Fu et al, Genome-wide Dynamics of Alternative Polyadenylation in Rice

Supplemental Materials

Distribution of PACs in different genic and intergenic regions

Results showed that more than half of the PACs were located within the 3'UTR across tissues, while remaining PACs were defined in intergenic regions, introns, or promoters, as well as a small amount in 5'UTR and coding sequences (CDS). For each tissue, however, PAC locations within 3'UTR differed slightly from their intergenic distribution. For example, 50%~57% of PACs located within the 3'UTR across flower-related tissues, with the least percentage of PACs in pollen. On the other hand, 59%~68% of PACs located within the 3'UTR across leaf-related tissues, with the highest percentage in young shoot. From 63%~66% of PACs located within the 3'UTR across seed-related tissues, but their distribution reflected no difference. On the contrary, 28% of PACs located within intergenic regions across pollen were the most abundant among all tissues, but, as noted above, pollen was also the tissue with the least number of PACs located within the 3'UTR. A total of 17% of PACs located within the intergenic regions across seedling shoots, and it was the least percentage among these tissues. From 9~13% of PACs located within the intron across tissues, and the numbers of distribution in intron across tissues were similar. In addition, PACs in the 5'UTR were almost the same in percentile across tissues. However, PAC counts inCDS were significantly lower in certain tissues of young shoots, 20-day-old-leaf, mature pollen and pistil; PAC counts in CDS were very similar in seed-related tissues, 60-day-leaf, 60-day-stem, root-related tissues, husk and anther (Figure 1).

PAT results are quantitative as validated byRNA-seq

The validityof the poly(A) sites tallied herein was cross-referenced with previously published independent datasets generated using classical ESTs, as detailed in Shen et al (2008). To this end, poly(A) site data were downloaded and remapped to the annotation of MSU 7 (Kawahara et al. 2013), and 57,846 poly(A) sites were obtained. Supplemental Fig. 1 shows that 40,481 (70%) poly(A) sites of the EST data overlapped identical sites in the PAT-seq data, indicating a significant match of these two datasets. The remaining unmatched poly(A) sites could be explained by the strict filtration of raw data in this study, different materials used, or the fact that some poly(A) sites were not used in terms of spatiotemporal patterns.

To examine whether PAT-seq results reflected relative gene expression level, the corresponding rice RNA-seq data (20-day-old-leaf, anther, pistil, dry seed, embryo and endosperm) were obtained from NCBI Short Reads Archive (accession number: SRP008821)(Davidson et al. 2012), and the Pearson correlation of gene expression level from PATs and FPKM (expected number of fragments per kilobases of transcript sequence per millions of base pairs sequenced) across tissues were calculated. Pearson correlation across tissues was 0.59-0.83 between $\log_2$(PAT) and $\log_2$(FPKM). These results were very similar to those obtained in previous studies (Ulitsky et al. 2012; Lianoglou et al. 2013), suggesting the reasonable calculation of relative gene expression levels by PAT-seq (Supplemental Fig. 2), a finding similar to that of Wu et al. (2011).

KEGG pathway analysis of pollen specific isoforms

Pollen-specific isoforms function in some important metabolic pathways. For example, the metabolism of starch and sucrose provides enough nutrients and energy for pollen tube growth. Carotenoid biosynthesis and N-glycan biosynthesis provide reserves for sporopollenin synthesis, while flavonoid biosynthesis supplies raw materials for anthocyanin synthesis, offering special colors and defense compounds for pollen. As an important material in pollen and an indicator of fertility, proline metabolism was also active. Other basic pathways were over represented, such as basal transcription factors, carbon metabolism, amino sugar and nucleotide sugar metabolism. Therefore, pollen-specific APA isoforms might give pollen special traits and provide a potential regulatory role for pollen common functions (Supplemental Fig. 4).

Specific PACs across 14 samples

As another way to interrogate specificity, the degree of PAC distribution among the different samples was investigated, and only about 40% of PACs located in the 3’UTR were expressed in all 14 tissues tested herein (Supplemental Fig. 5). This also means that most 3’UTR PACs (60%) showed some level of sample specificity. In contrast, only a few PACs located in other genomic regions were expressed in all tissues. For example, only about 16% of intron-associated PACs and 10% CDS-associated PACs were expressed across all 14 tissues. Among them, 62% of CDS-associated PACs, 74% in intron and 83% in 5'UTR, were differentially expressed. This again demonstrates significant specificity of PACs among different tissues, suggesting potential roles of APA in rice.

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. Germinating seeds (~24 hr imbibed), leaves, and roots from the seedling stage (~5 days after imbibition) were collected after germinating in the laboratory (24°C, 16-h/8-h light/dark). Leaves from tillering stage (~20 days after transplant) were collected. Leaf, root, stem, husk, pistil, and anther from booting stage (~45 days after transplant) were collected from the field. Mature pollen was shucked off on a piece of paper and then collected using a blade. After collecting, all the tissues were immediately fixed in liquid nitrogen and stored at -80 °C until RNA isolation. Each sample had 3 replicates.

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). Briefly, 2 μg of DNA-free total RNA were fragmented into 200-400 nt by heating (94°C for 2min) with 5x first strand buffer (SuperScript® III Reverse Transcriptase, Invitrogen). Fragments with poly(A) attached were enriched using oligo(dT)$_{25}$ beads and treated with T4 polynucleotide kinase, both from New England Biolabs. Fragments were then ligated to the DNA/RNA hybrid adapter (5'-CGGTCTCGGCATTCCTGCTGAArCrCrGrCrUrCrUrUrCrCrGrArUrCrU-3'), using T4 RNA ligase I (New England Biolabs). Reverse-transcription was performed using barcoded oligo(dT) primers, purified by AMPURE® XP beads (Beckman) and eluted in DEPC-treated water. First single-strand product was amplified by PCR. To reduce bias, 18 cycles of PCR were performed with Phire II (Thermo Fisher Scientific) to generate the final PAT-seq libraries, which were tested by Agilent2100 before Illumina HiSeq2000 sequencing was performed at Novogen (Beijing, China).

Poly(A) site analysis

Raw reads were first filtered using FASTX-Toolkit (Version 0.0.14, parameters "-q 10 -p 50 -v -Q 33"), and then Ts at the beginning of the reads were trimmed by a Perl script. After that, clean reads were mapped to Nipponbare rice genome downloaded from MSU 7 (MSU Rice Genome Annotation Project Release 7) using Bowtie 2 (Version 2.1.0, parameters"-L 25 -N 0 -i S,1,1.15 --no-unal"), and only uniquely aligned reads were used in the following analysis. The resulting datasets were then processed using a series of custom Perl scripts. To reduce the extent of false poly(A) sites, internal priming was removed in the genome sequence based on a previous protocol (Loke et al. 2005) . Because cleavage sites are heterogeneous in plants, adjacent poly(A) sites in the range of 24 nt were pooled together and defined as a poly(A) cluster (PAC)(Vinciguerra and Stutz 2004), and the one with most read support was chosen to represent the poly(A) site of the cluster. To facilitate the assignments of PACs to annotated genes, genes with annotated 3' UTRs were extended for 300 nt, and genes without annotated 3' UTRs were extended by 648 nt, the average length of the annotated 3' UTRs 348 nt plus the extended length 300 nt. Each PAC was then annotated with information about its genomic locus, such as gene name and location (intron, CDS, 3'UTR, 5'UTR, and intergenic region). To remove possible artifacts with very low number of PATs and filter robustly expressed PACs, we required a PAC to have at least one of the following criteria: (1) total number of supported PATs ≥ 30; (2) PACs in the intergenic region with ≥1 PAT in ≥ 15 librariesor ≥3 PATs in ≥ 6 libraries; (3) PACs located in a genomic region with ≥ 10 PATs and ≥3 PATs in at least one library, accounting for ≥ 15% of PATs in its gene or ≥ 3% of PATs in ≥ 15 libraries. The expression levels of each PAC across different experiments were further normalized by applying a normalization factor derived from DESeq (Version 1.16.0, Anders and Huber 2010). To assess the variability of PAC expression model across samples, we performed PCA and hierarchical clustering using PAT. PCA was performed using the prcomp command with default parameters in the R software package. Hierarchical clustering was conducted based on Euclidean distances. 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.

Detection of novel motifs

Regions around the polyadenylation sites (±100nt) were predicted for sequence motifs using MEME (Bailey et al. 2009).

Correlation between PAT and RNA-seq

For better demonstration of the correlation between PATs and RNA-seq, we obtained rice RNA-seq data from the Short Read Archive in NCBI (accession number SRP00882)(E et al. 2014), including leaf_20_days (two replicates), anther, pistil, dry seed (5days and 10days), embryo and endosperm (two replicates). Then FPKM (expected number of fragments per kilobases of transcript sequence per millions of base pairs sequenced) gene expression level was obtained using TopHat (Version 2.0.14) software with these parameters: tophat -g 1 -a 10 -i 30 -I 500000 --segment-length 20 --segment-mismatches 2 cufflinks -I 500000 --min-intron-length 30. To verify that PAT-seq was quantitative at the level of mRNA abundance, Pearson's correlation coefficient was calculated for the comparison of mRNA abundance levels of samples obtained by PAT-seq ($\log_2$ PAT) and RNA-seq ($\log_2$ FPKM).

3' UTR length analysis

A strategy similar to that of aprevious studywas adopted to calculate the PAT-weighted 3' UTR length of each gene (Ulitsky et al. 2012). The 3' UTR length of each PAC is the distance from annotated stop codon to the location of this PAC. For each gene, the 3' UTR length was defined as the average 3' UTR length of all PACs weighted by the number of supported PATs. To compare 3' UTR length among different groups of genes, we divided genes into four classes: single-DE (differentially expressed genes with single PAC), single-NDE (genes with single PAC and not differentially expressed), APA-DE (genes with at least one differentially expressed PAC) and APA-NDE (genes with multiple PACs, but no differentially expressed PAC). Wilcoxon test was performed to test the statistical difference of 3' UTR length between two groups. We then applied the DESeq2 package (Version 1.4.5) (Love et al. 2014) package to identify DE genes, and genes with adjusted p-value less than 0.01 and $|\log_2(\text{Fold Change})| > 2$ were considered as DE genes. We adopted DEXSeq package (Version 1.10.8) (Anders et al. 2012) package to identify differential PAC usage between conditions, and PACs with adjusted p-value less than 0.1 were considered as differentially expressed. Genes were also grouped on the basis of their expression levels. To determine highly expressed genes, we first calculated the Z score of log2 transformed expression level of each gene denoting its relative gene expression levels. Genes with Z score ≥ 2 were defined as highly expressed genes.

Identification of APA site switching genes

To discover APA site switching genes with significant 3' UTR shortening or lengthening, genes with at least two PACs in 3' UTRs were considered. We applied a strategy similar to that of a previous study (Fu et al. 2011) to identify APA site switching genes by detecting a trend association for two-way tables with ordered levels. The chi-squared test for trend in proportions was performed using R function prop.trend.test to obtain the p-value. P-values were adjusted using the Benjamin method with R function p.adjust, and genes with adjusted p-values smaller than a given threshold were the genes with significant 3' UTR shortening or lengthening. Pearson correlation was also calculated for each gene falling between -1 and 1. As correlation increases in absolute value, 3' UTR shortening or lengthening proportionately increases. In addition, we adopted a method (Mangone et al. 2010; Wu et al. 2011) to detect APA site switching genes involving PACs in non-3' UTR regions (introns or CDS). First, for genes withmore than two PACs, the top two PACs supported by the greatest number of PATs were used and denoted as PA1 and PA2. Genes with both PA1 and PA2 located in 3' UTRs were discarded. Then genes passing through the following filtering criteria were considered as APA switching instances: (1) distance between PA1 and PA2 of at least 50 nt; (2) total read count >10 for a gene; (3) PA1:PA2 read count ratio more than 1.2-fold in one tissue and PA2:PA1 also larger than 1.2-fold in another tissue; (4) difference in read counts between PA1 and PA2 >5 within each tissue in which switching occurred.

References

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol* **11**(10): R106-R106.

Anders S, Reyes A, Huber W. 2012. Detecting differential usage of exons from RNA-seq data. *Genome Res* **22**(10): 2008-2017.

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. *Nucl Acids Res* **37**: W202-208.

Davidson RM, Gowda M, Moghe G, Lin H, Vaillancourt B, Shiu SH, Jiang N, Robin Buell C. 2012. Comparative transcriptomics of three Poaceae species reveals patterns of gene expression evolution. *Plant J* **71**(3): 492-502.

E ZG, Huang S, Zhang Y, Ge L, Wang L. 2014. Genome-Wide Transcriptome Profiles of Rice Hybrids and Their Parents. *Internat J Mol Sci* **15**(11): 20833-20845.

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 Research* **21**(5): 741-747.

Kawahara Y, de la Bastide M, Hamilton JP, Kanamori H, McCombie WR, Ouyang S, Schwartz DC, Tanaka T, Wu J, Zhou S et al. 2013. Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. *Rice* **6**(1): 1-10.

Lianoglou S, Garg V, Yang JL, Leslie CS, Mayr C. 2013. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Develop* **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.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.*Genome Biol* **15**(12): 550.

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.

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. *Nucleic Acids Res* **36**(9): 3150-3161.

Vinciguerra P, Stutz F. 2004.mRNA export: an assembly line from genes to nuclear pores. *Curr Opin Cell Biol* 16(3): 285-292.

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 Natl Acad Sci USA* **108**(30): 12533-12538.