Journal of Animal & Plant Sciences, 35(1): 2025, Page: 158-173
ISSN (print): 1018-7081; ISSN (online): 2309-8694
https://doi.org/10.36899/JAPS.2025.1.0012

GENOME-WIDE ANALYSIS OF POLYADENYLATION SITES IN *Glycine max*

W. Shah[1], M. Sajjad[1], N. Akhtar[2] and M. N. Akhtar[1]*

[1]Department of Biosciences, COMSATS University Islamabad, Islamabad 45550, Pakistan;
[2]Department of Health Informatics, University of Hail, Hail, Saudi Arabia

*Corresponding author's email: nadeemakhtar@comsats.edu.pk

ABSTRACT

Alternative polyadenylation (APA) is a critical cellular process that dynamically regulates gene expression and contributes to transcriptome and proteome diversity by impacting about 70% genes in animals and plants. However, the lack of extensive 3'-sequencing data limits comprehensive understanding of polyadenylation in *Glycine max.* This study aimed to address this by identifying high quality polyadenylation clusters (PACs) using 12 billion reads from the 568 RNA-Seq samples. This study identified 75,556 PACs in the *Glycine max* genome, primarily in 3'-UTRs but also in 5'-UTRs, introns, and intergenic regions. Intergenic PACs and RNA-Seq evidence extended the 3'-ends of many genes, revealing annotation gaps. APA was observed in 65% of the genes, much higher than 19% noted in Ensemb l annotations. APA genes depicted complex PAC expression, with dominant PACs linked to diverse cellular processes including translation, stability, transport, cellular organization, and stress response. Using a uniform criterion, the nucleotide composition and motifs in *Glycine max* were extensively compared with plants including *Oryza sativa*, *Arabidopsis thaliana, Medicago truncatula*, and *Zea mays*. The results highlighted preference for AAUAAA and its variant motifs, which were less frequent in all plants. However, *Glycine max* top 3'-UTRs motifs showed conservation and appeared consistently as top motifs across other plants. Additionally, nucleotide composition in AAUAAA region was conserved, but far upstream region diverged between monocotyledonous and dicotyledonous plants groups. Genes with AAUAAA were involved in metabolic processes consistent with *Zea mays* indicating evolutionary constraints. Taken together, our results offer a comprehensive resource for understanding polyadenylation mediated gene regulation in *Glycine max*.

Keywords: Alternative Polyadenylation; Incomplete 3'-UTR; AAUAAA; Monocotyledonous; Dicotyledonous; Transcriptome; RNA-Seq; Annotation; Far Upstream Region; Soybean

Published first online December 26, 2024 Published final February 18, 2025

INTRODUCTION

Polyadenylation is a crucial post-transcriptional modification in eukaryotes that defines the 3'-end of mRNA transcripts. This two-step process involves cleavage of pre-mRNA at specific site (termed as polyadenylation site or poly(A) site) followed by the addition of multiple adenosine residues. The site of cleavage is determined by the highly coordinated sets of transcription factors that recognize specific sequence motifs located around the poly(A) site (Rodríguez-Molina and Turtola, 2023).

In plants, poly(A) sites are mainly AT rich and contain three main sequence elements that are essential for the accurate polyadenylation (Sun *et al*., 2020; Liu *et al*., 2022). These elements are far upstream element (FUE), the near-upstream element (NUE), and the cleavage element (CE) These motifs, together with the surrounding nucleotide composition direct the polyadenylation machinery to the correct poly(A) site (De Felippes and Waterhouse, 2023; Torres-Ulloa *et al*., 2023). The NUE element is one of the most highly conserved polyadenylation signals, which includes the canonical AAUAAA motif and its single or multi-base variants. The NUE element is generally located between 10-30 nts upstream of the poly(A) site in A-rich region (Bell *et al*., 2016; Biłas *et al*., 2016; Zhao *et al*., 2019; Bernardes and Menossi, 2020). The region upstream of the NUE contains FUE element that is located between 30 to 100 nts upstream of poly(A) site within the U-rich region (Song *et al*., 2021). The third element is CE, which is a U-rich region located on either side of the poly(A) site (De Felippes and Waterhouse, 2023). The processing efficiency and choice of poly(A) site vary significantly depending upon the presence and combination of these different sequence elements (Chakrabarti *et al*., 2018; Pereira-Castro and Moreira, 2021; Gorjifard *et al*., 2024).

APA adds further complexity to the polyadenylation process. APA refers to the use of multiple poly(A) sites within the same gene, resulting in mRNA isoforms with

varying 3'-UTRs or coding sequences (Neve and Furger, 2014). This mechanism is widespread in plants and affects transcript stability, localization, and translational efficiency (Wu *et al*., 2023). In several plant species, including *Arabidopsis thaliana*, *Oryza sativa*, and *Zea mays*, APA has been shown to be a significant regulatory mechanism that impacts approximately 70% of genes (Wu *et al*., 2011; Wu *et al*., 2014; Wu *et al*., 2015; Wu *et al*., 2019; Yan *et al*., 2021). The variations in 3'-UTR length is considered a dominant mechanism to regulate gene expression with approximately 83% of the APA sites found within the 3'-UTR regions of *Arabidopsis thaliana* (Peng *et al*., 2023). In different plants significant polyadenylation has been detected in CDS, introns, and intergenic regions that can contribute to transcriptome diversity (Wu *et al*., 2014; Wu *et al*., 2015; Fu *et al*., 2016; Wu *et al*., 2019; Yan *et al*., 2021). Previous studies highlight the central role of APA in plant development including leaf development (Yu *et al*., 2022), different cell types of roots (Bi *et al*., 2024), and response to biotic (Liu *et al*., 2024) and abiotic stresses (Téllez-Robledo *et al*., 2019; Wang *et al*., 2023). Given the significant role of APA, comprehensive characterization of polyadenylation is crucial for improving understanding about the polyadenylation related gene regulation.

Several experimental techniques have been designed to sequence 3'-end of the mRNAs including 3'-Reads, TAIL-seq, PAT-seq (Ye *et al*., 2023). Furthermore, rapidly growing RNA-Seq data provides an excellent alternative resource to identify poly(A) sites in organisms that lack 3'-specific sequencing data (Jafar *et al*., 2019). The sequencing data from these experimental techniques have been used to develop different databases with poly(A) specific data including PlantAPAdb, and PlantAPA. However, most databases provide poly(A) sites in limited numbers of plant species including *Oryza sativa*, *Arabidopsis thaliana*, *Chlamydomonas reinhardtii*, *Medicago truncatula*, *Trifolium pratense*, *Populus trichocarpa* and *Phyllostachys edulis* (Zhu *et al*., 2020). Understanding the polyadenylation and APA landscape in key crop species like G*lycine max* is essential due to its significant role in agricultural productivity and its complex genome. Despite its importance, the polyadenylation landscape in *Glycine max*, a major source of protein and oil globally (Du *et al*., 2023), has not been extensively studied. The exact proportion of genes that undergo APA, genomic distributions of APA sites, and motifs remains unknown. Given the significance of polyadenylation in regulating gene expression, understanding the distribution and characteristics of poly(A) sites across the *Glycine max* genome can offer new insights into both basic biological processes and applied biotechnological traits.

In this study, present a comprehensive mapping of poly(A) sites in the *Glycine max* genome, identifying 75,556 high quality PACs spanning multiple genomic regions using 12 billion reads from the 568 RNA-Seq samples. This data covers approximately 51% of annotated protein-coding genes in *Glycine max*, providing a detailed view of poly(A) site distribution and robust estimation of APA events. APA occurred in 65% of *Glycine max* genes, far higher than distinct transcript ends reported in Ensembl database. Additionally, this study highlights the potential for improving gene annotations. Our analysis focused on the sequence elements and nucleotide composition surrounding PACs and their conservation across plants. The findings reported here advance our understanding of polyadenylation in Glycine max, providing a valuable resource for future research on gene regulation and efforts to improve gene annotations.

MATERIALS AND METHODS

Data source and poly(A) site discovery: A total of 21 RNA-Seq datasets of William 82 cultivar of *Glycine max* consisting of 568 samples were retrieved from the NCBI Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra) database in the SRA format. The selection of William 82 was based on availability of large numbers of transcriptomic samples as compared to other *Glycine max* cultivars.

The poly(A) sites were identified in each sample as described previously (Jafar *et al*., 2019)**.** Briefly, the raw sequence reads were extracted from the SRA files in the FASTQ format using *fastq-dump* software available in the NCBI SRA tool kit (2.10.9) (Leinonen *et al*., 2020)**.** For the paired-end data, read one and read two were extracted from the SRA files into separate FASTQ files using "-split--files" parameter in the *fastq-dump* program. The extracted reads were filtered to remove low quality bases and adapter sequences using *Trimmomatic* software and reads that were less than 25 nts in length were discarded (Bolger *et al*., 2014)**.** The 5'-end or 3'-end of the filtered reads were checked for the presence of poly (A) tails using *FindTail* program depending upon the library type (stranded or unstranded). For the unstranded data, reads with eight or more A-residues at the 3'-end or T-residues at the 5'-end with a maximum of two non-A or T-residues were considered as polyadenylated reads. For the stranded data, the first reads with eight or more T-residues at the 5'-end with a maximum of two non-T-residues were considered as the polyadenylated reads. Considering the post-transcriptional nature of the polyadenylation process, the selected polyadenylated reads should not map to the *Glycine max* genome. To ensure that the poly (A) or poly (T) tails didn't come from the genomic A- or T-rich regions, the selected reads were mapped to the reference genome (*Glycine_max_v2.1*) of *Glycine max* retrieved from the Ensembl database using *bowtie2* software (version 2.2.9) (Langmead and Salzberg, 2012). The alignment was performed in the "end-to-end" fashion with sensitive settings and the un-mapped reads were retrieved

in the SAM format (Li *et al*., 2009) using "–unal" parameter. The uncalled (N) residues in the tails were randomly replaced with A, T, C, or G residues using *the ReplaceN* program. The un-mapped reads with tail identity <90% were discarded to allow imperfect poly (A) tails using *FilterIdentity* program. Finally, the tails of the filtered polyadenylated reads were trimmed and re-mapped to the *Glycine max* reference genome using *bowtie2* program in the "end-to-end" fashion using sensitive settings. Only reads that mapped uniquely to the genome were retained further. The *CleanSam.Pl* script was used to define cleavage sites after discarding reads with internal priming that was defined as the presence of nine or more A-residues in a 15nt window around the read end. Finally, cleavage sites from all the samples were pooled together for the analysis.

The stochastic nature of the polyadenylation process generates considerable heterogeneity in the cleavage sites in which multiple cleavage sites occur within few nucleotides of each other (Tian *et al*., 2005). To reduce heterogeneity, the cleavage sites with two or more than two supporting reads and located within 24 nts of each other on the same chromosome and strand were clustered. For each cluster, the cleavage site with the highest number of supporting reads was considered as the representative and denoted as the PAC (Poly(A) Cluster).

Annotation of the PAC and transcript assembly: Initially Ensembl transcript annotations (release 51) of *Glycine max* were used to annotate PACs. According to Ensembl annotations, 57,147 genes encoded a total of 89,662 transcripts. Each PAC was classified as the 5'-UTR, CDS, intronic, 3'-UTR, or intergenic PAC based on the selected Ensembl transcript annotations. The PACs that were not located within any transcript on the same strand were treated as intergenic PACs. The distance between the intergenic PACs and the 3'-end of the closest upstream gene was recorded. For the PACs situated in different genomic regions of more than one transcript, genomic region was assigned to PAC in the following order: 3'-UTR > intron > CDS > 5'-UTR.

The absence of well-characterized 3'-UTRs can lead to inaccurate classification of the large numbers of potential 3'-UTR PACs into the intergenic PACs. An extension in the 3' ends of the annotated transcripts improves PAC annotation in the species with incomplete 3'-UTR annotations as described previously (Zhao *et al*., 2014; Wang *et al*., 2016)**.** All intergenic PACs mapped within 50 nts downstream of the Ensembl gene models were considered as the "extended" 3'-UTRs. Intergenic PACs that mapped downstream of the Ensembl gene models were considered as the extended 3'-UTR PAC if the connectivity was supported by the RNA-Seq assembled transcripts.

For the transcript assembly, the RNA-Seq samples representing multiple cell types and seed development stages were retrieved from the SRA database (SRP006767). The adapter clipping, read quality, and read length filtering were performed using *Trimmomatric* software (Bolger *et al*., 2014). For each sample, the filtered reads were mapped to the *Glycine max* genome using *hisat2* software (version 2.2.1) with default parameters. The resulting alignments in SAM format were sorted and converted into the BAM format using SamTools (version 1.17). StringTie was used to assemble transcripts from each BAM file using Ensembl annotation as a reference with default parameters. Finally, transcript assemblies were merged to generate non-redundant transcriptome across all 21 RNA-Seq samples using StringTie (version 2.1.1). Only those assembled genes that overlapped with the Ensembl annotated genes were retained. The 3'-most transcript termini were considered as the genes end coordinates to check the connectivity of the intergenic PACs.

Comparison with known transcript ends: Based on the Ensembl transcript annotation (release 51), a total of 55,897 protein-coding genes encoded 88,412 protein-coding transcripts with non-unique 3'-ends. To make a meaningful comparison between known transcript ends and PACs, only those Ensembl genes and transcripts were selected where at least one PAC was located either within the gene or within 1 kb downstream of the Ensembl-annotated 3'-end of gene. Additionally, transcripts lacking annotated 3'-UTRs, regardless of length, were excluded from further analysis. To eliminate redundancy in 3'-end transcript annotations, if multiple transcript 3'-ends of the same gene were located within 24 nts of one another, only one representative transcript end was retained. The distance between PACs and the selected 3'-ends of known transcripts was then measured to assess the accuracy of the reported PACs.

Comparison with PACs of related plants species: *Glycine max* PACs were compared with previously reported PACs in *Oryza sativa*, *Arabidopsis thaliana*, *Medicago truncatula*, and *Zea mays* to study the similarities and differences in PAC regions across different plants. These plants were selected due to the availability of sufficient amount of PAC data with classification into different genomic regions. The 400 nts FASTA sequence of *Oryza sativa*, *Arabidopsis thaliana*, and *Medicago truncatula* PACs were downloaded from the PlantAPAdb (accessed on May 24, 2024; http://www.bmibig.cn/plantAPAdb/APAsequence.php). The *Zea mays* PACs were obtained from the previous study (Jafar *et al*., 2019)**.** For each species, both 3'-UTR and extended 3'-UTR PAC were retrieved and pooled together. A total of 39,837, 31928, 13,090, and 57,963 PACs in *Oryza sativa*, *Arabidopsis thaliana*, *Medicago truncatula*, and *Zea mays* were used to perform the nucleotide composition and NUE motif comparisons with *Glycine max*, respectively.

Nucleotide composition analysis : The 400 nts [-300: +100] sequence surrounding each PAC was extracted from the *Glycine max* genome. The position-specific single nucleotide counts were computed using the SignalSleuth2.pl program with -gap =0 and -k=1 parameters in the 400 nts region (Zhao *et al*., 2014). The counts were converted into relative frequency using the total nucleotide sum for each position. A similar procedure was adopted to compute the nucleotide composition for the *Oryza sativa*, *Arabidopsis thaliana*, *Medicago truncatula*, and *Zea mays* species.

Identification of NUE motif: The significant 6-nt motifs in the NUE regions (-40:-11) of the different PAC types were identified using the "oligo-analysis" program in the RSAT package (version 1.169; https://rsat.eead.csic.es/plants/oligo-analysis_form.cgi). The program detected the over-represented motifs in the PACs sequences by comparing their observed frequency with the expected occurrence by chance estimated from the background sequences. For the background model, a 2nd-order markov model was used. The motifs were counted only on the provided single strand and sequence purge option was disabled. The program reported different significant values based on binomial distribution including occurrence probability and E-value of occurrence. Here, motifs with E-value < 0.05 were considered as significant motifs. An iterative procedure was employed to avoid improper selection of the less significant motifs that share NUE region with more frequent motifs (Beaudoing *et al*., 2000). In each iteration, the PAC sequences containing the most significant NUE motif reported by RSAT were removed and the procedure was repeated until no significant motif was detected or the top 20 most significant motifs were detected for each PAC type. The later criterion was adopted considering the degenerate nature of the plant NUE motifs. Similar procedure was used to identify the significant NUE motifs in *Oryza sativa*, *Arabidopsis thaliana*, *Medicago truncatula*, and *Zea mays*.

APA Analysis: Genes having at least two PACs were considered APA genes. Total reads mapped on a given gene were used to estimate the relative expression (RE) of each PAC within the gene. The PACs were ranked in descending order of the RE values. The total number of reads mapped on ranked-wise PACs across all genes were used to analyze the global distribution of PAC choice and preferences within the APA genes. Furthermore, to quantify the number of genes that showed any preference towards the PAC selection, APA genes having a PAC with RE value greater than 0.70 were considered as genes with "dominant" PACs, while genes that lacked any "dominant" PACs were considered "medium" genes (Zhao *et al*., 2014).

Gene Ontology (GO) analysis: The GO analysis was performed using the Statistical overrepresentation test in PANTHER (https://www.pantherdb.org/; version 18.0) to investigate the significant GO-Slim biological processes (BP) within the genes of interests compared to the complete set of *Glycine max* genes in the PANTHER database. A Fisher's exact test was used, and Bonferroni correction was employed to account for multiple hypothesis testing. The GO terms with adjusted p-value < 0.05 were considered significant.

RESULTS

Poly(A) Sites in Glycine max Genome: To assess the landscape of poly(A) sites in *Glycine max*, the 568 RNA-Seq samples from 21 datasets of William 82 cultivar were retrieved from the SRA database (Supplementary Table S1). Collectively, these samples were comprised of 12,609,303,263 raw reads. Subsequently, 11,966,785,637 reads survived the read quality and length filtering criteria using Trimmomatic software. Depending upon strand specificity, 9,139,866 reads that exhibited untemplated stretches of A-residues at their 3′ termini or T-residues at 5′ termini with the reference genome were considered as polyadenylated reads. The tails of these reads were trimmed and re-aligned with the reference genome. After filtering out potential cases of internal priming within the genome we considered only 3,802,711 uniquely mapped reads and these reads were pooled across all samples to identify a total of 454,042, distant cleavage sites in *Glycine max*. Here, 209,370 cleavage sites that were supported by at least two reads amongst all libraries are reported. The distance between cleavage sites supported by at least two reads was calculated which indicated that the large numbers of cleavage sites were adjacent to each other in the genome (Figure 1A). The adjacent cleavage sites that fell within 24 nts of each other were iteratively clustered and dominant cleavage site was selected as cluster representative (termed as Poly (A) Cluster or PAC). In total, 75,556 high confidence PACs supported by at least two reads were identified (Supplementary Table S2).

Polyadenylation processing is imprecise in nature in which cleavage can occur at multiple favorable nucleotides that are few base pairs apart *(*Jafar *et al., 2019)*. About 46.83% (n = 35,380) of the total PACs showed heterogeneity with an average of 2.8 and standard deviation of 3.96 cleavage sites per PAC**.** The Pearson correlation between the number of cleavage sites per PAC and numbers of supporting reads for each representative site in the PACs was calculated. A strong correlation (0.76; p-value < 2.2×10^{-16}) was observed indicating that PACs with higher expression tends to show more heterogeneous 3'end formation in *Glycine max* (Figure 1B).

Over 75,000 PACs were assigned to genomic features based on Ensembl annotated transcripts (release 51). A vast majority (75.5%, n = 57,039) of the PACs were distributed in annotated protein-coding genes and only 31 were in non-coding RNA genes. The PACs found in non-Coding RNAs which mainly included snoRNA (n = 7), snRNA (12), and SRP RNA (9) were excluded from further analysis. Importantly, the highest numbers of protein coding PACs were distributed in 3'-UTRs (54,446 out of 57,039), whereas only 258, 421, and 1,914 were located in 5'-UTR, CDS, and introns, respectively (Figure 1C). A similar trend was observed in terms of proportion of mapped reads where 88.58% of protein coding reads mapped onto 3'-UTR regions (Figure 1D). The mapping of large numbers of PACs and supporting reads closer to the expected 3'-end of transcripts in 3'-UTR region highlights the quality of the detected PACs.

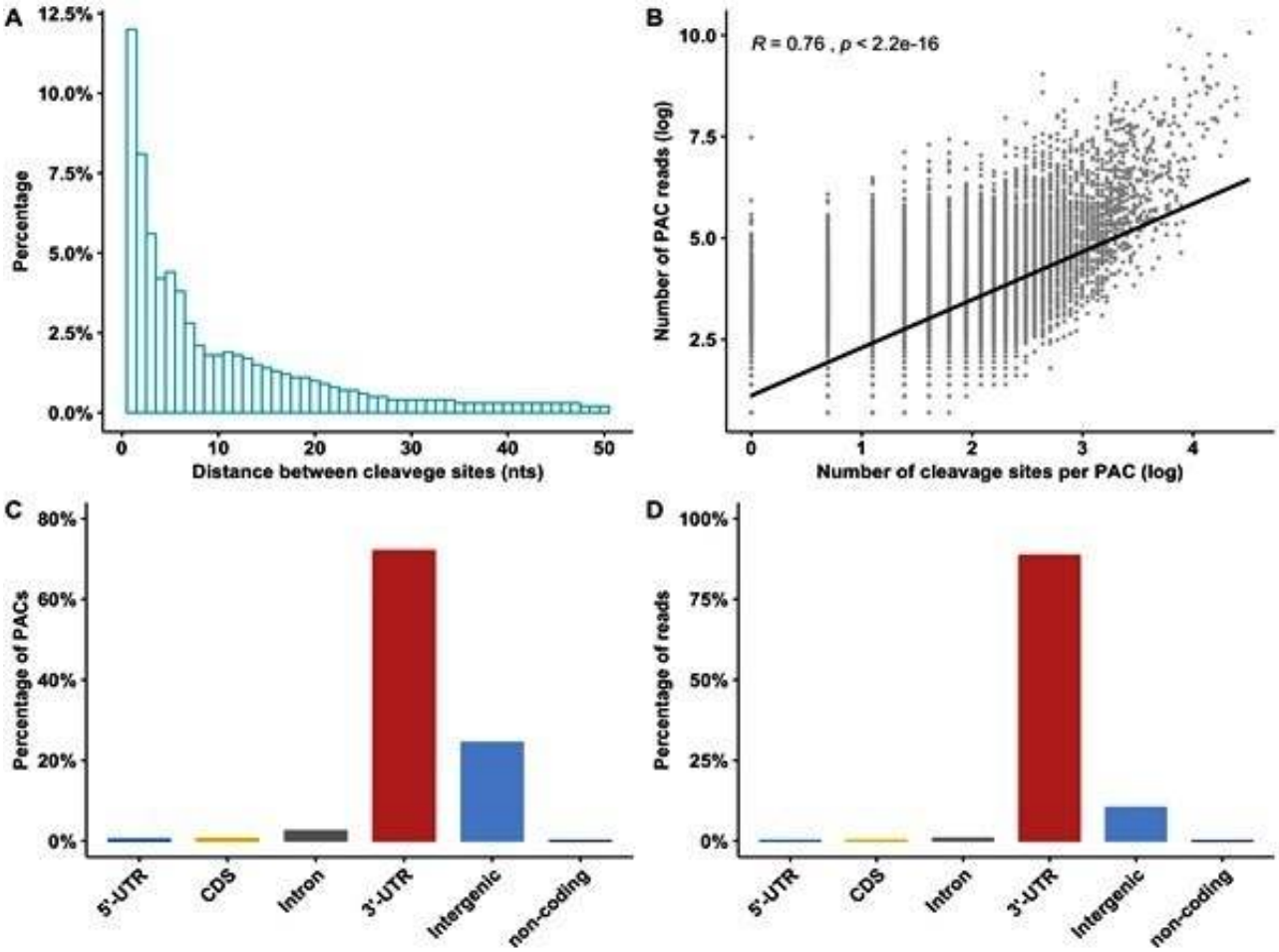

Figure 1: Heterogeneity of cleavage sites and distribution of PACs in the *Glycine max* genome. (A) The distribution of distances between the adjacent cleavage sites. (B) Pearson correlation between the natural log of the cleavage sites per PAC and the number of supporting reads for each representative site in the PACs (R = 0.76, p-value < 2.2 x 10^{-16}). (C) Genomic distribution of PACs regarding the genic and intergenic regions in the *Glycine max* genome. (D) The distribution of mapped polyadenylated reads to various genomic regions, including the 5′ UTR, CDS, Introns, 3′ UTR, intergenic regions, and non-coding RNAs.

Extensions in 3' gene ends: A total of 18,486 PACs mapped outside of the Ensembl transcript annotations and were termed as intergenic PACs (Figure 1C). For further examination of the intergenic PACs, the distance distribution of the 13,833 PACs that were within 2Kb of the longest transcript of the nearby upstream gene was analyzed. The vast majority of these PACs were located immediately downstream of the existing Ensembl genes than further downstream in the intergenic region. About 54.5% and 81% of all intergenic PACs were mapped within 75 nts and 300 nts of the Ensembl genes, respectively (Figure 2A).

To further explore these, Ensembl genes with at least one PAC located within 2Kb region were classified into two groups, those with and without annotated 3'-UTRs (Figure 2A). The comparison of the distribution of

intergenic PACs in genes with and without Ensembl annotated 3'UTR indicated contrasting trends. Overall, much more intergenic PACs (71.4%) were situated downstream of the genes that had annotated 3'-UTRs relative to genes without annotated 3'-UTRs. In genes with annotated 3'UTR, the vast majority of the intergenic PACs were concentrated immediately downstream of the gene ends with about 60% and 77% in 50 nts and 100 nts, respectively. On the other hand, genes without annotated 3'UTRs showed more gradual accumulation of intergenic PACs over a much greater distance with only 3.6% and 10% of PACs located within 50 nts and 100 nts, respectively. Previous studies in other plants have shown that the large fraction of these PACs could be due to incomplete 3'-end annotations of the already annotated genes (Wu *et al*., 2014; Zhao *et al*., 2014; Fu *et al*., 2016; Jafar *et al*., 2019).

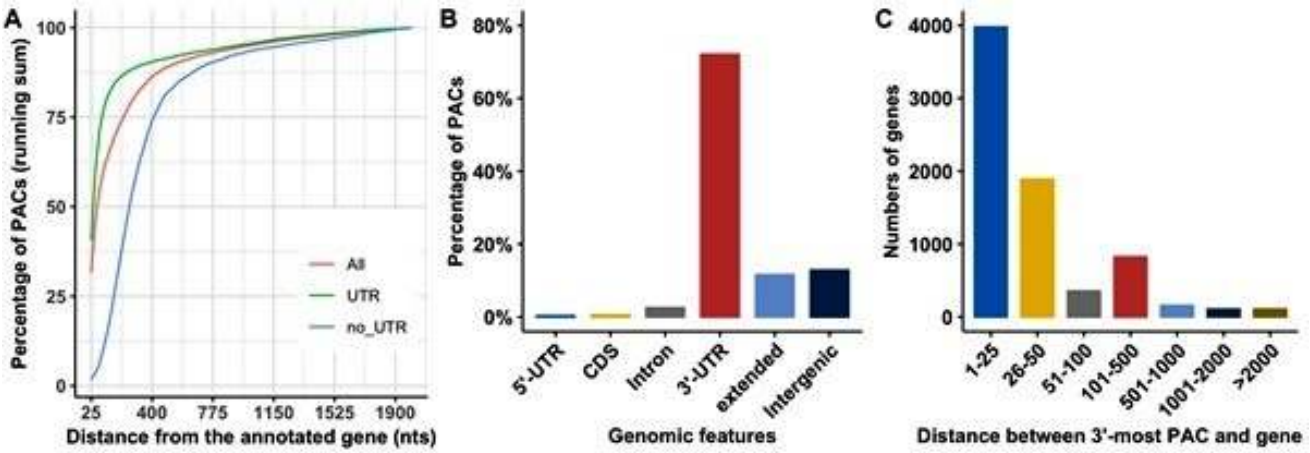

Figure 2: Extension in the 3′ UTR based on Ensembl annotated gene ends. (A) The distribution of distances from the annotated genes with or without annotated 3′ UTR and the intergenic PACs. Most PACs were located near the annotated genes. (B) The percentage of PACs after 3′ UTR extension in various genomic regions of the *Glycine max*. (C) The distribution of the extended gene ends based on 3'-most PAC.

To further improve 3'-end annotation of the Ensembl genes, RNA-seq data was used as additional evidence to connect intergenic PACs with the upstream transcript models. Briefly, RNA-seq reads were mapped to genome using hisat2 software and transcripts were assembled in reference-guided manner for which Ensembl transcripts were used using StringTie software. Assembled transcripts from the individual samples were pooled to generate a single transcript file in reference-guided manner. RNA-seq assembled genes that overlapped with the Ensembl annotated genes were used to extend the 3'-ends of the Ensembl genes. A total of 4,437 (24%) of the intergenic PACs that were within the extended Ensembl gene coordinates were considered as "extended" 3'-UTR regions and part of upstream Ensembl gene. Any remaining intergenic PAC that was not supported by RNA-seq based extensions were considered "extended" if they were within 50nts of the Ensembl annotated 3'most gene termini. As indicated in the Figure 2A, most genes that can benefit from this second 50 nts criteria were genes with already annotated 3'-UTR. Based on this second criterion, an additional 4,333 (23%) intergenic PACs were classified as part of "extended" 3'-UTRs regions. The remaining 9,716 PACs that were not supported by either of the above two extension criteria were considered as intergenic PACs. Figure 2B and Supplementary Table S2 provide the final breakdown and complete list of the PACs in different genomic regions of *Glycine max*.

The Figure 2C summarizes the numbers of 3'-end extensions of Ensembl gene models based on the distance between the Ensembl annotated gene ends and 3'-most extended PAC. Overall, 3'-ends of 7,422 protein coding genes were extended by an average of 176 nts (SD = 1133; median = 25 nts). The large majority of the genes (79%) were extended by 1-50 nts, and more than thousand genes were extended by >100 nts.

Comparison of PACs with known transcripts: To evaluate how well PACs identified in this study agree with the known Ensembl transcript 3'ends, a total of 27,998 protein-coding genes with at least one PAC or PAC within 1Kb downstream of the 3' end boundary was identified. These genes collectively encoded 45,376 protein-coding transcripts with annotated 3'-UTRs (regardless of the 3'-UTR size). As genes can have multiple transcripts with same 3'-ends or with 3' ends in close proximity to each other, a 24 nts proximity rule was applied to identify unique 3' transcript ends for each gene. A total of 32,373 unique 3' transcript ends were compared with associated PACs in each gene. The analysis of the distance between these transcript ends and PACs revealed a significant

majority of the transcripts (69%) had a supporting PAC within 100 nts (Figure 3). A strong majority of these transcript ends (55.5%) had supporting PACs within 50 nts, highlighting the quality of the detected PACs in this study.

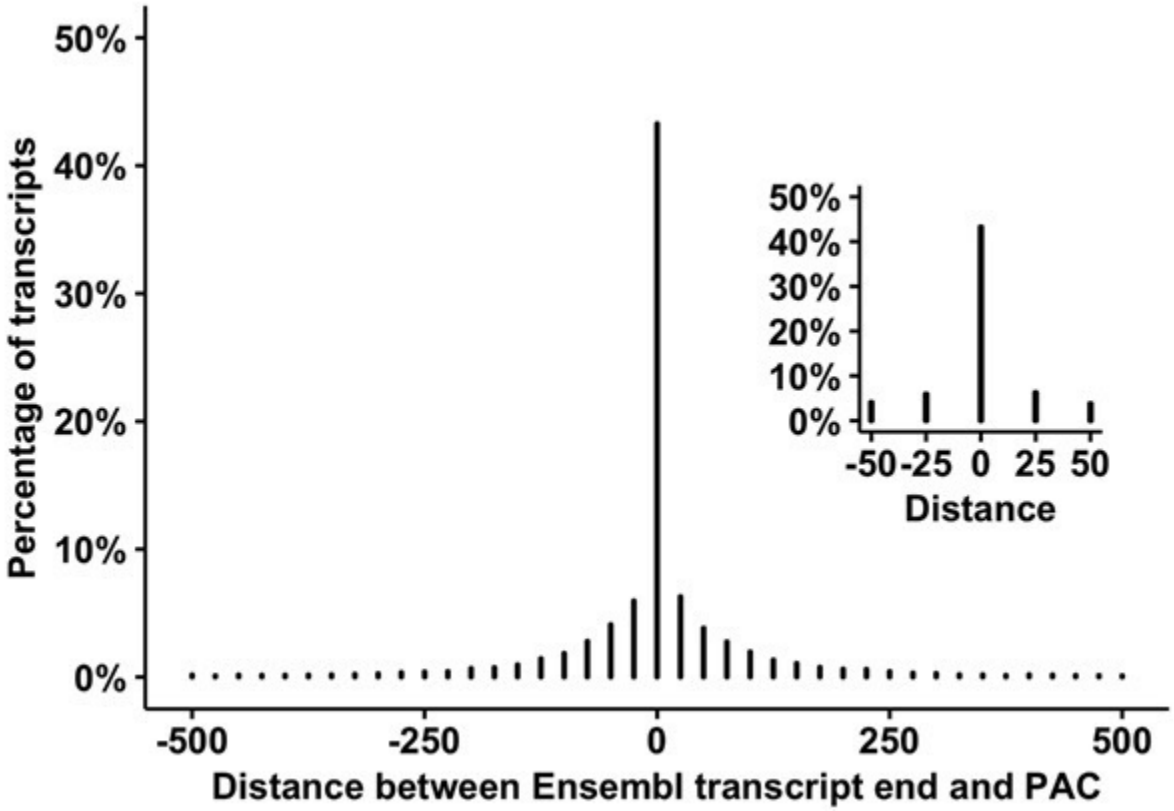

Figure 3: Comparison of PACs with known Ensembl transcripts ends. The distance between the Ensembl transcript 3'-ends and closest PAC was measured and plotted. The plots show that the most detected PACs were within few nts of the annotated Ensembl 3'-transcript end. The inset plot highlights excellent agreement between the PACs and transcript ends.

Single nucleotide composition exhibits hallmark of polyadenylation region: The 3' sequence elements that reside both upstream and downstream within approximately 150 nts PACs drive efficient processing of mRNA. Within this 150 nt region, different 3' sequence elements exhibit both nucleotide composition and positional preferences (Gorjifard *et al*., 2024). The single nucleotide composition analysis is useful in studying gradual transitions in nucleotide composition across regions of different 3' sequence elements. To assess the single nucleotide composition in Glycine max PAC, 150 nts (-100 to +50 with CS at +1) were selected as regions longer than this generally tends to show richness of certain residues but no transitions mainly in nucleotide compositions (Jafar *et al*., 2019).

The nucleotide composition in all PACs of *Glycine max* showed U and A-richness with clear transitions between nucleotides in different regions (Figure 4A). In the FUE region (-100 to -30) of PACs, a clear U-richness was observed (average U content was 39%) exhibiting a nucleotide pattern of U>A>G>C in this region. Moving further towards PAC from this region, A-richness (37%) was prominent in the 10 nts long region (-25 to -15) with a peak around -20 (39%). This region showed a gradual drop in T, C, and G content until -18 positions, resulting in small grooves in three contents. The -10 to -1 region immediately upstream of the cleavage site showed U-richness (46%). As expected, U and A peaks were observed at positions -1 and 0, respectively. The downstream region (+3 to +30) demonstrated U-richness with an average value of 41%. The average values of the enriched residue in different regions of *Glycine max* PACs were compared with average values of PACs in 3'-UTR of *Arabidopsis thaliana*, *Medicago truncatula*, *Oryza sativa*, and *Zea mays* (Supplementary Table S3). In all regions surrounding PACs, enrichment of similar residues was observed in all plants despite variations in the average values.

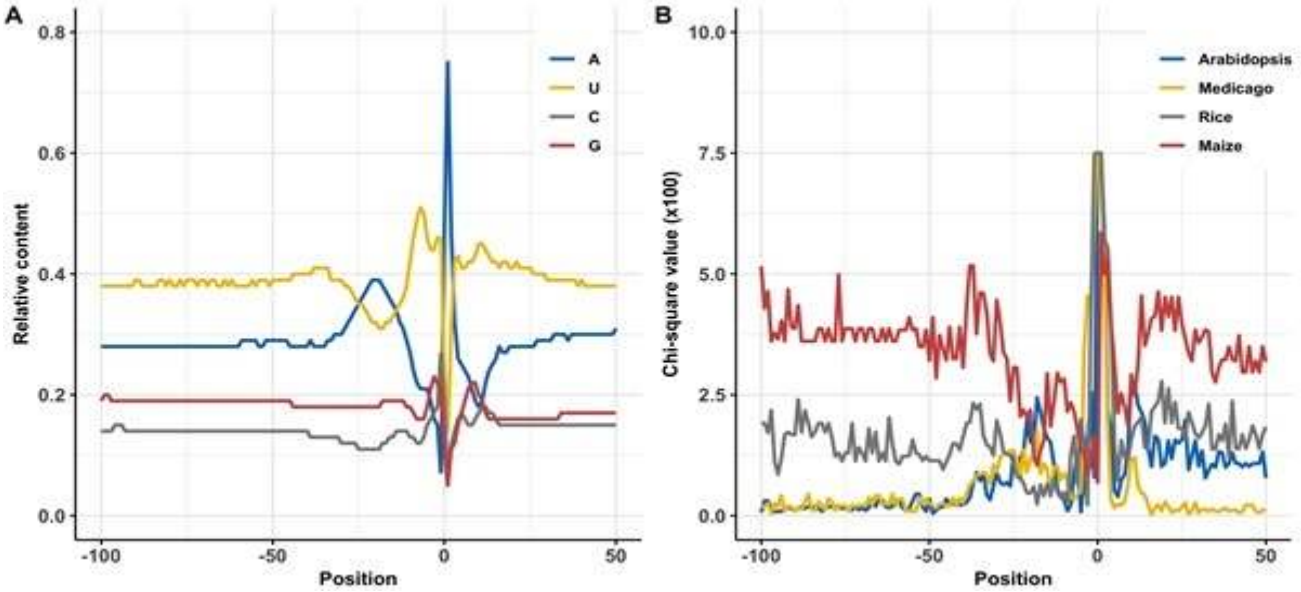

Figure 4: Analysis of single nucleotide composition. (A) Analysis of single nucleotide composition in all *Glycine max* PACs surrounding -100 to +50 base pairs. The x-axis represents PAC location and Y-axis represents position-wise nucleotide composition. In the graph, the positive values indicate a downstream region, while negative numbers indicate an upstream region from the PAC. (B) The comparison of single nucleotide profiles of *Glycine max* with the 3'-UTR regions of *Arabidopsis thaliana*, *Medicago truncatula*, *Oryza sativa*, and *Zea mays* using the x^2 values. The values are multiplied with 100 and maximum value is set to 7.5 to improve comparison.

The single nucleotide profiles of *Glycine max* were compared in position-by-positions manner with single nucleotide compositions observed in 3'-UTR regions of *Arabidopsis thaliana*, *Medicago truncatula*, *Oryza sativa*, and *Zea mays* using the x^2 values (Figure 4B). In the U-rich FUE region (-100 to -30), the nucleotide composition of *Glycine max* was more similar to other dicotyledonous plants relative to the monocotyledonous plants. The average U-content was slightly higher in dicotyledonous plants than monocotyledonous (Supplementary Table S3). Interestingly, the differences in the nucleotide compositions of the monocotyledonous and dicotyledonous plants deceased in NUE region and both sides of the cleavage sites (Figure 4B). In these regions nucleotide profiles of both dicotyledonous and monocotyledonous plants were almost similar.

NUE hexanucleotide motifs: The significant motifs in the NUE region of PACs located across various genomic regions were detected using the RSAT program as described in the methods section. The results for 3'-UTR and all regions are summarized in Table 1 and Supplementary Table S4, respectively. The analysis of the significant NUE motifs in PACs located across different genomic region indicated the preference for the diverse and distinct set of NUE motif sequences that occur in low percentages. Among the top 20 motifs in the 3'-UTR and intergenic PACs, only four (AAUAAA, AAUGAA, AAUAGA, and AUGAUA) were common. Fewer significant motifs were detected for the 5'-UTR (n = 1), CDS (2), and intronic (2) PACs (Figure 5A). The AAUAAA was the most frequent and significant motif in 3'-UTR, intronic, and intergenic region accounting for 7%, 8.05%, and 8.43% of the PACs, respectively. However, this motif was not significantly detected in PACs of 5'-UTR and CDS regions. Similarly, the UAUAUA motif was the second most common and significant motif in the 3'-UTR (4.5%) and intronic regions (4.6%), however, this motif was not among the top 20 significant motifs of the intergenic region. This motif initially appeared in list for first little iteration but was less significant. Similarly, ranked three AAUGAA motifs in 3'-UTR region was at 11 in intergenic region. These motifs were less significant and appeared in the same sequences along with more significant motifs which ultimately lead to either decrease in rank or not selection in top 20.

The NUE motifs in the 3'-UTR PACs of Glycine max were compared with NUE motifs in 3'-UTR PACs of related plant species including *Arabidopsis thaliana, Oryza sativa, Medicago truncatula*, and *Zea mays*. The results are summarized in Table 1 and Supplementary Table S4. Interestingly, unlike the PACs in different genomic regions in Glycine max that showed little overlaps of NUE motifs, the same genomic region across different species showed more conserved presence of NUE among top 20 motifs. The 13 out of first 15 Glycine

max NUE were detected in at-least three other plant species as well regardless of the rank (Figure 5B). The canonical AAUAAA motif was the top ranked motif in all species. The UAUAUA was the second most significant

Table 1: The significant NUE motifs detected in the 3'-UTR region of *Glycine max* and their rank-wise comparison with other plants.

Hexamer	Number of PAC	% PAC	E-value	Position ± SD	Arabidopsis thaliana	Oryza sativa	Medicago truncatula	Zea mays
AAUAAA	4416	6.99	6.0×10^{-251}	-20 ± 5.9	1	1	1	1
UAUAUA	2824	4.47	8.2×10^{-96}	-22 ± 6.5	2	2	6	2
AAUGAA	2783	4.4	2.3×10^{-71}	-20 ± 6.0	6	7	2	12
AAGAAA	1641	2.6	8.7×10^{-75}	-20 ± 6.8	7	5	12	10
AAUAUA	2108	3.33	4.5×10^{-70}	-20 ± 6.2	9	4	5	4
AAUAAU	2065	3.27	4.8×10^{-51}	-21 ± 6.6	-	6	4	3
UUAAUU	2505	3.96	1.1×10^{-51}	-22 ± 6.6	-	-	-	6
UUAUAA	1645	2.6	3.1×10^{-64}	-21 ± 6.2	12	13	7	9
AAUAGA	855	1.35	1.1×10^{-53}	-20 ± 6.0	15	10	9	14
AUGAUA	1247	1.97	8.7×10^{-55}	-21 ± 6.4	-	-	10	-

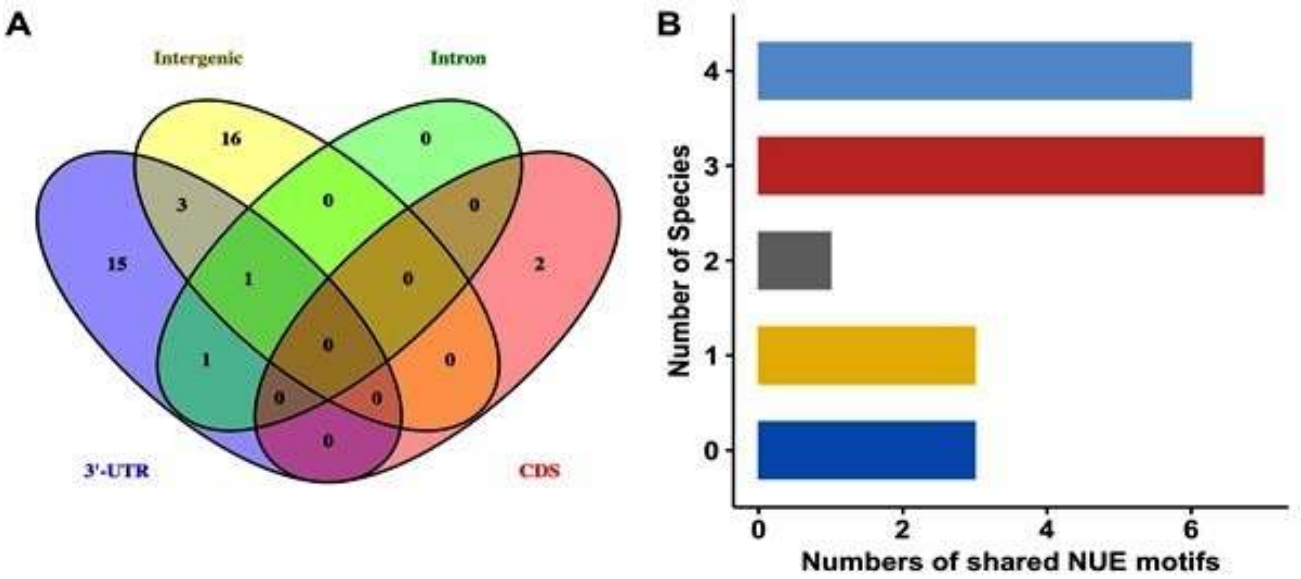

Figure 5: Analysis of *Glycine max*'s NUE motifs. (A) Shared and unique NUE motifs in different PAC types of *Glycine max* (B) Comparison of *Glycine max* NUE motifs with NUE motifs in the 3′ UTR PACs of related plant species.

and frequent motif in all species except *Medicago truncatula* where it was ranked at number 3. Except for the top two NUE motifs, all other *Glycine max* motifs consistently changed their ranks in other species (Table 1). The 12 variants of AAUAAA that were significant among at least three other plant species were either single base (n = 6), two bases (n = 2), or three or more than three base variants of AAUAAA (n = 4) (Figure 6A). In these variants, position 5 (64%) and 3 (42%) were the most degenerate positions where variants had different bases relative to the AAUAAA motif. The consensus for these variants was AAuANA (where upper-case base represents > 66% and lower > 58%). The remaining seven motifs of *Glycine max* were shared with either only two other species (ranked 14 in *Glycine max*), single species (3, 19, 20), or were not overlapped with any species (16, 17, 18) (Figure 5B).

A

# of variant bases	Variants of AAUAAA common in plants
2	UAUAUA
1	AAGAAA
1	AAUAUA
1	AAUAAU
5	UUAAUU
4	UUAUAA
1	AAUAGA
3	AUGAUA
1	AAUACA
1	AAUAAG
2	AAUGGA
5	UAAUGU
Consensus	AAuANA

B

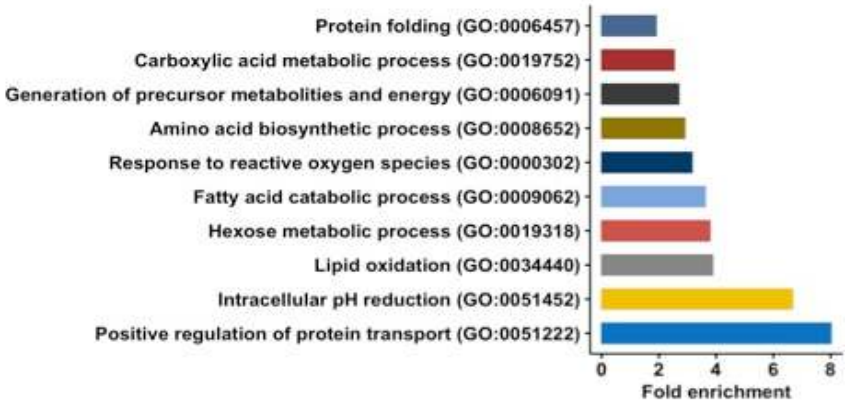

Figure 6: Different variants of AAUAAA motif and Gene ontology (GO) analysis. (A) Significant variants of AAUAAA motif in *Glycine max* and other related plant species. The analysis revealed that the position 5 in the AAUAAA motif were the most degenerate positions. (B) The GO enrichment analysis of genes with AAUAAA motif in the NUE region of *Glycine max*.

A total of 4,187 *Glycine max* genes had at least one PAC containing AAUAAA motif in the NUE region. A vast majority (94.7%) of the genes had only a single PAC with this motif. The GO enrichment analysis was performed to understand the biological processes associated with these genes that still use AAUAAA motif compared to the overall *Glycine max* genes using PANTHER database (Supplementary Table S5). The results highlighted that these genes were likely involved in a variety of cellular activities related to metabolism including breakdown and synthesis of small molecules, carbohydrates, amino acids, fatty acids, protein folding, and response to oxidative stress (Figure 6B).

APA in Glycine max genome: The distribution of PACs in the extended Ensembl genes models was analyzed to quantify the extent of APA in the Glycine max genome. Overall, 65,809 PACs were distributed among 28,838 distinct protein coding genes, which constitutes about 51.6% of the total protein coding genes in Glycine max (release 51). Figure 7A shows the number of PACs observed per gene in this study. A total of 34.5% (n=9,940) of the genes had a single PAC, while 65.5% (n = 18,898) showed APA leading to formation of different 3'-ends. This proportion is significantly higher than the current number of genes (19%) that undergo APA according to the Ensembl transcript models regardless of the 3'-UTR annotation status (Figure 7B).

The global analysis of relative expression levels of multiple PACs in APA genes indicated that some PACs were processed more efficiently than the others (Figure 7C). The processing efficiency of the most abundant PACs in the APA genes was significantly higher than the rest of the PACs (Wilcoxon one tailed greater), whereas all other PACs in the genes were utilized at low levels. These abundant PACs contributed 85% of the reads mapped to APA genes (Figure 7D). The second most abundant PACs contributed 11% reads to APA genes.

To further quantify the numbers of APA genes that express dominant PACs, the relative expression level (RE) of each PAC was computed. The APA genes showed huge diversity in terms of the dominant PACs. About 8,330 (44.1%) genes were classified as dominant genes that contained a dominant PAC (expressed at RE > 0.7) and within these 2,550 had RE values of more than 0.90. These dominant APA genes contributed 72.9% of all reads mapped on APA genes in *Glycine max*. The numbers of genes with strong PACs even further increased significantly when a less strict RE cutoff of 0.5 was considered. Using RE 0.5 or 50% read criterion, a vast majority of the APA genes (n = 14,466, 76.5%) expressed at least one dominant PAC.

GO enrichment analysis was performed on APA genes with a dominant PAC (RE > 0.7; n = 8,330) and those lacking a dominant PAC (RE ≤ 0.4; n = 1,372). The results of GO analysis are provided in Supplementary Table S6. The two groups shared only a limited number of common biological processes, primarily related to general cellular functions such as metabolism (e.g., organic substance metabolic process, nitrogen compound metabolic process), protein localization, and RNA processing (e.g., RNA metabolic process, RNA processing). However, a significantly larger number of biological processes were enriched in the group of genes with a dominant PAC. These processes included translation and ribosome biogenesis (e.g., ribosomal subunit assembly, translation initiation, elongation, and termination), regulation of gene expression (e.g., negative regulation of gene expression, negative regulation of translation, mRNA stability, epigenetic regulation), protein modification and transport (e.g., protein

glycosylation, folding, and localization to specific cellular compartments), cellular organization (e.g., microtubule nucleation, chromatin organization, and organelle assembly), cell cycle regulation, and response to stress.

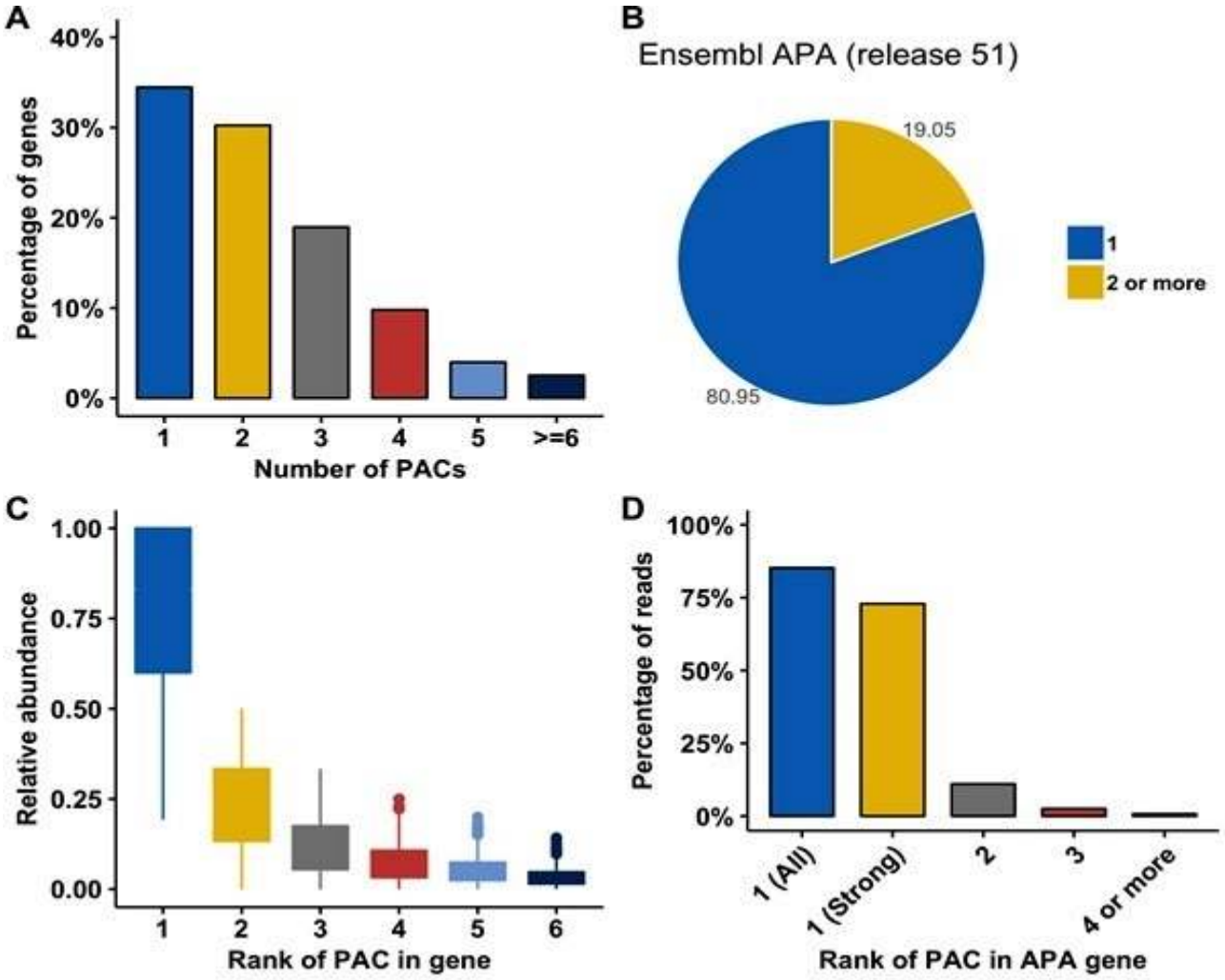

Figure 7: Analysis of APA distribution in *Glycine max*. (A) APA in *Glycine max* based on identified PACs in 28,838 genes. (B) APA in *Glycine max* according to the Ensembl annotation release 51. (C) Rank-wise RE of the PACs in the APA genes. Within each gene, PACs were arranged in the decreasing order of RE. (D) Rank-wise breakdown of the total polyadenylated reads mapped on to APA genes and the dominant PACs (RE > 0.70) of the APA genes contributed a vast majority these reads.

DISCUSSION

This study provides a detailed landscape of poly(A) sites in the *Glycine max* genome with identification of over 75,000 PACs across multiple genomic regions, covering about 51% of the *Glycine max* annotated protein coding genes. The PAC data generated here can be used to develop and optimize new plant biotechnology traits by identifying better expression elements, generating new expression elements from the existing ones, and in analyzing terminator strengths in plants (To *et al*., 2021; Gorjifard *et al*., 2024). Furthermore, it can also be used as evidence to improve existing gene annotations (Schulz *et al*., 2023). The large majority of identified PACs in the *Glycine max* genome mapped on to known 3'UTRs, with 88% of all polyadenylated reads mapping to these regions. In addition to 3'-UTRs, limited occurrences of PACs are also observed in 5'-UTRs, CDS, and introns. This is consistent with previous findings in other plants species, where the 3'-UTR remained the dominant region for PAC accumulation among all regions (Wu *et al*., 2014; Zhao *et al*., 2014; Fu *et al*., 2016; Chakrabarti *et al*., 2018; Jafar *et al*., 2019).

About 18,486 intergenic PACs were outside of the annotated genes and more than 7500 and 11,000 of these PACs are located within 75 nts and 300 nts of the known

genes, respectively. The presence of these intergenic PACs near gene termini suggests that many may represent extended 3'end rather than novel genes, thus challenging the current completeness of gene annotations in *Glycine max*. Using transcript assembled from RNA-Seq as additional evidence and immediate proximity of PACs to gene termini, 3'-ends of the Ensembl genes were extended which led to the classification of 47% intergenic PACs as extended 3'-UTRs of already known genes. These extensions are not specific to *Glycine max* only and are in line with studies that have revealed extended 3'-UTRs due to incomplete gene annotations in other model plants including *Arabidopsis thaliana*, *Oryza sativa*, *Zea mays*, and green algae (Zhao *et al*., 2014; Wu *et al*., 2015; Fu *et al*., 2016; Jafar *et al*., 2019). The 3'- extended PACs, averaging 176 nts, can offer new insights into gene regulation, as longer 3'-UTRs are often associated with additional regulatory elements like microRNA binding sites or RNA-binding protein motifs (Brümmer and Hausser, 2014; Lo Giudice *et al*., 2023). This aligns with findings in *Arabidopsis thaliana*, where extended 3'-UTRs have been linked to post-transcriptional regulation under stress conditions, especially drought and salinity stress (Hardy and Balcerowicz, 2024). A significant portion of *Glycine max* PACs are still considered as intergenic PACs (53%) in this study and further studies are needed to improve gene features and in understanding the role of PACs in post-transcriptional regulation in *Glycine max*.

This study reveals that a significant numbers of *Glycine max* genes exhibit APA (65.5%), a much higher percentage than previously reported in Ensembl models (19%). This suggests that APA is a significant regulatory mechanism in *Glycine max* that contributes to transcriptome diversity. This is consistent with reports in other plant species, where APA affects approximately 70% of the genes (Zhao *et al*., 2014; Jafar *et al*., 2019; Gorjifard *et al*., 2024). Large numbers of genes (44%) express at least one dominant PAC, with the most abundant PACs accounting for 85% of the mapped reads in APA genes, indicating preferential usage. Genes with dominant PACs (RE > 0.7) are enriched in a wider range of biological processes, particularly those related to translation, gene expression regulation, cellular organization, cell cycle, stress response, and mRNA stability. These processes for the dominant genes are consistent with previous findings in *Zea mays* (Jafar *et al*., 2019). The enrichment of these processes highlights the potential role of APA in adapting gene expression under dynamic conditions (such as during growth, division, or stress). In contrast, genes without dominant PACs (RE ≤ 0.4) shared biological processes with dominant PACs genes and these processes are associated with fewer specialized functions including metabolic and RNA processing functions.

Our analysis of nucleotide composition around PACs in *Glycine max* reveals conserved U-rich and A-rich patterns, aligning with previously reported polyadenylation patterns in dicot and monocot plants (Wu *et al*., 2011; Wu *et al*., 2014; Wu *et al*., 2015; Fu *et al*., 2019; Wu *et al*., 2019; Jafar *et al*., 2019). Specifically, a distinct U-A-U-A-U transition was observed in the PAC regions of *Glycine max*, consistent with earlier studies in related plants (Li and Du, 2017). When comparing the nucleotide composition of *Glycine max* PACs in position-by-position manner with other species such as *Arabidopsis thaliana*, *Zea mays*, *Oryza sativa*, and *Medicago truncatula*, no significant differences were found in the NUE and CE regions. However, the FUE region (-100 to -30) in *Glycine max* shows higher U-richness, resembling that of other dicotyledonous plants. This observation is consistent with the previous study in which dicot genomes generally have shown lower GC content than monocots, contributing to species-specific differences in 3' end processing (Gorjifard *et al*., 2024). The higher U-content in the FUE region of dicots, including *Glycine max*, supports the idea that nucleotide composition affects polyadenylation efficiency in a position-specific manner (Gorjifard *et al*., 2024) observed that U-rich sequences upstream of the cleavage site increased terminator strength in dicot systems like *Nicotiana tabacum* but not in monocot systems such as *Zea mays*. This differential association between U-content and terminator strength highlights the importance of species-specific regulatory elements in defining polyadenylation signal strength.

The canonical AAUAAA motif remains the most frequent and significant NUE across various PAC regions in *Glycine max*, including 3'-UTR, intronic, and intergenic regions. This finding reinforces the established role of AAUAAA as a key polyadenylation signal, as observed across multiple species like *Arabidopsis thaliana*, *Oryza sativa*, *Zea mays*, and *Medicago truncatula* (Wu *et al*., 2011; Wu *et al*., 2014; Fu *et al*., 2019; Jafar *et al*., 2019). The functional analysis of the genes containing AAUAAA indicated enrichment of biological processes related to metabolism of small molecules, carbohydrates, amino acids, fatty acids, protein folding, and response to oxidative stress, which is consistent with previous observations in *Zea mays* (Jafar *et al*., 2019). However, in *Glycine max*, there is minimal overlap in the top 20 significant motifs across different genomic regions, with only four motifs shared between the 3'-UTR and intergenic regions. This suggests that while AAUAAA is crucial, other motifs may vary in their usage based on genomic context, likely driven by sequence-specific constraints.

Interestingly, the presence of variant motifs such as AAUAAG and AAUAAU in the upstream NUE region of *Glycine max* PACs, showing efficiencies comparable to the canonical AAUAAA in *Zea mays* and *Nicotiana tabacum*, indicates a potential role for these variants in modulating polyadenylation (Gorjifard *et al*., 2024).

While the AAUAAA motif is most efficient, these variants may offer alternative regulatory mechanisms in different plant species. The degenerate nature of the fifth position of AAUAAA in *Glycine max*, consistent with trends observed in human polyadenylation, further underscores the flexibility of polyadenylation signals across eukaryotes (Beaudoing *et al*., 2000).

The comparison of 3'-UTR regions across different plants shows that despite the conserved nucleotide composition and overlap in top NUE motifs, there are significant variations in the order of preference among motifs. These differences may arise from species-specific genomic compositions, suggesting that while the polyadenylation machinery is broadly conserved, its fine-tuning may be influenced by local genomic features unique to each plant.

CONCLUSION

This study presents a comprehensive map of poly(A) sites in the *Glycine max* genome, highlighting the widespread occurrence and regulatory significance of APA. About 65% genes were found to have more than one PAC. The majority of PACs map to 3'-UTR, but also a considerable number PACs were also detected in other genomic regions, including 5'UTRs, CDS, and introns. Notably, the discovery of a substantial number of intergenic PACs, many of which represent extensions of existing 3'UTRs, challenging current gene annotations and suggesting that a significant portion of the *Glycine max* transcriptome remains to be fully characterized. While this study has significantly advanced our knowledge of polyadenylation in *Glycine max*, further research is needed to fully elucidate the functional roles of the remaining intergenic PACs and to investigate the specific regulatory elements associated with APA in this important crop species. The data generated here will serve as a valuable resource for future studies aimed at improving gene annotation and understanding gene regulation. The present study identified at least one PAC for 51% of the Soybean genes indicating that standard RNA-Seq data can be used to detect these events in absence of 3'-specific sequencing datasets. However, still no polyadenylation event was reported for 49% genes, highlighting the need to use polyadenylation specific sequencing reads in future research.

Supplementary Materials: The following supporting information can be downloaded at: Supplementary_materials.zip, Table S1: summary of the SRA datasets and reads; Table S2: Identified PACs in *Glycine max*; Table S3: Comparison of average base composition between plants in different region around PACs; Table S4: NUE motifs in different PAC types of *Glycine max* and 3'-UTRs of different plants; Table S5: GO biological processes associated with genes containing AAUAAA motif; Table S6: GO biological processes associated with APA genes with dominant and without dominant PACs.

Author Contributions: Conceptualization, MS and MNA; Data curation, WS and MNA; Data analysis, WS, MS, NA and MNA; Funding acquisition, MS and MNA; Methodology, MS and MNA; Resources, MNA; Supervision, MS and MNA; Validation, WS and NA; Visualization, WS, NA and MNA; Writing – original draft, WS and NA; Writing – review & editing, WS, MS, NA and MNA. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by HEC-NRPU (grant number: 5385), doctorate fellowship to W.S by HEC (grant number: 2BS5-56.

Data Availability Statement: The original data analyzed in this study are openly available in SRA database at https://www.ncbi.nlm.nih.gov/sra. The accession numbers are available in the article. The original contributions presented in this study are included in the article and supplementary material.

Acknowledgments: The authors would like to acknowledge the Dr. Awais Rasheed at the Quaid-i-Azam University Pakistan for helping with the analysis and discussion.

Conflicts of Interest: The authors declare no conflicts of interest.

REFERENCES

Beaudoing, E., S. Freier, J.R. Wyatt, J.M. Claverie and D. Gautheret (2000). Patterns of variant polyadenylation signal usage in human genes. Genome Res. 10(7): 1001-1010. DOI: 10.1101/gr.10.7.1001

Bell, S.A., C. Shen, A. Brown and A.G. Hunt (2016). Experimental genome-wide determination of RNA polyadenylation in *Chlamydomonas reinhardtii*. PLoS One. 11(1): e0146107. https://doi.org/10.1371/journal.pone.0146107

Bernardes, W.S. and M. Menossi (2020). Plant 3'regulatory regions from mRNA-encoding genes and their uses to modulate expression. Front. in Plant Sci. 11: 1252. https://doi.org/10.3389/fpls.2020.01252

Bi, X., S. Zhu, F. Liu, and X. Wu (2024). Dynamics of alternative polyadenylation in single root cells of Arabidopsis thaliana. Front. in Plant

Sci. *15*:1437118. https://doi.org/10.3389/fpls.2024.1437118

Biłas, R., K. Szafran, K. Hnatuszko-Konka and A.K. Kononowicz (2016). Cis-regulatory elements used to control gene expression in plants. Plant Cell, Tissue and Organ Culture (PCTOC). 127: 269-287. https://doi.org/10.1007/s11240-016-1057-7

Bolger, A.M., M. Lohse and B. Usadel (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinfo. 30(15): 2114-2120. https://doi.org/10.1093/bioinformatics/btu170

Brümmer, A. and J. Hausser (2014). MicroRNA binding sites in the coding region of mRNAs: extending the repertoire of post-transcriptional gene regulation. Bioessays. 36(6): 617-626. https://doi.org/10.1002/bies.201300104.

Chakrabarti, M., R.D. Dinkins and A.G. Hunt (2018). Genome-wide atlas of alternative polyadenylation in the forage legume red clover. Scientific rep. 8(1): 1-14. https://doi.org/10.1038/s41598-018-29699-7

De Felippes, F.F. and P.M. Waterhouse (2023). Plant terminators: the unsung heroes of gene expression. J. of Exp. Botany. 74(7): 2239-2250. https://doi.org/10.1093/jxb/erac467

Du, H., C. Fang, Y. Li, F. Kong and B. Liu (2023). Understandings and future challenges in soybean functional genomics and molecular breeding. J. of Integrative Plant Bio. 65(2): 468-495. https://doi.org/10.1111/jipb.13433

Fu, H., D. Yang, W. Su, L. Ma, Y. Shen, G. Ji and Q.Q. Li (2016). Genome-wide dynamics of alternative polyadenylation in rice. Genome Res, 26(12): 1753-1760. DOI: 10.1101/gr.210757.116

Gorjifard, S., T. Jores, J. Tonnies, N.A. Mueth, K. Bubb, T. Wrightsman and C. Queitsch (2024). Arabidopsis and maize terminator strength is determined by GC content, polyadenylation motifs and cleavage probability. Nat. Com. 15(1): 5868. https://doi.org/10.1038/s41467-024-50174-7.

Hardy, E.C. and M. Balcerowicz (2024). Untranslated yet indispensable—UTRs act as key regulators in the environmental control of gene expression. J. of Exp. Botany: erae073. https://doi.org/10.1093/jxb/erae073.
https://doi.org/10.1186/1471-2164-15-615

Jafar, Z., S. Tariq, I. Sadiq, T. Nawaz and M.N. Akhtar (2019). Genome-Wide Profiling of Polyadenylation Events in Maize Using High-Throughput Transcriptomic Sequences. G3: Genes, Genomes, Genetics. 9(8): 2749-2760. https://doi.org/10.1534/g3.119.400196

Langmead, B. and S.L. Salzberg (2012). Fast gapped-read alignment with Bowtie 2. Nat. met. 9(4): 357-359. https://doi.org/10.1038/nmeth.1923

Leinonen R., H. Sugawara, M. Shumway and International Nucleotide Sequence Database Collaboration. (2010). The sequence read archive. Nucleic acids Res. 39(suppl_1): D19-D21. https://doi.org/10.1093/nar/gkq1019

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, ... and 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. Bioinfo. 25(16):2078-2079. https://doi.org/10.1093/bioinformatics/btp352

Li, X.Q. and D. Du (2017). RNA Polyadenylation Site Regions: Highly Similar in Base Composition Pattern but Diverse in Sequence—A Combination Ensuring Similar Function but Avoiding Repetitive-Regions-Related Genomic Instability. Somatic Genome Variation in Animals, Plants, and Microorganisms: 267-290. https://doi.org/10.1002/9781118647110.ch11

Liu, J., X. Lu, S. Zhang, L. Yuan and Y. Sun (2022). Molecular insights into mRNA polyadenylation and deadenylation. Int. J. of Mol. Sci. 23(19): 10985. https://doi.org/10.3390/ijms231910985

Liu, S., S. Luo, D. Yang, J. Huang, X. Jiang, S. Yu, J. Fu, D. Zhou, X. Chen, H. He and H. Fu (2024). Alternative polyadenylation profiles of susceptible and resistant rice (Oryza sativa L.) in response to bacterial leaf blight using RNA-seq. BMC Plant Bio. 24(1):145. https://doi.org/10.1186/s12870-024-04839-6

Lo Giudice, C., F. Zambelli, M. Chiara, G. Pavesi, M.A. Tangaro, E. Picardi and G. Pesole (2023). UTRdb 2.0: a comprehensive, expert curated catalog of eukaryotic mRNAs untranslated regions. Nucleic Acids Res. *51*(D1): D337-D344. https://doi.org/10.1093/nar/gkac1016

Neve, J. and A. Furger (2014). Alternative polyadenylation: less than meets the eye? Biochemical Society Transactions, 42(4): 1190-1195. https://doi.org/10.1042/BST20140054

Peng, Z., S. Yu, J. Meng, K. H. Jia, J. Zhang, X. Li, W. Gao and S. Wan (2023). Alternative polyadenylation regulates acetyl-CoA carboxylase function in peanut. BMC Gen. 24(1):637. https://doi.org/10.1186/s12864-023-09696-5

Pereira-Castro, I. and A. Moreira (2021). On the function and relevance of alternative 3′-UTRs in gene expression regulation. Wiley Interdisciplinary Reviews: RNA. 12(5): e1653. https://doi.org/10.1002/wrna.1653

Rodríguez-Molina, J.B. and M. Turtola (2023) Birth of a poly (A) tail: mechanisms and control of mRNA polyadenylation. *FEBS Open Bio*. 13(7): 1140-1153. https://doi.org/10.1002/2211-5463.13528

Schulz, A.J., J. Zhai, T. AuBuchon-Elder, M. El-Walid, T.H. Ferebee, E.H. Gilmore and S.K. Hsu (2023). Fishing for a reelGene: evaluating gene models with evolution and machine learning. bioRxiv: 2023-09. https://doi.org/10.1101/2023.09.19.558246

Song, P., J. Yang, C. Wang, Q. Lu, L. Shi, S. Tayier and G. Jia (2021). Arabidopsis N6-methyladenosine reader CPSF30-L recognizes FUE signals to control polyadenylation site choice in liquid-like nuclear bodies. Mol. Plant. 14(4): 571-587. https://doi.org/10.1016/j.molp.2021.01.014

Sun, Y., K. Hamilton and L. Tong (2020). Recent molecular insights into canonical pre-mRNA 3'-end processing. Transcription. 11(2): 83-96. https://doi.org/10.1080/21541264.2020.1777047

Téllez-Robledo, B., C. Manzano, A. Saez, S. Navarro-Neila, J. Silva-Navas, L. de Lorenzo, M.P. González-García, R. Toribio, A.G. Hunt, R. Baigorri and I. Casimiro (2019). The polyadenylation factor FIP1 is important for plant development and root responses to abiotic stresses. The Plant Journal.;99(6):1203-19. https://doi.org/10.1111/tpj.14416

Tian, B., J. Hu, H. Zhang and C.S. Lutz (2005). A large-scale analysis of mRNA polyadenylation of human and mouse genes. Nucleic Acids Res. 33(1): 201-212. https://doi.org/10.1093/nar/gki158

To, J.P., I.W. Davis, M.S. Marengo, A. Shariff, C. Baublite, K. Decker and T.D. Elich (2021). Expression elements derived from plant sequences provide effective gene expression regulation and new opportunities for plant biotechnology traits. Front. in Plant Sci. 12: 712179. https://doi.org/10.3389/fpls.2021.712179

Torres-Ulloa, L., E. Calvo-Roitberg E, A.A Pai (2024). Genome-wide kinetic profiling of pre-mRNA 3' end cleavage. RNA. 30(3): 256-270. http://www.rnajournal.org/cgi/doi/10.1261/rna.079783.123

Wang, H., R. Li, X. Zhou, L. Xue, X. Xu and B. Liu (2016). Genome-wide analysis and functional characterization of the polyadenylation site in pigs using RNAseq data. Scientific Reports. 6(1): 36388. https://doi.org/10.1038/srep36388

Wang, T., W. Ye, J. Zhang, H. Li, W. Zeng, S. Zhu, G. Ji, X. Wu and L. Ma (2023). Alternative 3′-untranslated regions regulate high-salt tolerance of Spartina alterniflora. Plant Physio. 191(4):2570-87. https://doi.org/10.1093/plphys/kiad030

Wu, J., L. Ma and Y. Cao (2023). Alternative polyadenylation is a novel strategy for the regulation of Gene expression in response to stresses in plants. Int. J. of Mol. Sci. 24(5): 4727. https://doi.org/10.3390/ijms24054727

Wu, X., A.G. Hunt and Q.Q. Li (2019). Genome-wide determination of poly (A) sites in *Medicago truncatula*: evolutionary conservation of alternative poly (A) site choice. The Model Legume Medicago truncatula: (pp. 911-920). John Wiley Publishing. https://doi.org/10.1002/9781119409144.ch116

Wu, X., B. Gaffney, A.G. Hunt and Q.Q. Li (2014). Genome-wide determination of poly (A) sites in *Medicago truncatula*: evolutionary conservation of alternative poly (A) site choice. BMC Gen. 15(1): 1-11.

Wu, X., M. Liu, B. Downie, C. Liang, G. Ji, Q.Q. Li and A.G. Hunt (2011). Genome-wide landscape of polyadenylation in Arabidopsis provides evidence for extensive alternative polyadenylation. Proc. of the National Academy of Sci. 108(30): 12533-12538. https://doi.org/10.1073/pnas.1019732108

Wu, X., Y. Zeng, J. Guan, G. Ji, R. Huang and Q.Q. Li (2015). Genome-wide characterization of intergenic polyadenylation sites redefines gene spaces in *Arabidopsis thaliana*. BMC Gen. 16(1): 1-14. https://doi.org/10.1186/s12864-015-1691-1

Yan, C., Y. Wang, T. Lyu, Z. Hu, N. Ye, W. Liu and H. Yin (2021). Alternative polyadenylation in response to temperature stress contributes to gene regulation in *Populus trichocarpa*. BMC Gen. 22: 1-10. https://doi.org/10.1186/s12864-020-07353-9

Ye, W., Q. Lian, C. Ye and X. Wu (2023). A survey on methods for predicting polyadenylation sites from DNA sequences, bulk RNA-seq, and single-cell RNA-seq. Gen. Prot. & Bioinfo. 21(1):67-83. https://doi.org/10.1016/j.gpb.2022.09.005

Yu, Z., L. Hong and Q.Q. Li (2022). Signatures of mRNA alternative polyadenylation in Arabidopsis leaf development. Front. in Genet. *13*:863253. https://doi.org/10.3389/fgene.2022.863253

Zhao, Z., X. Wu, G. Ji, C. Liang and Q.Q. Li (2019). Genome-wide comparative analyses of polyadenylation signals in eukaryotes suggest a possible origin of the AAUAAA signal. Int. J. of Mol. Sci. 20(4): 958. https://doi.org/10.3390/ijms20040958

Zhao, Z., X. Wu, P. K. R. Kumar, M. Dong, G. Ji, Q.Q. Li and C. Liang (2014). Bioinformatics analysis of alternative polyadenylation in green alga *Chlamydomonas reinhardtii* using transcriptome sequences from three different sequencing platforms. G3: Genes, Genomes, Genetics. 4(5): 871-883. https://doi.org/10.1534/g3.114.010249

Zhu S., W. Ye, L. Ye, H. Fu, C. Ye, X. Xiao, Y. Ji, W. Lin, G. Ji and X. Wu (2020). PlantAPAdb: a comprehensive database for alternative polyadenylation sites in plants. Plant phys. 182(1):228-42. https://doi.org/10.1104/pp.19.00943.