poly(A) sites. Additional work will be needed in the future to design an objective comparative study for a better evaluation of diverse poly(A) site identification pipelines.

With accurate identification and comprehensive annotation of APA sites from a large volume of 3′ seq data in plants, as well as conserved APA configuration across various species, PlantAPAdb will be valuable for annotating gene 3′ ends and understanding APA-mediated gene regulation. Future work will involve adding APA sites from more species and more diverse cell/tissue types upon the availability of new 3′ seq data. Recently, a wide range of computational tools for identifying APA sites from RNA-seq data have continued to emerge, and an increasing number of APA sites have been found beyond annotated sites from 3′ seq data (Chen et al., 2019). APA sites were also found from single-molecule sequencing (Wang et al., 2017, 2018a). At present, PlantAPAdb mainly focuses on 3′ seq data; we expect to leverage the merit of unprecedented RNA-seq and single-molecule sequencing data to expand our compendium of APA sites from more diverse species and biological samples in the future.

MATERIALS AND METHODS

A Uniform Pipeline to Identify Poly(A) Sites with 3′ Seq Data

Publicly available 3′ seq data were mainly downloaded from the NCBI Sequence Read Archive (www.ncbi.nlm.nih.gov/sra). The latest genome assemblies for all species except for moso bamboo (Phyllostachys edulis; Peng et al., 2013) were downloaded from Ensembl Plants (http://plants.ensembl.org). We designed a bioinformatics pipeline to facilitate the uniform processing of 3′ seq data obtained from diverse experimental protocols (Supplemental Fig. S4A). First, a quality control check was performed to examine the data quality. Barcodes of each sample were then examined according to the description in the relevant study and/or by an in-house Perl script, and the file was split into subfiles if barcodes were present. Next, A/T stretches in the raw file were trimmed off. For samples generated with protocols that assign a 5′ adapter sequence in the reads, reads with a T stretch at the 5′ end were filtered and the T stretch was trimmed off. For data with 3′ adapter, reads with an A stretch at the 3′ end were filtered and the A stretch was trimmed off. Then Trimomatic (v0.38; Bolger et al., 2014) was adopted for quality trimming and reads shorter than 25 nucleotides were discarded. Remaining reads were then aligned to the corresponding genome assembly using STAR (v2.6.0a; Dobin et al., 2013). Only uniquely mapped reads were retained, and coordinates of candidate cleavage sites were obtained. Finally, those sites representing potential internal priming artifacts, with six consecutive adenines or more than six adenines within the −10 to +10 nucleotide window from the cleavage site, were filtered out. Next, qualified 3′ end cleavage sites were grouped into PACs to reduce the impact of microheterogeneity. For each individual sample or the pool of samples, cleavage sites within 24 nucleotides of each other were merged into PACs. Finally, relevant information of each PAC was recorded, including the start and end coordinates of the cluster, the coordinate of the dominant cleavage site, and the number of supported reads in each sample. Detailed commands and relevant scripts were provided in our PlantAPAdb web site.

To quantify the usage of each PAC, we calculated the raw expression level (number of reads) and normalized expression level in tags per million for each individual sample. The mean tags per million value across all samples for each PAC was also calculated. We also calculated for each PAC the percentage of samples expressed, which is defined as the ratio of the number of samples with raw expression level ≥ 2 to the total number of samples. We defined high-confidence PACs as those PACs with normalized read count ≥ 1 in two or more experiments. Both raw PACs and high-confidence PACs are available for download in PlantAPAdb, whereas most analysis results, such as poly(A) signals and 3′ UTR switching results, are based on high-confidence PACs.

Annotation and Quantification of Poly(A) Sites

Poly(A) sites were annotated with respective genes and genomic locations based on the latest genome annotations (Supplemental Fig. S4B). An R script based on the GenomicFeatures R package was implemented to uniformly process the genome annotation file in GFF3 format. Annotation for both protein-coding genes and noncoding genes was parsed. PACs were then annotated based on their genomic locations (i.e. 3′ UTR, coding sequence, and intron for protein-coding genes, exon for noncoding genes, and intergenic region). To resolve annotation ambiguity owing to multiple transcripts from the same gene, we annotated a PAC based on the following priority: 3′ UTR, CDS, intron. Particularly if a PAC is located in an intergenic region, we recorded the neighboring genes and its distance from the 5′ end of the nearby 5′ gene and the distance from the 5′ end of the nearby 3′ gene. This strategy to annotate intergenic PACs allows annotation of PACs with higher flexibility to recruit PACs falling within extended 3′ UTRs.

Quantification of Poly(A) Site Usage

We adopted four PAC-level metrics and two gene-level metrics to quantify the dynamics of APA across samples. The PAC-level metrics include NSE (number of samples expressed), PSE (percentage of samples expressed), Ratio, and Sample Specificity. NSE of a PAC is calculated as the number of samples in which the PAC is expressed. PSE of a PAC is calculated as the ratio of NSE to the total number of samples. Ratio is the relative usage of a PAC in a gene, which is calculated as the ratio of the expression level of a PAC to the total expression level of the respective gene. Sample Specificity (Ni et al., 2013; Ji et al., 2018b) of a PAC is a Shannon entropy score denoting the overall sample specificity:

\[ H = - \sum_{p} p \log_2 (p), \]

where \( n \) is the number of samples and \( p \) is the ratio of the expression level of the PAC in sample \( s \) to the total expression level of this PAC in all samples. Then the specificity of a PAC for sample \( s \) can be calculated as:

\[ Q = H - \log_2 (p) \]

A lower \( H \) or \( Q \) score means higher sample specificity.

The gene-level metrics include RUD (relative usage of distal PAC) and WUL (weighted 3′ UTR length). RUD (Ji et al., 2009) of a gene in a sample is calculated as the ratio of NSE to the total number of samples. Ratio is the relative usage of a PAC in a gene, which is calculated as the ratio of the expression level of a PAC to the total expression level of the respective gene. Sample Specificity (Ni et al., 2013; Ji et al., 2018b) of a PAC is a Shannon entropy score denoting the overall sample specificity:

\[ H = - \sum_{p} p \log_2 (p), \]

where \( n \) is the number of samples and \( p \) is the ratio of the expression level of the PAC in sample \( s \) to the total expression level of this PAC in all samples. Then the specificity of a PAC for sample \( s \) can be calculated as:

\[ Q = H - \log_2 (p) \]

A lower \( H \) or \( Q \) score means higher sample specificity.

The gene-level metrics include RUD (relative usage of distal PAC) and WUL (weighted 3′ UTR length). RUD (Ji et al., 2009) of a gene in a sample is calculated as the ratio of NSE to the total number of samples. Ratio is the relative usage of a PAC in a gene, which is calculated as the ratio of the expression level of a PAC to the total expression level of the respective gene. Sample Specificity (Ni et al., 2013; Ji et al., 2018b) of a PAC is a Shannon entropy score denoting the overall sample specificity:

\[ H = - \sum_{p} p \log_2 (p), \]

where \( n \) is the number of samples and \( p \) is the ratio of the expression level of the PAC in sample \( s \) to the total expression level of this PAC in all samples. Then the specificity of a PAC for sample \( s \) can be calculated as:

\[ Q = H - \log_2 (p) \]

A lower \( H \) or \( Q \) score means higher sample specificity.

Detection of 3′ UTR Shortening/Lengthening Events

We adopted two strategies for detecting 3′ UTR shortening/lengthening events (also called 3′ UTR switching or APA site switching) from samples with or without replicates (Supplemental Fig. S4C). The first strategy is based on the \( \chi^2 \) test for linear trend in proportions, which is applicable for samples without
replicates (replicates were averaged first). This method has the advantage to consider both abundance and 3′ UTR length of all 3′ UTR PACs in a gene, which was adopted in several previous APA studies (Fu et al., 2011, 2016; Ye et al., 2018; Zhou et al., 2019). Briefly, PACs in a gene are sorted by the respective 3′ UTR length (denoted as score). A contingency table of read count is then created with rows representing the indexes of samples and columns denoting the scores. Next the \( \chi^2 \) test for trend in proportions is performed with the function prop.trend.test, and the Pearson correlation \( r \) is obtained using the read count in the table as the value and the score as the coordinate. The correlation \( r \) ranges from \(-1\) to \(1\), with larger absolute value indicating a higher extent of 3′ UTR shortening/lengthening. Finally, genes with adjusted \( P \) value smaller than a given cutoff (e.g. 0.05) are considered as genes with significant 3′ UTR shortening/lengthening.

The second strategy is applicable for samples with replicates, which extends the differential expression (DE) results from DESeq2 to identify genes with 3′ UTR shortening/lengthening events. To detect DE PACs, genes with only one shortening/lengthening. To detect 3′ UTR shortening/lengthening events. To detect DE PACs, genes with only one PAC were discarded. Then DESeq2 (Anders and Huber, 2010) was used for DE PAC identification. For each PAC in APA genes, both \( P \) value and adjusted \( P \) value were obtained and PACs with adjusted \( P \) value below a given cutoff were considered as DE PACs. To detect 3′ UTR shortening/lengthening events, 3′ UTR APA genes with at least one DE PAC were filtered. Given a pair of 3′ UTR PACs \( i \) and \( j \) between two samples \( a \) and \( b \), the respective expression levels are denoted as \( E_{ia} \), \( E_{jb} \) and \( E_{ia} \), \( E_{ja} \). The relative change for this pair of PACs is calculated by

\[
R_{ij} = \log\left(\frac{E_{ia}}{E_{ja}}\right) - \log\left(\frac{E_{ib}}{E_{ja}}\right)
\]

A Fisher’s exact test was also performed to test the significance of differential usage of PACs \( i \) and \( j \) between samples \( a \) and \( b \), and \( P \) value was obtained. Genes with \( |R_{ij}| \) larger than a given threshold (e.g. 1) and \( P \) value below a given cutoff (e.g. 0.05) were considered as genes with 3′ UTR shortening/lengthening events.

To facilitate users to obtain DE results between samples stored in PlantAPAdb, we have precalculated all genes with 3′ UTR shortening/lengthening events between each two samples with default parameters (e.g. adjusted \( P < 0.05 \) and \( |R_{ij}| > 1 \)). Users are free to download the 3′ UTR switching list with selected samples and custom parameters to meet their own needs.

Identification of Poly(A) Signals

Poly(A) signals for poly(A) sites in different genomic locations of each species were identified (Supplemental Fig. S4D). According to previous studies (Loke et al., 2005; Shen et al., 2008b), we divided the six species into two groups: one is Chlamydomonas (Chlamydomonas reinhardtii) and the other contains the remaining species. In Chlamydomonas, four signal elements were reported: FUE (far upstream element), NUE, CE (cleavage element), and downstream element (Shen et al., 2008b). In the other five species with similar single-nucleotide profiles surrounding poly(A) sites, we used the signal model of Arabidopsis (Arabidopsis thaliana). Loke et al. (2003) that contains FUE, NUE, and CE. The choice of signal region range and the respective length of signal patterns are based on the observations in previous studies (Loke et al., 2005; Shen et al., 2008b). To identify statistically significant signal patterns in a given poly(A) signal region, we applied an oligo analyzer called RSAT (Thomas-Chollier et al., 2008). We classified poly(A) sites of each species into four groups based on their genomic locations (3′ UTR, CDS, intron, and 5′ UTR) and then obtained 50 top-ranked signal patterns by RSAT for each signal element in each group of poly(A) sites.

Conservation of Poly(A) Sites

The strategy to identify conserved poly(A) sites is shown in Supplemental Figure S4E. We used the Arabidopsis genome as the reference and downloaded the pairwise genome alignment chain files from Ensembl Plants. Then synthetic regions between other genomes and the reference genome were obtained. Next, coordinates of PACs of all other species were converted to coordinates of the reference genome. We adopted the reciprocal best match method (Wang et al., 2013b) to determine conserved PACs. Briefly, two PACs from two species were considered as orthologous if their distance was smaller than 24 nucleotides based on the whole-genome alignment. For each PAC in each species, the information of conservation is recorded, including the species with conserved sites and the corresponding PACs.

Database and Web Site Design

To ensure the smooth access of PlantAPAdb, we have built two mirror web sites, which can be accessed through http://bmi.xmu.edu.cn/plantAPAdb or http://www.bmbig.cn/plantAPAdb. Pipelines of poly(A) site identification and annotation were implemented by a series of Perl scripts and R scripts, which are available for download in PlantAPAdb. The JBrowse genome browser (Skinner et al., 2009) was embedded for browsing of poly(A) sites associated with genome annotations at the genome-wide level. We used JavaScript (https://www.jquery.com/) to construct interactive web pages with simplified JavaScript and XML programming. To ensure asynchronous data transmission between the web application and the server, Asynchronous Javascript and XML (AJAX) technology was extensively employed in PlantAPAdb. The one-click search was implemented based on Sphinx (http://sphinxsearch.com/), an open-source full-text search engine. The back-end database was implemented by MySQL, which uses tables to store gene annotation files and poly(A) site data. We also utilized several integration plugins, such as Plotly (https://plot.ly/) and DataTable (https://datatables.net/), to enable the interactive display of data.

Supplemental Data

The following supplemental materials are available.

Supplemental Figure S1. The sample-centric search in PlantAPAdb.

Supplemental Figure S2. Poly(A) signals in Chlamydomonas.

Supplemental Figure S3. The bulk download module.

Supplemental Figure S4. Procedures for data processing in PlantAPAdb.

Supplemental Data Set S1. Full information of samples used in PlantAPAdb.

Received July 31, 2019; accepted November 18, 2019; published November 25, 2019.

LITERATURE CITED

Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11: R106

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–2120

Brockman JM, Singh P, Liu D, Quinlan S, Salisbury J, Graber JB (2005) PACdb: PolyA cleavage site and 3′-UTR database. Bioinformatics 21: 3691–3693

Chen M, Ji G, Fu H, Lin Q, Ye C, Ye W, Su Y, Wu X (2019) A survey on identification and quantification of alternative polyadenylation sites from RNA-seq data. Brief Bioinform 20: 1–16

Dhaini A, Garrett-Engle P, Maciasac KD, Stevens RC, Sriman S, Chen R, Rohl CA, Johnson JM, Babak T (2012) A quantitative atlas of polyadenylation in five mammals. Genome Res 22: 1173–1183

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: Ultrafast universal RNA-seq aligner. Bioinformatics 29: 15–21

Fu H, Wang P, Wu X, Zhou J, Ji G, Shen Y, Gao Y, Li QQ, Liang J (2019) Distinct genome-wide alternative polyadenylation during the response to silicon availability in the marine diatom Thalassiosira pseudonana. Plant J 99: 67–80

Fu H, Yang D, Su W, Ma L, Shen Y, Ji G, Ye X, Wu X, Li QQ (2016) Genome-wide dynamics of alternative polyadenylation in rice. Genome Res 26: 1753–1760

Fu Y, Sun Y, Li Y, Li J, Rao X, Chen C, Xu A (2011) Differential genome-wide profiling of tandem 3′ UTRs among human breast cancer and normal cells by high-throughput sequencing. Genome Res 21: 741–747

Gruber AJ, Schmidt R, Gruber AR, Martin G, Ghosh S, Belmadani M, Keller W, Zavolan M (2016) A comprehensive analysis of 3′ end sequencing data sets reveals novel polyadenylation signals and the repressive role of heterogeneous ribonucleoprotein C on cleavage and polyadenylation. Genome Res 26: 1145–1159

Gruber AJ, Zavolan M (2019) Alternative cleavage and polyadenylation in health and disease. Nat Rev Genet 20: 599–614
Shen Y, Martin G, Müller P, Schmidt A, Gruber AJ, Gumienny R, Mittal N, Jayachandran B, Pieters J, Keller W, et al (2014) Global 3' UTR shortening has a limited effect on protein abundance in proliferating T cells. Nat Commun 5: 5465

Hong L, Ye C, Lin J, Fu H, Wu X, Li QQ (2018) Alternative polyadenylation is involved in auxin-based plant growth and development. Plant J 93: 246–258

Hoque M, Ji Z, Zheng D, Luo W, Li W, You B, Park JY, Yehia G, Tian B (2013) Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing. Nat Methods 10: 133–139

Ji G, Chen M, Ye W, Zhu S, Ye C, Su Y, Peng H, Wu X (2018) TSAPA: Identification of tissue-specific alternative polyadenylation sites in plants. Bioinformatics 34: 2123–2125

Ji G, Guan J, Zeng Y, Li QQ, Wu X (2015) Genome-wide identification and predictive modeling of polyadenylation sites in eukaryotes. Brief Bioinform 16: 304–313

Ji Z, Lee JY, Pan Z, Jiang B, Tian B (2009) Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. Proc Natl Acad Sci USA 106: 7028–7033

Lee JY, Yeh I, Park JY, Tian B (2007) PolyA_DB: mRNA polyadenylation sites in vertebrate genes. Nucleic Acids Res 35: D165–D168

Li Y, Sun Y, Fu Y, Li M, Huang G, Zhang C, Liang J, Huang S, Shen G, Yuan S, et al (2012) Dynamic landscape of tandem 3' UTRs during zebrafish development. Genome Res 22: 1899–1906

Lin J, Xu R, Wu X, Shen Y, Li QQ (2017) Role of cleavage and polyadenylation specificity factor 100: Anchoring poly(A) sites and modulating transcription termination. Plant J 91: 829–839

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

Mayr C (2017) Regulation by 3'-untranslated regions. Annu Rev Genet 51: 171–194

Mayr C, Bartel DP (2009) Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. Cell 138: 673–684

Müller S, Rycak L, Aforo-Gruz F, Winter P, Zawada AM, Damrath E, Scheider J, Schmäh J, Koch I, Kahl G, et al (2014) APADB: A database for alternative polyadenylation and microRNA regulation events. Database (Oxford) 2014: bau076

Ni T, Yang Y, Hafer D, Yang W, Kiesewetter K, Wakabayashi Y, Ohler U, Peng W, Zhu J (2013) Distinct polyadenylation landscapes of diverse human tissues revealed by a modified PA-seq strategy. BMC Genomics 14: 615

Peng Z, Lu Y, Li L, Zhao Q, Feng Q, Gao Z, Lu H, Hu T, Yao N, Liu K, et al (2013) The draft genome of the fast-growing non-timber forest species moso bamboo (Phyllostachys edulis). Nat Genet 45: 456–461

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, Li QQ (2008b) Unique features of nuclear mRNA poly(A) signals and alternative polyadenylation in Chlamydomonas reinhardtii. Genetics 179: 167–176

Shen Y, Venu RC, Nobuta K, Wu X, Notibala V, Demirci C, Meyers BC, Wang GL, Ji G, Li QQ (2011) Transcriptome dynamics through alternative polyadenylation in developmental and environmental responses in plants revealed by deep sequencing. Genome Res 21: 1478–1486

Skinner ME, Usilov AV, Stein LD, Mangul CJ, Holmes IH (2009) JBrowse: A next-generation genome browser. Genome Res 19: 1630–1638

Smibert P, Miura P, Westholm JO, Shenker S, May G, Duff MO, Zhang D, Eads BD, Carlson J, Brown JB, et al (2012) Global patterns of tissue-specific alternative polyadenylation in Drosophila. Cell Rep 1: 277–289

Thomas PE, Wu X, Liu M, Gaffney B, Ji G, Li QQ, Hunt AG (2012) Genome-wide control of polyadenylation site choice by CPSF30 in Arabidopsis. Plant Cell 24: 4376–4388

Thomas-Chollier M, Sand O, Turatsinze JV, Janky R, Defrance M, Vervisch E, Brohée S, van Helden J (2008) RSAT: Regulatory sequence analysis tools. Nucleic Acids Res 36: W119–W127

Tian B, Graber JH (2012) Signals for pre-mRNA cleavage and polyadenylation. Wiley Interdiscip Rev RNA 3: 385–396

Tian B, Manley JL (2013) Alternative cleavage and polyadenylation: The long and short of it. Trends Biochem Sci 38: 312–320

Tian B, Manley JL (2017) Alternative polyadenylation of mRNA precursors. Nat Rev Mol Cell Biol 18: 18–30

Ulitsky I, Shkumatava A, Jan CH, Subtelny AO, Koppstein D, Bell GW, Sive H, Bartel DP (2012) Extensive alternative polyadenylation during zebrafish development. Genome Res 22: 2054–2066

Wang R, Regulski M, Tseng E, Olson A, Goodwin S, McCombie WR, Ware D (2018a) A comparative transcriptional landscape of maize and sorghum obtained by single-molecule sequencing. Genome Res 28: 921–932

Wang R, Nambiari R, Zheng D, Tian B (2018b) PolyA_DB 3 catalogs cleavage and polyadenylation sites identified by deep sequencing in multiple genomes. Nucleic Acids Res 46: D315–D319

Wang T, Wang H, Cai D, Gao Y, Zhang H, Wang Y, Lin C, Ma L, Gu L (2017) Comprehensive profiling of rhizome-associated alternative splicing and alternative polyadenylation in moso bamboo (Phyllostachys edulis). Plant J 91: 684–699

Wu X, Gaffney B, Hunt AG, Li QQ (2014) Genome-wide determination of poly(A) sites in Medicago truncatula: Evolutionary conservation of alternative poly(A) site choice. BMC Genomics 15: 615

Wu X, Liu M, Downing B, Liang C, Ji G, Li QQ, Hunt AG (2011) Genome-wide landscape of polyadenylation in Arabidopsis provides evidence for extensive alternative polyadenylation. Proc Natl Acad Sci USA 108: 12533–12538

Wu X, Zhang Y, Li QQ (2016) PlantAPA: A portal for visualization and analysis of alternative polyadenylation in plants. Front Plant Sci 7: 889

Xing D, Li QQ (2011) Alternative polyadenylation and gene expression regulation in plants. Wiley Interdiscip Rev RNA 2: 445–458

Ye C, Long Y, Ji G, Li QQ, Wu X (2018) APAtrap: Identification and quantification of alternative polyadenylation sites from RNA-seq data. Bioinformatics 34: 1841–1849

You L, Wu J, Feng Y, Fu Y, Guo Y, Long L, Zhang H, Luan Y, Tian P, Chen L, et al (2015) APASdb: A database describing alternative poly(A) sites and selection of heterogeneous cleavage sites downstream of poly(A) signals. Nucleic Acids Res 43: D59–D67

Zhang H, Hu J, Recce M, Tian B (2005) PolyA_DB: A database for mammalian mRNA polyadenylation. Nucleic Acids Res 33: D116–D120

Zhou Q, Fu H, Yang D, Ye C, Zhu S, Lin J, Ye W, Ji G, Ye X, Wu X, et al (2019) Differential alternative polyadenylation contributes to the developmental divergence between two rice subspecies, japonica and indica. Plant J 98: 260–276