Genome-wide dynamics of alternative polyadenylation in rice

Haihui Fu, Dewei Yang, Wenyue Su, et al.

Genome Res. published online October 12, 2016
Access the most recent version at doi:10.1101/gr.210757.116

Genome-wide Dynamics of Alternative Polyadenylation in Rice

Running title: Alternative polyadenylation in rice

Haihui Fu[1,5], Dewei Yang[2,5], Wenyue Su[1], Liuyin Ma[1], Yingjia Shen[1], GuoliJi[3], Xingfu Ye[2,6], Xiaohui Wu[3,6], Qingshun Q. Li[1,2,4,6],

[1]Key Laboratory of the Ministry of Education for Coastal and Wetland Ecosystems, College of the Environment and Ecology, Xiamen University, Xiamen, Fujian, China; [2]Rice Research Institute, Fujian Academy of Agricultural Sciences, Fuzhou, Fujian, China; [3]Department of Automation, Xiamen University, Xiamen, Fujian, China; [4]Graduate College of Biomedical Sciences, Western University of Health Sciences, Pomona, CA 91766, USA.

[5]These authors contributed equally to this work.

[6]Corresponding authors

Corresponding author: Dr. Q. Quinn Li, Graduate College, Western University of Health Sciences, 309 E. 2nd Street, Pomona, CA 91769, USA; Phone: 909-469-8523;Fax: 909-469-8750; email: qqli@westernu.edu

Keywords: alternative polyadenylation; mRNA processing; post-transcriptional regulation; transcriptome; rice.

Abstract

Alternative polyadenylation (APA), in which a transcript uses one of the poly(A) sites to define its 3'-end, is a common regulatory mechanism in eukaryotic gene expression. However, the potential of APA in determining crop agronomic traits remains elusive. This study systematically tallied poly(A) sites of 14 different rice tissues and developmental stages using the Poly(A) Tag Sequencing (PAT-Seq) approach. The results indicate significant involvement of APA in developmental and quantitative trait loci (QTL) gene expression. About 48% of all expressed genes use APA to generate transcriptomic and proteomic diversity. Some genes switch APA sites allowing differentially expressed genes to use alternate 3'UTR. Interestingly, APA in mature pollen are distinct where differential expression level of a set of poly(A) factors and different distribution of APA sites were found, indicating a unique mRNA 3'-end formation regulation during gametophyte development. Equally interesting, statistical analyses showed that QTL tends to use APA for regulation of gene expression of many agronomic traits, suggesting a potential important role of APA in rice production. These results provide thus far the most comprehensive and high-resolution resource for advanced analysis of APA in crops, and shed light on how APA are associated with trait formation in eukaryotes.

Messenger RNA (mRNA) polyadenylation is a key step during gene expression in eukaryotes, which involves 3'-end cleavage of pre-mRNA and an addition of a poly(A) tail. Only mRNA with a poly(A) tail can be exported out of the nucleus to play out its roles in the cytoplasm, including mRNA stability, localization, and formation of a translational initiation complex (Millevoi and Vagner 2010).

The formation of a poly(A) tail at the correct location involves a complex and finely tuned biochemical process consisting of interactions among many polyadenylation factors and *cis*-elements, e.g., AAUAAA (Beaudoing et al. 2000). Before forming a poly(A) tail, a large number of poly(A) factors, including the cleavage and polyadenylation specificity factor (CPSF), poly(A) polymerase, cleavage stimulation factor (CstF) and Cleavage factors CF I and CF II, assemble near the 3'-end of pre-mRNA to achieve cleavage and poly(A) tail addition (Colgan and Manley 1997; Shi et al. 2009). In plants, a similar set of poly(A) factors were identified, albeit with some differences in the number of encoding genes and domain structures(Hunt et al. 2008; Hunt et al. 2012). Genetic and biochemical analysis of these factors also revealed similar, but distinct, and, sometimes, unique, functions to their animal counterparts (Zhao et al. 2009; Zhao et al. 2011; Thomas et al. 2012; Xing et al. 2013; Liu et al. 2014).

Previous studies showed that many eukaryotic genes harbor more than one poly(A) site (Ji and Tian 2009; Jan et al. 2011; Sherstnev et al. 2012), suggesting that the locations of poly(A) sites can be chosen in different regions of transcripts. This is termed alternative polyadenylation, or APA. For example, studies showed that ~50% and 82% of genes in rice had more than one poly(A) site by the Expressed Sequence Tags (EST) database, restriction enzyme digestion-mediated massively parallel signature sequencing (MPSS),or Illumina sequencing approaches, respectively (Shen et al. 2008; Shen et al. 2011). Using high-throughput sequencing, it was discovered that more than

70% of genes in *Arabidopsis thaliana* had two or more poly(A) sites, in which ~17% was located within 5'UTRs, coding sequences (CDS), and introns (Wu et al. 2011). The number of APA genes in Arabidopsis was significantly more than that found in a previous study using a different genome-scale approach (Sherstnev et al. 2012).

The transcripts resulting from APA can be used to regulate gene expression, as demonstrated by previous studies showing that errors associated with polyadenylation could result in human diseases (Danckwardt et al. 2008; Mayr and Bartel 2009) or developmental defects in plants (Thomas et al. 2012). For example, *FCA*, a gene with both proximal and distal poly(A) sites, is associated with flowering time control in plants. In general, the distal poly(A) site is used, and it produces a long transcript. However, when the proximal poly(A) site is used, it results in a short transcript without function, thus affecting flowering time(Boss et al. 2004). *CPSF30*, encoding a poly(A) factor in Arabidopsis, acts on APA selection in a broad spectrum of genes since itsmutant reflects changes in a large number of gene poly(A) sites and the use of *cis*-elements in 3'UTR (Thomas et al. 2012; Liu et al. 2014). Otherwise, APA is a common post-transcriptional regulatory mechanism in other eukaryotes, leading to transcripts with variable 3'UTR lengths (Mangone et al. 2010; Ozsolak et al. 2010) and impacting physiological properties of proteins (Takagaki et al. 1996) and mRNA stability (Mayr and Bartel 2009).

Rice is one of the most important crops and also a model monocot. Although some APA analyses in rice were performed, the roles of APA in development and tissue specification remain to be elucidated. Previous work evaluated rice poly(A) to the nearest restriction enzyme site next to a poly(A) site, resulting in inaccuracy at the nucleotide level (Shen et al. 2011). Hence, we used a poly(A) tag sequencing approach (PAT-seq) to investigate the genome-wide landscape of APA in 14 different tissues and developmental stages of rice.

Results

Profiling of rice poly(A) sites

First, raw reads were processed and low-quality reads discarded, followed by mapping the remaining reads to the *Oryza sativa japonica* reference genome by Bowtie 2 (Langmead and Salzberg 2012) (Supplemental_Table_S1.pdf). After removing potential internal priming (Wu et al. 2011), a total of 136,955,352 individual PATs and 68,220 poly(A) site clusters (PACs) were obtained (Supplemental_Table_S2A.pdf), which were dispersed in 28,032 genes (listed in Supplemental_Table_S2B.csv). Among them, 13,419 genes (47.9%) had more than one PAC, indicating that the phenomenon of APA occurs extensively in rice transcriptome. As shown in Figure 1, the distributions of these PACs in different genic and intergenic regions of the genome were mostly in a similar order (3'UTR>intergenic>intron>promoter>CDS>5'UTR) in different tissues except in pollen where more PACs apparently are located in the intergenic regions and less in the 3'UTR(more details in Supplemental_Materials.pdf).

The verification of the poly(A) sites shows that 40,480 (70%) poly(A) sites of the EST data overlapped identical sites in the PAT-seq data (Shen et al.2008), indicating a significant match of these two datasets (more details see Supplemental_Materials.pdf; Supplemental_Fig_S1.pdf). Pearson correlation across tissues was 0.59-0.83 between $\log_2$(PAT) and $\log_2$(FPKM), suggesting the reasonable representation of relative gene expression level by PAT-seq (Supplemental_Fig_S2.pdf). Pearson correlation >0.9 showed that PACs had good reproducibility in 3 replicates (Supplemental_Fig_S3.pdf).

Poly(A) sites in mature pollen are distinct

To gain more insight into PACs lying within different genomic regions across 14 tissues, single nucleotide profiles ranging from100 nt upstream to 100 nt downstream of

cleavage sites were studied as described (Loke et al. 2005；Wu et al.2011). The results of other tissues were very similar, except the introns of pollen (Fig. 2A). Content of A at the cleavage site (position “0”) of pollen was noticeably higher than that of leaf, indicating potential preferences in *cis*-element. At the same time, however, A content was distinctly lower in the NUE region of pollen introns compared to that of other samples (Fig. 2A). Such change hints at different poly(A) signals used in these introns. Therefore, we used MEME (Bailey et al. 2009) analyses to find different signals in pollen introns. As shown in Figure 2B, A-rich motif was ranked at the top in the flanking region -35 bp to -10 bp of the cleavage site from leaf, while T-rich motif was particularly enriched in the same region from the introns of pollen. Since polyadenylation requires the interaction of polyadenylation factors and *cis*-elements, the selection of certain poly(A) signals may reflect difference of expression levels of polyadenylation factors between pollen and leaf.

To examine this possibility, we used the PAT-seq transcriptome data collected herein to measure the expression level of many poly(A) factors, and the results showed that *FY* was significantly up-regulated in pollen compared to 20-day-old leaf, while *CPSF30* showed an inverse trend (Fig. 2C). Interestingly, the orthologs of these two factors in human were recently revealed to recognize the poly(A) signals AAUAAA (NUE in plants)(Chan et al. 2014), and genetic evidence from Arabidopsis also demonstrated that *CPSF30* is responsible for NUE recognition (Thomas et al. 2012). Taken together, these results indicate that the change of expression of *FY* and *CPSF30* could be responsible for the difference of poly(A) site selection in pollen.

Significant portion of poly(A) clusters show tissue specificities

To further explore the possibility of tissue-specific PACs among all samples, we used two measures: a PAC only expressed in one tissue, but not in any other tissues tested herein, or a PAC in one tissue having significantly higher expression level (32-fold) than another tissue. Such tissue-specific PACs are rare across all 14 samples, pollen again showed the highest tissue specificity with 447 PACs (Fig. 3A). Of these, 47% were from the 3'UTR, while about 44% were from intronic regions, as expected, based on the unique poly(A) signals described above. A distant second, 197 tissue-specific PACs were also found in 60-day-old root. Of these, 90% were mostly from the 3'UTR, and only 9% were from introns, compared to pollen. Similarly, a previous study showed that tissue-specific PACs across different tissues in human were also very rare (Lianoglou et al. 2013). Pollen-specific isoforms were enriched in the development-related pathways, including energy, defense, and core protein biosynthesis (Supplemental_Materials.pdf; Supplemental_Fig_S4.pdf). As another way to interrogate specificity, the degree of PAC distribution among the different samples was investigated (Supplemental_Materials.pdf; Supplemental_Fig_S5.pdf).

Expression patterns of PACs across 14 tissues

Unsupervised hierarchal clustering analysis and principal component analysis revealed that different tissues cluster in distinct branches (Fig. 3B,C). Furthermore, different developmental stages of the same tissues have similar expression patterns of transcript isoforms. However, cluster analyses showed that mature pollen differed from other tissues by exhibiting a distinct transcriptome pattern. PAT ratio of differentially expressed (DE) PACs among 14 tissues revealed that some PACs that expressed across tissues were different from each other, especially those highly expressed in pollen (Fig. 3D). Therefore, when pollen was compared to other tissues, a large number of DE PACs

were identified ($|\log_2$FoldChange$|$ >2; p-adjust<0.01; Fig. 3E). These results showed the diversity of PAC expression and further showed that pollen is very special compared to other tissues.

3'UTR length variation in rice

As shown in Supplemental_Table_S3.pdf, our results indicate that seed-related tissues had the shortest median 3'UTR length among all 14 tissues. On the other hand, 20-day-old leaves had the longest 3'UTR, being 17nt longer than the shortest 3'UTR in APA genes. In particular, with the development of rice, the length of 3'UTR became incrementally longer (Supplemental_Fig_S6.pdf). Even though the difference in median 3'UTR length was less than 20nt, statistical significance was reached (Supplemental_Table_S4.pdf).

The length of those genes with 3'UTR with single PAC, whether differentially expressed or not, was shorter than the length of 3'UTR of APA genes (see Fig. 4A). While among APA genes, in contrast, the 3'UTR of differentially expressed genes was longer than that of APA genes not differentially expressed. This result indicated that differentially expressed genes tended to have longer 3'UTR, notwithstanding the use of proximal or distal PAC. Moreover, highly expressed APA genes tend to use shorter 3'UTR compared with those not highly expressed (Fig. 4B), andhighly expressed APA genes used significantly more proximal PACs (Fig. 4C). Previous studies have also shown the use of shorter 3'UTR in highly expressed cancer cell genes (Ozsolak et al. 2010). Interestingly, the use of proximal PACs in highly expressed APA genes was reversed in mature pollen where highly expressed genes used significantly fewer proximal PACs than genes with low expression. However, it was shown that shorter

3'UTR length was used in highly expressed genes in testis (Ulitsky et al. 2012). The reason for such discrepancy in pollen requires further investigation.

APA site switching alters gene expression profile

APA site switching is a phenomenon whereby two or more APA sites of a gene change their usage frequency in different tissues or developmental stages (e.g., Fig. 5A). The heat-map showed that significant APA site switching occurs at different developmental stages/tissues (Fig. 5C,D). Most APA site switching genes (371~529) were observed between pollen and other tissues in pair-wise comparisons (Fig. 5C). Those with APA switching in non-3'UTR, other than 3'UTR, are still observed in significant amounts, but, again, dominant in pollen (Fig. 5D). In addition, upon investigating APA site switching dynamics in pollen compared to other tissues (Fig. 5B), results showed that pollen possesses longer 3'UTR compared with embryo, endosperm and dry seed, but used shorter 3'UTR when compared with the rest of the tissues tested, except pistil where they were almost equal. This is an example that the dynamic of 3'UTR length control could be broader than previously anticipated.

To confirm APA switching gene in different stages, we performed qRT-PCR for four 3'UTR lengthening and one 3'UTR shortening genes using primers that amplify a common region (PA1+PA2) and the longer 3'UTR (PA2). The qRT-PCR results match what were tallied from the PAT-seq large-scale data (Supplemental_Fig_S7.pdf).

Intriguingly, with the development of rice, more genes used longer 3'UTR than shorter 3'UTR between two early stages (Supplemental_Fig_S8.pdf), thus contributing to overall 3'UTR lengthening (Supplemental_Table_S3.pdf). Accordingly, our analysis found that some proteins would obviously become shorter in varying degrees as a result of APA leading to potential functional changes (Supplemental_Fig_S8.pdf). As predicted

based on what was showed, there was a sharp change from anther to pollen, but the number of genes used shorter 3'UTR was more than that used longer 3'UTR.This was similar as described above that many PACs were located within introns. Gene Ontology analysis of switching genes in 3'UTR showed that, switching genes between anther and mature pollen mostly function in biological processes, including single organism signaling, cell communication, etc. (Supplemental_Fig_S9.pdf). Switching genes between antherand mature pollenthat are located in the non-3'UTR areas, however, affect a different array of pathways like carbohydrate metabolic process and cellular process (Supplemental_Fig_S9.pdf). Above analysis illustrated that development of rice might be regulated by different APA switching patterns.

APA genes are highly enriched at Quantitative Trait Loci

To explore if APA plays any role in regulating the expression of QTL, we first examined whether APA genes or APA site switching genes are enriched in QTL regions compared with randomly selected genomic regions (see Methods). The coincidence of QTL regions with APA genes was significantly higher than expected ($p = 2.36E-12$). Similar analyses were done on APA site switching genes, which were located in 150 of the 1078 QTL regions, but in this case, only 77 randomly selected regions possessed APA switching genes. Therefore, the coincidence of QTL regions with APA site switching genes was also significantly higher than expected ($p = 3.58E-07$). We further subdivided QTL regions into different groups according to their trait categories to study which categories were enriched with APA genes. Results showed that 14 out of 17 kinds of QTLs had high statistical confidence, with 3 exceptions being biochemical, leaf senescence and biotic stress (Supplemental_Table_S5.pdf). To rule out the possibility that such relationships derived from difference in the number of genes between QTL and random

selected regions, we performed a control experiment by computing the total number of genes with or without APA located on the two defined genomic regions, with or without QTL. The results, however, showed no significant difference in gene numbers, with or without APA, in either genomic region. This result suggests that QTL tends to use APA genes, implying that APA genes may play an important role in determining agronomic traits of rice.

Next, we investigated the relationship between APA gene expression level and QTL across all 14 tissues. To this end, highly expressed APA genes of each tissue (Z score of $\log_2$PAT >2) were selected, and their genomic distributions were studied. As shown in Figure 6 (only top 30 shown; full list in Supplemental_Table_S6.pdf), highly expressed APA genes were mainly enriched in two groups of QTL traits consistently on the top tier across all samples. The first group includes the top 5: root dry weight, 1000-seed weight, days to heading, plant height, and spikelet number. The second group is slightly less enriched and includes leaf length, panicle number, biomass yield, total biomass yield, panicle length, grain yield, and tiller number. Importantly, most of these traits were also in the three main categories of QTL: yield, vigor, and anatomy, which were shown to be significantly relevant to APA and switching genes (Supplemental_Table_S7.pdf). These results further suggest that APA genes may play an essential role in the determination of agronomic traits in rice.

Discussion

Rice transcriptomes have been extensively studied using high-throughput sequencing methods. Up to now, however, only few reports have investigated 3'-end information on a genome-wide scale. The present work reports a comprehensive and high-resolution map of genome-scale poly(A) sites systematically characterizing the role of APA gene

expression in 14 rice tissues that have agronomic importance,and the dynamic profile of 3'-end of genome-wide information at different stages of rice were also revealed.

Many poly(A) sites in the intergenic regions were identified. While their authenticity as potentially unannotated gene transcripts remains to be tested, their existence at such a scale (~23%) requires further investigation. Similarly，previously studies also showed that there were many poly(A) sites in the intergenic regions (Lopez et al. 2006; Shepard et al. 2011; Derti et al. 2012). Concerning the rice data presented herein, additional evidence imply that these are authentic poly(A) sites. Firstly, a large number of poly(A) sites of PAT-seq data were confirmed by previously collected rice EST data (often sequenced by low through-put and Sanger method). There are about 70% of the intergenic PACs are confirmed by EST. Secondly, single nucleotide profiles of these intergenic PACs surrounding the poly(A) sites (Supplemental_Fig_S10.pdf) are similar to classical poly(A) signal profile in annotated 3'-UTRs, indicative of their authenticity to be regular poly(A) sites. However, the question that if they are all from complete genes, or even coding for protein or RNA, is the subject for future studies because full-length transcript sequences are needed to reach a conclusion.

We also identified poly(A) sites located in regions other than 3'UTR. Their functions, as well as potential in altering gene expression, are largely unknown. About 9%~13% of PACs were located within introns across 14 tissues, suggesting that many truncated transcripts might be generated. These transcripts might escape miRNA targeting to regulate gene expression at different stages of rice development, or the alternative form may lose function, as in the case of Arabidopsis *FCA* (McKnight et al. 2002). Of particular interest, APA sites in coding sequences were also found (1.95%). This rate of APA in CDS is different from that seen in Arabidopsis (11%), using a similar PAT-seq

protocol (Wu et al. 2011). However, this figure was more similar to a previous study of Arabidopsis (1.52%), using a different method of direct RNA sequencing (Schurch et al. 2014). Two different results of APA in CDS in Arabidopsis probably resulted from the use of different methods of library production, sequencingand different ways tallying APA. However, the difference between Arabidopsis and rice might stem from different species.

The rate of APA genes in the rice genome issimilar to those of ESTs (Shen et al. 2008), but less than that of MPSS-DGE and SBS-DGE (Shen et al. 2011). As pointed out by the authors, MPSS-DGE and SBS-DGE couldnot provide exact cleavage sites information, and many poly(A) sites were double counted compared with EST data (Shen et al. 2011). The most tissue-specific PACs were in pollen, as seen in testis of zebra fish. Relative to the control of gamete-specific gene expression, some intersection between plants and animals has been reported (Ulitsky et al. 2012). About 44% of PACs were from introns of pollen. Previous studies also showed that pollen-specific splicing patterns are common in Arabidopsis (Loraine et al. 2013). Many APA events were found that potentially produce truncated proteins in rice, meaning that some proteins lost function and thus changed the biological functions of some species.

The single nucleotide profile of NUE regions for those intronic poly(A) sites found in pollen, particularly enriched in U (or T in DNA), was significantly different from the same region of other tissues, suggesting a different poly(A) signal in the introns of pollen. Recent studies have shown that several poly(A) factors have distinct preference in the choice of *cis*-element. For instance, AAUAAA hexamer was recognized by CPSF30 and WDR33 (plant homologue FY), but not CPSF160, as suggested in the past (Chan et al. 2014). In this study, *CPSF30* was down-regulated in pollen compared with leaf, while *FY* was distinctly up-regulated. This is in good agreement with the notion that reduced

CPSF30 is accompanied with reduction of A-rich motif recognition (Thomas et al. 2012). Such an alteration of poly(A) factors expression levels is known to cause a change of APA site selections (Li et al. 2015).

It was shown that 3'UTR length varies along the progression of development in animals (Ji et al. 2009). Here, we also found that seed-related tissues, especially in dry seeds, had the shortest average 3'UTR length, suggesting that some important *cis*-elements, or miRNA target sites, may be lost from 3'UTR, indicating that some mRNAs were unable to stabilize by escaping from miRNA-mediated regulation or other regulatory elements in 3'UTR. Previous studies have shown that many mRNAs were stored in seed until sowing (Sano et al. 2015). Therefore, in order to maintain the stability of storage mRNA in dry seed, it is assumed that the shortest average 3'UTR be used. Interestingly, with the development of rice, the average 3'UTR length gradually becomes longer, potentially subject to additional miRNA regulation. Moreover, we found that 3'UTR length in APA genes was longer than that of non-APA genes, suggesting the possible role of APA in regulating gene expression by increasing additional diversity and complexity.

In this study, we found many APA site switching genes among different tissues or different developmental stages of rice, indicating that APA site switching events may play an important role during the growth of rice. Discrepancy among the expression levels of switching genes was higher than that of non-switching genes. The fact that highly expressed PACs tend to use proximal poly(A) site is intriguing. Both of these imply that APA site switching is able to change the expression level, as previously identified by RNA-seq data and shown in Supplemental_Table_8.pdf, and that proximal PAC usage may have the capacity to escape miRNA targeting or other regulatory mechanisms, thus maintaining high expression. For APA switching genes that involve

switching between other genomic regions (CDS, introns, 5'UTR), however, the impact could be greater and lead to the production of truncated proteins. Previous studies have also shown that miRNA was related to rice leaf senescence, and many miRNAs were identified in rice (Xu et al. 2014). We supposed that the APA site switching event may be a requirement for several genes in order to carry out further miRNA regulation during different developmental stages or tissues of rice. Consequently, we speculate that APA site switching events may regulate gene expression via two methods. One involves indirect regulation of gene expression level via miRNA. The other involves the production of null proteins or those with reduced function. Indeed, the functions of the switching genes may substantially affect rice growth and development. As listed in Supplemental_Table_S7.pdf, many genes reported to play an important role during the growth and development of rice were APA site switching genes also shown in this study. These cases indicated that APA genes may regulate many aspects during the development of rice, including, for example, photosynthesis, tillering, as well as salt and drought tolerance.

Furthermore, we provide evidence that QTLs tend to have APA genes, indicating that APA genes may participate in the regulation of some QTLs in rice. APA can generate several isoforms and increase the diversity of gene expression regulation, and this model may be able to meet the regulatory requirement of QTLs controlled by many genes. Particularly, APA regulation should be good for fine-tuning the expression of those genes with minor effects, just like QTLs. In addition, since rice has many QTLs, APA would be a good way to increase transcripts for QTL regulation, but not increase gene number. We also found that some QTLs were enriched to specific tissues, illustrating that the choice of some QTLs occurred in these rice tissues. These results

could therefore lay a theoretical foundation for understanding the formation of agronomic traits and may be helpful for molecular-assisted breeding in rice in the near future.

Methods

Plant Materials

Rice (*Oryza sativa* L. subsp *japonica* cultivar Nipponbare) was grown in the experimental field of the Rice Research Institute, Fuzhou, Fujian Academy of Agricultural Sciences. 14 samples were collected from different developmental stages of rice (Supplemental_Materials.pdf and Supplemental_Table_S9.pdf).

PAT-seq library construction and sequencing

Total RNAs were first isolated by TRIzol reagent and used after DNase I digestion (Qiagen). The PAT-seq libraries were constructed as described (Liu et al. 2014；see Supplemental_Materials.pdf).

Data analyses

Poly(A) site analysis was performed as described (Wu et al. 2011). A strategy similar to that of a previous study was adopted to calculate the PAT-weighted 3' UTR length of each gene (Hunt et al. 2015). 3'UTR switching and non-3'UTR switching for each gene were detected as described previously (Fu et al. 2011; Mangone et al.2010). Correlation analysis between PAT and RNA-seq was as described (Lianoglou et al.2013). Details are in Supplemental_Materials.pdf and the custom scripts are included in the Perl_scripts.rar package in the supplemental documents.

APA genes and QTL analysis

The physical positions of 8217 rice QTLs were downloaded from Gramene (www.gramene.org)，covering 9 categories and 237 QTL traits. Only QTLs< 500K nt were further analyzed, and 3468 QTLs were gained. To examine PAC distributions, PACs across tissues were mapped to these 3468 QTLs and a total of 1590 QTLs with distinct information of chromosome, strand, and the start and end coordinates were retained. To examine whether APA genes or APA site switching genes were enriched in QTL regions, we compared the number of studied genes contained in QTLs with that in control regions. First, overlapped QTLs were reduced, and then 1078 QTLs were obtained. Then the same number of control regions as that in QTL was randomly selected from genomic regions not hosting QTL, but still required to preserve the same distributions of chromosomes and length as QTL. Ten trials were then run, each consisting of a random selection of control regions. Next, the number of studied genes in QTL and in control regions was calculated. The average value from ten trials was used as a control. Then a two-way contingency table recording the numbers of studied genes in or outside of QTL and control regions was obtained. Fisher's exact test was performed using fisher.test function in R to test whether the difference was statistically significant. To study the relationship of APA and QTL across different tissues, highly expressed (Z score of $\log_2$PAT >2) APA genes of each tissue were selected. For each QTL trait, the number of PACs in the highly expressed APA genes for each tissue was calculated.

qRT-PCR analysis of APA switching gene

qRT-PCR was performed on an iCycler (BioRad) using SYBR green PCR master mix. Primer sequences have shown in Supplemental_Table_S10.pdf. All data were normalized to actin.

Data access

The sequencing (PAT-seq) data from this study have been deposited to the NCBI Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra/) under accession number SRP073467. PACs can be visualized (Wu et al. 2015) and their coordinates downloaded at http://bmi.xmu.edu.cn/plantapa/; and the poly(A) sequence DNA datasets can be downloaded from the supplemental data file Rice_PAC_datasets.zip.

Acknowledgments

The authors thank other lab members for their discussion and suggestions, Haidong Qu and Wenjia Lu for technical assistance, and David Martin for language editing. Funding support of the project was from 100 Talent Plan of Fujian Province and Xiamen City, Xiamen University, and, in part, from U.S. NSF (IOS-154173) all to Q.Q.L.

Author Contributions

Q.Q.L. conceived the ideas. D.Y. and X.Y. provided materials. H.F., W.S., and L.M. performed the wet experiments. Q.Q.L., X.W., H.F., G.J. and Y.S. contributed to data analysis. Q.Q.L., H.F. and X.W. wrote the manuscript. All authors reviewed and approved the manuscript.

Disclosure Declaration

The authors declare that they have no competing financial interests.

References

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nuc Acids Res* **37**: W202-208.

Beaudoing E, Freier S, Wyatt JR, Claverie JM, Gautheret D. 2000. Patterns of variant polyadenylation signal usage in human genes. *Genome Res* **10**(7): 1001-1010.

Boss PK, Bastow RM, Mylne JS, Dean C. 2004. Multiple pathways in the decision to flower: Enabling, promoting, and resetting. *Plant Cell* **16**: S18-S31.

Chan SL, Huppertz I, Yao C, Weng L, Moresco JJ, Yates JR, 3rd, Ule J, Manley JL, Shi Y. 2014. CPSF30 and Wdr33 directly bind to AAUAAA in mammalian mRNA 3' processing. *Genes Dev* **28**(21): 2370-2380.

Colgan DF, Manley JL. 1997. Mechanism and regulation of mRNA polyadenylation. *Genes Dev* **11**(21): 2755-2766.

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

Derti A, Garrett-Engele P, Macisaac KD, Stevens RC, Sriram S, Chen R, Rohl CA, Johnson JM, Babak T. 2012. A quantitative atlas of polyadenylation in five mammals. *Genome Res***22**(6): 1173-1183.

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**(5): 741-747.

Hunt AG, Xing D, Li QQ. 2012. Plant polyadenylation factors: conservation and variety in the polyadenylation complex in plants. *BMC Genomics* **13**: 641.

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

Jan CH, Friedman RC, Ruby JG, Bartel DP. 2011. Formation, regulation and evolution of *Caenorhabditiselegans* 3'UTRs. *Nature* **469**(7328): 97-101.

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 Nat Acad Sci USA* **106**(17): 7028-7033.

Ji Z, Tian B. 2009. Reprogramming of 3 ' Untranslated Regions of mRNAs by Alternative Polyadenylation in Generation of Pluripotent Stem Cells from Different Cell Types. *PloS One* **4**(12).

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**(4): 357-359.

Lopez F, Granjeaud ST, Ghattas B, Gautheret D. 2006. The disparate nature of "intergenic" polyadenylation sites. *RNA* **12**(10): 1794-1801.

Li W, You B, Hoque M, Zheng D, Luo W, Ji Z, Park JY, Gunderson SI, Kalsotra A, Manley JL, Tian B. 2015. Systematic profiling of poly(A)+ transcripts modulated by core 3' end processing and splicing factors reveals regulatory rules of alternative cleavage and polyadenylation. *PLoS Genet* **11**(4):e1005166.

Lianoglou S, Garg V, Yang JL, Leslie CS, Mayr C. 2013. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev* **27**(21): 2380-2396.

Liu M, Xu R, Merrill C, Hong L, Von Lanken C, Hunt AG, Li QQ. 2014. Integration of developmental and environmental signals via a polyadenylation factor in Arabidopsis. *PloS One* **9**(12): e115779.

Loke JC, Stahlberg EA, Strenski DG, Haas BJ, Wood PC, Li QQ. 2005. Compilation of mRNA polyadenylation signals in Arabidopsis revealed a new signal element and potential secondary structures. *Plant Physiol* **138**(3): 1457-1468.

Loraine AE, McCormick S, Estrada A, Patel K, Qin P. 2013. RNA-seq of Arabidopsis pollen uncovers novel transcription and alternative splicing. *Plant Physiol* **162**(2): 1092-1109.

McKnight R, Duroux M, Laurie R, Dijkwel P, Simpson G, Dean C. 2002. Functional significance of the alternative transcript processing of the Arabidopsis floral promoter FCA. *Plant Cell* **14**(4): 877-888.

Mangone M, Manoharan AP, Thierry-Mieg D, Thierry-Mieg J, Han T, Mackowiak SD, Mis E, Zegar C, Gutwein MR, Khivansara V et al. 2010. The landscape of *C. elegans* 3'UTRs. *Science* **329**(5990): 432-435.

Mayr C, Bartel DP. 2009. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* **138**(4): 673-684.

Millevoi S, Vagner S. 2010. Molecular mechanisms of eukaryotic pre-mRNA 3' end processing regulation. *Nuc Acids Res* **38**(9): 2757-2774.

Ozsolak F, Kapranov P, Foissac S, Kim SW, Fishilevich E, Monaghan AP, John B, Milos PM. 2010. Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell* 143(6): 1018-1029.

Sano N, Ono H, Murata K, Yamada T, Hirasawa T, Kanekatsu M. 2015. Accumulation of long-lived mRNAs associated with germination in embryos during seed development of rice. *J Exp Bot* **66**(13): 4035-4046.

Schurch NJ, Cole C, Sherstnev A, Song J, Duc C, Storey KG, McLean WH, Brown SJ, Simpson GG, Barton GJ. 2014. Improved annotation of 3' untranslated regions and complex loci by combination of strand-specific direct RNA sequencing, RNA-Seq and ESTs. *PloS One* **9**(4): e94270.

Shen Y, Ji G, Haas BJ, Wu X, Zheng J, Reese GJ, Li QQ. 2008. Genome level analysis of rice mRNA 3'-end processing signals and alternative polyadenylation. *Nuc Acids Res* **36**(9): 3150-3161.

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**(9): 1478-1486.

Shepard PJ, Choi EA, Lu J, Flanagan LA, Hertel KJ, Shi Y. 2011. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA* 17(4): 761-772.

Sherstnev A, Duc C, Cole C, Zacharaki V, Hornyik C, Ozsolak F, Milos PM, Barton GJ, Simpson GG. 2012. Direct sequencing of Arabidopsis thaliana RNA reveals patterns of cleavage and polyadenylation. *Nat Struct Mol Biol* **19**(8): 845-852.

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

Takagaki Y, Seipelt RL, Peterson ML, Manley JL. 1996. The polyadenylation factor CstF-64 regulates alternative processing of IgM heavy chain pre-mRNA during B cell differentiation. *Cell* **87**(5): 941-952.

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**(11): 4376-4388.

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**(10): 2054-2066.

Wu X, Liu M, Downie B, Liang C, Ji G, Li QQ, Hunt AG. 2011. Genome-wide landscape of polyadenylation in Arabidopsis provides evidence for extensive alternative polyadenylation. *Proc Nat Acad Sci USA* **108**(30): 12533-12538.

WuX, Zhang Y, Li QQ. 2016. PlantAPA: a portal for visualization and analysis of alternative polyadenylation in plants. *Front Plant Sci* 7:889.

Xing D, Wang Y, Xu R, Ye X, Yang D, Li QQ. 2013. The regulatory role of Pcf11-similar-4 (PCFS4) in Arabidopsis development by genome-wide physical interactions with target loci. *BMC Genomics* **14**: 598.

Xu XB, Bai HQ, Liu CP, Chen EY, Chen QG, Zhuang JY, Shen B. 2014. Genome-Wide Analysis of MicroRNAs and Their Target Genes Related to Leaf Senescence of Rice. *PloS one* **9**(12): e114313.

Zhao H, Xing D, Li QQ. 2009. Unique features of plant cleavage and polyadenylation specificity factor revealed by proteomic studies. *Plant Physiol* **151**(3): 1546-1556.

Zhao H, Zheng J, Li QQ. 2011. A novel plant in vitro assay system for pre-mRNA cleavage during 3'-end formation. *Plant Physiol* **157**(3): 1546-1554.

Figure legends

Figure 1. Distribution of PACs in genomic regions in different samples. CDS: Coding sequences, 3'UTR: three primer untranslated region, 5'UTR: five primer untranslated region.

Figure 2. Motif around the poly(A) sites. A, Single nucleotide profiles around the poly(A) sites found in different genic regions. CDS : Coding sequence; NUE: Near upstream element; CE : Cleavage element ; FUE: Far upstream element. B, Motif analysis by MEME. Upper: introns of the pollen; lower: leaf_20days. Letter heights indicate the frequency. C, Relative expression levels of poly(A) factors (Hunt et al. 2008)between pollen and leaf_20days. PAP (LOC_Os06g21470), FY (LOC_Os01g72220), Symplekin (LOC_Os07g49320), PCSF (LOC_Os09g3927), Fip(I) (LOC_Os03g19570), CstF64 (LOC_Os05g43780), CPSF73(I) (LOC_Os03g63590), CPSF30 (LOC_Os06g46400), CPSF160 (LOC_Os04g18010), CPSF100 (LOC_Os09g39590), CFIm68 (LOC_Os07g08960), CLPS (LOC_Os02g12570). Color blocks indicate total PAT counts for each poly(A).

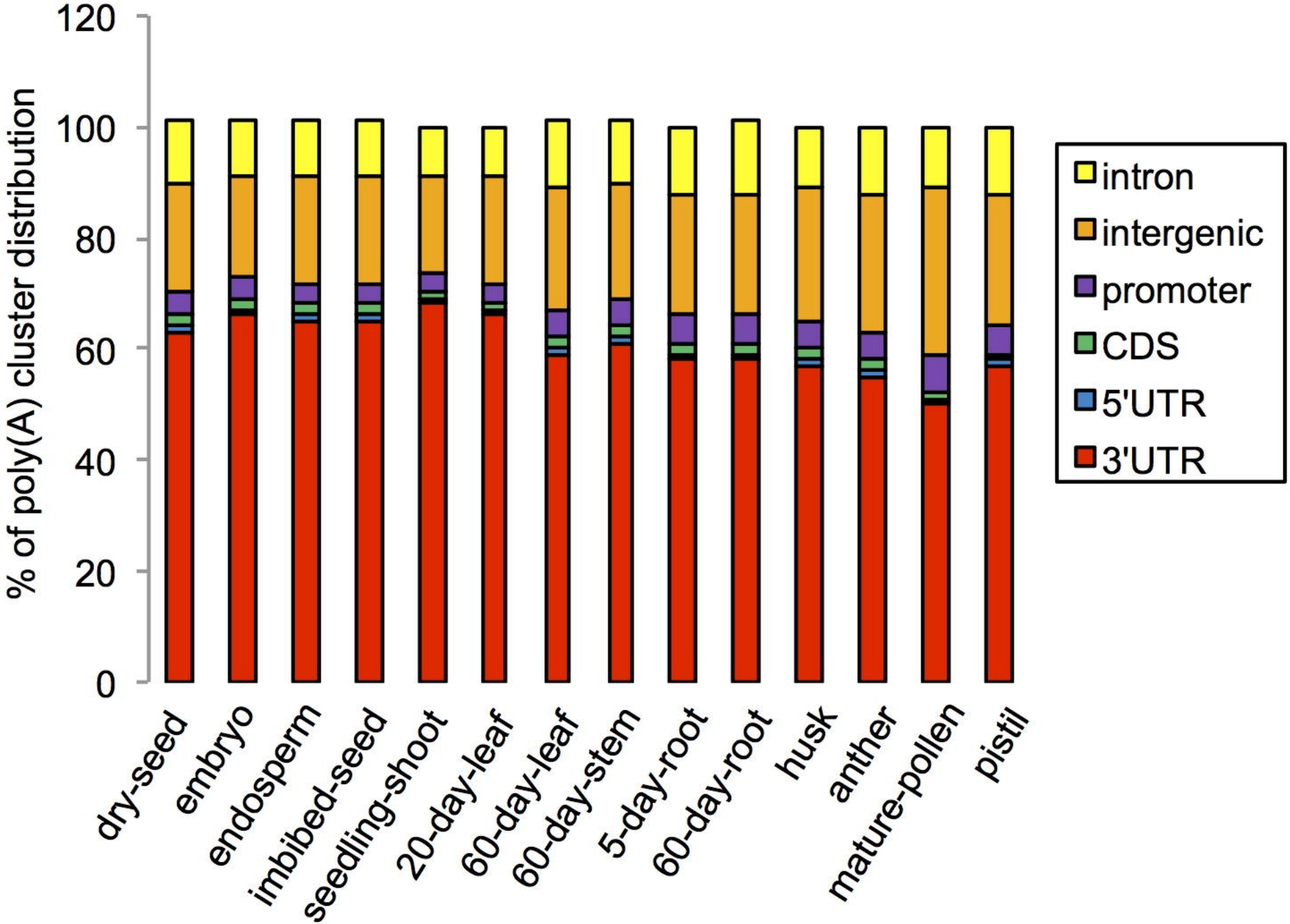
Figure 3. Expression patterns of PACs across 14 tissues. A, Number of tissue specific PACs in different samples. B, The Euclidean distances between the samples as calculated from the regularized log transformation. C, Principal component analysis of PAC expression pattern across 14 tissues. 1, anther; 2, dry seed; 3, embryo; 4, endosperm; 5, husk; 6,imbibed seed; 7,leaf_20days; 8, leaf_60days; 9, pistil; 10, pollen; 11, root_5days; 12, root_60days; 13, seedling shoot; 14, stem_60days. D, PAT ratio of differentially expressed PACs across 14 tissues. DE PACs were identified by DEXSeq. For each DE PAC, the PAT ratio is calculated as the ratio of number of PATs in a respective tissue to the total number of PATs in all tissues. E, The number of

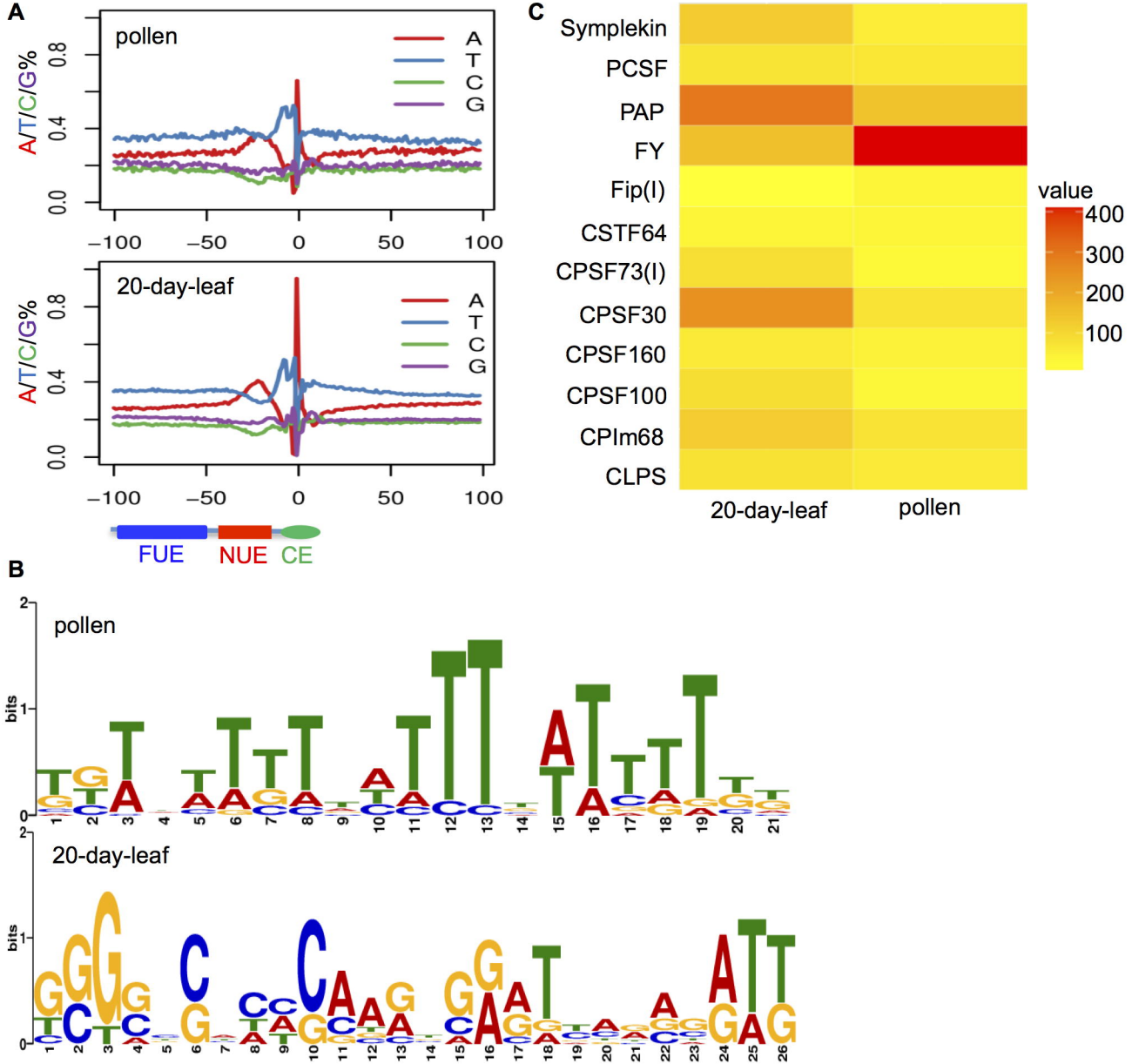
differentially expressed PACs between pollen and other tissues. Red bar represents up-regulated PAC and blue bar represents down-regulated PAC.

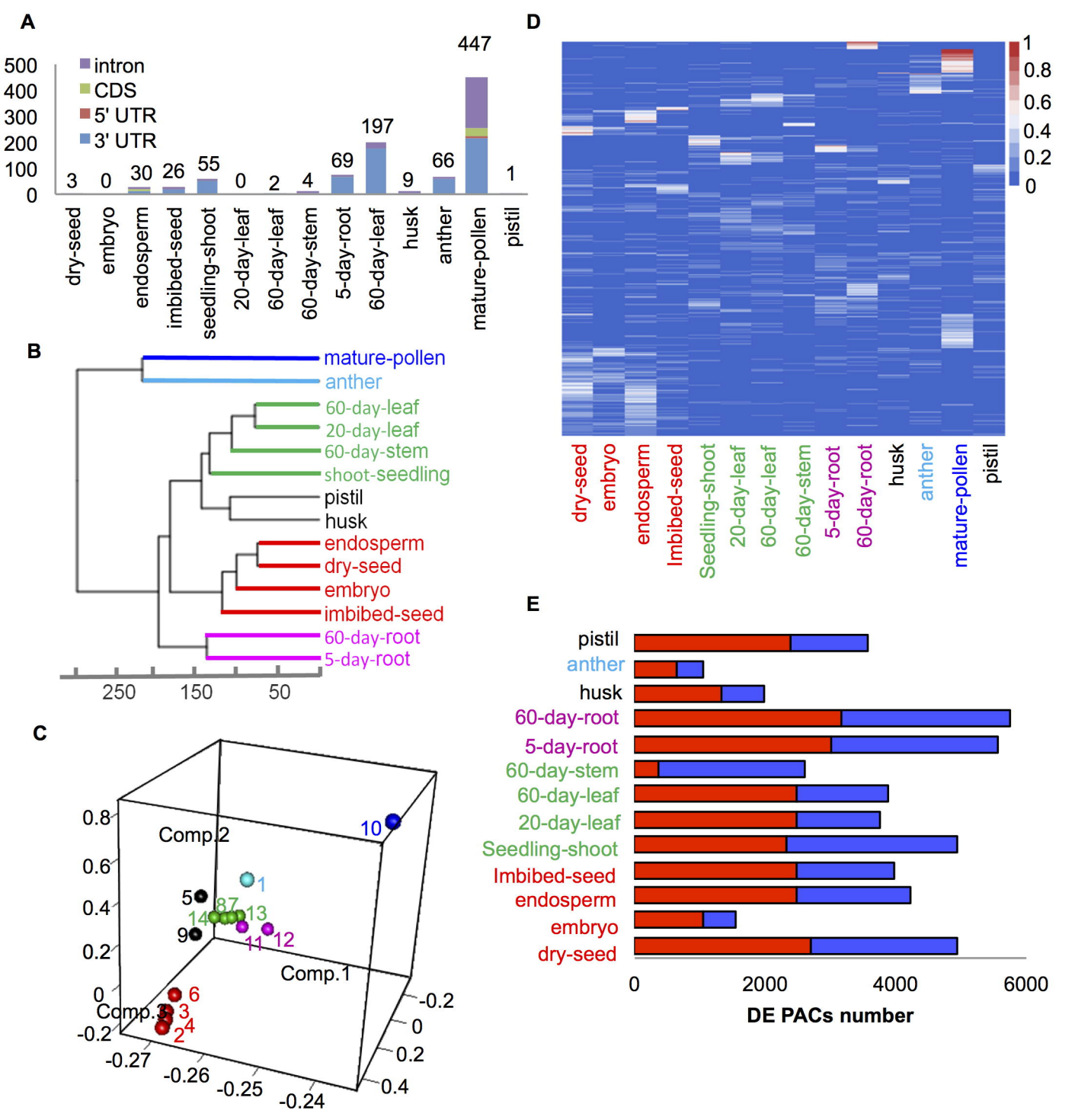
Figure 4. 3'UTR length analyses. A, Median 3' UTR length of distal PACs in different groups of genes. APA-DE, genes with at least one differentially expressed (DE) PAC; APA-NDE, genes with multiple PACs but none is DE PAC; single-DE, DE genes with single PAC; single-NDE, genes with single PAC and not DE. Vertical line is the median length of proximal PACs (204 nt for APA-NDE and 189 nt for APA-DE). B, The weighted 3'UTR median length of the highly and not highly expressed APA genes. C, Usage of the proximal PAC in highly expressed genes of APA across tissues. The red line in mature pollen marks a reverse trend where highly expressed gene use less proximal PACs. Asterisk presents statistically significant (p-value<0.05).

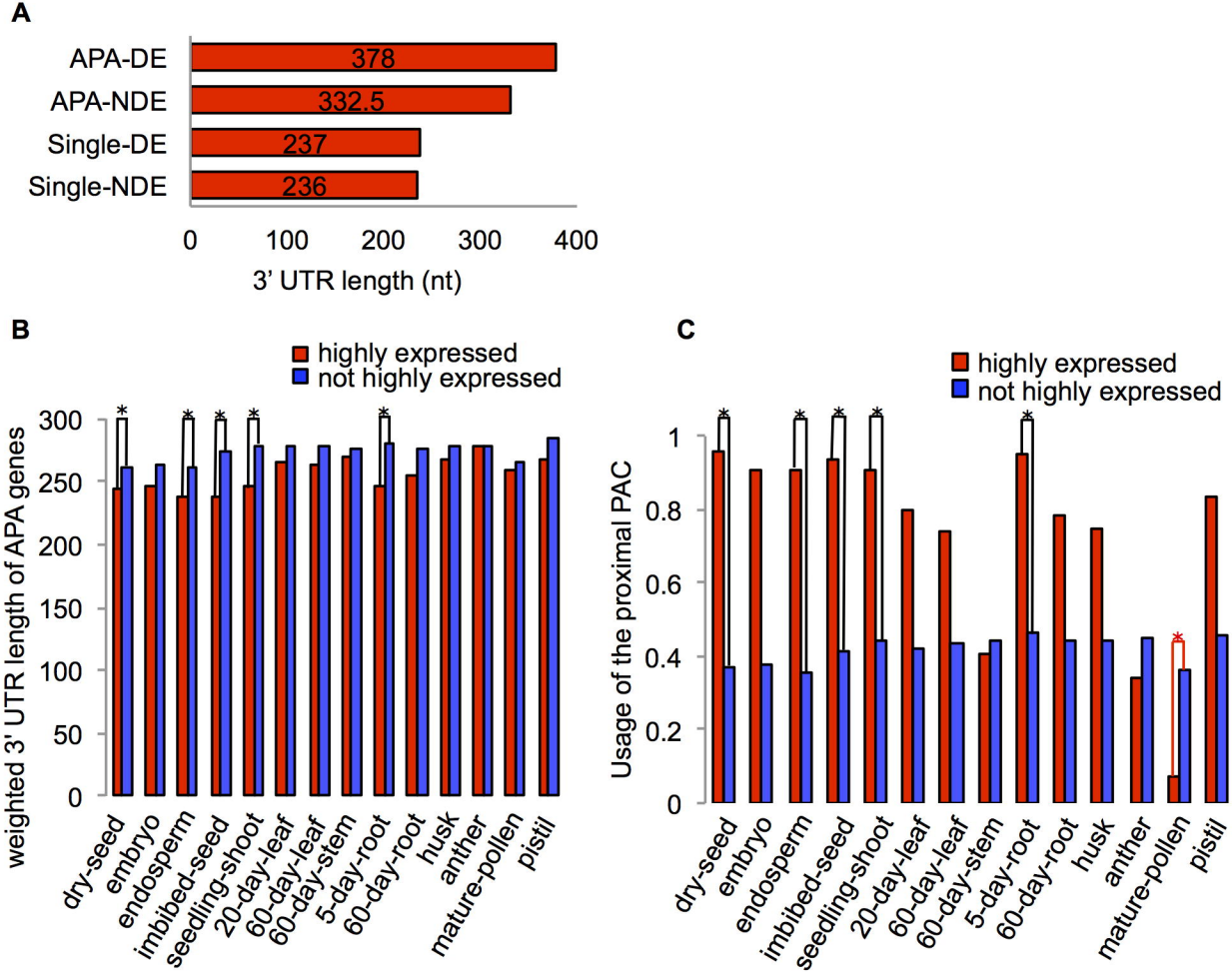
Figure 5. APA site switching genes and their distributions among different samples. A, An example of a switching gene. Bottom is gene body, big blue block is exon and small blue block is untranslated region; top indicates two poly(A) sites; middle is 3'reads in the two tissues; top indicates two poly(A) sites; middle is 3'reads. B, Switching genes between mature pollen and other tissues. C, Switching genes with 3' UTR PACs in pair-wise comparisons. D, APA site switching genes with non 3' UTR PACs in pair-wise comparisons.

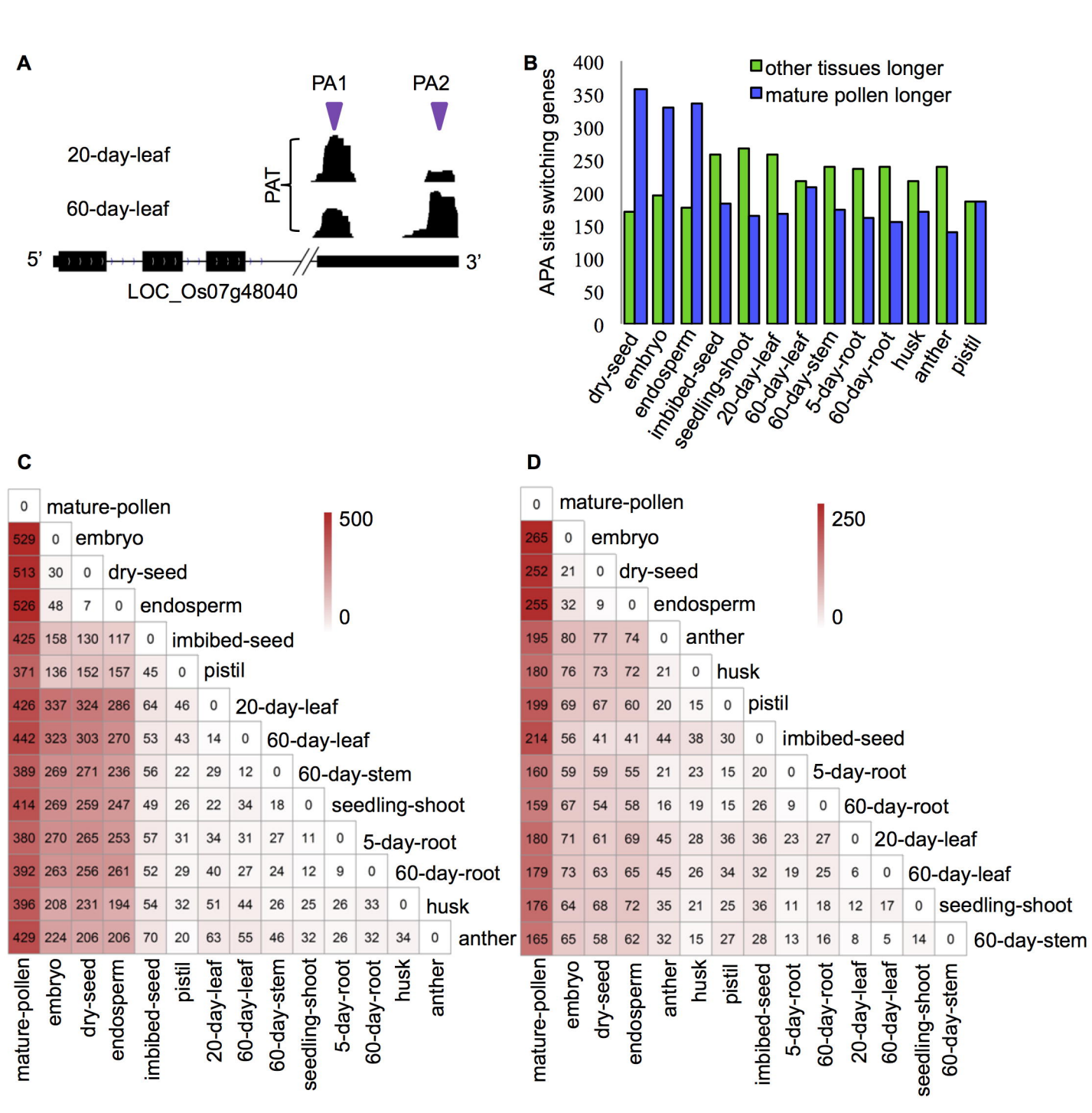
Figure 6. Number of highly expressed APA genes are enriched in QTLs (top 30) across 14 tissues. Each color block represents PAC amount (scale bar on the left) of QTL in each tissue.

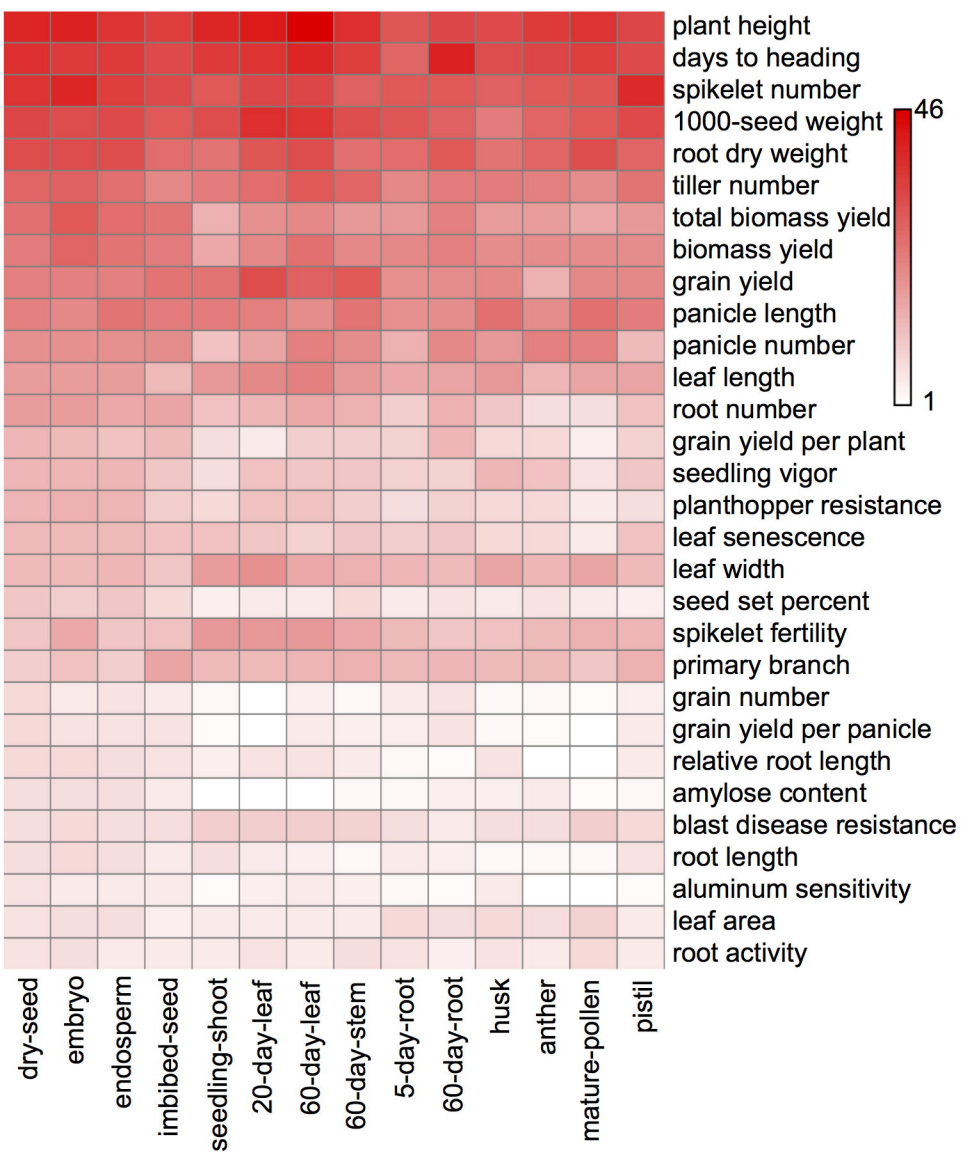

120
100
80
60
40
20
0
% of poly(A) cluster distribution
intron
intergenic
promoter
CDS
5'UTR
3'UTR
dry-seed
embryo
endosperm
imbibed-seed
seedling-shoot
20-day-leaf
60-day-leaf
60-day-stem
5-day-root
60-day-root
husk
anther
mature-pollen
pistil


A
pollen
A/T/C/G%
0.0
0.4
0.8
A
T
C
G
−100
−50
0
50
100
20-day-leaf
FUE
NUE
CE
B
pollen
bits
20-day-leaf
C
Symplekin
PCSF
PAP
FY
Fip(I)
CSTF64
CPSF73(I)
CPSF30
CPSF160
CPSF100
CPIm68
CLPS
20-day-leaf
pollen
value
400
300
200
100


A
intron
CDS
5' UTR
3' UTR
500
400
300
200
100
0
3
0
30
26
55
0
2
4
69
197
9
66
447
1
dry-seed
embryo
endosperm
imbibed-seed
seedling-shoot
20-day-leaf
60-day-leaf
60-day-stem
5-day-root
60-day-leaf
husk
anther
mature-pollen
pistil
B
mature-pollen
anther
60-day-leaf
20-day-leaf
60-day-stem
shoot-seedling
pistil
husk
endosperm
dry-seed
embryo
imbibed-seed
60-day-root
5-day-root
250
150
50
C
Comp.2
Comp.1
Comp.3
0.8
0.6
0.4
0.2
0
-0.2
-0.27
-0.26
-0.25
-0.24
-0.2
0
0.2
0.4
10
1
5
8
7
14
13
9
11
12
6
3
2
4
D
1
0.8
0.6
0.4
0.2
0
dry-seed
embryo
endosperm
Imbibed-seed
Seedling-shoot
20-day-leaf
60-day-leaf
60-day-stem
5-day-root
60-day-root
husk
anther
mature-pollen
pistil
E
pistil
anther
husk
60-day-root
5-day-root
60-day-stem
60-day-leaf
20-day-leaf
Seedling-shoot
Imbibed-seed
endosperm
embryo
dry-seed
0
2000
4000
6000
DE PACs number


A
APA-DE
APA-NDE
Single-DE
Single-NDE
378
332.5
237
236
0
100
200
300
400
3' UTR length (nt)
B
highly expressed
not highly expressed
weighted 3' UTR length of APA genes
300
250
200
150
100
50
0
dry-seed
embryo
endosperm
imbibed-seed
seedling-shoot
20-day-leaf
60-day-leaf
60-day-stem
5-day-root
60-day-root
husk
anther
mature-pollen
pistil
C
highly expressed
not highly expressed
Usage of the proximal PAC
1
0.8
0.6
0.4
0.2
0
dry-seed
embryo
endosperm
imbibed-seed
seedling-shoot
20-day-leaf
60-day-leaf
60-day-stem
5-day-root
60-day-root
husk
anther
mature-pollen
pistil


A
PA1
PA2
20-day-leaf
60-day-leaf
PAT
5'
3'
LOC_Os07g48040
B
other tissues longer
mature pollen longer
APA site switching genes
400
350
300
250
200
150
100
50
0
dry-seed
embryo
endosperm
imbibed-seed
seedling-shoot
20-day-leaf
60-day-leaf
60-day-stem
5-day-root
60-day-root
husk
anther
pistil
C
500
0
mature-pollen
embryo
dry-seed
endosperm
imbibed-seed
pistil
20-day-leaf
60-day-leaf
60-day-stem
seedling-shoot
5-day-root
60-day-root
husk
anther
D
250
0
mature-pollen
embryo
dry-seed
endosperm
anther
husk
pistil
imbibed-seed
5-day-root
60-day-root
20-day-leaf
60-day-leaf
seedling-shoot
60-day-stem


plant height
days to heading
spikelet number
1000-seed weight
root dry weight
tiller number
total biomass yield
biomass yield
grain yield
panicle length
panicle number
leaf length
root number
grain yield per plant
seedling vigor
planthopper resistance
leaf senescence
leaf width
seed set percent
spikelet fertility
primary branch
grain number
grain yield per panicle
relative root length
amylose content
blast disease resistance
root length
aluminum sensitivity
leaf area
root activity
46
1
dry-seed
embryo
endosperm
imbibed-seed
seedling-shoot
20-day-leaf
60-day-leaf
60-day-stem
5-day-root
60-day-root
husk
anther
mature-pollen
pistil