nature plants

LETTERS

https://doi.org/10.1038/s41477-020-0688-1

Check for updates

Post-transcriptional splicing of nascent RNA contributes to widespread intron retention in plants

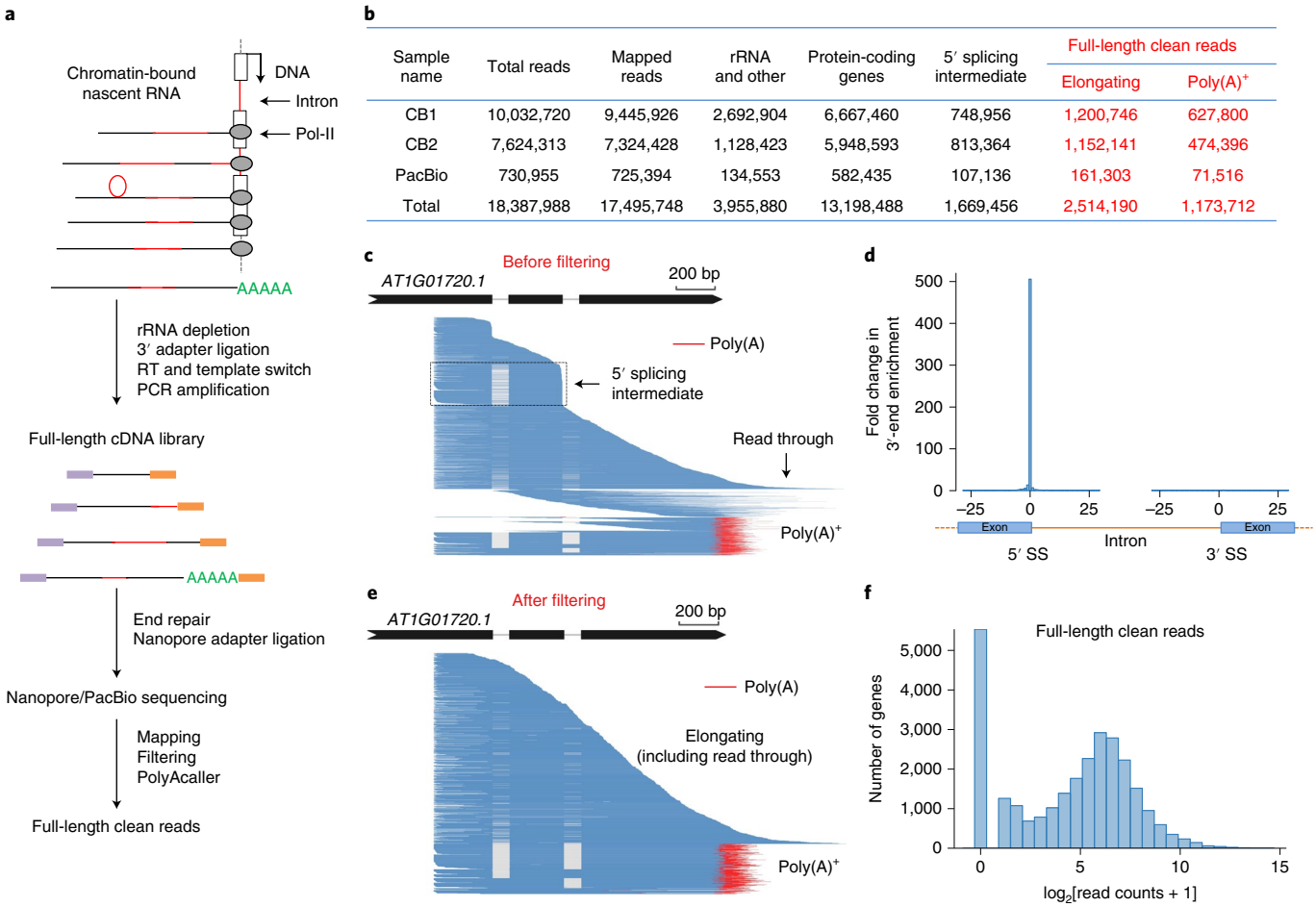
Jinbu Jia[1,2,7], Yanping Long[1,3,7], Hong Zhang[1], Zhuowen Li[1], Zhijian Liu[1], Yan Zhao[1], Dongdong Lu[1], Xianhao Jin[1], Xian Deng[4], Rui Xia[2], Xiaofeng Cao[4,5,6] and Jixian Zhai[1]✉

In eukaryotes, genes are transcribed by RNA polymerase-II (Pol-II) and introns are removed by the spliceosome largely cotranscriptionally[1–3]; analysis using long-read sequencing revealed that splicing occurs immediately after Pol-II passes introns in yeast[4,5]. Here, we developed a Nanopore-based method to profile chromatin-bound RNA that enables the simultaneous detection of splicing status, Pol-II position and polyadenylation at the genome-wide scale in *Arabidopsis*. We found that more than half of the introns remain unspliced after Pol-II transcribes 1 kb past the 3′ splice site, which is much slower than the rate of splicing reported in yeast[4,5]. Many of the full-length chromatin-bound RNA molecules are polyadenylated, yet still contain unspliced introns at specific positions. These introns are nearly absent in the cytoplasm and are resistant to nonsense-mediated decay, suggesting that they are post-transcriptionally spliced before the transcripts are released into the cytoplasm; we therefore termed these introns post-transcriptionally spliced introns (pts introns). Analysis of around 6,500 public RNA-sequencing libraries found that the splicing of pts introns requires the function of splicing-related proteins such as PRMT5 and SKIP, and is also influenced by various environmental signals. The majority of the intron retention events in *Arabidopsis* are at pts introns, suggesting that chromatin-tethered post-transcriptional splicing is a major contributor to the widespread intron retention that is observed in plants, and could be a mechanism to produce fully spliced functional mRNAs for rapid response.

Pre-mRNA splicing is a fundamental process in eukaryotic mRNA maturation, and the majority of introns are spliced cotranscriptionally[4,6,7]. Previous studies using synthesized pre-mRNA to reconstruct splicing events in vitro have yielded a wealth of knowledge on the mechanisms of splice-site recognition and spliceosome assembly[8,9]. Applications of high-throughput short-read sequencing technology in characterizing nascent RNA have greatly facilitated the global quantification of splicing events and accurate tracking of active Pol-II transcription in animals[6,10–14] and plants[15–19]. However, studying the coordination between splicing and transcription remains challenging owing to difficulties in simultaneously characterizing splicing events and transcriptional progression on the same transcript. Recent research in yeast applied long-read sequencing technology to address this challenge and found that introns are rapidly spliced after intron exit from Pol-II[4,5], but this remains to be investigated in multicellular organisms, which often have more complex patterns of splicing. In addition to cotranscriptional splicing, abundant post-transcriptional splicing was also reported in animal cells; intron-containing, polyadenylated transcripts are retained at the chromatin[20,21]. In plants, the splicing kinetics, the link between splicing and transcription, and the relationship between splicing and polyadenylation (PA) remain largely unknown.

To examine the full spectrum of nascent RNA molecules in *Arabidopsis*, we developed a method to profile both the elongating and the polyadenylated fractions using full-length sequencing technology (Fig. 1a). Two biological replicates of chromatin-bound RNAs of *Arabidopsis* seedlings (aged 12 d) were extracted and converted into double-stranded cDNA for long-read Nanopore sequencing by ligating a 3′ linker and using template-switch reverse transcription PCR (Fig. 1a, Supplementary Fig. 1). Nanopore MinION sequencing yielded 10 million and 7.6 million raw reads for the first and second biological replicate, respectively, and ~95% of the reads were mapped to the *Arabidopsis* genome (Fig. 1b, Supplementary Fig. 2). A quick glance at the reads that mapped to the genic region showed that our libraries were of high-quality—the majority of long reads started at the 5′ transcriptional start site and ended along the course of transcription; read-through transcripts moved past the polyadenylated site, as well as a large number of polyadenylated full-length, yet incompletely spliced, transcripts (Fig. 1c). We also observed very strong enrichment of RNA 3′ ends exactly at the 5′ splice site (Fig. 1d). This is most likely due to the capturing of splicing intermediates produced by the first transesterification reaction during the splicing reaction, and has also been reported in other recent studies using direct RNA sequencing (RNA-seq) to profile nascent RNAs from humans and *Drosophila*[22]. The precise 3′ end peak at the last nucleotide of the exon also suggests that there was little RNA degradation during the process of library construction (Fig. 1d). Furthermore, we also detected thousands of reads of entire spliced introns (Supplementary Fig. 3c). To obtain an accurate estimation of the cotranscriptional splicing kinetics, the splicing intermediates, as well as reads missing 5′ ends (probably due to incomplete reverse transcription), were removed from further analysis (Fig. 1e). After filtering (Supplementary Fig. 2), we obtained a total of ~3.5 million full-length clean reads covering 23,844 protein-coding genes, with a median of ~50 transcripts per gene (Fig. 1b,f). We also developed a data analysis pipeline to distinguish between elongating RNAs (without a polyadenylated tail) and polyadenylated RNAs (with a polyadenylated tail), and estimated the length of the polyadenylated tail (a method that we named PolyAcaller; see Methods;

[1]Institute of Plant and Food Science, Department of Biology, Southern University of Science and Technology, Shenzhen, China. [2]College of Horticulture, South China Agricultural University, Guangzhou, China. [3]Institute for Advanced Studies and College of Life Science, Wuhan University, Wuhan, China. [4]State Key Laboratory of Plant Genomics and National Center for Plant Gene Research, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China. [5]University of Chinese Academy of Sciences, Beijing, China. [6]Center for Excellence in Molecular Plant Sciences, Chinese Academy of Sciences, Beijing, China. [7]These authors contributed equally: Jinbu Jia, Yanping Long. ✉e-mail: zhaijx@sustech.edu.cn



Sample name	Total reads	Mapped reads	rRNA and other	Protein-coding genes	5′ splicing intermediate	Full-length clean reads	
						Elongating	Poly(A)+
CB1	10,032,720	9,445,926	2,692,904	6,667,460	748,956	1,200,746	627,800
CB2	7,624,313	7,324,428	1,128,423	5,948,593	813,364	1,152,141	474,396
PacBio	730,955	725,394	134,553	582,435	107,136	161,303	71,516
Total	18,387,988	17,495,748	3,955,880	13,198,488	1,669,456	2,514,190	1,173,712

Fig. 1 | Full-length nascent RNA-seq. a, Schematic of full-length nascent RNA-seq. **b**, The basic statistics of reads sequenced by Nanopore and PacBio. CB1 and CB2 are two biological replicates that were sequenced by Nanopore Technologies. **c**, An example of Nanopore reads aligned to the *AT1G01720* gene before filtering. **d**, The fold change in 3′-end enrichment around the 5′ splice site (5′ SS) and 3′ splice site (3′ SS) of introns. The position 0 indicates the final base of the exon or intron. **e**, An example of Nanopore reads aligned to the *AT1G01720* gene after filtering. **f**, The distribution of Nanopore full-length clean read counts per gene (only protein-coding genes).

Supplementary Figs. 2 and 4). Seventy percent of the full-length clean reads were elongating Pol-II products, whereas the remaining 30% were already polyadenylated (Fig. 1b). In contrast to other recent long-read nascent RNA-seq techniques that focus on the elongating RNA fraction[4,5], our method can be used to simultaneously investigate splicing, transcriptional elongation and PA.

As previously shown in other systems[4,5,22], capturing the full-length elongating transcripts enables us to study the coordination between splicing and elongation. Introns in *Arabidopsis* are relatively short compared with introns in *Drosophila* and humans (Supplementary Fig. 5), and it will be interesting to determine whether *Arabidopsis* has different splicing kinetics. Our libraries captured ~2.3 million full-length elongating transcripts, of which approximately 20% have at least one spliced intron (Fig. 2a). More than 50% of the introns remain unspliced when Pol-II already transcribed more than 1,100 nucleotides past the 3′ splice site (Fig. 2b). To further confirm our results from the analysis of Nanopore data, we sequenced a third biological replicate of the long-read chromatin-bound RNA library using PacBio, another popular long-read sequencing platform (Fig. 1a,b). PacBio utilizes a fundamentally different technology compared with the technology used by Nanopore, and the downstream signal processing and data analysis are also quite different between these two platforms (see Methods). The results from PacBio and Nanopore data are highly consistent, confirming that our method is robust across different platforms and analysis pipelines (Fig. 2b). The half-saturation value of the splicing in *Arabidopsis* contrasts sharply with the much faster splicing speed that was reported in yeast (~45 nucleotides)[4], but is similar to that of *Drosophila* and humans[22], demonstrating a big difference in the splicing kinetics between yeast and the three higher eukaryotes examined to date.

As our full-length long-read sequencing method can simultaneously track the status of multiple introns on the same transcript, we next examined the splicing order of multiple introns. For any two adjacent introns, we found that, in around 70% of cases, the upstream intron was spliced first (Fig. 2c). To track the splicing dynamics of multiple introns, we checked all of the genes containing five introns as an example, and found that the spliced ratios of upstream introns are higher than those of downstream introns; a considerable proportion of the upstream introns are spliced before Pol-II transcribed past the downstream introns (Fig. 2d). We also observed that certain introns on specific transcripts remain unspliced after the downstream introns have been spliced or the transcripts have been polyadenylated. These results showed that the splicing of multi-introns largely follows the order of transcription, but some introns are retained even after the transcripts are fully polyadenylated (Fig. 2d).

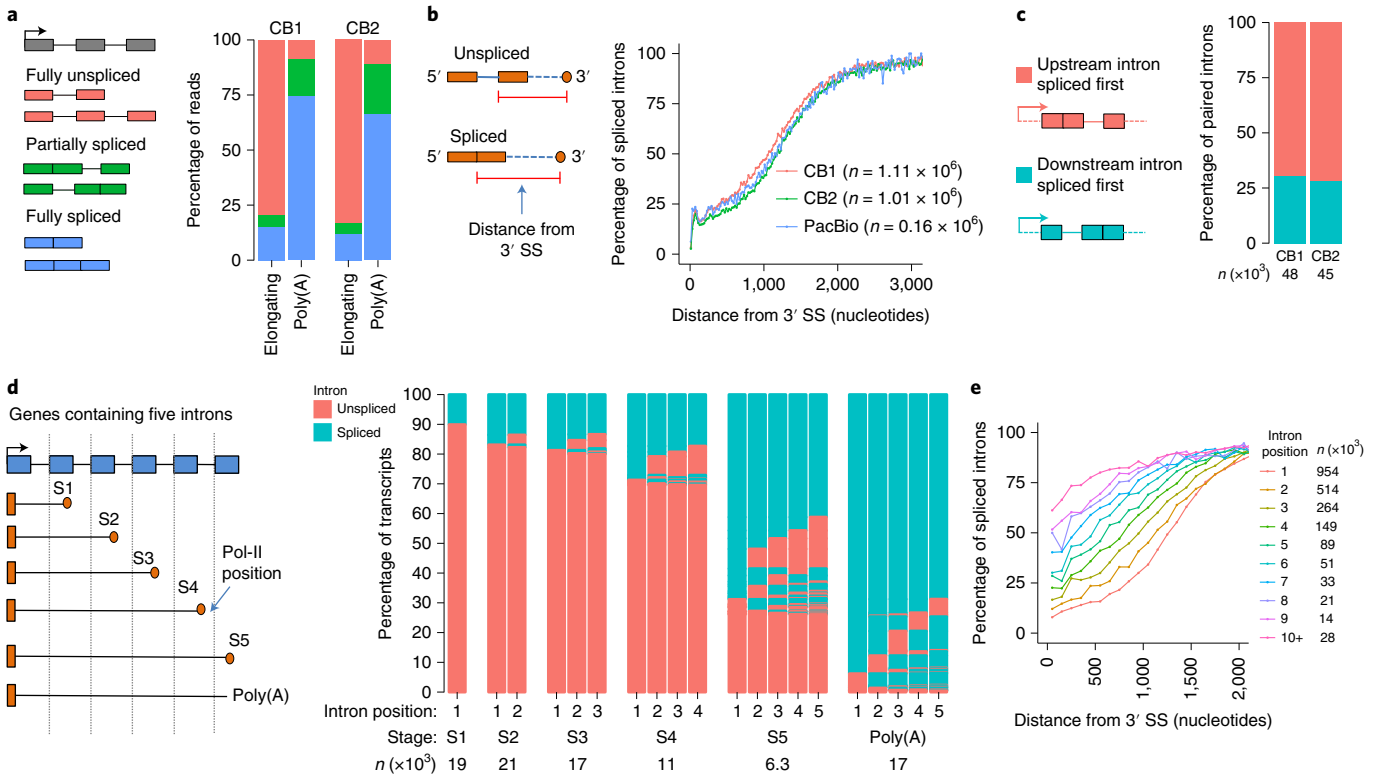


Fig. 2 | Splicing can occur cotranscriptionally in *Arabidopsis* and is coupled in multi-intron genes. a, All unspliced, partially spliced and fully spliced fractions in the elongating and polyadenylated transcripts on chromatin. **b**, Global analysis of the ratio of spliced introns versus the transcription distance from the 3′ splice site. **c**, The frequency of different splicing orders in adjacent intron pairs with different splicing states. **d**, Visualization of the splicing status of multi-introns in each nascent transcript derived from genes with five introns. The elongating transcripts were separated into five groups (S1, S2, …, S5) on the basis of the number of introns that were transcribed. Transcripts of each group were ordered by the splicing status of the introns. In the S2 group, for example, the first two introns had been transcribed; there were therefore four categories of splicing status as follows: both spliced (13.6%), the first spliced and the second unspliced (3.3%), the first unspliced and the second spliced (0.8%) and both unspliced (82.3%). These four types of transcripts were stacked on top of each other in the figure. **e**, Global analysis of the ratio of spliced introns versus the transcription distance of introns in different positions; '1' indicates the first introns, '2' indicates the second introns and so on; the read numbers (*n*) are indicated.

We also observed a strong association among the splicing status of introns on the same transcript. For example, the splicing status of the first and second introns is closely linked (Fig. 2d). The cooperative splicing of neighbouring introns was also observed in yeast, *Drosophila* and humans[5,22]. This association is probably due to previous splicing events recruiting necessary protein complexes and therefore facilitating the splicing of other introns on the same transcript. As a result, the speed of splicing of the downstream intron was generally faster than that of the upstream intron (Fig. 2e). For example, more than 50% of the tenth introns were spliced within 100 nucleotides, whereas only 10% of the first introns were spliced within 100 nucleotides. These results show that early splicing events could help to create a more favourable environment for additional splicing of the same transcript, therefore enabling efficient splicing of multiple-intron genes.

Unexpectedly, we also detected a high percentage of polyadenylated transcripts on chromatin, and ~30% of these transcripts remained incompletely spliced (Figs. 1b, 2a and Fig. 3a). Both the ratio of incompletely splicing transcripts and the ratio of unspliced introns are highly consistent between the two Nanopore libraries (r=0.83 for genes and r=0.85 for introns; Supplementary Figs. 6 and 7). The splicing results are also highly consistent between the PacBio library and the two Nanopore libraries (r=0.87 and r=0.92 for genes, and r=0.88 and r=0.91 for introns; Supplementary Figs. 6 and 7). We also constructed Illumina RNA-seq libraries of chromatin-bound polyadenylated RNA with three biological replicates (Supplementary Table 1). The ratios of unspliced introns are also highly consistent between Illumina RNA-seq and Nanopore data (r=0.91), and are also consistent across three Illumina RNA-seq biological replicates (r=0.94, r=0.95 and r=0.96; Fig. 3b and Supplementary Fig. 7). Furthermore, most of the incompletely spliced polyadenylated transcripts contained only one or a few unspliced introns, whereas the other introns had been spliced efficiently (Figs. 2d, 3a and 3c), indicating the specificity of unspliced introns. To further analyse the differences between introns with high and low unspliced ratios in chromatin-bound polyadenylated RNA, we separated introns into two groups using a cut-off value of 0.1. For introns in the group with an unspliced ratio of chromatin-bound polyadenylated RNA of higher than 0.1, the cotranscriptional splicing rates were much slower (Fig. 3d), consistent with the result showing that these introns remain unspliced after PA.

We next examined whether these incompletely spliced transcripts can be released from the chromatin. We constructed next-generation sequencing (NGS) RNA-seq libraries using polyadenylated RNAs from different cell fractions, and found that the introns with an unspliced ratio of higher than 0.1 in chromatin-bound polyadenylated RNA were barely detectable in the nucleoplasm and cytoplasm (Fig. 3e), suggesting that the intron-containing transcripts were still bound to chromatin after PA. Moreover, the RNA accumulation level (fragments per kilobase of transcript per million mapped reads (FPKM) value) of genes with a higher incompletely spliced ratio in chromatin-bound polyadenylated RNA was notable higher

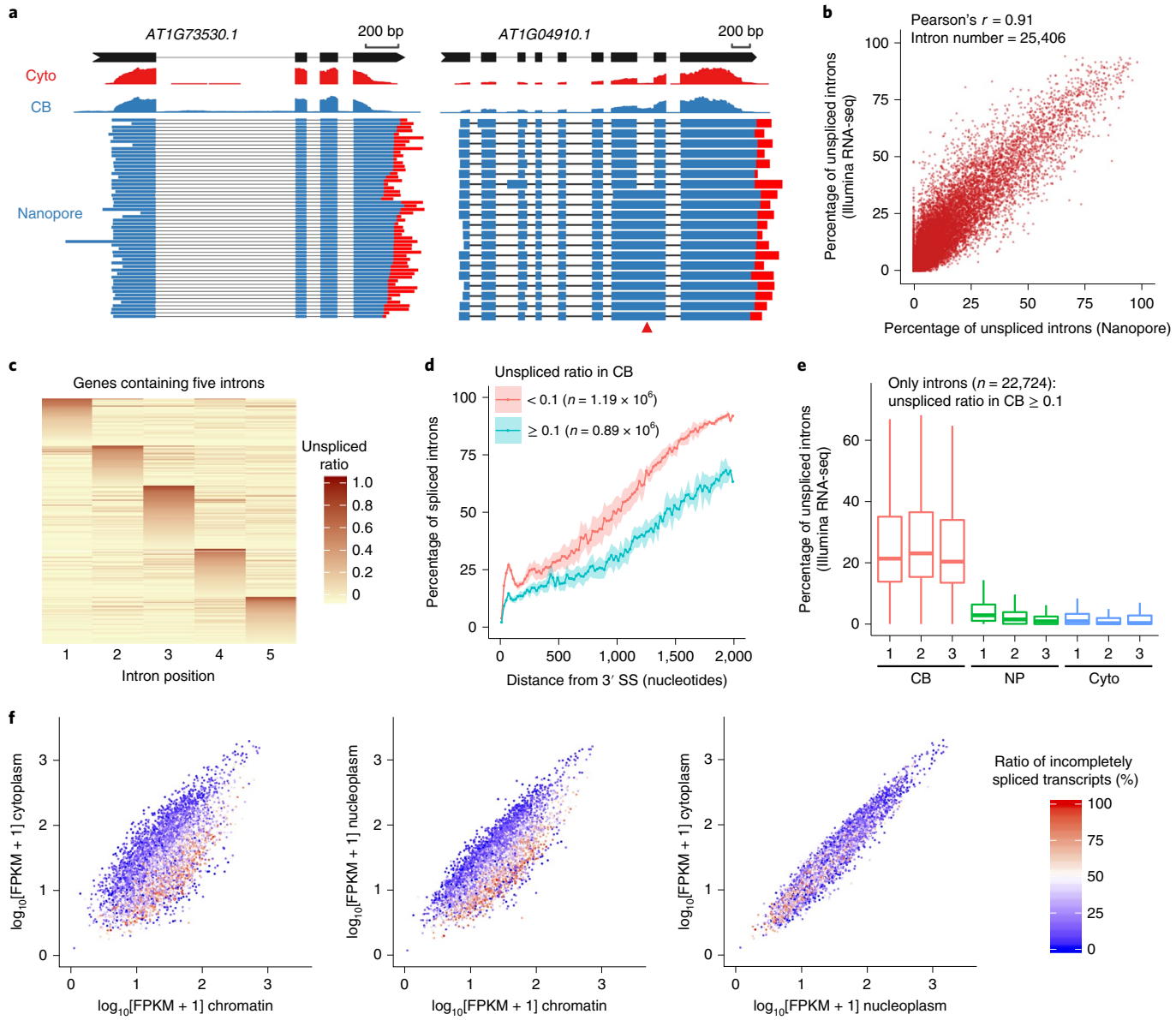


Fig. 3 | High accumulation of polyadenylated transcripts with unspliced introns at specific positions on chromatin. a, Examples of genes with fully (left) and partially (right) spliced introns in chromatin-bound polyadenylated transcripts (CB). **b**, Correction of the ratios of unspliced introns between Illumina RNA-seq and Nanopore. The ratio of unspliced introns from RNA-seq data was estimated on the basis of the percentage of intron retention (PIR) value (see Methods). **c**, Heat-map visualization of the ratio of unspliced introns of genes with five introns. Each row represents a gene. **d**, Global analysis of the spliced ratio versus the transcription distance of introns with different unspliced ratios in chromatin-bound polyadenylated RNA. The shaded regions represent the s.d. across biological replicates. The points and lines represent the mean values of two biological replicates. **e**, The distribution of ratios of unspliced introns of polyadenylated transcripts in chromatin (CB), nucleoplasm (NP) and cytoplasm (Cyto), with three biological replicates per sample. For the box plots, the centre lines show the median, the box limits show the interquartile range and the whiskers extend to the furthest value within 1.5× the interquartile range from the quartiles. **f**, The difference in transcript abundance between chromatin, nucleoplasm and cytoplasm. Each point represents one gene.

in the chromatin fraction than in the nucleoplasm or cytoplasm, strongly suggesting that the incompletely spliced transcripts were retained on the chromatin (Fig. 3f). Furthermore, if these intron-containing transcripts were to be released from chromatin, they would probably be targeted by the nonsense-mediated decay (NMD) pathway, but we found that the accumulation level of most of these unspliced introns remained the same in the *up frameshift 1* (*upf1*) *upf3* double mutant, which disrupts the NMD pathway (Fig. 4b), suggesting that most of these incompletely spliced transcripts are not degraded by NMD, consistent with previous reports that most transcripts with intron retention (IR) are resistant to NMD in plants[23]. Thus, our results suggest that chromatin retention of incompletely spliced transcripts represents an important step in mRNA maturation and has a critical role in post-transcriptional gene regulation. Misannotation of introns are frequent and would appear as IR events when analysing the RNA-seq data. To obtain introns that are being actively spliced, we followed a recent study that characterized splicing kinetics in *Drosophila* and humans, and we used only constitutively spliced introns for our splicing analysis (see Methods)[22]. To simplify the analysis, we hereafter define pts introns as constitutively spliced introns with an unspliced ratio of at least 0.1 in the chromatin-bound polyadenylated RNA fraction.

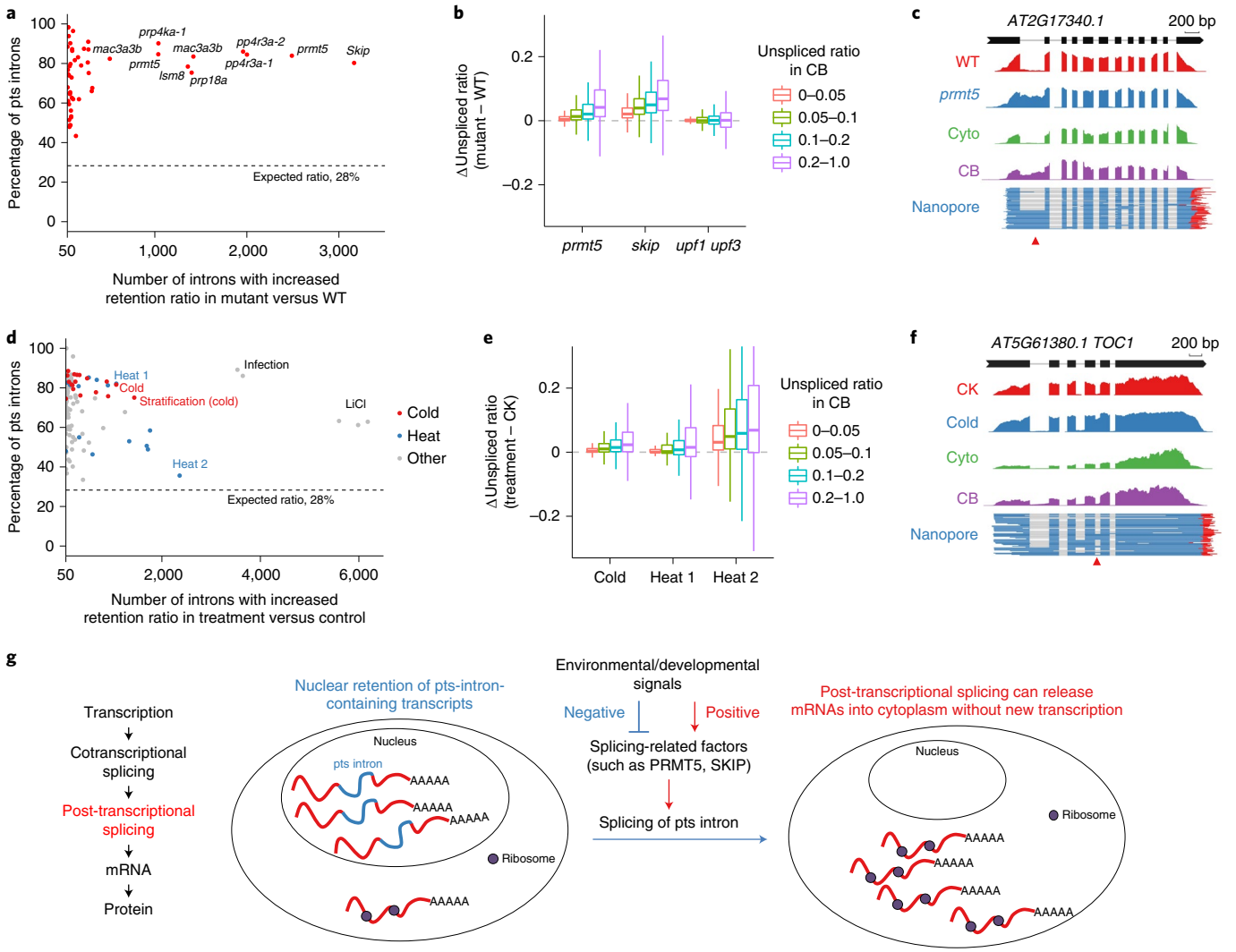


Fig. 4 | The splicing of pts introns was regulated by splicing-related factors and various environmental signals. a, The pts intron (unspliced ratio in chromatin-bound polyadenylated RNA ≥ 0.1) ratio of introns with an increased retention ratio in the mutant compared with the WT identified in public RNA libraries. Only mutants with more than 50 upregulated IRs are shown. **b**, The Δunspliced ratios (mutant – WT) distribution of introns with different unspliced ratios in chromatin-bound polyadenylated RNA. Only introns with total junction read numbers detected in both mutant and WT samples of more than 10 were used. The numbers of four groups of introns (labelled as 0–0.05, 0.05–0.1, 0.1–0.2 and 0.2–1) were 42,826, 11,686, 9,724 and 10,618, respectively. For the box plots, the centre lines show the median, the box limits show the interquartile range and the whiskers extend to the furthest value within 1.5× the interquartile range from the quartiles. **c**, An example of an intron that is affected in *prmt5*. WT: Sequence Read Archive (SRA), ERX1663062; *prmt5*: SRA, ERX1663065. **d**, The pts intron ratio of introns with an increased retention ratio in treatment compared with the mock/control identified in public RNA libraries. Only treatments with more than 50 upregulated IRs are shown. **e**, The Δunspliced ratios (treatment – control/mock) distribution of introns with different unspliced ratios in chromatin-bound polyadenylated RNA. Only introns with total junction read numbers detected in both treatment and WT samples of more than 10 were used. The numbers of four groups of introns (labelled as 0–0.05, 0.05–0.1, 0.1–0.2 and 0.2–1) are 34,317, 9,215, 7,447 and 7,907, respectively. For the box plots, the centre lines show the median, the box limits show the interquartile range and the whiskers extend to the furthest value within 1.5× the interquartile range from the quartiles. **f**, Example of cold-regulated intron in the clock gene *TOC1*. CK: SRA, SRX3203983; cold: SRA, SRX3203987. **g**, A model for nuclear retention and post-transcriptional splicing as an important layer of post-transcriptional regulation.

An alternative explanation for the observed chromatin-bound intron-containing, yet polyadenylated, transcripts is that PA of these transcripts is still in process, and splicing will be completed when the PA step is finished, and what we observed is a reflection of the delay in this process. Given the technical difficulties in measuring the completion of the PA step, these alternative explanations remain to the resolved until more data on the dynamics of PA process are available in the future.

To determine whether the splicing of pts introns is affected in certain mutants or regulated by environment signals, we selected 6,521 high-quality, publicly available RNA-seq libraries that account for 1,512 matched groups with at least two biological replicates of each sample, including 713 groups of mutant versus wild type (WT) and 799 groups of treatment versus mock/CK, to search for differentially retained introns at genome-wide scale (Supplementary Tables 2 and 3). From the mutant versus WT analysis, we identified ten mutant samples with an increased retention ratio (that is, increased unspliced ratio) in more than 500 introns compared with their matched WT controls (see Methods). The corresponding genes for these mutants were all previously reported to be involved in splicing, such as *PROTEIN ARGININE METHYLTRANSFERASE 5* (*PRMT5*) and *SKI-INTERACTING PROTEIN* (*SKIP*)[24–27] (Fig. 4a,

Supplementary Table 2). Although pts introns account for only 28% of all introns, ~80% of enhanced IR events in these ten groups were found at pts introns (Fig. 4a). The *prmt5* and *skip* mutants strongly suppress the splicing of pts introns while having small impacts on the non-pts introns (Fig. 4b). Moreover, the cotranscriptional splicing rates of introns affected in either the *prmt5* or the *skip* mutants are also slower than those that are unaffected by the mutations (Supplementary Fig. 8a,b). PRMT5 is implicated in various developmental processes, such as flowering time control, stress response and circadian rhythm, by promoting the recruitment of NineTeen Complex to the spliceosome and modulating pre-mRNA splicing of diverse genes[24,25,27]. We found that most of the signature IRs that were previously identified in *prmt5* (refs. [24,25,27]) were at pts introns, such as the first intron in *AT2G17340* (Fig. 4c). These results showed that the splicing of pts introns was susceptible to regulation by PRMT5, SKIP and other splicing-related proteins, whereas splicing of the normal introns was less dependent on these factors. Furthermore, these two groups of introns, which were impacted in *prmt5* and *skip* mutants, were largely distinct from each other (Supplementary Fig. 8c,d), suggesting that pts introns can be further classified into subgroups of which the splicing is regulated by distinct pathways.

Moreover, most of the introns that showed an increase in retention ratio under various stress treatments compared with the controls were also pts introns (Fig. 4d, Supplementary Table 3), suggesting that pts introns provide an important basis for observed IR events in plants. Several reports have shown that rapid induction of expression could lead to an increased proportion of nascent RNA compared with mature mRNA[28,29] and, therefore, a higher fraction of intron-derived reads. We examined whether the increased IR that we observed was correlated with upregulation of gene expression, and found that increased IRs were not enriched in genes that showed upregulated expression levels, suggesting that the increase in IRs that we observed was not due to the potential rapid production of nascent RNA during transcriptional upregulation (Supplementary Fig. 9). Thus, the splicing of pts introns can be regulated by various environmental signals. For example, the splicing of a large number of pts introns was inhibited after cold and heat treatment, which is suggestive of temperature-regulated splicing (Fig. 4d,e). Furthermore, temperature-dependent IR has an important role in regulating the expression of circadian clock genes during temperature transitions[30], and these retained introns on clock genes, such as *TIMING OF CAB EXPRESSION 1 (TOC1)*, are also mostly pts introns (Fig. 4f). The dynamic regulation on the splicing of pts introns may therefore have an important role in plant responses to environmental challenges.

Two recent studies used Illumina sequencing to profile chromatin-bound nascent RNAs and showed that cotranscriptional splicing is widespread in *Arabidopsis*[16,17], which fits well with our observation using long-read sequencing methods. In contrast to the Illumina-based method, one unique advantage of our approach is that it could simultaneously track the status of splicing and the progress of transcriptional elongation, which enabled us to directly calculate the splicing kinetics in *Arabidopsis*. Furthermore, our method can also detect polyadenylated-tail length and the splicing order of multiple introns, which cannot be observed using the regular RNA-seq method. Our method also captured a large number of splicing intermediates, which can be used to estimate the splicing reaction rates[31,32], and our data show that pts introns have lower splicing efficiency compared with the non-pts introns (Supplementary Fig. 10). Moreover, the intron-containing, yet polyadenylated, transcripts in our library are well positioned to examine the relationship between splicing and PA site usage. We found that, for those genes with increased IR ratio under cold or heat stress, only a small fraction undergo alternative PA (APA), and the ratio is similar to other genes without increased IR (Supplementary Fig. 11a). For genes with multiple PA sites, we found that the spliced ratios of their introns are similar between the major PA isoform and the minor PA isoform ($r = 0.91$; Supplementary Fig. 11b). There are a handful of introns with a different spliced ratio between the major and minor PA isoforms (Supplementary Fig. 11c), suggesting that PA site usage could potentially influence splicing, or vice versa. In the future, it will be interesting to further investigate the causal relationship between these two processes in regulating these genes.

Our data highlight that chromatin-tethered post-transcriptional splicing is an important stage in the maturation of a large number of mRNAs. We speculate that this stage can be regulated by various splicing factors and can act as an important layer in gene regulation in response to a wide range of environment signals. Future profiling of the chromatin-bound RNAs, as well as cytoplasmic RNAs, in mutants that influence the splicing of pts introns, such as *prmt5* and *skip*, would be critical in validating this concept (Fig. 4g). Splicing-based, post-transcriptional regulation can have some key advantages as a regulatory step during mRNA maturation. The nuclear-retained, polyadenylated, incompletely spliced transcripts are only one step away from being fully functional by splicing of the pts intron, and these transcripts can be stored in the nucleus, safe from being degraded by the NMD pathway in the cytoplasm, until splicing is activated by the environmental or developmental triggers. Many previous studies in various systems support this model. In regard to whether intron-containing, yet fully polyadenylated, RNAs can be stored in the nucleus and released after environmental or developmental signals, it was previously demonstrated that a large amount of intron-containing, polyadenylated pre-mRNA is stored in the microspore of *Marsilea vestita* (named intron-retaining transcripts (IRTs)), and the maturation of IRTs to mRNA required the function of the spliceosome machinery, but not the transcription machinery; these introns are removed from IRTs at specific timing during development, enabling translation to proceed[33]. The chromatin-tethered polyadenylated, yet incompletely spliced, transcripts were also reported in various cell types and tissues in animals, including mouse embryonic stem cells[21], mouse inflamed macrophages[20], meiotic spermatocytes[34] and neocortex[35]. These intron-containing RNAs can be further spliced in response to a stimulus. For example, a group of intron-containing polyadenylated transcripts retained in the nucleus of the mouse neocortex is rapidly spliced within minutes in response to neuronal stimulation[35]. There are also studies that support that incompletely spliced transcripts were tethered in the nucleus. In human cells, a study combining fluorescence in situ hybridization and fluorescence recovery after photobleaching showed that slower splicing leads to an increase in retention of the full-length polyadenylated pre-mRNA at the transcription site[36]. In plants, there is direct evidence, obtained using imaging of living cells, showing that intron-containing transcripts from two genes, *ARGININE/SERINE-RICH ZINC KNUCKLE-CONTAINING PROTEIN* (*RS2Z33*) and *SERRATED LEAVES AND EARLY FLOWERING* (*SEF*), are retained in the nucleus so that they are not subject to NMD before being fully spliced and exported to the cytoplasm[37]. Nuclear retention and post-transcriptional splicing may therefore serve as a conserved layer for post-transcriptional regulation in both plants and animals.

Methods

Extraction of chromatin-bound RNA. Chromatin-bound RNAs were extracted using a previously reported method[18,38]. *Arabidopsis* seedlings (Col-0) were grown on 1/2 MS plates at 22 °C (16 h light–8 h dark) for 12 d before collection. Seedlings (2 g) were ground with liquid nitrogen into fine powder and transferred into an ice-cold 50 ml tube with 10 ml Honda buffer (0.44 M sucrose, 1.25% (w/v) Ficoll, 2.5% (w/v) dextran T40, 20 mM HEPES-KOH pH 7.4, 10 mM $MgCl_2$, 0.5% (w/v) Triton X-100, 1 mM dithiothreitol (DTT), 1× protease inhibitor (Roche) and 100 ng μl^{-1} tRNA). After homogenization by vortexing, the homogenate was filtered through one layer of Miracloth. Another 10 ml of Honda buffer was added to the filter to wash the remaining plant material. After centrifugation at 4 °C and 3,500*g*

for 5 min, the supernatant was collected as the cytoplasmic fraction and the pellet was resuspended and washed once with another 20 ml Honda buffer. The pellet was then resuspended in 1 ml Honda buffer (0.44 M sucrose, 1.25% (w/v) Ficoll, 2.5% (w/v) dextran T40, 20 mM HEPES-KOH pH 7.4, 10 mM $MgCl_2$, 0.5% Triton X-100, 1 mM DTT, 500 ng µl^{-1} tRNA, 2× proteinase inhibitor (Roche) and 10 µl RNase inhibitor (RNase out, Thermo Fisher Scientific)) and centrifuged at 4 °C and 8,000*g* for 1 min. The supernatant was completely removed and the nuclei pellet was weighed. One volume of nuclei resuspension buffer (50% glycerol, 0.5 mM EDTA, 1 mM DTT, 25 mM Tris-HCl pH 7.5, 100 mM NaCl, 1× RNase Out and 200 ng µl^{-1} tRNA) was added and the pellet was stirred to mix with a pipette tip. Two volumes of washing buffer (25 mM Tris-HCl pH 7.5, 300 mM NaCl, 1 M urea, 0.5 mM EDTA, 1 mM DTT, 1% Tween-20, RNase Out and 200 ng µl^{-1} tRNA) were then added, and the pellet was washed by pipetting up and down 30 times and centrifuged at 4 °C and 8,000*g* for 1 min. The supernatant was collected as the nucleoplasmic fraction. For the second wash, the pellet was resuspended in one volume of resuspension buffer, and washed by pipetting up and down seven times with one volume of washing buffer, and was centrifuged at 4 °C and 8,000*g* for 1 min. The supernatant was removed, and the pellet was retained as the chromatin fraction.

For RNA extraction, the pellet was resuspended in 1 ml TRIzol. The cytoplasmic and nucleoplasmic fractions were mixed with three volumes of TRIzol LS, vortexed to mix thoroughly and kept at room temperature for 10 min. Chloroform (0.2 ml) was then added, vortexed for 10 s and kept at room temperature for 5 min. The mixture was centrifuged at 4 °C and 14,000 r.p.m. for 10 min. The supernatant was transferred to a new tube, and one volume of 100% ethanol was added. The tube was mixed by inverting, and the solution was transferred to a ZYMO RNA column. The extraction procedures were performed according to the manufacturer's instructions (ZYMO, R2070). The RNA samples were quantified using a Nanodrop and stored at −80 °C.

To verify the purity of each fraction, the total protein and the cytoplasmic, nucleoplasmic as well as chromatin protein fractions were subsequently analysed using western blot. For immunoblot analysis, antibodies against UGPase (Agrisera, AS05086, 1:1,500) and Histone H3 (ABclonal, A2348, 1:5,000) were used for cytoplasmic and chromatin fraction-specific markers, respectively.

Removal of rRNA. Before ribosomal RNA (rRNA) depletion, RNA samples (6 µg) were concentrated using the ZYMO RNA Clean & Concentrator-5 kit (ZYMO, R1013). rRNA was then eliminated using the RiboMinus Plant Kit for RNA-Seq (Invitrogen, A10838-08) twice according to the manufacturer's instructions. RNA was eluted with 6.5 µl RNase-free H_2O. The RNA samples were quantified using a Qubit 3.0 Fluorometer using the Qubit RNA BR Assay Kit (Life Technologies).

Adapter ligation and cDNA synthesis. A 50 pmol 3′ adapter (5′-rAppCTGTAGGCACCATCAAT–NH^2-3′), where 'rApp' is a modified base, was added to the rRNA-depleted RNA, mixed by pipetting and incubated at 65 °C for 5 min, and then placed on ice for >1 min. Then, 2 µl 10× T4 RNA ligase reaction buffer (NEB, M0242), 10 µl 50% PEG 8000 (NEB, M0242), 1 µl RNaseOUT (40 U µl^{-1}) and 1 µl T4 RNA ligase 2, truncated K227Q (NEB, M0242) were added to the RNA tube and mixed thoroughly by pipetting. After a brief centrifugation, the reaction was incubated at 16 °C for 10 h. The RNA was then purified using the ZYMO RNA Clean & Concentrator-5 kit (ZYMO, R1013) and eluted with 6 µl RNase-free H_2O. Double-stranded cDNA synthesis was performed using the SMARTer PCR cDNA Synthesis Kit (Takara, 634926) with minor modifications. For first-strand cDNA synthesis, according to the NEB Universal miRNA Cloning Linker, the SMART CDS Primer II A was replaced by 5′-AAGCAGTGGTATCAACGCAGAGTACATTGATGGTGCCTACAG-3′.

Library construction and sequencing using Nanopore and PacBio. To minimize PCR bias, PCR cycle number optimization was performed according to the instructions of Procedure & Checklist-Iso-Seq Template Preparation for Sequel Systems (Pacbio, PN 101-070-200 Version 05). The PCR products were recovered by gel purification and analysed by TA cloning (Vazyme, C601) and Sanger sequencing for quality control. A large-scale double-stranded DNA library was amplified from the remaining cDNA using the optimized cycle number and then purified twice with the VAHTS DNA Clean Beads (Vazyme, N411). For the Nanopore method, 100–200 fmol double-stranded DNA was used to construct a library and was sequenced using a MinION sequencer according to the 1D Lambda Control Experiment protocol (SQK-LSK109). For the PacBio method, 1 µg double-stranded DNA was used to prepare a library according to the protocol of SMRTbell Template Prep Kit 1.0-SPv3, and sequenced using a PacBio Sequel II System.

NGS RNA library construction and sequencing. The mRNA libraries were prepared and sequenced at the HaploX Bioinformatics Institute. In brief, polyadenylated RNAs from cytoplasmic, nucleoplasmic and chromatin fractions were purified using poly(A) beads and converted to strand-specific libraries using the VAHTSTM mRNA-seq V2 Library Prep Kit for Illumina (Vazyme). The libraries were then sequenced using an Illumina HiSeq X Ten platform.

NGS RNA-seq data analysis and identification of differentially retained introns between paired samples. Paired-end reads were aligned to the *Arabidopsis* TAIR10 genome[39] by HISAT2 (v.2.1.0) using the following parameters: --min-intronlen 20 --max-intronlen 12000. The PCR duplicates were then removed using Picard (v.2.18.22-SNAPSHOT) and the FPKM values were calculated using StringTie (v.1.3.6) with the parameters '-e --rf -B' using the Araport11 annotation file[40]. For public RNA-seq data, StringTie was also used, but without the parameter '--rf', and DEseq2 (v.1.19.31) was used to calculate the fold change in the level of gene expression between different samples[41].

The ratio of unspliced introns was estimated on the basis of the PIR value, similar to that described previously[42]. In brief, for a specific intron (*I*) and its two adjacent exons (E1 and E2), three types of junction reads were extracted from the alignment file: (1) the junction reads connecting E1 and E2 and *I* (E1I and E2I) represented intron retention events. (2) The junction reads connecting E1 and E2 (E1E2) represented splicing events. (3) The junction reads represented other alternative splicing events (*O*). Each feature (such as E1, E2 and *I*) connected by these junction reads was required to be covered by at least four bases. The PIR and the percentage of splicing in (PSI) values were defined as follows: $PIR = 100 \times (E1I + E2I)/(E1I + E2I + 2 \times E1E2 + 2 \times O)$; $PSI = 100 \times (2 \times E1E2)/(E1I + E2I + 2 \times E1E2 + 2 \times O)$. For samples with biological replicates, the junction reads from all of the biological replicates were combined and used to calculate the PIR or PSI of the sample. Introns with a total junction read number of less than 10 were discarded.

For all of the intron-associated analyses (including NGS and Nanopore data), we selected only constitutively spliced introns in the represented transcripts as described previously[22]. Constitutively spliced introns were defined as introns with a PSI value in the cytoplasmic sample of more than 0.8.

To identify differentially retained introns in mutants and environmental stresses, we collected more than 20,000 public RNA-seq libraries deposited at NCBI by performing a search using the keyword '((*Arabidopsis thaliana*[Organism]) AND "transcriptomic"[Source]) AND "rna seq"[Strategy])' and organized the library information manually. To reduce the interference of experimental batches, we filtered for matched samples that have either 'mutant versus wildtype' or 'treatment versus mock/control' under the same research project, and each sample has at least two replicates. The resulting 6,000+ RNA-seq libraries—representing 713 mutants and 799 treatments, as well as their corresponding controls—were selected and analysed as described above. The counts of E1I, E2I, E1E2 junction reads were then calculated from each biological replicate, and only introns that satisfied the following two criteria were used for further differential analysis: (1) the PIR value of at least one sample was greater than 0.2; and (2) the total junction read number of each sample was more than 10. The IR count (E1I + E2I) and the splicing count (E1E2) of each biological replicate were then submitted to a generalized linear model in the DESeq2 package (v.1.19.31)[41] as used in IRFinder[43] to perform the differential IR test. The *P* values were adjusted using the Benjamini–Hochberg method to decrease the false-discovery rate. The introns with an adjusted $P < 0.05$ and with a PSI fold change higher than 2 (mutant/WT, or treatment/mock) were identified as introns with increased retention ratio in mutant or treatment samples.

Nanopore sequence basecalling and mapping. The Nanopore data analysis workflow is provided in Supplementary Fig. 2. First, raw signal FAST5 files were basecalled using Guppy (v.3.1.5+781ed57) with the default parameters (--c dna_r9.4.1_450bps_hac.cfg). The reads with a mean quality score of less than 7 were filtered out. The remaining reads were mapped to the *Arabidopsis* TAIR10 genome[39] using Minimap2 (v.2.17-r943-dirty)[44] with the following parameters: -ax splice --secondary=no. The reads mapped to rDNA, mitochondria and chloroplast genomes were removed, and only the reads from protein-coding genes were used for further analysis.

The 3′ end cDNA adapter can indicate the 3′ end integrity of cDNA, which is the basis for the identification of PA and Pol-II position, and can also be used to determine the strand direction of mRNA. Thus, only reads with the 3′ end adapter were used for further analysis. To find the cDNA adapter, the 5′- and 3′-end unmapped region and their nearest 20-nucleotide mapped regions were aligned against the cDNA 3′ end adapter sequence using BLASTn (v.2.9.0+)[45] with the following parameter: -word_size 6. The length of the 3′-end and the 5′-end adapters was 42 nucleotides and 30 nucleotides, respectively. The first 15 nucleotides of the 3′-end adapter and the first 3 nucleotides of 5′ end adapter is a specific sequence, but the final 27 nucleotides are consistent between the adapters. A 3′-end-adapter alignment must meet the following criteria: (1) the alignment start position of the 3′ end adapter is ≤12; and (2) the alignment length is ≥15.

After adapter alignment, the following reads were filtered out: (1) reads without the 3′ end adapter; (2) the antisense transcripts (the strand direction of mRNA was determined by the position of 3′ end adapter); and (3) reads missing the 5′ end (the 5′ end is located downstream of the annotated first exon, probably due to incomplete reverse transcription). Further poly(A) identification and read filtering are described below.

3′ poly(A) identification and length estimation using PolyAcaller. As the DNA passes through the pores of the Nanopore, the changes in the base sequence cause

transitions in the raw current signal. However, for long polymer regions, the raw signal value fluctuates less, making it difficult to determine the region length. Therefore, the base-calling software may even recognize 100 consecutive A bases as one A. To solve this problem, we developed a method called PolyAcaller to infer the poly(A) length from the duration of the measured signal (Supplementary Fig. 4). As the double-stranded cDNA libraries can be sequenced from either the 5′ end or from the 3′ end, we first determined whether to find poly(A) or poly(T) on the basis of the 3′ adapter position. For example, to find poly(A), the base sequence and the duration of the signal (in units of the event) corresponding to each base were extracted from the Guppy output file. To reduce errors, only the unmapped region of the read between the genome mapping region and the 3′ adapter region was used to find poly(A). The final ten bases of the genome mapping region and the first five bases of the adapter region were also included. The bases near A and with event lengths of greater than 20× mean event length were converted to A. A local score system (A, 1× event length; non-A, −1.5× event length) was then used to search for the local poly(A) region with the highest score. The poly(A) length was calculated by dividing the total event length of this region by the mean event length.

We then separated full-length clean reads into two groups. The 3′ end of the reads in group I was located upstream of the last exon, and therefore belonged to elongating transcripts (which may also include a few polyadenylated transcripts owing to the APA). The other reads belonged to group II, and included elongating transcripts and polyadenylated transcripts. The results showed that the calculated poly(A) length of reads in group I had one peak at 3 bases, and 98% of reads were less than 15 bases, while the poly(A) length (logarithm) of reads in group II had two peaks, one at 3 bases and one at ~130 bases (Supplementary Fig. 4d). Thus, we chose a cut-off of 15 to distinguish between elongating transcripts and polyadenylated transcripts.

Filtering splicing intermediates and reads with low mapping accuracy at the 3′ end. For the reads representing elongating transcripts, the genome mapping position indicated the Pol-II transcription position, and the 3′ adapter was expected to be exactly next to the genome mapping region. However, sequencing errors of a few bases at the 3′ end may result in inaccurate alignment and, therefore, inaccurate determination of the Pol-II transcription position. Thus, if the distance between the start position of the 3′ adapter and the end position of the genome mapping region was more than 5, the read was removed. Furthermore, we also filtered out splicing intermediates because their Pol-II positions could not be determined. Considering the potential inaccuracy of 3′ end alignment, all reads with the 3′ end between 10 nucleotides upstream and 10 nucleotides downstream of the 5′ splice site of an intron were considered to be splicing intermediates and were removed.

Determining intron splicing status. The introns with a mapping ratio of at least 80% in a read were identified as unspliced, and the others were identified as spliced. To reduce the interference of alternative splicing, only constitutively spliced introns were used for further analysis.

PacBio data analysis. The highly accurate circular consensus sequences (CCS reads) were extracted from PacBio subreads using ccs (v.4.0.0) with the parameter '--min-rq 0.9'. The adapter was identified and removed using lima (v.1.11.0) with the parameter '--isoseq'. The adapter-removed CCS reads were mapped to the *Arabidopsis* genome using the same method as described in the Nanopore data analysis. The final 20 bases of the genome mapping region and the unmapped region of the 3′ end read were used to find poly(A) using a local alignment score method (A score, 1; non-A score, −1.5). The read including a polyadenylated region of longer than or equal to 15 bases was identified as a polyadenylated read. The splicing intermediates, as well as reads missing 5′ ends were removed as described above.

Analysis of APA. The polyadenylated RNA reads terminating at each high confidence PA site annotated in PlantAPAdb[46] were counted. The PA site with the most reads in a gene was referred to as the major PA site, and the other sites were referred to as the minor PA site. To identify the introns that were differentially spliced in different PA isoforms, the counts of spliced and unspliced reads were submitted to a generalized linear model in the DESeq2 package (v.1.19.31)[41] using the same method as described for the identification of differentially retained introns from RNA-seq data.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All data generated in this study were deposited at NCBI under the accession number PRJNA591665.

Code availability

The code used to perform Poly(A) tail analysis is available at https://github.com/zhailab/polyACaller.

Received: 2 December 2019; Accepted: 8 May 2020;
Published online: 15 June 2020

References

1. Merkhofer, E. C., Hu, P. & Johnson, T. L. Introduction to cotranscriptional RNA splicing. *Methods Mol. Biol.* **1126**, 83–96 (2014).
2. Naftelberg, S., Schor, I. E., Ast, G. & Kornblihtt, A. R. Regulation of alternative splicing through coupling with transcription and chromatin structure. *Annu. Rev. Biochem.* **84**, 165–198 (2015).
3. Wissink, E. M., Vihervaara, A., Tippens, N. D. & Lis, J. T. Nascent RNA analyses: tracking transcription and its regulation. *Nat. Rev. Genet.* **165**, 535–519 (2019).
4. Oesterreich, F. C. et al. Splicing of nascent RNA coincides with intron exit from RNA polymerase II. *Cell* **165**, 372–381 (2016).
5. Herzel, L., Straube, K. & Neugebauer, K. M. Long-read sequencing of nascent RNA reveals coupling among RNA processing events. *Genome Res.* **28**, 1008–1019 (2018).
6. Khodor, Y. L. et al. Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in *Drosophila*. *Gene Dev.* **25**, 2502–2512 (2011).
7. Bentley, D. L. Coupling mRNA processing with transcription in time and space. *Nat. Rev. Genet.* **15**, 163–175 (2014).
8. Hoskins, A. A. et al. Ordered and dynamic assembly of single spliceosomes. *Science* **331**, 1289–1295 (2011).
9. Wahl, M. C., Will, C. L. & Lührmann, R. The spliceosome: design principles of a dynamic RNP machine. *Cell* **136**, 701–718 (2009).
10. Chen, W. et al. Transcriptome-wide interrogation of the functional intronome by spliceosome profiling. *Cell* **173**, 1031–1044 (2018).
11. Burke, J. E. et al. Spliceosome profiling visualizes operations of a dynamic RNP at nucleotide resolution. *Cell* **173**, 1014–1030 (2018).
12. Mayer, A. et al. Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. *Cell* **161**, 541–554 (2015).
13. Churchman, L. S. & Weissman, J. S. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* **469**, 368–373 (2011).
14. Nojima, T. et al. Mammalian NET-seq reveals genome-wide nascent transcription coupled to RNA processing. *Cell* **161**, 526–540 (2015).
15. Hetzel, J., Duttke, S. H., Benner, C. & Chory, J. Nascent RNA sequencing reveals distinct features in plant transcription. *Proc. Natl Acad. Sci. USA* **113**, 12316–12321 (2016).
16. Zhu, D. et al. The features and regulation of co-transcriptional splicing in *Arabidopsis*. *Mol. Plant* **13**, 278–294 (2020).
17. Li, S. et al. Global co-transcriptional splicing in *Arabidopsis* and the correlation with splicing regulation in mature RNAs. *Mol. Plant* **13**, 266–277 (2020).
18. Wu, Z. et al. Quantitative regulation of FLC via coordinated transcriptional initiation and elongation. *Proc. Natl Acad. Sci. USA* **113**, 218–223 (2016).
19. Zhu, J., Liu, M., Liu, X. & Dong, Z. RNA polymerase II activity revealed by GRO-seq and pNET-seq in *Arabidopsis*. *Nat. Plants* **4**, 1112–1123 (2018).
20. Bhatt, D. M. et al. Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. *Cell* **150**, 279–290 (2012).
21. Boutz, P. L., Bhutkar, A. & Sharp, P. A. Detained introns are a novel, widespread class of post-transcriptionally spliced introns. *Gene Dev.* **29**, 63–80 (2015).
22. Drexler, H. L., Choquet, K. & Churchman, L. S. Splicing kinetics and coordination revealed by direct nascent RNA sequencing through nanopores. *Mol. Cell* **77**, 985–998 (2020).
23. Kalyna, M. et al. Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in *Arabidopsis*. *Nucleic Acids Res.* **40**, 2454–2469 (2012).
24. Deng, X. et al. Arginine methylation mediated by the *Arabidopsis* homolog of PRMT5 is essential for proper pre-mRNA splicing. *Proc. Natl Acad. Sci. USA* **107**, 19114–19119 (2010).
25. Deng, X. et al. Recruitment of the NineTeen complex to the activated spliceosome requires AtPRMT5. *Proc. Natl Acad. Sci. USA* **113**, 5447–5452 (2016).
26. Wang, X. et al. SKIP is a component of the spliceosome linking alternative splicing and the circadian clock in *Arabidopsis*. *Plant Cell* **24**, 3278–3295 (2012).
27. Sanchez, S. E. et al. A methyl transferase links the circadian clock to the regulation of alternative splicing. *Nature* **468**, 112–116 (2010).
28. La Manno, G. et al. RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
29. Gaidatzis, D., Burger, L., Florescu, M. & Stadler, M. B. Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation. *Nat. Biotechnol.* **33**, 722–729 (2015).
30. James, A. B. et al. Alternative splicing mediates responses of the *Arabidopsis* circadian clock to temperature changes. *Plant Cell* **24**, 961–981 (2012).
31. Gahura, O. et al. Prp45 affects Prp22 partition in spliceosomal complexes and splicing efficiency of non-consensus substrates. *J. Cell. Biochem.* **106**, 139–151 (2009).

32. Siatecka, M., Reyes, J. L. & Konarska, M. M. Functional interactions of Prp8 with both splice sites at the spliceosomal catalytic center. *Gene Dev.* **13**, 1983–1993 (1999).
33. Boothby, T. C., Zipper, R. S., van der Weele, C. M. & Wolniak, S. M. Removal of retained introns regulates translation in the rapidly developing gametophyte of *Marsilea vestita*. *Dev. Cell* **24**, 517–529 (2013).
34. Naro, C. et al. An orchestrated intron retention program in meiosis controls timely usage of transcripts during germ cell differentiation. *Dev. Cell* **41**, 82–93 (2017).
35. Mauger, O., Lemoine, F. & Scheiffele, P. Targeted intron retention and excision for rapid gene regulation in response to neuronal activity. *Neuron* **92**, 1266–1278 (2016).
36. Brody, Y. et al. The in vivo kinetics of RNA polymerase II elongation during co-transcriptional splicing. *PLoS Biol.* **9**, e1000573 (2011).
37. Göhring, J., Jacak, J. & Barta, A. Imaging of endogenous messenger RNA splice variants in living cells reveals nuclear retention of transcripts inaccessible to nonsense-mediated decay in *Arabidopsis*. *Plant Cell* **26**, 754–764 (2014).
38. Yeom, K. H. & Damianov, A. Methods for extraction of RNA, proteins, or protein complexes from subcellular compartments of eukaryotic cells. *Methods Mol. Biol.* **1648**, 155–167 (2017).
39. Lamesch, P. et al. The *Arabidopsis* Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* **40**, D1202–D1210 (2012).
40. Cheng, C. Y. et al. Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *Plant J.* **89**, 789–804 (2017).
41. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
42. Braunschweig, U. et al. Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Res.* **24**, 1774–1786 (2014).
43. Middleton, R. et al. IRFinder: assessing the impact of intron retention on mammalian gene expression. *Genome Biol.* **18**, 51 (2017).
44. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
45. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinform.* **10**, 421 (2009).
46. Zhu, S. et al. PlantAPAdb: a comprehensive database for alternative polyadenylation sites in plants. *Plant Physiol.* **182**, 228–242 (2020).

Acknowledgements

We thank K.-H. Yeom and Z. Wu for advice on chromatin-bound RNA isolation. The group of J.Z. is supported by the National Key R&D Program of China Grant (2019YFA0903903); an NSFC to J.Z. (grant no. 31871234); the Program for Guangdong Introducing Innovative and Entrepreneurial Teams (2016ZT06S172); and the Shenzhen Sci-Tech Fund (KYTDPT20181011104005). J.J. is supported by China Postdoctoral Science Foundation (2018M640787). The group of X.C. is supported by the National Natural Science Foundation of China (grant nos. 31788103 and 91540203, to X.C.); the Chinese Academy of Sciences (Strategic Priority Research Program, XDB27030201 and QYZDY-SSW-SMC022, to X.C.); the Youth Innovation Promotion Association of CAS (grant no. 2018131, to X.D.); and the State Key Laboratory of Plant Genomics.

Author contributions

J.J., Y.L., D.L., X.J. and X.D. performed the experiments. J.J., Y.L., H.Z., Z.Li, Z.Liu and Y.Z. analysed the data. R.X., X.C. and J.Z. oversaw the study. J.J., Y.L. and J.Z. wrote the manuscript, and all of the authors revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41477-020-0688-1.

Correspondence and requests for materials should be addressed to J.Z.

nature research

Corresponding author(s): Jixian Zhai

Last updated by author(s): Apr 25, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	All used public data were collected from NCBI. And the accession numbers were recorded in supplementary tables. The data were downloaded by a FTP software Filezilla (version 3.47.2.1) from ftp.ncbi.nih.gov.
Data analysis	We used hisat2 (version 2.1.0), stringtie (version v1.3.6), guppy (version 3.1.5+781ed57), DESeq2 (version 1.19.31), minimap2 (version 2.17-r943-dirty), picard (version 2.18.22-SNAPSHOT), ccs (v4.0.0), lima (1.11.0) softwares as described in manuscript.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Raw reads were deposited to the Short Read Archive (SRA) at the NCBI under the BioProject ID: PRJNA591665.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample-size was determined based on previous studies and experiments. Sample size of all experiments was sufficient to result in statistical significance and reproducibility. The initial three biological replicates are typically used for RNA-seq experiments. And the results from different biological replicates and from data generating by different methods (Nanopore, PacBio and Illumina) in this study are highly consistent, confirming our method is robust and the results are reliable and reproducible.
Data exclusions	No data exclusion.
Replication	All experimental findings were reliably reproduced for multiple times. To reproduce the experiments, all the experiments were performed independently between different biological replicates. For each repeat, plants were grown in the same condition as described in method. The results from different biological replicates are highly consistent, confirming our results are reproducible.
Randomization	Samples were randomly chosen in related experiments.
Blinding	Not applicable, as samples were processed identically through standard and in some cases automated procedures (DNA sequencing, Next Generation Sequencing, Nanopore sequencing, DNA/RNA isolation) that should not have bias outcomes.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	UGPase (Agrisera, AS05086, 1:1500, Lot: 1807) and Histone H3 (ABclonal, A2348, 1:5000, Lot: 3507357001)
Validation	Both antibodies were validated for Arabidopsis (https://abclonal.com/catalog-antibodies/HistoneH3RabbitpAb/A2348, https://www.agrisera.com/en/artiklar/ugpase-udp-glucose-pyrophosphorylase-marker-of-cytoplasm.html).