the plant journal

SEB Society for Experimental Biology

The Plant Journal (2017) doi: 10.1111/tpj.13597

Comprehensive profiling of rhizome-associated alternative splicing and alternative polyadenylation in moso bamboo (*Phyllostachys edulis*)

Taotao Wang[1,2,†], Huiyuan Wang[1,2,†], Dawei Cai[1,2,†], Yubang Gao[1], Hangxiao Zhang[1], Yongsheng Wang[1], Chentao Lin[3], Liuyin Ma[1,*] and Lianfeng Gu[1] *

[1]*Basic Forestry and Proteomics Research Center, Fujian Provincial Key Laboratory of Haixia Applied Plant Systems Biology, Haixia Institute of Science and Technology, Fujian Agriculture and Forestry University, Fuzhou 350002, China,*
[2]*College of Forestry, Fujian Agriculture and Forestry University, Fuzhou 350002, China, and*
[3]*Department of Molecular, Cell and Developmental Biology, University of California, Los Angeles, CA 90095, USA*

Received 30 January 2017; revised 27 April 2017; accepted 3 May 2017.
*For correspondence (e-mails lma223@fafu.edu.cn; lfgu@fafu.edu.cn).
†These authors contributed equally to this work.

SUMMARY

Moso bamboo (*Phyllostachys edulis*) represents one of the fastest-spreading plants in the world, due in part to its well-developed rhizome system. However, the post-transcriptional mechanism for the development of the rhizome system in bamboo has not been comprehensively studied. We therefore used a combination of single-molecule long-read sequencing technology and polyadenylation site sequencing (PAS-seq) to re-annotate the bamboo genome, and identify genome-wide alternative splicing (AS) and alternative polyadenylation (APA) in the rhizome system. In total, 145 522 mapped full-length non-chimeric (FLNC) reads were analyzed, resulting in the correction of 2241 mis-annotated genes and the identification of 8091 previously unannotated loci. Notably, more than 42 280 distinct splicing isoforms were derived from 128 667 intron-containing full-length FLNC reads, including a large number of AS events associated with rhizome systems. In addition, we characterized 25 069 polyadenylation sites from 11 450 genes, 6311 of which have APA sites. Further analysis of intronic polyadenylation revealed that LTR/Gypsy and LTR/Copia were two major transposable elements within the intronic polyadenylation region. Furthermore, this study provided a quantitative atlas of poly(A) usage. Several hundred differential poly(A) sites in the rhizome-root system were identified. Taken together, these results suggest that post-transcriptional regulation may potentially have a vital role in the underground rhizome-root system.

Keywords: *Phyllostachys edulis*, rhizome system, single molecular real-time (SMRT) sequencing, alternative splicing, alternative polyadenylation.

INTRODUCTION

Moso bamboo (*Phyllostachys edulis*) is one of the most important non-timber rhizomatous forestry plants in the world, sustaining the livelihood of approximately 2.5 billion people (Peng *et al.*, 2013). Lateral buds on moso bamboo rhizome form new shoots which grow rapidly after emerging from soil and achieve an average culm height of 13 m within 38 days (Li *et al.*, 1998; Song *et al.*, 2016). The fast growth of the new shoots is entirely dependent on the well-developed rhizome-root system, which has very wide horizontal spread and connects the young culms with other mature bamboo plants (Li *et al.*, 2000; Embaye *et al.*, 2005; Zhou *et al.*, 2005; Song *et al.*, 2016). The rhizome system has important functions in energy storage, transportation and vegetative reproduction (Li *et al.*, 1998). The roots of bamboo are also generated from the nodes of the rhizome system (Isagi *et al.*, 1997). However, the mechanisms underlying this complex underground rhizome-root system in moso bamboo are not well understood at the whole transcriptome level.

Alternative splicing (AS) and alternative polyadenylation (APA) are two common ways for post-transcriptional regulation to generate a transcriptome and subsequently

proteomic diversity in eukaryotic organisms (Kalsotra and Cooper, 2011; Elkon *et al.*, 2013; Reddy *et al.*, 2013). Multiple transcripts are generated from a single gene locus via AS (Nilsen and Graveley, 2010). Analysis of short-read RNA sequencing (RNA-seq) has shown that about 40–60% of intron-containing transcripts are alternatively spliced in different tissues and developmental stages, linking AS with the regulation of development in *Arabidopsis thaliana*, *Oryza sativa*, *Zea mays* and *Glycine max* (Zhang *et al.*, 2010, 2017; Marquez *et al.*, 2012; Shen *et al.*, 2014b; Thatcher *et al.*, 2014; Deng and Cao, 2017). Alternatively, transcripts with a diverse 3′ end can be generated by APA (Di Giammartino *et al.*, 2011; Proudfoot, 2011; Tian and Manley, 2013). More than 70% of Arabidopsis, yeast and mammalian genes contain multiple APA sites (Ozsolak *et al.*, 2010; Wu *et al.*, 2011; Derti *et al.*, 2012; Hoque *et al.*, 2013). The Pacific BioSciences (PacBio) sequencing platform, a single-molecule sequencing technology, offers great improvement over current second-generation sequencing (SGS) technologies on read lengths (Rhoads and Au, 2015) and avoids the transcriptome assembly that is required for SGS (Steijger *et al.*, 2013). Recently, single-molecule sequencing technology has been used to characterize the complexity of transcriptome in *Salvia miltiorrhiza* (Xu *et al.*, 2015), *Z. mays* (Wang *et al.*, 2016), *Fragaria vesca* (Li *et al.*, 2017) and *Sorghum bicolor* (Abdel-Ghany *et al.*, 2016), but not yet in bamboo (Zachariah *et al.*, 2016). The draft genome of bamboo has been built and 31 987 protein-coding genes have been annotated (Peng *et al.*, 2013), which makes it an ideal system for investigating the mechanism of fast growth in the plant. Taking advantage of RNA-seq, bamboo transcriptome dynamics have been investigated for flowering (Gao *et al.*, 2014; Ge *et al.*, 2016), internode elongation (He *et al.*, 2013) and shoot development (Peng *et al.*, 2013; Li *et al.*, 2016). However, it is still difficult to obtain full-length cDNA to annotate the transcriptome because SGS only generates an incomplete profile of the transcriptome due to the limitation of read length. Compared with other members of Poaceae, the transcriptome diversity due to both AS and APA in particular remains largely unclear in bamboo.

Here, we adapted single-molecule long-read sequencing to generate full-length non-chimeric (FLNC) reads that significantly improved current annotations. Compared with previous annotations, we identified 8091 novel transcribed regions and corrected 2041 mis-annotated gene models of the bamboo genome, demonstrating that single-molecule long-read sequencing is superior for investigating poorly assembled loci or novel genes. Moreover, the genome-wide identification of multiple AS events and APA genes generated a comprehensive map of the post-transcriptional regulation network, which will facilitate future research towards elucidating the mechanisms underlying the fast growing rhizome-root system in bamboo.

RESULTS

General properties of single-molecule long-read sequencing

In this study, we combined in equal amounts total RNA from five tissues including underground rhizome tip, lateral rhizome bud, new shoot tip, root initiated from rhizome and leaf to acquire accurate full-length transcripts for single-molecule long-read sequencing (Figure 1). In total, three cDNA libraries of different sizes (1–2 kb, 2–3 kb and >3 kb) were constructed and sequenced using a PacBio RSII sequencing system. Seven single molecular real-time (SMRT) cells generated 122 787, 116 762 and 48 763 reads of inserts (ROIs) from three libraries, respectively (Table 1). In total, 288 312 ROIs were generated and more than 50% (147 362) of them were FLNC reads with the entire transcript region from the 5′ to the 3′ end based on the inclusion of barcoded primers and the 3′ poly(A) tails. The distribution of PacBio read lengths was consistent with the size of the three libraries (see Figure S1 in the online Supporting Information). To avoid and correct the high error rates compared with the Illumina platform (Abdel-Ghany *et al.*, 2016), we therefore generated 198 827 816 paired-end (PE) reads using the Illumina platform to correct the single-molecule long-reads (Figure 1). In total, we obtained 146 225 error-corrected FLNC reads. These FLNC reads cover 52% of annotated genes and 11 902 genes were supported by at least two PacBio reads (Figure S2).

In order to compare the expression level using PacBio reads with that by RNA-seq, we first calculated the fragments per kilobase of exons per million fragments mapped (FPKM) values for rhizome tip, new shoot tip and lateral bud using Illumina short read sequencing data. Hierarchical clustering analysis of RNA-seq data from different samples and biological repeats clustered together as expected (Figure S3). Real-time PCR (qPCR) and RNA-seq results further showed a high correlation (Figure S4). Primer information is listed in Data S1. Compared with RNA-seq, we found that the gene expression level determined by PacBio long reads was comparable to that quantified with FPKM of Illumina short reads (Figure S5), which suggested that with large numbers of PacBio FLNC reads it is possible to quantify gene expression with statistical confidence to sequence much deeper.

Correcting previous mis-annotated gene models

We compared mapped FLNC reads with previous annotations (Peng *et al.*, 2013) and revealed some adjacent annotated genes overlapping single contiguous FLNC reads, referred to here as mis-annotated genes (Figure 2a). In total, 2241 genes were mis-annotated as multiple split genes, which could be merged into 1092 new loci according to FLNC reads (Data S2). In order to further validate the observation, we performed RT-PCR with primers crossing

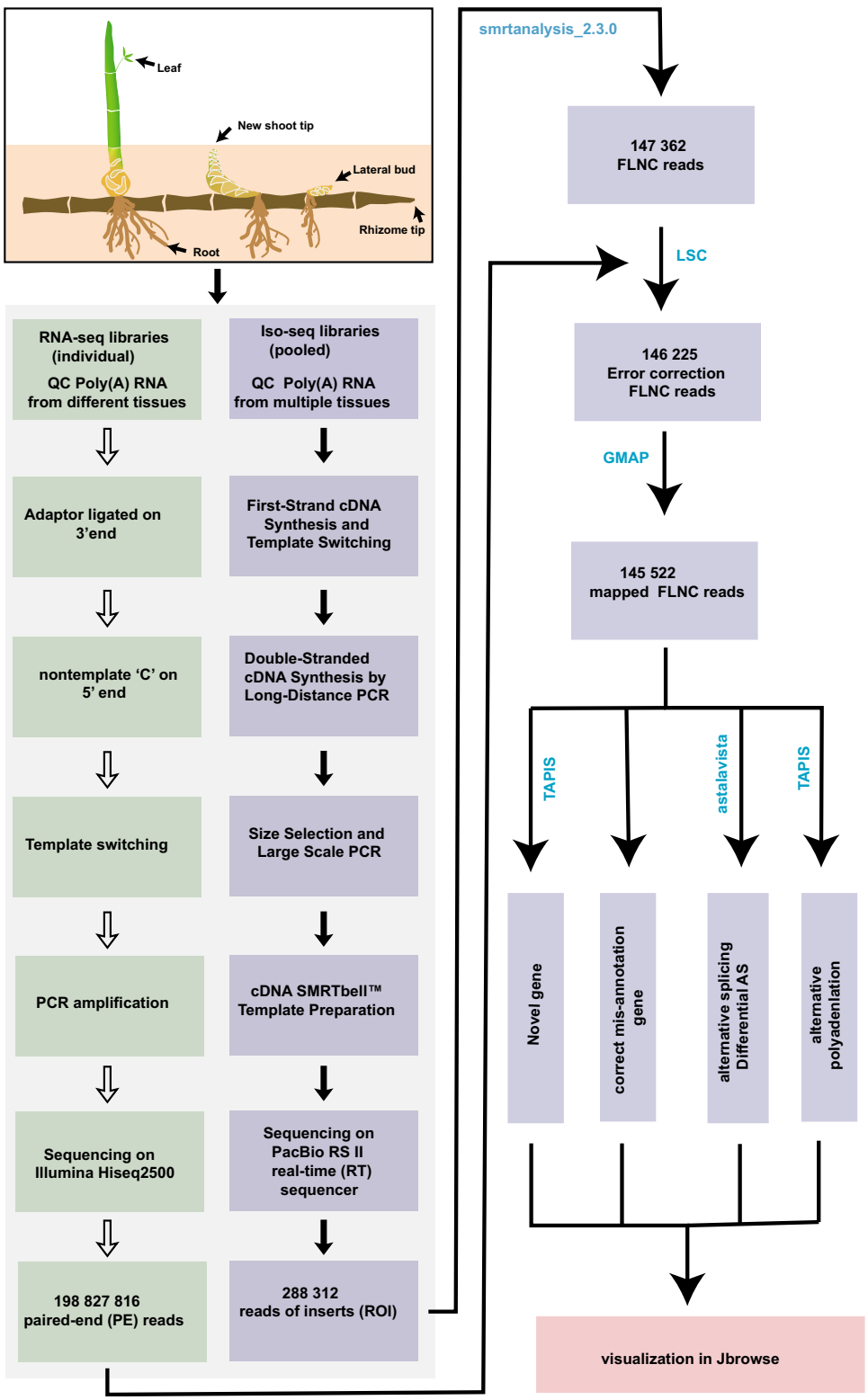

Figure 1. Flowchart of the experimental design and analysis for PacBio sequencing and RNA sequencing.
Schematic graph of tissues collected for PacBio sequencing and Illumina sequencing. Both PacBio and RNA-seq libraries were constructed and PacBio long reads from Pacific BioSciences were corrected by Illumina reads using LSC 1.alpha. In total, 145 522 mapped full-length non-chimeric (FLNC) reads were used for the downstream analysis, including identification of miss-annotated split genes, novel genes, alternative splicing (AS) and alternative polyadenylation (APA). Finally, all the datasets were visualized using Jbrowse. QC, quality control.

Table 1 Summary of PacBio single-molecule long-read sequencing

	1–2 kb	2–3 kb	>3 kb
No. of reads of insert	122 787	116 762	48 763
No. of 5′ reads	76 844	70 226	29 666
No. of 3′ reads	80 191	73 977	29 518
No. of poly(A) reads	78 192	73 008	28 595
No. of filtered short reads	7965	4138	590
No. of non-full-length reads	50 472	52 911	23 955
No. of full-length reads	64 350	59 713	24 218
No. of full-length non-chimeric reads	63 846	59 471	24 045
Average full-length non-chimeric read length	1717	2719	3710

the mis-annotated split gene pairs, and the results were consistent with those based on FLNC reads (Figures 2b and S6). Given that mis-annotated split genes originated from the same loci, they should have the same promoter and similar expression profiles due to coordinate expression in different tissues. In agreement with this hypothesis, Pearson's correlation coefficient (PCC) for each pair of mis-annotated split genes has a strong bias to a positive correlation, which strongly supports that they indeed originated from one locus (Figure 2c). More interestingly, mis-annotated split genes are much longer than other genes, making them more difficult to reconstruct using transcriptome assembly by SGS technologies (Figure 2d).

Discoveries of unannotated transcripts

We further assessed the completeness of current gene annotation. In total, 35 447 FLNC reads have no overlap with any annotated regions, which indicates that these unannotated transcripts might come from novel loci due to the lower expression level. These FLNC reads were combined into 8091 consensus clusters (Data S3). Because the reads were longer than SGS lengths, we generated an atlas of previously inaccessible bamboo transcripts. Though the expression level of these unannotated loci was low, RT-PCR did validate these new loci based on FLNC reads, indicating that they were *bona fide* loci (Figure 3a). Analysis of FPKM values of these novel loci based on Illumina short read sequencing indicated a bimodal profile, suggesting that the low expression of these novel loci might be one of the reasons that why these transcripts were excluded from previous annotation (Figure 3b). These novel transcripts displayed fewer introns than other transcripts (Figure 3c). Among the 8091 new loci, 3740 did not match any entry in SWISSPROT (Figure 3d). In total, there were 1989 loci showing homology with plant long non-coding RNA (lncRNA) databases. Non-coding prediction software predicted 3096 loci as being from the non-coding region, which further suggested that some of these new loci originate from non-coding regions. Cluster analysis of these new loci showed that these loci exhibited a tissue-specific expression pattern, which was consistent with a previous report (Derrien *et al.*, 2012), and might be another reason why these novel genes were not distinguished before (Figure 3e). These results revealed that PacBio sequencing was reliable for identification of novel loci due to the advantage of sequencing FLNC reads.

Identification of full-length splicing isoforms and differential AS events

At present, reconstructing full-length splice isoforms using an Illumina-based transcriptome assembly is still a challenge (Au *et al.*, 2013; Sharon *et al.*, 2013), thus the study of the bamboo transcriptome is far from complete, in particular the annotation of full-length splice isoforms in bamboo. We obtained a total of 128 667 intron-containing FLNC reads that collapsed into 42 280 distinct full-length splice isoforms representing 135 372 unique introns. The consensus donor sites and acceptor sites of bamboo are (A)nG|GTAAG(T)n and (T)nGCAG|GTTTGTT, respectively (Figure 4a). The consensus donor and acceptor sites were consistent with those in rice (Campbell *et al.*, 2006). The median number of introns in intron-containing genes in Arabidopsis is four (Reddy, 2007), whereas that for each unique splicing isoform in bamboo is seven.

In order to assess whether SGS reads could reconstruct these full-length isoforms, we assembled 161 379 311 aligned PE reads from our RNA-seq by genome-based assembly. In total, 110 006 (81%) unique introns based on FLNC reads could also be identified by assembling SGS reads. Surprisingly, only 10 471 (25%) distinct splice isoforms with multiple introns can be reconstructed by Cufflinks-based assemblies, suggesting that PacBio-generated long reads produce more comprehensive annotation of splicing isoforms with superior contiguity (Figure 4b). Isoforms with large number of introns are more difficult to reconstruct by Cufflinks-based assemblies as they are constrained by short read lengths. Specifically, transcriptome assembly rarely reconstructed splicing isoforms with more than 30 introns.

In total, 21 154 AS events were identified without transcriptome assembly, so any possible artificial result can be avoided (Data S4). AS genes showed higher abundance than non-AS genes (Figure 4c), which is consistent with our previous results from other species (Wu *et al.*, 2014). Gene Ontology (GO) enrichment analysis showed that these AS genes are highly enriched in cellular carbohydrate metabolic process, microtubule-based process, blue light signaling pathway, post-translational protein modification, chromatin modification and so on. All the enriched GO terms are listed in Data S5. Among these AS events 9848 have multiple AS types, suggesting that the complexity of AS types could be identified using PacBio reads. We further classified the remaining 11 306 AS events into four

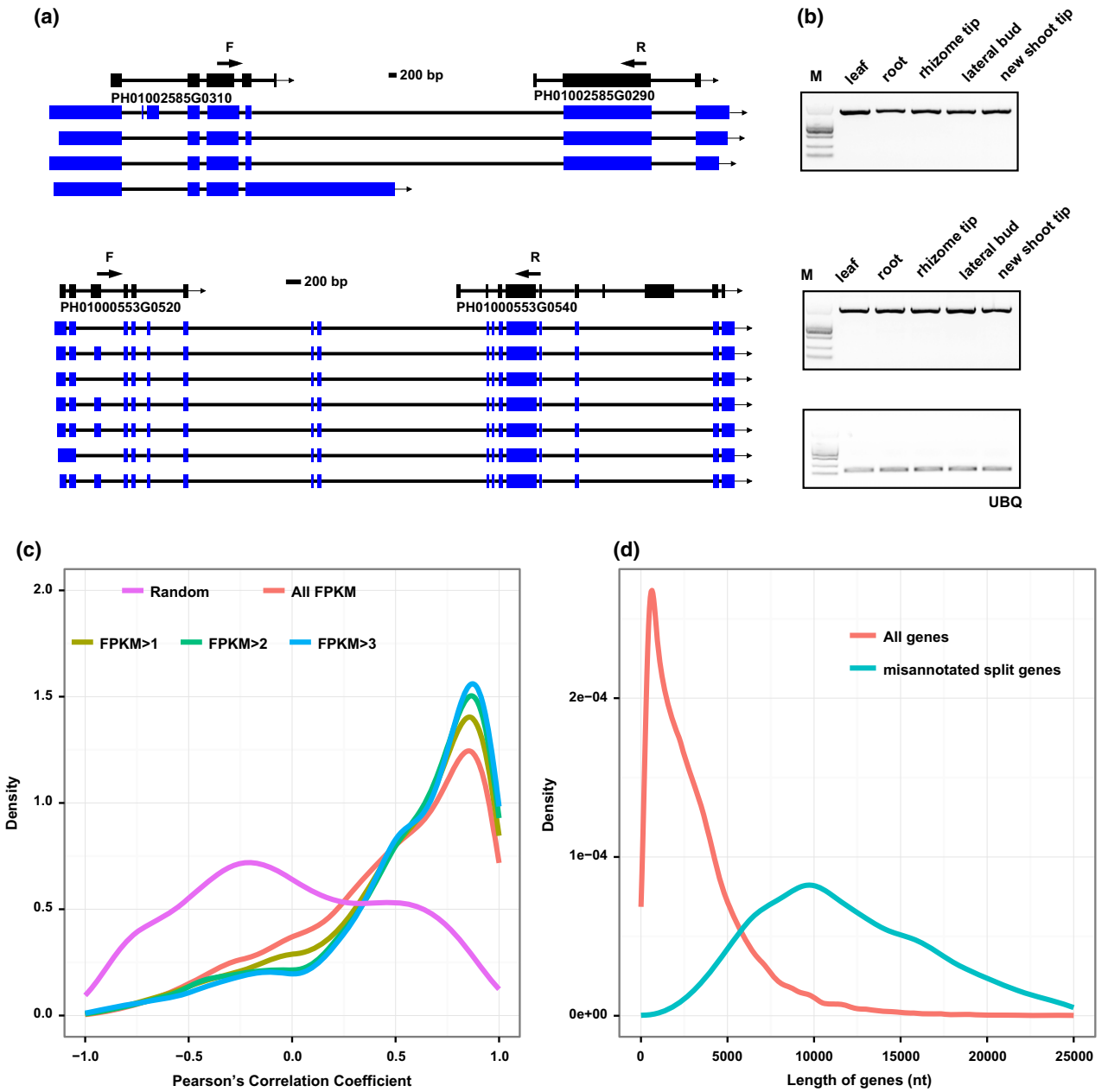

Figure 2. Identification of mis-annotated split genes.
(a) Gene structure shown mis-annotated split genes. The blue box indicates exons and the solid lines represent introns.
(b) Mis-annotated split genes are validated by RT-PCR.
(c) The distribution of Pearson's correlation coefficients by pair-wise mis-annotated split genes together with random sampling. FPKM, fragments per kilobase of exon per million fragments mapped.
(d) The length distribution of mis-annotated split genes and all genes.

distinct types: 6300 intron retention, 1427 alternative 5′ donor, 2902 alternative 3′ donor and 677 exon skipping events. Consistent with a previous study, intron retention comprised the majority of AS events (Campbell *et al.*, 2006). Since AS is a highly tissue-specific regulation, we quantified differential splicing events by pair-wise comparisons of, respectively, rhizome tip, new shoot tip and lateral bud combining Illumina sequencing data. In total, 820, 503 and 729 differential AS events, respectively, were identified (Figure 4d, Data S6). Differential AS events could be validated by RT-PCR (Figures 4e and S7). For the pair-wise comparison of new shoot tip and lateral bud, differentially spliced genes were enriched in RNA 3′-end processing (correction $P = 6.1835 \times 10^{-3}$, hypergeometric

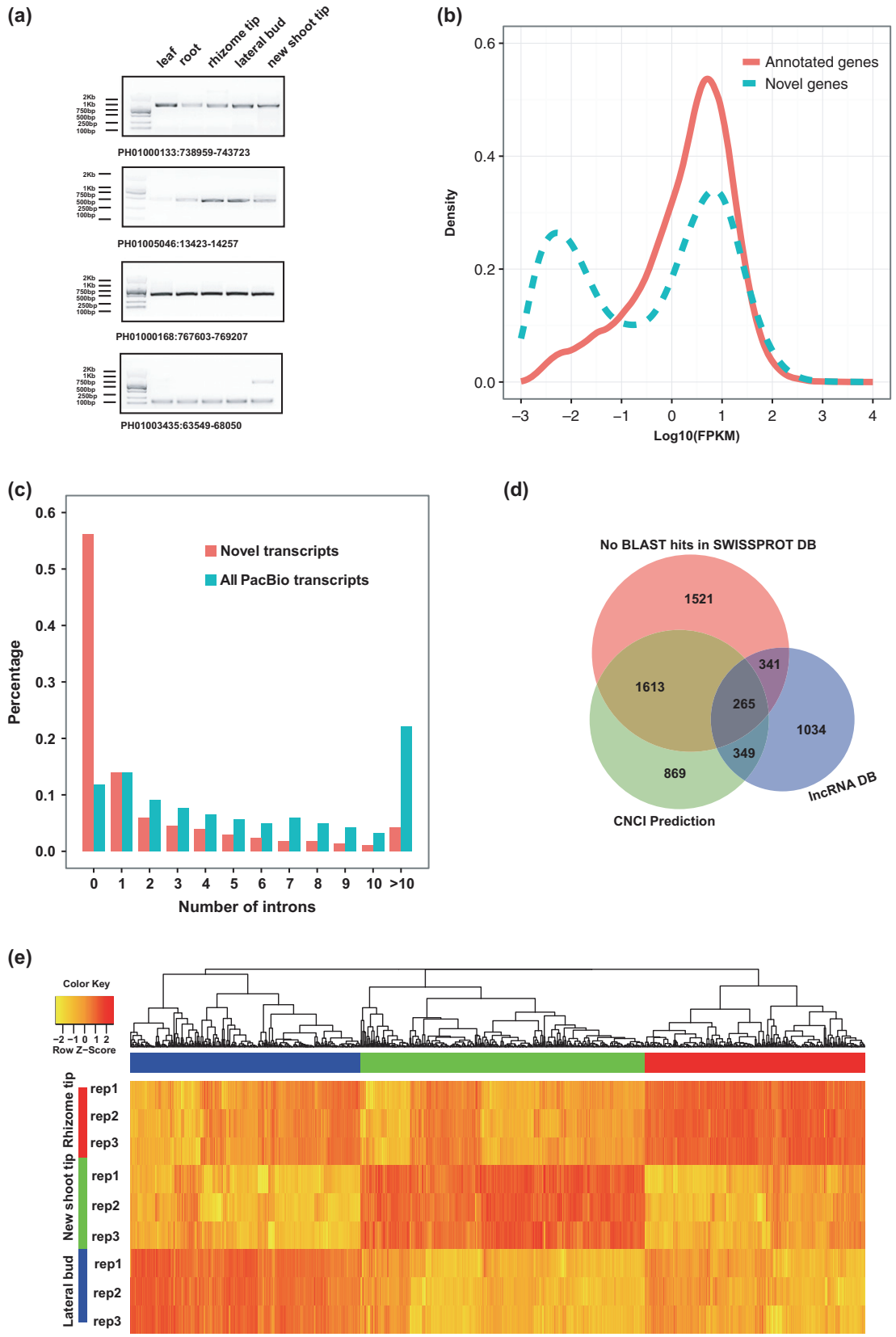

Figure 3. Summary of unannotated loci.
(a) The expression of new loci among leaf, root, rhizome tip and bud was validated by RT-PCR.
(b) Comparison of the distribution of fragments per kilobase of exon per million fragments mapped (FPKM) values from novel genes (blue line) and annotated genes (red line).
(c) Intron distribution among novel transcripts and all the transcripts.
(d) Venn diagram illustrating the overlap between the long non-coding RNA (lncRNA) database, the number of BLAST hits in the SWISPROT database and Coding–Non-Coding Index (CNCI) non-coding prediction.
(e) Heat map showing expression changes (false discovery rate <0.01 and fold change >2) in rhizome tip, new shoot tip and bud by hierarchical clustering. Three major clustering groups were identified.

test), suggesting that AS might be involved in the regulation of 3′-end processing during the fast growth of new shoot tips compared with lateral buds.

Profiling of global APA sites and differential APA

Establishment of the exact polyadenylation cleavage sites to define the 3′-end position for each gene will definitely improve gene annotation. Polyadenylation site sequencing (PAS-seq) has been used to specifically sequence and define the 3′ end of plant genes (Wu *et al.*, 2011; Fu *et al.*, 2016). However, the PAS-seq strategy has an internal priming issue since the oligo (dT) primer can mispair with the A-rich regions and produce fake 3′ ends, as reported previously (Nam *et al.*, 2002; Sherstnev *et al.*, 2012). Since PacBio sequencing of full-length of cDNA is suitable for identifying APA events (Abdel-Ghany *et al.*, 2016), here we applied single-molecule long-read sequencing to report the first genome-wide map of global polyadenylation events at single-nucleotide resolution. A total of 25 069 polyadenylation sites from 11 450 genes were profiled (Figure S8, Data S7). Motifs near bamboo polyadenylation locations included UGUA and AAUAAA (Figure S9), similar to previous reported patterns in *Sorghum bicolor* (Abdel-Ghany *et al.*, 2016).

We next sought to identify APA events, which are among the most important types of post-transcriptional regulation. In total, 6311 genes were identified with two or more polyadenylation sites (Figure 5a, Data S8). APA events were validated by RT-PCR and 3′ rapid amplification of cDNA ends (3′-RACE) (Figure S10). Nucleotide compositions around proximal and distal poly(A) cleavage sites were consistent with previous reports (Sherstnev *et al.*, 2012; Abdel-Ghany *et al.*, 2016). Proximal and distal poly (A) cleavage sites showed a very similar distribution, indicating that proximal sites are genuine poly(A) cleavage sites and therefore single-molecule long-read sequencing allowed us to facilitate comprehensive profiling of APA in bamboo (Figure 5b). Thus the PacBio full-length-reads reflected genuine cleavage sites and provided a combined regulation of AS and APA by sequencing full-length cDNA. The length and expression level of APA genes were correlated with the number of poly(A) sites (Figure S11). A total of 19 cellulose synthase (*CesA*) and 38 cellulose synthase-like (*Csl*) genes have been identified in the bamboo genome (Peng *et al.*, 2013). Among these genes, there were 11 *CesA*, 11 *Csl* and two lignin genes regulated by APA. This result suggests that APA might be involved in regulating the formation of the cell wall structure and the secondary cell wall. Enrichment of GO terms for APA genes also included cellulose biosynthetic process (Figure 5c), suggesting a potential regulation network between APA and the fast growth of woody bamboo.

To further identify the pattern of APA-containing genes, we evaluated APA in autonomous factor *FPA* (PH01000191G0930). *FPA* is a well-known flowering pathway gene and shows a striking similarity to other species in its use of poly(A) sites (Hornyik *et al.*, 2010; Xing and Li, 2011). It has two major poly(A) sites, with the proximal poly(A) sites being located in the last intron of the longer transcripts; this can be validated by 3′ RACE (Figure 5d). We performed genome-wide investigation of *FPA*-like intronic APA. A total of 362 intronic polyadenylation sites were identified. Interestingly, these introns were longer than that without polyadenylation sites (Figure 5e). We then identified the transposable elements (TEs) from the 319 introns. The 328 intronic TEs including retrotransposons, transposons and miniature inverted-repeat transposable elements (MITEs) inside these long introns (Figure 5f). Gypsy and Copia are two major retrotransposons found in intronic TEs. DNA/MuDR comprises the majority of transposons. Ten of the MITEs are *Tourist* elements. These TEs located at the long intron-containing polyadenylation cleavage sites might regulate intronic heterochromatin and thus influence the choice of polyadenylation site.

At present, it is still difficult to quantify APA in the different growth stages by single-molecule long-read sequencing due to the low sequencing depth. Thus, we quantified differential poly(A) sites by combining the strength of PacBio-generated full-length reads for definition of poly(A) sites with the advantage of SGS for quantification due to the high depth, using the recent development of strand-specific PAS-seq in plants to reveal differential poly(A) sites in the rhizome-root system in bamboo (Ma *et al.*, 2014b; Zhang *et al.*, 2015). In total, 1224 differential poly(A) genes were identified by pair-wise comparisons of rhizome tip, new shoot tip and lateral bud, respectively (Data S9, Figure 6a, b). For the pairwise comparison of new shoot tip and rhizome tip, there were 590 differential poly(A) genes (Figure 6a). Of these, 197 differential poly(A) genes

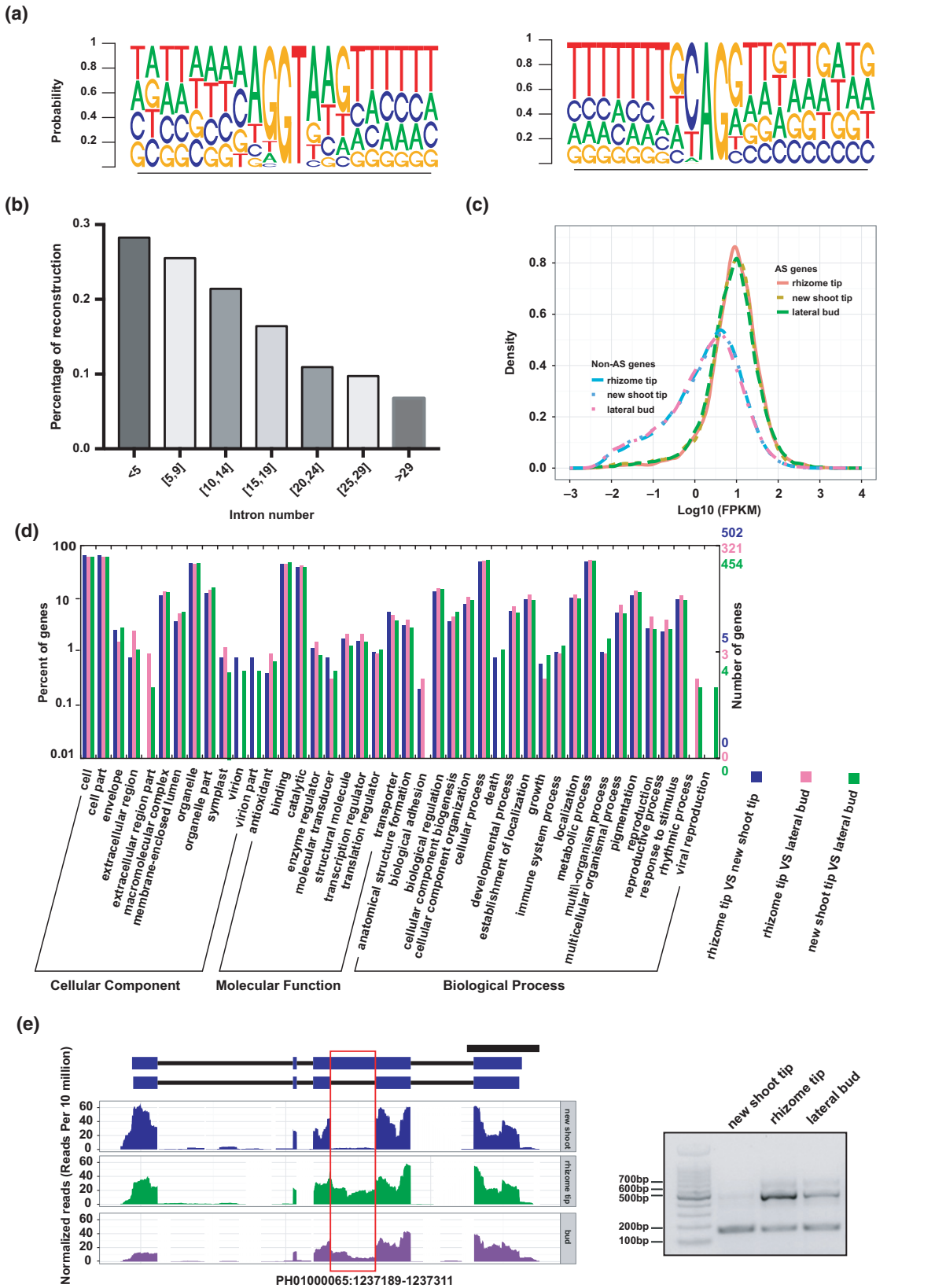

Figure 4. Genome-wide identification of splicing isoforms.
(a) The nucleotide distributions flanking the donor or acceptor sites are shown as a pictogram. The sequence logo shows the frequency distribution of nucleotides at the 5′ donor site and 3′ acceptor sites.
(b) The percentage of splicing isoforms reconstructed by Cufflink assembly with different intron numbers.
(c) Distribution of expression levels between alternatively spliced (AS) genes and non-AS genes.
(d) Distribution of Gene Ontology (GO) terms (the second level) including the percentage of the differentially spliced genes and the numbers of genes attributed to the GO terms.
(e) Wig plots of the differential splicing events. The *y*-axis represents the normalization value (reads per 10 million). The right panel presents the validation of differential AS events using RT-PCR.

can also be identified between new shoot tip and lateral bud (blue dot in Figure 6a). As shown in the Figure 6(c), the use of the proximal poly(A) site was apparently promoted in new shoot tip compared with other tissues. Among 770 differential poly(A) genes in the comparisons of lateral bud and rhizome tip (Figure 6b), 295 were also identified in the comparison between new shoot tip and rhizome tip (green dot in Figure 6b). GO analysis showed that these genes were involved in the categories molecular transducer, transcription regulator, translation regulator and transporter (Figure 6d). Take together, these PAS-seq results provides a quantitative atlas of poly(A) sites in the rhizome system of bamboo.

DISCUSSION

Our current knowledge about the bamboo transcriptome is mainly based on gene expression from SGS technologies. Thus the bamboo transcriptome had not been fully characterized due to lack of full-length cDNA. In this work we re-annotated bamboo using PacBio third-generation sequencing and identified 2241 loci that were mis-annotated as multiple genes based on FLNC reads spanning long distances. Also, 8091 previously unannotated transcripts might derive from low-abundance transcripts or lncRNA that exhibited highly tissue-specific expression. Those new loci might play an important role during the development of rhizome tip, new shoot tip and lateral bud. Previous studies used seven bamboo tissues to annotate the bamboo genome, which may be one of the reasons why these novel genes were not previously distinguished as the rhizome system was not used (Peng *et al.*, 2013). The bamboo annotation was significantly refined based on FLNC reads, highlighting the potential of single-molecule long-read sequencing for genomic annotation.

The bamboo transcriptome is complex due to post-transcriptional regulation. Though Cufflinks could perform transcript model reconstructions, our results showed that it was difficult for transcript assembly to detect transcripts possessing multiple introns. Thus these complex transcripts were missing from current annotation due to the intractable problem of assembly of short reads from SGS technologies. Here, PacBio sequencing has enabled the identification of the complexity of AS in bamboo at a genome-wide level because FLNC reads were superior to those from Cufflinks-based assemblies and could required the accurate identification of AS. In this study we detected several hundred differential splicing events in the rhizome-dependent system. Core proteins (Sm) (Kambach *et al.*, 1999) encoded by *LSM8*, *Prp43-1* and *SmD3-b* showed obvious upregulation at the new shoot tip. *RNPA/B-6*, belonging to the hnRNP A/B family, showed upregulation at the rhizome tip and *RBP47c*, *hnRNP-G2* and *RBP4*5b encoding classic hnRNP proteins showed strong upregulation at the lateral bud (Figure S12). It was possible that these differential splicing-related genes took part in splicing regulation in the rhizome-dependent system.

Common RNA-seq cannot provide sufficient resolution to profile polyadenylation sites (Nagalakshmi *et al.*, 2008). Recently, PAS-seq has been developed to detect APA (Shepard *et al.*, 2011); however, the method may introduce artifacts due to potential internal priming issues (Nam *et al.*, 2002; Sherstnev *et al.*, 2012). Comprehensive annotation of the PAS has been conducted in yeast and human by using direct RNA-seq (DRS) technology (Ozsolak *et al.*, 2009). However, DRS reads cannot represent the full structures of transcripts (Sherstnev *et al.*, 2012). PacBio sequencing allowed the identification of cleavage sites for polyadenylation, which is important for gene annotation (Abdel-Ghany *et al.*, 2016). Thus, in this study, PacBio sequencing was used to obtain long reference sequences and profiling of APA from 6311 genes in bamboo. Compared with PAS-seq and DRS, single-molecule long-read sequencing generated full-length structures of each splicing isoform and defined exact polyadenylation sites. Our data present a comprehensive view of APA at genome-wide level.

The present total of 6311 APA genes might be an underestimate of the true number due to the low expression of part proximal poly(A) sites. For example, we could not detect the proximal poly(A) for *FCA* (PH01002230G0270), which could be detected in Arabidopsis (Wu *et al.*, 2011). One possibility is that the abundance of proximal poly(A) was low or the proximal poly(A) showed tissue-specific expression. Two *PAP* genes showed upregulation at the new shoot tip (Figure S13). *CPSF100* and *CFIm25* showed upregulation and *CPSF73-I* indicated downregulated expression at the lateral bud (Figure S13). These poly(A) factors might take part in the regulation of poly(A) cleavage.

In total, 362 introns from 319 genes included intronic polyadenylation which may encode shorter proteins

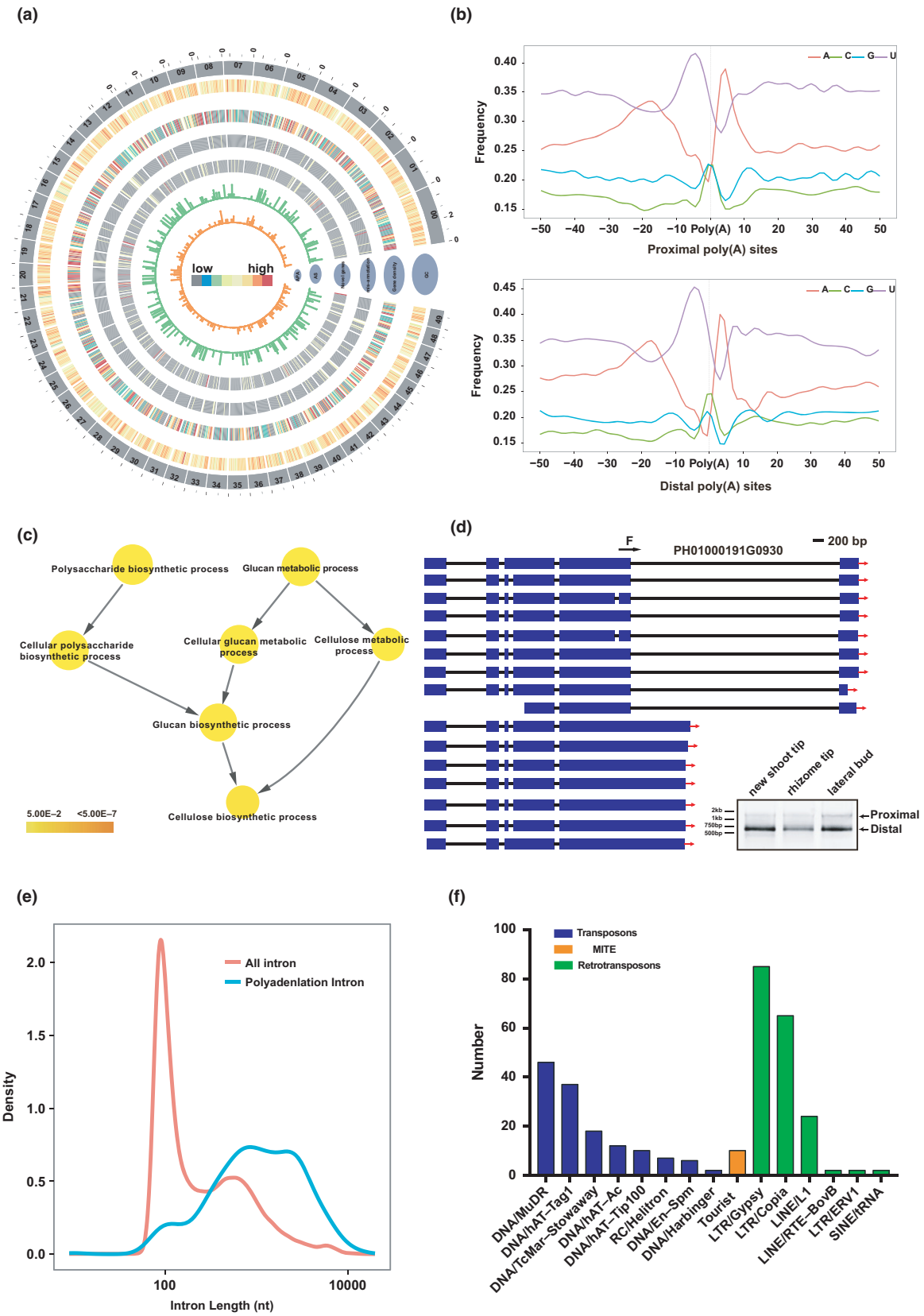

Figure 5. Characteristics of alternative polyadenylation (APA) in bamboo.
(a) Concentric circles diagram illustrating the genome-wide profile of novel genes, alternative splicing (AS) and APA. The largest 50 scaffolds are plotted at the outer circles. Outer to inner circles display GC content, gene density, mis-annotated split genes, novel genes, AS and APA, respectively. (b) The nucleotide composition profile around proximal and distal poly(A) sites. (c) Significantly enriched Gene Ontology (GO) terms for APA genes involving a cell wall-related function. The *x*-axis indicates the locations relative to middle polyadenylation cleavage sites and the *y*-axis represents the nucleotide composition percentages. (d) Schematic representation of gene structure for FPA which included distal poly(A) sites in the 3′-untranslated region and proximal poly(A) sites within the last intron in a strand-specific manner. The blue box indicates exons and the solid lines represent introns. (e) Length distribution introns with polyadenylation cleavage sites and all the introns. The *x*-axis is intron length and the *y*-axis density. (f) Plot diagram presents the distribution of transposable elements located in introns including polyadenylation cleavage sites. MITE, miniature inverted-repeat transposable element.

(Ozsolak *et al.*, 2010). Previous studies showed that association of intronic repeat elements with epigenetic regulation affects the intronic polyadenylation within long introns in Arabidopsis (Wang *et al.*, 2013; Lei *et al.*, 2014). As the intronic polyadenylation site is located within the TE-containing region, the modification of TE elements might control the use of such intronic polyadenylation and therefore regulate the choice of polyadenylation site. In this study we investigated all intronic polyadenylation and found that Gypsy/Copia from retrotransposons and MuDR from transposons made up the majority of TEs located in the long intron, including polyadenylation; this might be related to DNA methylation, as previously reported (Wang *et al.*, 2013; Lei *et al.*, 2014; Ma *et al.*, 2014a). More DNA methylation and histone modification data from bamboo will reveal the interplay between epigenetics and polyadenylation in future.

In this study there were 1224 genes that showed a differential poly(A) site, suggesting that differential poly(A) sites might potentially contribute to the regulation of the fast-growing rhizome-root system. In total, there were 31 splicing-related genes that showed differential poly(A), suggesting that APA might be involved in the regulation of splicing factors (Data S9). At the same time differential AS genes were enriched in RNA 3′-end processing (Data S6), suggesting that AS might be involved in the regulation of APA. In addition to splicing factor and poly(A) factors, there were 21 transcription factors that showed differential poly(A) (Data S9) and 20 with differential AS (Data S6). Transcription was coupled to pre-mRNA splicing and polyadenylation. It may be expected that APA and AS are also regulated by transcription factors to form a feedback loop. Together, this work provides useful information for future studies of the interplay between APA and AS, which should help to fully reveal the molecular mechanism of the fast-growing rhizome-root system in bamboo.

EXPERIMENTAL PROCEDURES

Plant materials

The underground rhizome tip, lateral bud, new shoot tip and root initiated from rhizome, and leaf of moso bamboo (*P. edulis*) were collected from the botanical garden of bamboo at Fujian Agriculture and Forestry University (E119°14′; N26°05′) in late November. The sampled plants originated from the same rhizome system. After peeling off the shoot sheath leaves and cutting off the radical part, the tips of each bud were quickly frozen in liquid nitrogen and stored at −80°C. Tips from three buds with similar growth vigor were combined to form a sample and three biological replicates were used for each tissue. Total RNA was extracted using an RNAprep Pure Plant Kit (Tiangen, cat. no. DP441, http://www.tiangen.com/en/) with DNase I treatment to remove genomic DNA contamination. The concentration and integrity of total RNA were assessed with Quant-IT RiboGreen (Invitrogen, cat. no. R11490, http://www.invitrogen.com/) and an Agilent 2200 TapeStation (Agilent Technologies, http://www.agilent.com/), respectively. Total RNAs with an RNA integrity number (RIN) value larger than 8 were mixed equally from different tissues.

Construction of a PacBio library with different library sizes

The isoform sequencing (Iso-seq) libraries were constructed using 1 μg of equally mixed RNAs. First-strand cDNA was synthesized using a Clontech SMARTer PCR cDNA synthesis Kit (cat. no. 634926, http:/www.clontech.com/) with anchored oligo(dT)30 as the primer. Double-strand cDNA was amplified by long-distance PCR (LD-PCR) using an Advantage 2 PCR kit (Clontech, cat. no. 639206). The 1–2-, 2–3- and >3-kb cDNA fractions were generated with a BluePinppin size selection system (Sage Science, http://www.sagescience.com/). The three libraries were constructed with a Pacific Biosciences SMRTbell Template Prep Kit 1.0 (part 100-259-100, http://www.pacb.com/) following the vendor's protocol. The libraries were subsequently sequenced on the PacBio RS II real-time (RT) sequencer platform using SMRT Cell 8 Pac v3 (part 100-171-800) with a total of seven SMRT cells: the 1–2- and 2–3-kb libraries were sequencing using three SMRT cells, respectively; while the >3-kb library used one SMRT cell.

Analysis of PacBio single-molecule long-reads

Bamboo genome sequences and annotated gene models were downloaded from NCGR (Peng *et al.*, 2013). ConsensusTools from the smrtanalysis_2.3.0 (Pacific Biosciences) was used from the command line to get reads of the insert. Then, the command-line wrapper script pbtranscript.py from smrtanalysis_2.3.0 was used for identification of the full-length read. The 5′ and 3′ primers observed with the Clontech kit as well as a poly(A) tail signal preceding the 3′ primer were used to produce strand-specific full-length reads (https://github.com/PacificBiosciences/cDNA_primer/wiki/RS_IsoSeq-(v2.3)-Tutorial-%231.-Getting-full-length-reads). Then SMRT long reads from Pacific BioSciences were corrected by Illumina reads using LSC 1.alpha with Bowtie 2 v.2.2.1 alignment (Au *et al.*, 2012). Then, PacBio sequencing was mapped to the bamboo genome using GMAP with the following option: -B 5 -K 8000 -t 40 -f 2 -n 1 (Wu *et al.*, 2016). The above GFF3 format was transferred into GTF format using gffread – one of the Cufflinks tools (Trapnell *et al.*, 2012). Then AS events were identified based on the above GTF file using the ASTALAVISTA algorithm (Foissac and Sammeth, 2007). New genes and APA sites were identified by TAPIS (Abdel-Ghany *et al.*, 2016). Since we removed non-full-length reads, we thus regarded the terminal

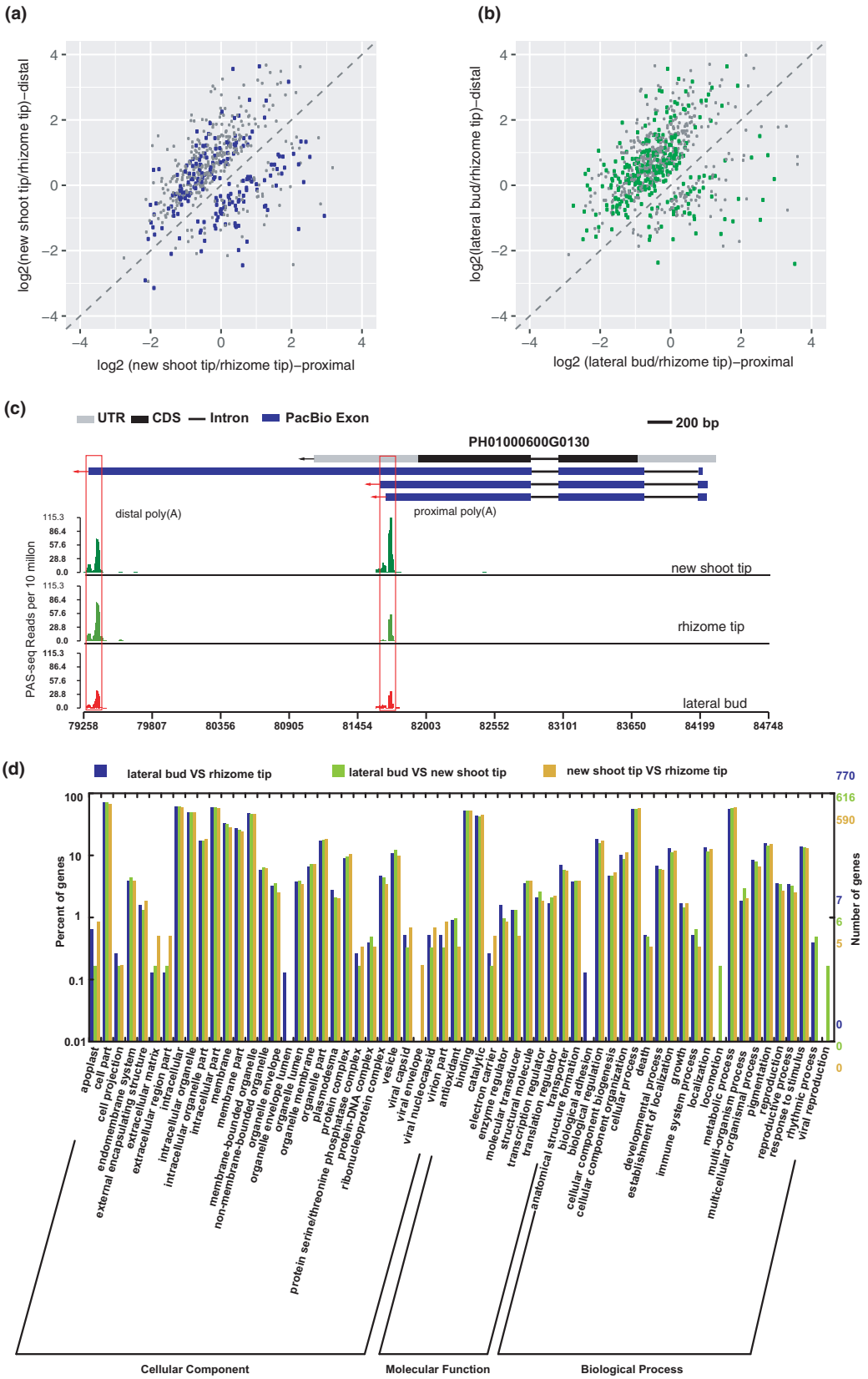

Figure 6. Quantitative maps of differential poly(A) sites.
(a) Scatterplots showing the pairwise comparison of new shoot tip and rhizome tip. The *y*- and *x*-axes are $\log_2$(new shoot tip/rhizome tip) values for distal and proximal poly(A) sites, respectively. Those that also were regulated in the comparison of new shoot tip and lateral bud are shown in the blue plot.
(b) Scatter plots representing pairwise comparison of lateral bud and rhizome tip. The *y*- and *x*-axes are $\log_2$(lateral bud/rhizome tip) values for distal and proximal poly(A) sites, respectively. Those that are also regulated in the comparison of new shoot tip and rhizome tip are shown in the green plot.
(c) One example (PH01000600G0130) showing differential alternative polyadenylation sites. Annotated untranslated regions (UTRs) and coding sequences (CDS) are denoted by grey and black rectangles, respectively. PacBio transcripts are represented as blue rectangles. Wiggle plots represent the distribution of polyadenylation site sequencing (PAS-seq) reads for new shoot tip, rhizome tip and lateral bud, respectively.
(d) Distribution of Gene Ontology (GO) terms (the third level) including the percentage of the differentially poly(A) genes and the numbers of genes attributed to the GO terms.

sites of single reads as genuine poly(A) sites. When many reads occurred at one poly(A) site, a short sequence window of 20 nucleotides (nt) was used to cluster micro-heterogeneity sites according to the default option of TAPIS (Abdel-Ghany *et al.*, 2016). MEME was used for motif searches on the upstream and downstream sequences of polyadenylation sites (Bailey *et al.*, 2009). Sequences of new genes were used to search the homolog with the plant lncRNA database GreeNC (Gallart *et al.*, 2016) and CANTATAdb (Szcześniak *et al.,* 2016) using NCBI blast-2.2.27 + with a threshold E-value of $<1 \times 10^{-6}$.

Preparation and sequencing of the RNA-seq library

The Illumina HiSeq 2500 platform was used to generate paired-end (PE) reads to correct PacBio reads and quantify splicing. Total RNA quality was assessed using an Agilent 2100 with an RIN value greater than 8. Strand-specific RNA-seq libraries were constructed using 5 μg of total RNA from the rhizome tip, lateral bud and new shoot tip with three replicates by a dUTP strand-specific library protocol (Levin *et al.*, 2010), which was described in our previous study (Wu *et al.*, 2014). Then dUTP strand-specific libraries were sequenced using the HiSeq 2500 sequencing system as 125-nt paired-end reads.

Preparation and sequencing of the strand-specific PAS-seq library

The PAS-seq libraries were constructed using total RNA with a modified protocol based on our previous studies (Ma *et al.*, 2014b; Zhang *et al.*, 2015). Briefly, 2 μg of total RNA was fragmented using SMARTscribe reverse transcription reaction buffer at 94°C for 6 min. the polyadenylated mRNA fragments were separated using a NEBNext® Poly(A) mRNA Magnetic Isolation Module (New England Biolabs, cat. no. E7490L, https://www.neb.com/). The cDNA fragments were then synthesized using a Clontech SMARTer cDNA synthesis Kit (Clontech, cat. no. 634926) with primers HITS-3′ and HITS-5′ (Data S1). The fragments were further size-selected using AMPure XP beads (Beckman Coulter, cat. no. A63880, https://www.beckmancoulter.com/) using the recommended selection condition for 150-bp insertion size form the vendor's protocol for the NEBNext® Ultra™ DNA Library Prep Kit (NEB.E7370L). The fusion DNA polymerase (ThermoFisher, cat. no. F-530L, https://www.thermofisher.com/) was used to amplify fragments for 16 cycles with PA1.0 and indexed PA primers (Data S1). The fragments were size-selected again with AMPure XP beads to reach a library size range of 200–400 bp. The qualifying libraries for rhizome tip, lateral bud and new shoot tip were sequenced using the Illumina HiSeq 2500 sequencing system to generate 50-nt single-end reads for quantification of polyadenylation sites.

Identification of differential alternative splicing

The raw RNA-seq reads generated from the Illumina sequence platform were filtered using the ht2-filter from the HTQC package (1.92.1) with the default parameters to remove low-quality reads (Yang *et al.*, 2013). Then filtered reads were aligned to the bamboo genome using Tophat (v.2.0.11) with options '-r 50 -a 8′ (Trapnell *et al.*, 2009). The gene expression levels were measured and normalized as FPKM. Mapped reads were assembled with Cufflinks v.2.1.1 using the option -F 0.05 -A 0.01 -I 100000 –min-intron-length 30, to generate a genome-based Cufflinks assembly (Trapnell *et al.*, 2012). The differential alternative splicing events were identified by rMATS.3.2.2 using the option -t paired -len 125 -a 8 -c 0.0001 -analysis U (Shen *et al.*, 2014a). In total, AS events with a false discovery rate (FDR) < 0.05 were regarded as differential AS events.

Identification of differential alternative polyadenlation

The bioinformatics analysis of PAS-seq followed our previous studies (Zhang *et al.*, 2015). Briefly, PAS-seq reads were reverse-complemented and aligned using Tophat (v.2.0.11) with option '-a 8′ to the bamboo genome (Trapnell *et al.*, 2009). The abundance of genuine poly(A) sites based on PacBio FLNC transcripts was calculated according to each PAS-seq library and Fisher's exact test was used for differential APA sites among rhizome tip, lateral bud and new shoot tip ($P < 0.001$).

GO enrichment analysis

The GO terms of bamboo were assigned to each gene based on BLAST2GO (Conesa *et al.*, 2005). GO enrichment analysis was performance using BINGO 3.0.2 with Benjamini–Hochberg multiple testing using $P < 0.05$ as the cutoff (Maere *et al.*, 2005).

Data presentation

All the statistical analyses and plots were carried out using R packages. Heat maps were generated by the gplots R package. All the genome features with PacBio tracks can be viewed at http://www.forestrylab.org/db/PhePacBio/.

Experimental validation of mis-annotated, novel genes, AS and APA

First-strand cDNA from different tissues was synthesized using the SMART cDNA library construction kit manual (Clontech, cat. no. 634901) with 1 μg total RNA and oligo (dT)30 primers. The AS, mis-annotated and novel genes were separately validated by RT-PCR using PrimerSTAR GXL DNA Polymerase (Takara, cat. no. R050Q, http://www.takara-bio.com/) with 5 × diluted cDNA as the template. The PCR products were visualized in 1% agarose gel stained by GelStain (TransGen, cat. no. GS101-1, http:/www.transbionovo.com). Gene expression from the RNA-seq result was validated by quantitative real-time RT-PCR (qPCR) with GoTaq qPCR Master Mix (Promega, cat. no. A6002, http://www.promega.com/). The APA events were validated using a SMARTer RACE 5′/3′ Kit (Clontech, cat. no. 638459) with two-step 3′ RACE methods.

Briefly, 1 μl of cDNA was amplified with gene-specific primer and CDS III primer for 20 cycles. The PCR products were diluted 50 times and 5 μl of diluted products was amplified with gene-specific and a modified CDS III primer without oligo (dT). The RACE products were analyzed as previously described. Primer lists used in the validation are listed in Data S1.

ACCESSION NUMBERS

Illumina HiSeq 2500 data have been submitted to GEO under accession number GSE90517. PacBio SMRT sequencing data have been submitted to the Sequence Read Archive (SRA) of NCBI under accession number SRP093919.

ACKNOWLEDGEMENTS

The authors thank Dr Michael Hamilton for advice on TAPIS. This work was supported by the National Key Research and Development Program of China (grant no. 2016YFD0600106 to LG), the National Natural Science Foundation of China Grant (grant no. 31570674 to LG and 31500258 to LM) and the Natural Science Foundation of Fujian Province of China (2015J01078 to LM).

AUTHOR CONTRIBUTIONS

CL LM and LG conceived and designed the study. LM, TW and DC performed most of the experiments. High-throughput sequencing data were analyzed by HW, YG, HZ and YW. LG analyzed the data as a whole and wrote the manuscript. All authors have read and approved the final version.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Distribution of reads length from three libraries with different insert sizes.

Figure S2. Number of genes with supported PacBio genes.

Figure S3. Hierarchical cluster RNA-seq samples of rhizome tip, new shoot tip and lateral bud based on the fragments per kilobase of exon per million fragments mapped (FPKM) value, which were quantified as described in the Experimental Procedures.

Figure S4. Scatterplot showing the correlation coefficients between RNA-seq and qPCR. Fold changes based on RNA-seq and qPCR were plotted on the *x*-axis and *y*-axis, respectively. The size of each point is proportional to the abundance of gene expression.

Figure S5. Density of fragments per kilobase of exon per million fragments mapped (FPKM) value for genes with different PacBio reads supported. The FPKM values were quantified using mean values from all the samples.

Figure S6. Validation of mis-annotated split genes. Left panel: the results of RT-PCR; right panel: the gene structures. The blue boxes indicate the exons and the black lines represent introns.

Figure S7. Validation of differential splicing events by RT-PCR.

Figure S8. Number of genes with different numbers of poly(A) sites.

Figure S9. Identification of motifs around polyadenylation sites and the distribution of two motifs around the poly(A) site. The AAUAAA motif is located around the poly(A) site. However, the second UGUA motif is located upstream from poly(A) site.

Figure S10. Validation of alternative polyadenylation by rapid amplification of cDNA ends (RACE).

Figure S11. Density of length and fragments per kilobase of exon per million fragments mapped (FPKM) values for genes with different number of poly(A) sites.

Figure S12. Heatmap of splicing factors in bamboo.

Figure S13. Heatmap of poly(A) factors in bamboo.

Data S1. Primers used for validation of mis-annotated genes, novel genes, alternative splicing, alternative polyadenylation, RNA sequencing and construction of polyadenylation site sequencing.

Data S2. The mis-annotated genes with supported PacBio SMRT reads.

Data S3. The coordinate information for all the novel loci.

Data S4. The AS events supported by PacBio single molecular real-time reads.

Data S5. Enriched Gene Ontology terms for the alternatively spliced genes.

Data S6. The differential alternative splicing events among rhizome tip, lateral bud and new shoot tip.

Data S7. All the poly(A) sites identified by the PacBio single molecular real-time reads reads.

Data S8. The alternative polyadenylation sites from the PacBio single molecular real-time reads.

Data S9. The differential alternative polyadenylation sites among rhizome tip, lateral bud and new shoot tip.

REFERENCES

Abdel-Ghany, S.E., Hamilton, M., Jacobi, J.L., Ngam, P., Devitt, N., Schilkey, F., Ben-Hur, A. and Reddy, A.S. (2016) A survey of the sorghum transcriptome using single-molecule long reads. *Nat. Commun.* **7**, 11706.

Au, K.F., Underwood, J.G., Lee, L. and Wong, W.H. (2012) Improving PacBio long read accuracy by short read alignment. *PLoS ONE*, **7**, e46679.

Au, K.F., Sebastiano, V., Afshar, P.T. *et al.* (2013) Characterization of the human ESC transcriptome by hybrid sequencing. *Proc Natl Acad Sci U S A*, **110**, 4821–4830.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W. and Noble, W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, 202–208.

Campbell, M.A., Haas, B.J., Hamilton, J.P., Mount, S.M. and Buell, C.R. (2006) Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis. *BMC Genom.* **7**, 327.

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M. and Robles, M. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.

Deng, X. and Cao, X. (2017) Roles of pre-mRNA splicing and polyadenylation in plant development. *Curr. Opin. Plant Biol.* **35**, 45–53.

Derrien, T., Johnson, R., Bussotti, G. *et al.* (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* **22**, 1775–1789.

Derti, A., Garrett-Engele, P., Macisaac, K.D., Stevens, R.C., Sriram, S., Chen, R., Rohl, C.A., Johnson, J.M. and Babak, T. (2012) A quantitative atlas of polyadenylation in five mammals. *Genome Res.* **22**, 1173–1183.

Di Giammartino, D.C., Nishida, K. and Manley, J.L. (2011) Mechanisms and consequences of alternative polyadenylation. *Mol. Cell*, **43**, 853–866.

Elkon, R., Ugalde, A.P. and Agami, R. (2013) Alternative cleavage and polyadenylation: extent, regulation and function. *Nat. Rev. Genet.* **14**, 496–506.

Embaye, K., Weih, M., Ledin, S. and Christersson, L. (2005) Biomass and nutrient distribution in a highland bamboo forest in southwest Ethiopia: implications for management. *For. Ecol. Manage.* **204**, 159–169.

Foissac, S. and Sammeth, M. (2007) ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets. *Nucleic Acids Res.* **35**, 297–299.

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

Gallart, A.P., Pulido, A.H., de Lagrán, I.A.M., Sanseverino, W. and Cigliano, R.A. (2016) GREENC: a Wiki-based database of plant lncRNAs. *Nucleic Acids Res.* **44**, 1161–1166.

Gao, J., Zhang, Y., Zhang, C., Qi, F., Li, X., Mu, S. and Peng, Z. (2014) Characterization of the floral transcriptome of Moso bamboo (Phyllostachys edulis) at different flowering developmental stages by transcriptome sequencing and RNA-seq analysis. *PLoS ONE*, **9**, e98910.

Ge, W., Zhang, Y., Cheng, Z., Hou, D., Li, X. and Gao, J. (2016) Main regulatory pathways, key genes and microRNAs involved in flower formation and development of moso bamboo (Phyllostachys edulis). *Plant Biotech J*, **15**, 82–96.

He, C.Y., Cui, K., Zhang, J.G., Duan, A.G. and Zeng, Y.F. (2013) Next-generation sequencing-based mRNA and microRNA expression profiling analysis revealed pathways involved in the rapid growth of developing culms in Moso bamboo. *BMC Plant Biol.* **13**, 119.

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

Hornyik, C., Terzi, L.C. and Simpson, G.G. (2010) The spen family protein FPA controls alternative cleavage and polyadenylation of RNA. *Dev. Cell*, **18**, 203–213.

Isagi, Y., Kawahara, T., Kamo, K. and Ito, H. (1997) Net production and carbon cycling in a bamboo Phyllostachys pubescens stand. *Plant Ecol.* **130**, 41–52.

Kalsotra, A. and Cooper, T.A. (2011) Functional consequences of developmentally regulated alternative splicing. *Nat. Rev. Genet.* **12**, 715–729.

Kambach, C., Walke, S., Young, R., Avis, J.M., de la Fortelle, E., Raker, V.A., Lührmann, R., Li, J. and Nagai, K. (1999) Crystal structures of two Sm protein complexes and their implications for the assembly of the spliceosomal snRNPs. *Cell*, **96**, 375–387.

Lei, M., La, H., Lu, K. *et al.* (2014) Arabidopsis EDM2 promotes IBM1 distal polyadenylation and regulates genome DNA methylation patterns. *Proc Natl Acad Sci U S A*, **111**, 527–532.

Levin, J.Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D.A., Friedman, N., Gnirke, A. and Regev, A. (2010) Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat. Methods*, **7**, 709–715.

Li, R., Werger, M., During, H. and Zhong, Z. (1998) Carbon and nutrient dynamics in relation to growth rhythm in the giant bamboo Phyllostachys pubescens. *Plant Soil*, **201**, 113–123.

Li, R., Werger, M., De Kroon, H., During, H. and Zhong, Z. (2000) Interactions between shoot age structure, nutrient availability and physiological integration in the giant bamboo Phyllostachys pubescens. *Plant Biol*, **2**, 437–446.

Li, L., Hu, T., Li, X., Mu, S., Cheng, Z., Ge, W. and Gao, J. (2016) Genome-wide analysis of shoot growth-associated alternative splicing in moso bamboo. *Mol. Genet. Genomics*, **291**, 1695.

Li, Y., Dai, C., Hu, C., Liu, Z. and Kang, C. (2017) Global identification of alternative splicing via comparative analysis of SMRT- and Illumina-based RNA-seq in strawberry. *The Plant J*, **90**, 164–176.

Ma, L., Guo, C. and Li, Q.Q. (2014a) Role of alternative polyadenylation in epigenetic silencing and antisilencing. *Proc Natl Acad Sci U S A*, **111**, 9–10.

Ma, L., Pati, P.K., Liu, M., Li, Q.Q. and Hunt, A.G. (2014b) High throughput characterizations of poly(A) site choice in plants. *Methods*, **67**, 74–83.

Maere, S., Heymans, K. and Kuiper, M. (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, **21**, 3448–3449.

Marquez, Y., Brown, J.W., Simpson, C., Barta, A. and Kalyna, M. (2012) Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. *Genome Res.* **22**, 1184–1195.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M. and Snyder, M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, **320**, 1344–1349.

Nam, D.K., Lee, S., Zhou, G., Cao, X., Wang, C., Clark, T., Chen, J., Rowley, J.D. and Wang, S.M. (2002) Oligo (dT) primer generates a high frequency of truncated cDNAs through internal poly (A) priming during reverse transcription. *Proc Natl Acad Sci U S A*, **99**, 6152–6156.

Nilsen, T.W. and Graveley, B.R. (2010) Expansion of the eukaryotic proteome by alternative splicing. *Nature*, **463**, 457–463.

Ozsolak, F., Platt, A.R., Jones, D.R., Reifenberger, J.G., Sass, L.E., McInerney, P., Thompson, J.F., Bowers, J., Jarosz, M. and Milos, P.M. (2009) Direct RNA sequencing. *Nature*, **461**, 814–818.

Ozsolak, F., Kapranov, P., Foissac, S., Kim, S.W., Fishilevich, E., Monaghan, A.P., John, B. and Milos, P.M. (2010) Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell*, **143**, 1018–1029.

Peng, Z., Zhang, C., Zhang, Y., Hu, T., Mu, S., Li, X. and Gao, J. (2013) Transcriptome sequencing and analysis of the fast growing shoots of moso bamboo (Phyllostachys edulis). *PLoS ONE*, **8**, e78944.

Proudfoot, N.J. (2011) Ending the message: poly(A) signals then and now. *Genes Dev.* **25**, 1770–1782.

Reddy, A.S. (2007) Alternative splicing of pre-messenger RNAs in plants in the genomic era. *Annu. Rev. Plant Biol.* **58**, 267–294.

Reddy, A.S., Marquez, Y., Kalyna, M. and Barta, A. (2013) Complexity of the alternative splicing landscape in plants. *Plant Cell*, **25**, 3657–3683.

Rhoads, A. and Au, K.F. (2015) PacBio sequencing and its applications. *Genomics Proteomics Bioinformatics*, **13**, 278–289.

Sharon, D., Tilgner, H., Grubert, F. and Snyder, M. (2013) A single-molecule long-read survey of the human transcriptome. *Nat. Biotechnol.* **31**, 1009–1014.

Shen, S., Park, J.W., Lu, Z.X., Lin, L., Henry, M.D., Wu, Y.N., Zhou, Q. and Xing, Y. (2014a) rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc Natl Acad Sci U S A*, **111**, 5593–5601.

Shen, Y., Zhou, Z., Wang, Z. *et al.* (2014b) Global dissection of alternative splicing in paleopolyploid soybean. *Plant Cell*, **26**, 996–1008.

Shepard, P.J., Choi, E.A., Lu, J., Flanagan, L.A., Hertel, K.J. and Shi, Y. (2011) Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA*, **17**, 761–772.

Sherstnev, A., Duc, C., Cole, C., Zacharaki, V., Hornyik, C., Ozsolak, F., Milos, P.M., Barton, G.J. and Simpson, G.G. (2012) Direct sequencing of Arabidopsis thaliana RNA reveals patterns of cleavage and polyadenylation. *Nat Struc Mol Biol*, **19**, 845–852.

Song, X., Peng, C., Zhou, G., Gu, H., Li, Q. and Zhang, C. (2016) Dynamic allocation and transfer of non-structural carbohydrates, a possible mechanism for the explosive growth of Moso bamboo (Phyllostachys heterocycla). *Sci. Rep.* **6**, 25908.

Steijger, T., Abril, J.F., Engström, P.G., Kokocinski, F., Hubbard, T.J., Guigó, R., Harrow, J., Bertone, P. and **RGASP Consortium** (2013) Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods*, **10**, 1177–1184.

Szcześniak, M.W., Rosikiewicz, W. and Makałowska, I. (2016) CANTATAdb: a collection of plant long non-coding RNAs. *Plant Cell Physiol.* **57**, e8.

Thatcher, S.R., Zhou, W., Leonard, A., Wang, B.B., Beatty, M., Zastrow-Hayes, G., Zhao, X., Baumgarten, A. and Li, B. (2014) Genome-wide analysis of alternative splicing in Zea mays: landscape and genetic regulation. *Plant Cell*, **26**, 3472–3487.

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

Trapnell, C., Pachter, L. and Salzberg, S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L. and Pachter, L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578.

Wang, X., Duan, C., Tang, K. *et al.* (2013) RNA-binding protein regulates plant DNA methylation by controlling mRNA processing at the intronic heterochromatin-containing gene IBM1. *Proc Natl Acad Sci U S A*, **110**, 15467–15472.

Wang, B., Tseng, E., Regulski, M., Clark, T.A., Hon, T., Jiao, Y., Lu, Z., Olson, A., Stein, J.C. and Ware, D. (2016) Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nat. Commun.* **7**, 11708.

Wu, X., Liu, M., Downie, B., Liang, C., Ji, G., Li, Q.Q. and Hunt, A.G. (2011) Genome-wide landscape of polyadenylation in Arabidopsis provides

evidence for extensive alternative polyadenylation. *Proc Natl Acad Sci U S A*, **108**, 12533–12538.

Wu, B., Suo, F., Lei, W. and Gu, L. (2014) Comprehensive Analysis of Alternative Splicing in Digitalis purpurea by Strand-Specific RNA-Seq. *PLoS ONE*, **9**, e106001.

Wu, T.D., Reeder, J., Lawrence, M., Becker, G. and Brauer, M.J. (2016) GMAP and GSNAP for genomic sequence alignment: enhancements to speed, accuracy, and functionality. *Methods Mol. Biol.* **1418**, 283–334.

Xing, D. and Li, Q.Q. (2011) Alternative polyadenylation and gene expression regulation in plants. *Wiley Interdisciplinary Reviews: RNA*, **2**, 445–458.

Xu, Z., Peters, R.J., Weirather, J. *et al.* (2015) Full-length transcriptome sequences and splice variants obtained by a combination of sequencing platforms applied to different root tissues of Salvia miltiorrhiza and tanshinone biosynthesis. *The Plant J*, **82**, 951–961.

Yang, X., Liu, D., Liu, F., Wu, J., Zou, J., Xiao, X., Zhao, F. and Zhu, B. (2013) HTQC: a fast quality control toolkit for Illumina sequencing data. *BMC Bioinformatics*, **14**, 33.

Zachariah, E.J., Sabulal, B., Nair, D.N., Johnson, A.J. and Kumar, C.S. (2016) Carbon dioxide emission from bamboo culms. *Plant Biol.* **18**, 400–405.

Zhang, G., Guo, G., Hu, X. *et al.* (2010) Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome Res.* **20**, 646–654.

Zhang, Y., Gu, L., Hou, Y. *et al.* (2015) Integrative genome-wide analysis reveals HLP1, a novel RNA-binding protein, regulates plant flowering by targeting alternative polyadenylation. *Cell Res.* **25**, 864–876.

Zhang, H., Lin, C. and Gu, L. (2017) Light regulation of alternative Pre-mRNA splicing in plants. *Photochem. Photobiol.* **93**, 159–165.

Zhou, B., Fu, M., Xie, J., Yang, X. and Li, Z. (2005) Ecological functions of bamboo forest: research and application. *J. For. Res.* **16**, 143–147.