




## Methods

## iFLAS: positive-unlabeled learning facilitates full-length transcriptome-based identification and functional exploration of alternatively spliced isoforms in maize

Feng Xu<sup>1</sup>, Songyu Liu<sup>1</sup>, Anwen Zhao<sup>1</sup>, Meiqi Shang<sup>1</sup>, Qian Wang<sup>1</sup> , Shuqin Jiang<sup>1</sup>, Qian Cheng<sup>1</sup>, Xingming Chen<sup>2</sup>, Xiaoguang Zhai<sup>2</sup>, Jianan Zhang<sup>2</sup>, Xiangfeng Wang<sup>1</sup>  and Jun Yan<sup>1</sup> 

<sup>1</sup>State Key Laboratory of Maize Bio-Breeding, National Maize Improvement Center, Frontiers Science Center for Molecular Design Breeding, College of Agronomy and Biotechnology, China Agricultural University, Beijing, 100094, China; <sup>2</sup>Molbreeding Biotechnology Co., Ltd, Shijiazhuang, Hebei Province, 051430, China

Authors for correspondence:

Jun Yan

Email: [yanjun@cau.edu.cn](mailto:yanjun@cau.edu.cn)

Xiangfeng Wang

Email: [xwang@cau.edu.cn](mailto:xwang@cau.edu.cn)

Received: 24 October 2023

Accepted: 6 January 2024

New Phytologist (2024) 241: 2606–2620

doi: 10.1111/nph.19554

**Key words:** alternative splicing, full-length transcriptome, isoform, maize, positive-unlabeled learning.

## Summary

- The advent of full-length transcriptome sequencing technologies has accelerated the discovery of novel splicing isoforms. However, existing alternative splicing (AS) tools are either tailored for short-read RNA-Seq data or designed for human and animal studies. The disparities in AS patterns between plants and animals still pose a challenge to the reliable identification and functional exploration of novel isoforms in plants.
- Here, we developed integrated full-length alternative splicing analysis (iFLAS), a plant-optimized AS toolkit that introduced a semi-supervised machine learning method known as positive-unlabeled (PU) learning to accurately identify novel isoforms. iFLAS also enables the investigation of AS functions from various perspectives, such as differential AS, poly(A) tail length, and allele-specific AS (ASAS) analyses.
- By applying iFLAS to three full-length transcriptome sequencing datasets, we systematically identified and functionally characterized maize (*Zea mays*) AS patterns. We found intron retention not only introduces premature termination codons, resulting in lower expression levels of isoforms, but may also regulate the length of 3'UTR and poly(A) tail, thereby affecting the functional differentiation of isoforms. Moreover, we observed distinct ASAS patterns in two genes within heterosis offspring, highlighting their potential value in breeding.
- These results underscore the broad applicability of iFLAS in plant full-length transcriptome-based AS research.

## Introduction

Alternative splicing (AS) is an important mechanism of transcriptional regulation in eukaryotes, which dramatically expands the diversity and complexity of the transcriptome in the context of a relatively limited genome background and gene repertoire (Chaudhary *et al.*, 2019). The resultant isoforms usually play crucial roles in various aspects, including cell differentiation, tissue-specific expression, and stress responses, enhancing an organism's adaptability (Marasco & Kornblihtt, 2023). In humans, up to 95% of multi-exon genes undergo AS (Pan *et al.*, 2008); while in *Arabidopsis* (*Arabidopsis thaliana*), the proportion is *c.* 61% (Marquez *et al.*, 2012). In-depth research on AS in plants not only provides a comprehensive understanding of complex physiological processes, but also helps the identification of key genes and regulatory pathways that can be targeted to accelerate plant breeding and improvement.

Advances in next-generation sequencing (NGS) have revolutionized AS studies at the transcriptome level (Sultan *et al.*, 2008). However, the ongoing constraint of limited read lengths in NGS data still poses a formidable technical hurdle accurately characterizing AS patterns from a full-length perspective (Wang *et al.*, 2019). Long-read transcriptome sequencing ingeniously alleviates this issue by directly sequencing full-length mRNA molecules, unlocking a deeper transcriptome understanding and greatly expediting functional study of AS and novel isoforms (Stark *et al.*, 2019). However, current full-length transcriptome sequencing technologies are plagued by high error rates and relatively low sequencing throughputs for a given cost, posing a challenge in developing an effective algorithm model capable of reliably identifying AS events and novel isoforms (Hu *et al.*, 2021).

Gene models and AS patterns differ between plants and animals. For example, plants contain more single-transcript genes



and shorter introns than animals (Jia *et al.*, 2020). While skipped exon (SE) events are more prevalent in animals (Keren *et al.*, 2010), plants tend to exhibit a higher frequency of intron retention (IR; Marquez *et al.*, 2012). Currently, cutting-edge AS analysis tools for plants, such as 3D RNA-seq (Guo *et al.*, 2021) and ASTOOLS (Qi *et al.*, 2022), primarily cater to short-read RNA-Seq data. Meanwhile, long-read AS tools, like FLAIR (Tang *et al.*, 2020) and IsoTools (Lienhard *et al.*, 2023), are predominantly designed and tested on human and animal data, potentially lacking a plant-specific perspective. Moreover, these tools are often tailored for specific data types or research interests, restricting their broader utility for comprehensive AS exploration. Thus, there is an urgent need for systematic method optimization and tool integration in the field of long-read AS research in plants.

In this study, we first applied a semi-supervised machine learning (ML) strategy, known as positive-unlabeled (PU) learning (Bekker & Davis, 2020), to reliably identify novel isoforms in three representative full-length maize (*Zea mays*) transcriptome datasets. We then investigated maize AS events and their functional implications from three perspectives: differential AS (DAS), AS-related differential poly(A) tail length, and allele-specific AS (ASAS). Finally, we developed and comprehensively evaluated the integrated full-length alternative splicing analysis (iFLAS) toolbox, which offers a 'one-stop' solution for plant full-length AS analysis by integrating and optimizing multiple methods and tools. iFLAS streamlines the processing of raw sequencing data, ensures accurate isoform and AS event identification, and provides versatile result exploration and visualization to meet diverse research interests.

## Materials and Methods

### Collection of full-length transcriptome datasets

We comprehensively evaluated iFLAS using three maize full-length transcriptome sequencing datasets. The maize cross panel (MCP) dataset consisted of four maize lines: B73, Ki11, and their reciprocal hybrid lines (Wang *et al.*, 2020). For each line, embryo and endosperm at 20 d after pollination (DAP) and root at 14 d after germination were collected, and then, PacBio long-read and Illumina pair-end 150 (PE150) transcriptome sequencing were performed on the mRNAs from each tissue. The maize direct RNA sequencing (MDRS) panel dataset focused exclusively on the B73 line. Kernel at 24 DAP and 14-d-old seedlings was collected, and total RNAs were sequenced using Nanopore long-read and Illumina PE150 transcriptome sequencing, respectively. The maize inbred panel (MIP) dataset comprised eight maize lines: B73, Chang7-2, Mo17, Huangzao4, PH207, PH4CV, PH6WC, and Zheng58. For each line, mRNAs were collected from the 20-d-old seedlings, followed by PacBio long-read and Illumina PE150 transcriptome sequencing. To further showcase the broad applicability of iFLAS across different plant species, we collected full-length transcriptome data and corresponding Illumina NGS data from leaf tissues of other four representative plants: rice (*Oryza sativa* cv. Nipponbare, monocotyledons),

Arabidopsis (col-0, dicotyledonous), potato (*Solanum tuberosum* cv. C88, autotetraploid), and wheat (*Triticum aestivum* cv. Chinese Spring, heterohexaploid). The demo dataset and result of maize have been deposited on GitHub repository (<https://github.com/CrazyHsu/iFLAS>), and the source datasets can be accessed with project ids listed in the Data availability section.

### Preprocessing of raw sequencing data

For the raw data generated by different sequencing platforms, we implemented different data preprocessing workflows to ensure high-quality FASTA/FASTQ reads. For PacBio bam data, we utilized CCS (v.4.2.0) to generate circular consensus sequence (CCS) reads, with parameters '--min-passes 2 --min-rq 0.9 --min-length 50', and then used LIMA (v.2.0.0) and ISOSEQ3 (v.3.3.0) to obtain full-length reads by removing primers, barcodes and poly(A) sequences from CCS reads. For Nanopore FAST5 data, we used GUPPY (v.3.4.5) and NANOPOLISH-POLYA (v.0.11.1; Workman *et al.*, 2019) to perform base calling and acquire poly(A) tail length information for each mRNA molecule. Regarding Illumina RNA-Seq data, we applied FASTP (v.0.20.1; S. Chen *et al.*, 2018) with the parameter set as '-l 150 -q 20' for read quality control. FMLRC2 (v.0.1.4; Wang *et al.*, 2018) was employed for long-read correction in a hybrid-correction strategy with parameters set to '-k 25 59', except for ASAS analysis, to avoid interference with allele genotype identification.

### Reads mapping, isoform collapsing, and junction refining

We employed MINIMAP2 (v.2.18-r1015; Li, 2018) and HISAT2 (v.2.2.0; Kim *et al.*, 2019) to map long-read and short-read transcriptome sequencing data to the maize RefGen\_v4 genome assembly (release 50), in which the maximum intron length was set to 10 000 to accommodate the relatively short intron length in maize. Cupcake (v.Py2\_v8.7x) was then utilized for isoform collapsing, and the parameters for MCP and MIP dataset were set to '--dun-merged-5-shorter --max\_5\_diff 1000 --max\_3\_diff 100 --fnc\_coverage 2 -i 0.9 -c 0.9 --max\_fuzzy\_junction 5'. Given a significant number of truncated reads in the Nanopore data, we fine-tuned the parameters with '--max\_5\_diff 500' and '--fnc\_coverage 5' for MDRS dataset.

Due to the high error rate of long-read sequencing, errors near splice junctions may introduce two types of error: exon-intron boundary shifts (Supporting Information Fig. S1a) and missing mini-exons (Fig. S1b). To address these issues, we obtained high-quality splice junctions (HQJ) by applying REGTOOLS (v.0.5.2; Feng *et al.*, 2018) on short-read alignment results, and revised conflicting junctions of full-length isoforms with the following criteria: (1) For exon-intron boundary shifts defined by long-read alignment, we revised the boundaries of conflicting junctions if HQJ data showed inconsistent splice junctions. The new splice junctions must follow canonical splice motifs (GT-AG, GC-AG, and AT-AC) with flanking exon lengths ranging from 80 to 120% of the original length (Fig. S1c). (2) For exons annotated in HQJ data but absent from long-read alignments, we incorporated these new exons into the isoform annotation if the



upstream and downstream splice sites were consistent with the conflicting junctions (Fig. S1d).

### Identification of isoforms based on PU learning

To demonstrate the ability of PU learning to identify reliable novel isoforms from a vast pool of unlabeled ones, we pooled all B73 tissue isoforms in the MCP dataset and constructed an isoform feature matrix based on 11 isoform-level and 7 splice-junction-level features for each isoform, which are inspired and adapted based on the features provided by the SQANTI3 toolkit (Tardaguila *et al.*, 2018; Table S1; Notes S1). We then categorized all isoforms into three sets: (1) the true-positive dataset, representing the annotated isoforms with coverage  $\geq 2$  and minimum splice junction RPKM  $\geq 0.05$ ; (2) the true-negative dataset, consisting of unannotated isoforms with isoform ratio  $< 0.05$  or those containing novel splice junctions and minimum splice junction RPKM  $< 0.05$ ; and (3) the unlabeled dataset containing the remaining isoforms. Subsequently, we randomly selected two-thirds of the true positive and true negative samples as training data, from which positive instances were sampled in increments of 500, 1000, ...,  $500 \times n$  (where  $n$  is the number of iterations), with their labels converted to negative. Finally, we calculated F1 scores using five-fold cross-validation for two positive-negative (PN) learning models, namely random forest (RF) and gradient boosting (GB), and two PU learning models (RF-PU, GB-PU). The F1 score is often used as a reliable indicator to evaluate model accuracy through average accuracy and recall score, and a higher F1 score indicates a better balance of the model.

To determine the optimal model of five PU learning models, we randomly selected 80% of PN samples as training sets to evaluate the area under the curve (AUC) value of each model using five-fold cross-validation, and the remaining 20% of the samples served as test set to assess model robustness (Fig. 1a). During the training process, we used a bagging strategy by constructing 100 classifiers, where each classifier randomly selected subsets of positive and unlabeled samples for training to predict the remaining unlabeled ones, and the final predicted value was obtained by averaging the predictions of all classifiers (Fig. 1b).

### Comparison of different isoform identification methods

We compared PU learning with two other popular isoform identification tools, SQANTI3 (v.5.0) and FLAIR (v.1.5.1; Tang *et al.*, 2020). For SQANTI3, the parameter in *sqanti\_gc.py* was set to '--fl\_count --coverage --expression', and two strategies in *sqanti\_filter.py* were used for isoform identification, namely SQANTI3-rule (rule-based) and SQANTI3-ml (ML based). For FLAIR, the parameters for *collapse* function were set to '--stringent --no\_end\_adjustment -s 2 --filter ginormous -w 1000'.

### Identification of alternative RNA processing events

We identified four types of AS events (SE, IR, A3SS, and A5SS), and alternative polyadenylation (APA) events using a

hybrid strategy. Briefly, for SE, A3SS, and A5SS events, the splicing boundaries of both inclusive and exclusive events in long-read alignment results should be supported by RNA-seq junctions. For IR events, two types of reads are needed: (1) reads completely covering the intron and the flanking exons and (2) reads spliced at the donor and acceptor sites of the junction in both datasets. SUPPA2 (Trincado *et al.*, 2018), a well-known short-read-based AS identification tool, was utilized independently with default parameters to validate the AS events. Regarding the definition of PA sites, the 3' end of each long-read alignment was defined as the cleavage site and clustered following a previous approach (Abdel-Ghany *et al.*, 2016). Briefly, cleavage sites within 24 bp were clustered into a PA cluster (PAC), with the cleavage site with the highest read coverage in each PAC defined as the PA site. An APA event was considered if multiple PA sites were found within a gene.

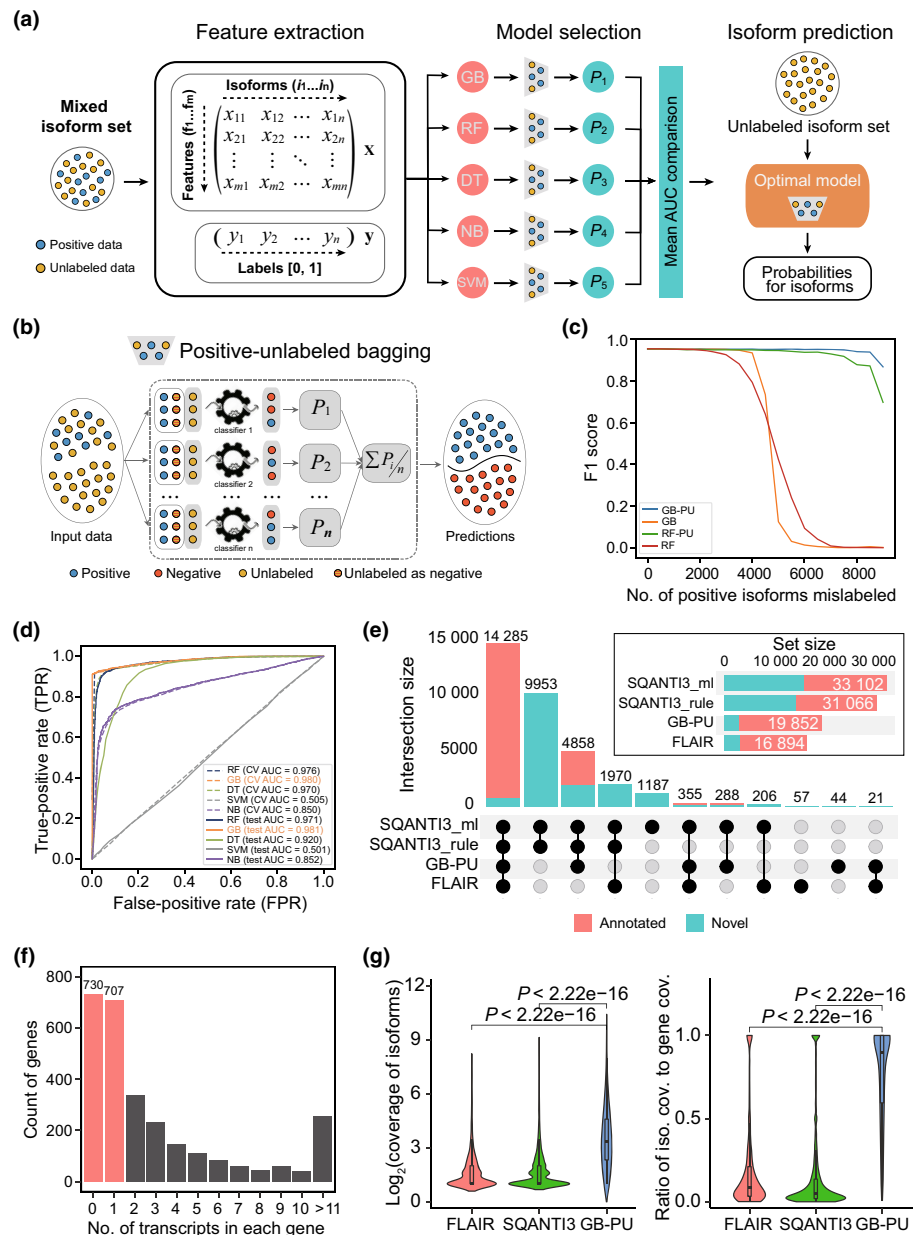
### Differential analyses

Isoform expression levels were quantified using short RNA-seq data for DAS, DEG, and AS-related expression pattern analysis. DAS analysis was performed using RMATS (v.3.1.0; Shen *et al.*, 2014) with the absolute value of delta percent spliced in ( $|\Delta\text{PSI}| \geq 0.1$  and false discovery rate (FDR)  $\leq 0.05$  as significance threshold. Read-count matrices were generated with FEATURE-COUNTS (v.2.0.1; Liao *et al.*, 2014), and DGE analysis was performed using DESEQ2 (v.1.26.0; Love *et al.*, 2014) with threshold of  $|\log_2(\text{fold change})| \geq 1.5$  and FDR  $\leq 0.01$ . Differential analysis of poly(A) tail length between splicing isoforms was performed using the Kruskal–Wallis test with a significance threshold of  $P \leq 0.001$ . ASAS analysis employed IsoPhase (Cupcake vPy2\_v8.7x; Wang *et al.*, 2020) pipeline to obtain parent-derived isoform haplotypes, followed by a joint analysis with AS events to determine ASAS events using chi-squared test with a threshold of  $P \leq 0.001$  (Notes S1).

### Functional annotation and data visualization

We merged maize Gene Ontology (GO) terms from Ensembl Plants database and those annotated using INTERPROSCAN (v.5.50, database version v.88.0; Jones *et al.*, 2014) based on sequence data, which served as enrichment background in CLUSTERPROFILER (v.3.14.0; Yu *et al.*, 2012). InterProScan was also used to predict protein domain for the target isoforms with an *E*-value  $\leq 0.001$ . The *translate* program in InterProScan was used for open reading frame (ORF) prediction, with the parameter set to '-find 1' to identify start and stop codons, and the longest predicted ORF was selected as the candidate ORF for each isoform. PHEATMAP (v.1.0.12; Kolde, 2012) was used for cluster analysis and FACTOMINER (v.2.5; Lê *et al.*, 2008) was used for principal component analysis (PCA). Mapped reads were visualized using IGV (v.2.4.19; Thorvaldsdóttir *et al.*, 2013), while SPLICEGRAPHER (v.0.2.7; Rogers *et al.*, 2012) and GVIZ (v.1.30.0; Hahne & Ivanek, 2016) were employed to display isoform structure and alignment details.





## Results

### PU learning for reliable identification of novel isoforms

To investigate maize transcriptome and isoform diversity, we first preprocessed data from 20 PacBio samples and 2 Nanopore samples, obtaining *c.* 6.85 million and 5.79 million raw long reads with average lengths of 2244 base pairs (bp) and 843 bp, respectively (Table S2). Approximately 85.54% of PacBio reads and 78.12% of Nanopore reads were successfully mapped to the maize reference genome, encompassing over 16 000 genes (Table S2). To increase the number of isoforms, we then pooled and collapsed all the reads from three tissues of B73 in MCP dataset, resulting in a total of 33 224 isoforms (Tables S3, S4). Of these, 16 871 matched known transcript annotations, while the remaining ones were categorized as novel.

We employed PU learning, a semi-supervised ML approach, to reliably identify novel isoforms (Materials and Methods section, Fig. 1a,b). For high-quality model construction and evaluation, we screened 13 914 true-positive and 7406 true-negative isoforms as training data through strict filtering criteria. Using the training data, we evaluated the performance of RF and GB by incrementally adding positive samples to unlabeled datasets under PN and PU training procedures. The F1 scores of all models were similar when a small number of positive samples were added to the unlabeled dataset. However, as the number of positive samples added to the unlabeled dataset increased to approximately half of the total number of positive samples, the F1 scores of the PN models declined sharply. The PU models maintained good performance until > 90% of positive samples were added to the unlabeled dataset (Fig. 1c), indicating that PU learning can still accurately identify positive samples ‘hidden’ in unlabeled



data even with limited positive samples, which is consistent with previous researches (Zheng *et al.*, 2019; Liu *et al.*, 2022). We further evaluated five PU learning models based on GB, RF, decision tree (DT), naive Bayes (NB), and support vector machine (SVM), with 80% of the training data allocated for cross-validation and 20% for testing. Among the five models, the GB-based model exhibited the best performance with AUC exceeding 0.98 on both cross-validation and testing data, followed by RF and DT (Fig. 1d). Consequently, we selected the GB-PU as the optimal model, resulting in a total of 19 851 reliable isoforms for B73 in MCP dataset, including 16 871 annotated and 2980 novel isoforms (Tables S3, S4).

We next compared GB-PU with two other long-read-based isoform identification applications, namely SQANTI3 and FLAIR (Materials and Methods section). In general, both SQANTI3 methods identified a much higher number of novel isoforms than either FLAIR or GB-PU, whereas GB-PU detected more annotated isoforms and fewer novel ones compared with FLAIR, despite a similar number of isoforms (Fig. 1e). Additionally, almost 99.8% of the isoforms identified by GB-PU were also captured by other three methods, with only 44 isoforms being unique to GB-PU (Fig. 1e). Although the novel isoforms identified by GB-PU were primarily located within genes that were either missing or represented as single-transcript genes in the reference annotation (Figs 1f, S2a,b), their coverage and ratio of isoform coverage to gene coverage were significantly higher than those identified by SQANTI3 and FLAIR (Figs 1g, S2c,d). The similar quantitative pattern was also observed by using short RNA-seq data, where GB-PU exhibited higher transcripts per million (TPM) values (Fig. S3a) and stronger correlation (Fig. S3b) compared with SQANTI3 and FLAIR. The higher level of expression and correlation coefficient indicates that GP-PU may provide more reliable results for identifying novel isoforms. Additionally, GP-PU requires less memory and CPU resources than FLARE and SQANTI3, though it takes more time to execute. (Fig. S4; Notes S1). In short, the GB-PU learning method proposed in this study may be a more suitable choice for the detection of major novel isoforms in plants.

To gain further insights into the biological functions of the novel isoforms, we identified tissue-specific isoforms in embryo, endosperm, and root in the MCP dataset, resulting 1348, 684, and 1084 isoforms, respectively (Fig. S5a). GO enrichment analysis of the major tissue-specific novel isoforms (coverage  $\geq 5$ , ratio of isoform coverage to gene coverage  $\geq 0.5$ ) revealed distinct functional enrichment patterns for each tissue (Fig. S5c), which was also observed in the MDRS dataset (Fig. S5b,d). Notably, the MIP dataset showed a substantial overlap among the novel isoforms identified in different maize lines (Fig. S5e). We then compared the expression patterns of genes corresponding to the overlapping isoforms based on the expression profile of 23 maize tissues collected from the MaizeGDB database (Walley *et al.*, 2016). Interestingly, a large proportion of these genes were highly expressed in leaf (Fig. S5f), consistent with the tissue source of the MIP dataset. In summary, the novel isoforms identified by GB-PU displayed relatively consistent patterns within the same tissue, while exhibiting significant tissue-specific

expression and functional divergence among different tissues, highlighting the reliability of novel isoforms identified by PU learning.

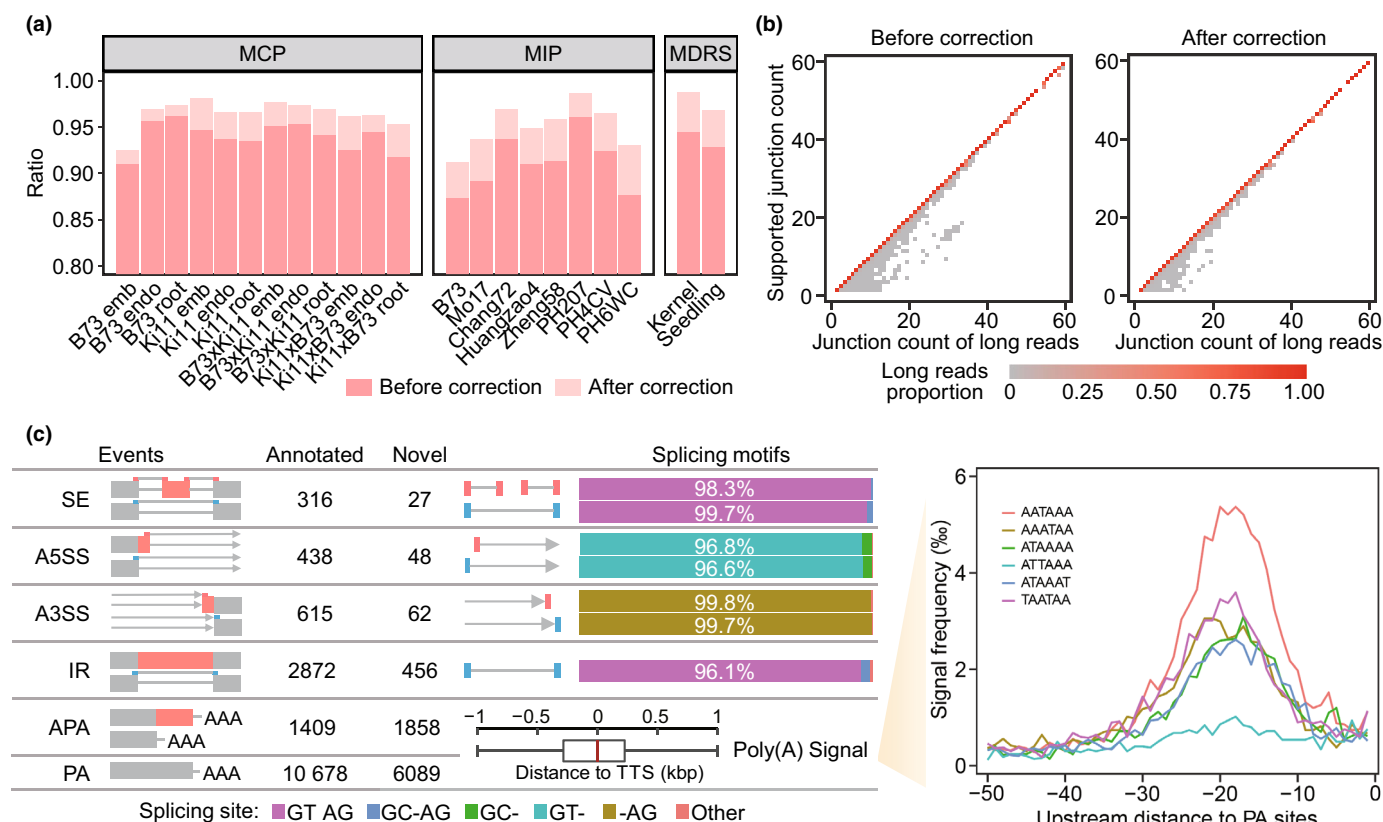
## Characterization of AS events

As Illumina transcriptome sequencing (RNA-Seq) has been proven efficient in defining the exon–intron structure at a local view, we used it as an independent control to assess the accuracy of long-read assay in delineating these events. In overall junction view, the splice junctions defined by filtered isoforms were largely supported by the RNA-Seq data from the same sample, with consistency rates ranging from 91.22 to 98.64% (Fig. 2a), indicating agreement between the two platforms in defining splicing events. Importantly, after PU filtration and junction refinement, we noticed that the consistency rate of junctions in each sample increased by 1–5% (Fig. 2a) compared with the raw mapping result. In the full-length isoform view, we also observed an increase in support ratio of all junctions per isoform for all samples (Fig. 2b), highlighting the necessity of the optimization strategies in elevating the quality of isoform structures.

Based on the optimized high-quality isoforms, we summarized the number of alternative RNA processing events identified in maize B73 from all three datasets, including IR, SE, alternative 5' or 3' splice site (A5SS or A3SS) and APA. In total, 4834 AS events were identified, of which 593 (12.3%) were novel according to the reference annotation (Fig. 2c). As observed in previous studies (Chaudhary *et al.*, 2019), the dominant AS event found in this study was IR, constituting 68.8% of the events, followed by A3SS (14.0%), A5SS (10.1%), and SE (7.1%). In strong support of the reliability of these AS events, the *cis* features near the events agreed with *a priori* knowledge of known regulatory mechanisms (Shang *et al.*, 2017). Briefly, >99% of the splice junctions were associated with a canonical GT-AG or GC-AG motif, even for novel IR events, indicating that these IR events may not represent random readout of nascent transcripts that have yet to undergo splicing at a particular intron. To further evaluate the quality of the identified AS events, we employed SUPPA2 for independent verification. Robust support for both known and novel A3SS, A5SS, and SE events by SUPPA2 was observed, with overlap rates ranging from 85 to 97% (Fig. S6a). Although only one-third of IR events could be validated by SUPPA2, the PSI score for all AS events fell within a reasonable range between 0.1 and 0.9 (Fig. S6b), suggesting generally good splicing levels for all events (Tapial *et al.*, 2017). These results highlight the limitations of traditional short-read-based AS analysis in accurately identifying common IR events in plants, emphasizing the importance of long-read-based tool optimized for plant gene splicing characteristics.

We also detected 3267 APA events, of which 1858 (56.9%) were novel events contributed by 6089 novel polyadenylation (PA) sites (Fig. 2c), suggesting that a large proportion of novel PA sites may originate from single-isoform genes. To evaluate the reliability of PA identification, we further calculated the distances between PA sites and the transcription termination sites (TTS) of the nearest genes (Fig. 2c). A distribution of distances with a





**Fig. 2** Genome-wide identification of maize alternative splicing (AS) events. (a) Proportion of long-read splicing junctions supported by short-read RNA-Seq data in each maize dataset before and after positive-unlabeled (PU) filtration and junction refinement. (b) Proportion of long-read splicing junctions supported by short-read RNA-Seq data for each isoform before and after PU filtration and junction refinement. For each isoform, the number of splice junctions (x-axis) and the numbers of junctions supported by RNA-Seq reads (y-axis) were counted. The density distribution of reads was then summarized and shown in tile form in the figure, with the density indicated by color ranging from gray to red. (c) Left panel: Statistical and characteristic evaluation of genome-wide AS and alternative polyadenylation (APA) events in maize. For the four categories of AS events, the frequencies of different splicing motifs are shown in bar plots in different colors according to the legend below. For APA events, the distances between PA sites and transcription termination sites (TTS) are summarized in the boxplot. Right panel: Frequencies of the top six poly(a) signals located within 50 bp upstream of novel PA sites.

median score near zero and a modest variance indicated that most of the PA sites were reliable. In line with this verification, the frequencies of the top six poly(A) signals within 50 bp upstream of the novel PA sites were consistent with known patterns (Proudfoot, 2011), with AATAAA being the most dominant signal (Fig. 2c). In this regard, the isoform optimization strategy used here provided a direct and efficient approach to identifying AS and APA events, laying a solid foundation for subsequent functional AS analysis.

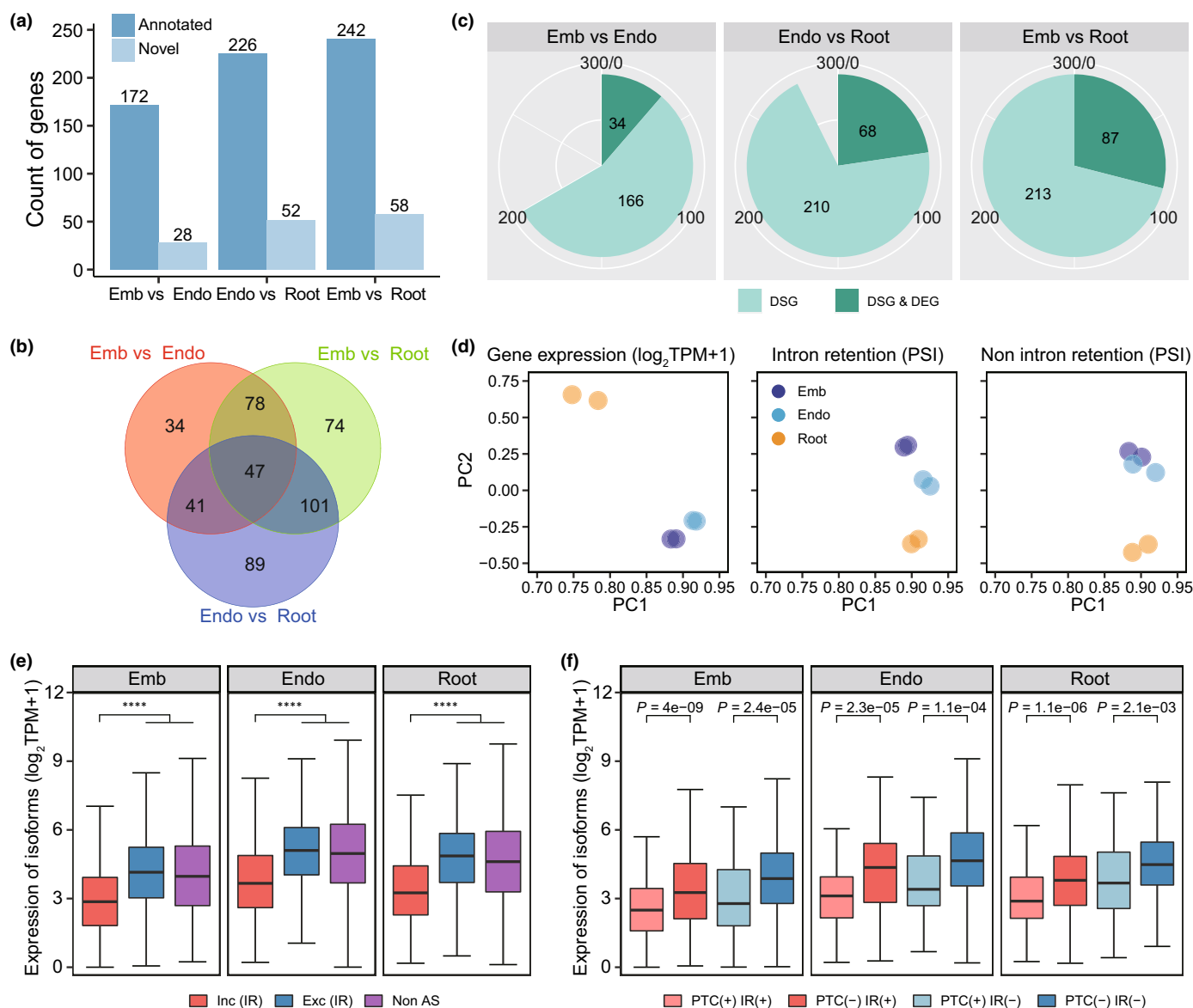
### Uncovering maize tissue splicing patterns and their relationship with gene expression

To explore splicing differences among tissues of maize, we performed a paired DAS analysis on different tissues of B73 in MCP dataset, and identified 200, 278, and 300 differentially spliced genes (DSGs) in the embryo vs endosperm, endosperm vs root, and embryo vs root comparison groups, respectively (Fig. 3a; Table S5). The numbers of total DSGs (Fig. 3a) and tissue-specific DSGs (Fig. 3b) in endosperm vs root and embryo vs root

comparisons were higher than the other one, with the proportion of DSGs containing novel AS events (Fig. 3a) also exceeded 12% of the background novel splicing events (Fig. 2c). GO enrichment analysis further showed that DSGs in different comparison groups were functionally different (Fig. S7), indicating that the splicing patterns of genes may be more divergent among tissues that possess greater morphological and functional differences.

To further investigate the relationship between AS and expression levels of genes, we analyzed differential gene expression in the three comparisons. The number of differentially expressed genes (DEGs) in the embryo vs endosperm comparison (6501) was significantly lower than that in endosperm vs root (10 939) and embryo vs root (10 079; Table S4), consistent with the trend for DSGs. Notably, only a small number of genes were both DSGs and DEGs, with 34 (17.0%), 68 (24.5%), and 87 (29.0%) genes identified respectively (Fig. 3c; Table S5), suggesting significant functional differences between genes affected by AS and transcriptional regulation. To gain a better understanding of these differences, we performed PCAs on DEGs using TPM values and on DAS events using PSI values, where TPM and PSI





**Fig. 3** Alternative splicing (AS) and gene expression patterns in maize MCP dataset. (a) Number of differentially spliced genes (DSGs) in each comparison group. (b) Venn diagram of DSGs between comparison groups. (c) Comparison of DSGs and differentially expressed genes (DEGs). (d) PCAs using gene expression and AS patterns. From left to right are the first two principal components (PCs) of global gene expression level, intron retention (IR), and nonintron retention AS events, respectively. (e) Expression levels of isoforms involved in IR events. Inc(IR) represents isoforms with retained introns in IR events, Exc(IR) represents isoforms without retained introns in IR events and Non-AS represents other isoforms without AS events. The significance of '\*\*\*\*' from left to right are  $P < 2.22 \times 10^{-16}$ ,  $P = 4.1 \times 10^{-11}$  and  $P < 2.22 \times 10^{-16}$ . (f) Effects of IR events and premature termination codons (PTCs) on isoform expression. PTC(+) and PTC(-) represent isoforms with and without PTCs, respectively; IR(+) and IR(-) represent isoforms with and without retained introns, respectively. The Kruskal–Wallis test was used to calculate the significance of differences between groups. Error bar represents the SD.

values were quantified from RNA-Seq data. The results showed that the clustering patterns based on gene expression and IR events effectively differentiate the three tissues, whereas non-IR events could not clearly distinguish between embryo and endosperm (Fig. 3d). This suggests that IR events, similar to gene expression, are more representative than other AS types in reflecting plant tissue specificity.

Consequently, we asked whether there is a correlation between IR and the expression level of isoforms. We compared the expression of isoforms involved in IR events with those not involved in

any AS events (non-AS). The expression levels of intron-inclusive isoforms from IR events were significantly lower than other isoforms, while no obvious expression differences were observed between the intron-exclusive isoforms in IR events and the non-AS isoforms (Fig. 3e). Considering that IR events often involve the retention of long introns, which are more likely to introduce the premature termination codons (PTCs; Pimentel *et al.*, 2016), we further investigated the relationship between isoform expression, IR events and PTCs. We determined that the expression levels of isoforms with PTCs were significantly lower than those



without PTCs, whereas the expression levels of isoforms containing both retained introns and PTCs were the lowest (Fig. 3f), indicating that retained introns may not only have a causal effect on premature termination, but also contribute a superposed effect to the reduction of isoform expression.

### Splicing isoform-specific poly(A) tail length profiling

Polyadenylation, like AS, is a post-transcriptional processing mechanism and results in the addition of a long sequence of adenosines to the 3' end of mRNAs (Lima *et al.*, 2017), which plays a crucial role in mRNA stability and translation efficiency (Subtelny *et al.*, 2014). To investigate the relationship between poly(A) tail length and AS in maize, we plotted poly(A) tail length distributions of isoforms involved in AS events using the MDRS dataset. We observed a longer median poly(A) tail length in the intron-exclusive isoform compared with the intron-exclusive isoforms, especially for IR events (Fig. 4a). One example involves *IAA28* (Aux/IAA-transcription factor 28, *Zm00001d037774*; Fig. 4b), which encodes a known auxin transcription factor. The intron-retained isoform of *IAA28* had a significantly longer poly(A) tail than the intron-exclusive isoform. Moreover, the intron-retained isoform is predicted to harbor a truncated Aux/IAA domain, implying that the retained intron introduces a PTC resulting in the premature termination of translation.

In IR events, long alternatively retained introns are often thought to change the length of the ORF or untranslated region (UTR) in isoforms (Jacob & Smith, 2017). We then compared the ORF and 3' UTR length distributions of isoforms pairs involved in AS events and observed a distinct pattern in IR events that intron-inclusive isoforms exhibited shorter ORF and longer 3' UTR compared with those without introns (Fig. 4c), whereas no differences were found for other AS types (Fig. S8). This contrast was unexpected, as the inclusive isoforms are typically assumed to introduce exon fragments, which would increase ORF length. In fact, the retained introns were more likely to introduce PTCs, leading to frameshift mutations resulting in shorter ORFs and longer 3' UTRs (Fig. 4d). After expanding the scope of our analysis to all isoforms, we observed a significant positive correlation between poly(A) tail length and 3' UTR length ( $R=0.25$ ,  $P<2.2e-16$ ), while ORF length was not that case (Fig. 4e), emphasizing the importance of the 3' UTR in shaping the extent of poly(A) tails.

Nevertheless, intron-retained isoforms tended to harbor longer poly(A) tails and exhibit relatively low levels of long-read abundance (Fig. 4f), consistent with our quantitative analysis of short-read RNA-seq data (Fig. 3e). To further explore the relationship between expression level and poly(A) tail length in maize, we plotted the median tail lengths of three isoform groups categorized by their relative abundances (Fig. 4g). As expected, most highly expressed isoforms had short tails, whereas the least abundant isoforms had longer tails. When we binned isoforms according to normalized median poly(A) tail lengths (Fig. 4h), we observed an inverse correlation between poly(A) length and isoform abundance, in which the isoforms with shorter poly(A) tails (40–120 nt) were more abundant than those with longer poly(A)

tails ( $>120$  nt). The only exception was the group with a median poly(A) tail length shorter than 40 nt, where the poly(A) tails of these isoforms may be too short to accommodate poly(A) binding protein, potentially leading to their degradation (Passmore & Collier, 2021).

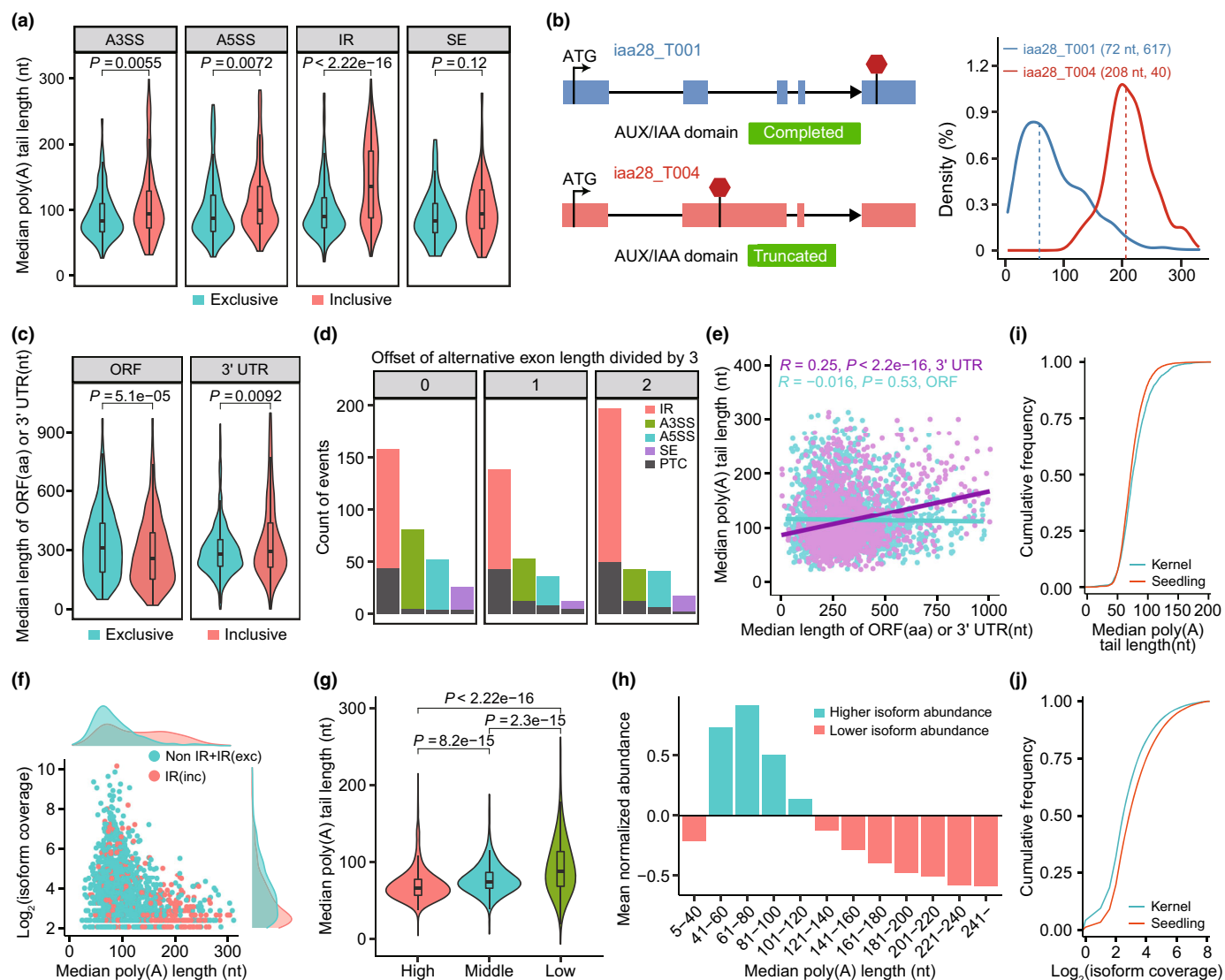
We also observed the same negative correlation between poly(A) tail length and isoform abundance in different maize tissues (Fig. 4i,j), where the median poly(A) tail length of isoforms in kernel was substantially higher than that in aboveground seedling tissues, whereas the isoform abundance showed the opposite pattern. Moreover,  $\sim 80\%$  of well-expressed isoforms (coverage  $\geq 20$ ) had median tail lengths ranging between 54 and 105 nt (Fig. S9a). To explore whether there were functional classes of isoforms associated with longer or shorter poly(A) tails, we classified the deciles of isoforms into two groups with short tails (median length  $\leq 54$  nt,  $n=1562$ ) and long tails (median length  $\geq 105$  nt,  $n=1719$ ). A subsequent GO enrichment analysis within each isoform category showed distinct functional and tissue specificity in kernel and seedling tissues (Fig. S9b). Interestingly, the median poly(A) tail length of 2905 well-expressed (coverage  $\geq 20$ ) genes showed a shorter distribution in seedlings than that in kernel (Fig. S9c), suggesting the poly(A) tail length of different transcripts within a given gene in maize may also undergo dynamic changes across tissues or development stages, as observed in zebrafish (Subtelny *et al.*, 2014).

### Allele-specific alternative splicing analysis in maize hybrid line

The MCP dataset contains transcriptomes of B73, Ki11, and their reciprocal hybrid offspring B73  $\times$  Ki11, enabling us to trace the parental origin of each isoform in B73  $\times$  Ki11 and perform systematic ASAS analysis. By employing the ASAS discovery pipeline (Materials and Methods section), we identified 41 ASAS genes in hybrid lines (Table S6), covering 505 single-nucleotide variants with around one-third being missense variants (Fig. S10a). Of the 41 ASAS genes, 29 contained more than one missense variant (Fig. S10b) and functional analysis showed their enrichment in GO terms related to signal transduction and DNA biosynthetic process (Fig. S10c), suggesting that the ASAS genes may play an important role in maintaining fundamental cellular activities in the hybrid offspring.

*AUXIN RESPONSE FACTOR 28* (*ARF28*; *Zm00001d023659*), as one of the representative ASAS genes, plays a crucial role in regulating plant growth and development (Xing *et al.*, 2011). *ARF28* has seven main isoforms, with *PB.26476.1*, *PB.26476.5*, and *PB.26476.6* being the isoforms predominantly expressed in B73  $\times$  Ki11 (Fig. 5a). Genotype analysis revealed that *PB.26476.1* and *PB.26476.6* belonged to the Ki11 haplotype, and *PB.26476.5* represented the B73 haplotype. The main difference between the two parental haplotypes lies in the retention of intron 14 that the intron is excluded in Ki11 but included in B73 (Fig. 5a). Another IR event at intron 12 also contributed to the difference between *PB.26476.1* and *PB.26476.6* in the Ki11 haplotype. The retention of intron 12 in *PB.26476.6* introduces a PTC, leading to the absence of the Aux/IAA domain (Fig. 5b). Notably, even though





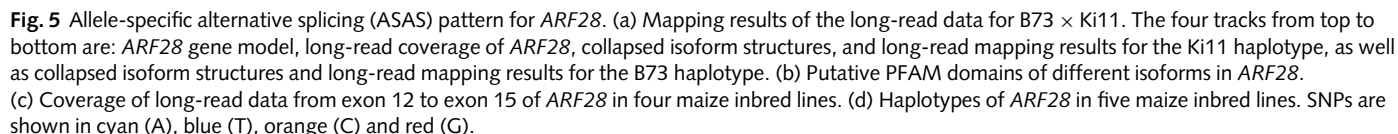
**Fig. 4** Analysis of poly(A) tail length of splicing isoforms in maize. (a) Median poly(A) tail length distribution of isoforms involved in each alternative splicing (AS) type. (b) Poly(A) tail length distribution pattern of different isoforms in *IAA28*. Structure models (left) and poly(A) tail length distribution (right) of the two splicing isoforms of *IAA28* are shown. The AUX/IAA domain is represented as green rectangle under each transcript and the red hexagon represents stop codon (left). The dashed line represents the median length of poly(A) tail (right). (c) Median length distribution of open reading frames (ORFs) and 3' untranslated regions (UTRs) of inclusive and exclusive isoforms associated with intron retention (IR) events. (d) Statistics for the offset of alternative exon length divided by 3 for each AS type. Premature termination codons (PTC; black) refer to the number of PTCs introduced by all AS events in each offset condition. (e) Correlations of ORF (cyan) and 3' UTR (purple) length with poly(A) tail length for each isoform. (f) Distributions of expression levels and poly(A) tail lengths for isoforms containing retained introns or involving other AS types. (g) Median poly(A) tail length distribution of isoforms with different expression levels. The three isoform abundance categories are the highest expressed isoform ( $n=500$ ), those closest to median expression ( $n=500$ ), and the lowest expressed isoform (coverage  $\geq 20$ ,  $n=500$ ). (h) Global relationship between isoform abundance and poly(A) tail length. Normalized abundance was first calculated as the  $\log_2$  of the fold change in the coverage of isoform over the median abundance of all isoforms, and then measured by plotting the mean normalized abundance of bins of isoforms divided by median tail lengths. (i) Cumulative curve distribution of median poly(A) tail length of isoforms in kernel and seedling. (j) Cumulative distribution of isoform abundance in kernel and seedling. The Kruskal–Wallis test was used to calculate the significance of differences between groups. The upper and lower horizontal lines of boxplot legend in (a, c, g) indicated the 75<sup>th</sup> and 25<sup>th</sup> percentiles, and the central bold line and the whiskers indicated the median and minimum–maximum values.

*PB.26476.5* retained an intron at its 3' end compared with the reference isoform T007, both isoforms were predicted to contain the same putative domains (Fig. 5b), suggesting polymorphism of Aux/IAA domains in different functional isoforms of *ARF28*.

*ARF28* also showed similar haplotype patterns in different maize lines in the MIP dataset. Specifically, *ARF28* in Zheng58

was consistent with the B73 haplotype, whereas the alternative haplotype was present in Mo17 and Chang7-2 (Fig. 5c). Notably, the haplotype of *ARF28* in Mo17 and Chang7-2 differ from Ki11 only by one synonymous mutation (GCG → GCA, Ala → Ala) located in the first exon (Fig. 5d), indicating that Mo17, Chang7-2 and Ki11 have the same *ARF28* haplotype in a

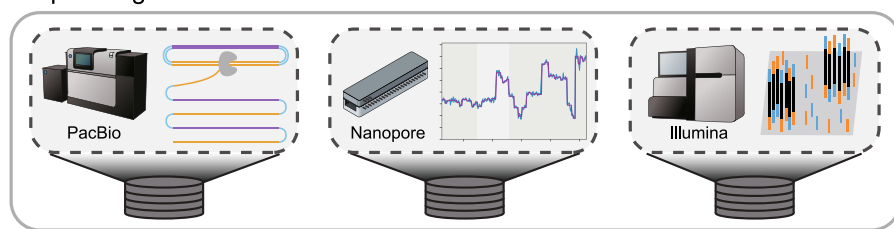




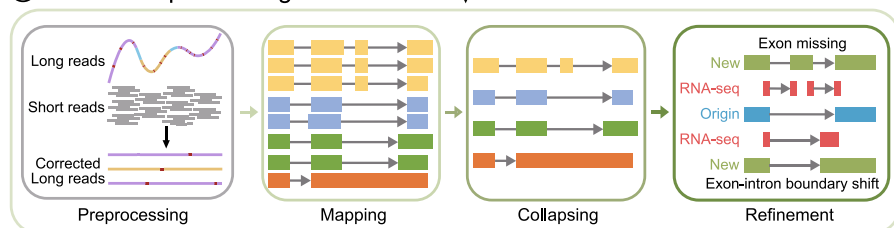
We comprehensively evaluated iFLAS with three real maize datasets, and confirmed its effectiveness and applicability for different AS applications in maize. We also processed the full-length transcriptomes of other four plants: Arabidopsis, rice, wheat, and potato (Notes S1). By comparing long- and short-read datasets across multiple plant species, iFLAS was shown to construct high-quality isoforms with high read coverage, expression levels, and correlation between datasets (Fig. S12). Notably, iFLAS not only consistently identifies AS events aligned with established short-read tools, but also captures challenging event types such as IR and APA (Fig. S13), demonstrating its wide application across various plant species (Tables S8, S9). iFLAS is full-featured and



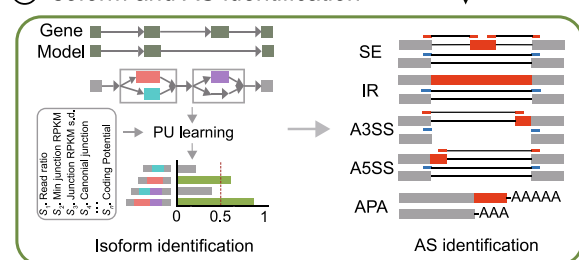
## Sequencing data



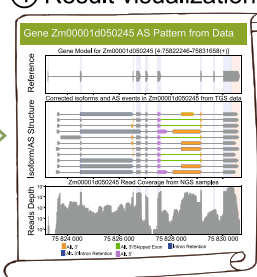
## ① Basic data processing workflow



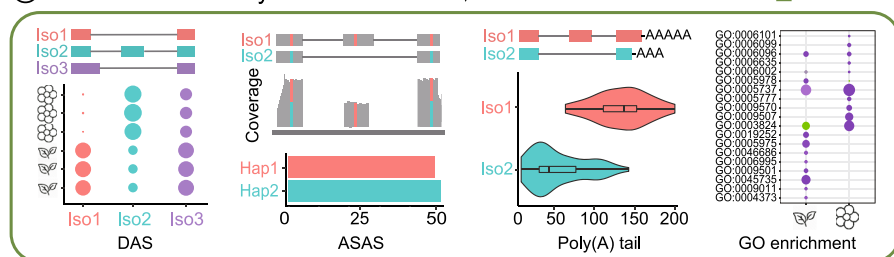
## ② Isoform and AS identification



## ④ Result visualization



## ③ Functional AS analyses



**Fig. 6** Workflow and functional modules of the integrated full-length alternative splicing analysis (iFLAS) toolkit. iFLAS supports ① transcriptome data from PacBio, Nanopore, and Illumina platforms, ② performs basic data processing (including data preprocessing, mapping, collapsing, and refinement), isoform and alternative splicing (AS) identification, ③ functional AS analyses (including DAS, ASAS, AS-related differential poly(A) tails, and ④ GO enrichment analyses) and result visualization.

easy-to-operate, which can be accessed from <https://github.com/CrazyHsu/iFLAS>.

## Discussion

In this study, we developed iFLAS to bridge the gap between long-read sequencing techniques and in-depth plant AS study (Amarasinghe *et al.*, 2020). iFLAS draws on the functional integration of other long-read based AS tools, such as NanoASPipe (Liu *et al.*, 2017) and NanoTrans (Wang *et al.*, 2022) for pipeline construction, SQANTI3 and FLAIR for quality control and isoform classification, TAPPAS (de la Fuente *et al.*, 2020) and ISOTools (Lienhard *et al.*, 2023) for functional analysis of AS and APA, FLEP-Seq (Long *et al.*, 2021) for poly(A) length specific AS analysis, and provided more comprehensive functions for plant research (Table 1). One notable feature of iFLAS is the utilization of PU learning to reliably identify novel isoforms,

addressing the challenge of high sequence noise encountered with long-read sequencing technologies. PU learning does not rely on labeled negative samples, making it a powerful tool to accurately classify numerous unknown samples based solely on positive samples, and has proven successful in various life science applications (Lan *et al.*, 2017; Kolosov *et al.*, 2021).

Integrated full-length AS analysis implements diverse analysis pipelines, enabling accurate AS identification and providing the potential for multi-dimensional functional elucidation of AS from an isoform-level perspective. By using iFLAS, we systematically explored the AS patterns in maize using three sets of full-length transcriptome sequencing data collected for different research purposes. First, we identified a notably lower number of AS events than previously reported (Mei *et al.*, 2017; Q. Chen *et al.*, 2018), which can be attributed to our use of PU learning and a hybrid AS defining strategy, ensuring reliable and conservative AS event identification. The finding is also supported by the fact that 98% of



**Table 1** Functional modules of iFLAS compared with other long-read-based alternative splicing (AS) tools.

	NanoASPipe (Liu <i>et al.</i> , 2017)	Flair (Tang <i>et al.</i> , 2020)	tappAS (de la Fuente <i>et al.</i> , 2020)	FLEP-seq (Long <i>et al.</i> , 2021)	NanoTrans (Wang <i>et al.</i> , 2022)	IsoTools (Lienhard <i>et al.</i> , 2023)	iFLAS
Supported data types							
PacBio (raw bam files)							✓
PacBio (processed FASTA/Q files)		✓		✓			✓
Nanopore (raw fast5 files)					✓		✓
Nanopore (processed FASTA/Q files)	✓	✓		✓	✓		✓
Processed BAM files						✓	
Processing							
Basecalling					✓		✓
Alignment	✓	✓		✓	✓		✓
Error correction and polishing	✓	✓		✓	✓	✓	✓
Generating consensus sequence		✓			✓	✓	✓
Isoform detection	✓	✓			✓	✓	✓
Isoform quality filtering		✓	✓		✓	✓	✓
AS events identification	✓	✓	✓	✓	✓	✓	✓
APA events identification			✓	✓			✓
Functional exploration							
Differential AS events analysis		✓	✓		✓	✓	✓
Poly(A) length specific AS analysis				✓			✓
Allele-specific AS Analysis		✓					✓
GO enrichment analysis			✓				✓
Functional evaluation of AS			✓			✓	
Visualization							
GO enrichment			✓				✓
AS events visualization		✓	✓	✓		✓	✓
Summary statistics			✓	✓	✓	✓	✓
Software type							
Pipeline	✓	✓		✓	✓		✓
Package						✓	
Standalone application			✓				
General LR-based AS analysis tools		✓			✓		✓
Publish time	2017	2019	2020	2021	2022	2023	

the splicing sites are canonical GT-AG motifs (Fig. 2c). Second, despite a weak correlation between DSGs and DEGs (Fig. 3c), our PCA results showed that IR events, like DEGs, were still able to effectively distinguish different tissues (Fig. 3d). Isoforms containing retained introns displayed significantly lower expression level than other isoforms (Fig. 3e,f), implying that genes with PTC may undergo IR to regulate isoform expression in different maize tissues, thereby mitigating the effect of PTCs.

As to poly(A) tail analysis, we found that variations in the poly(A) tail length of isoforms may also have functional implications in maize. On the one hand, poly(A) tail length closely correlates with AS events (Fig. 4a), particularly with IR events, potentially resulting in functional divergence between isoforms of

the same gene (Fig. 4b). Despite IR's potential to introduce long-range alternative exon changes affecting ORF and UTR lengths, we observed a positive correlation between isoform poly(A) tail length and 3' UTR length in this study, with no significant correlation found with ORF length (Fig. 4e), providing a new insight into the regulatory role of the 3' UTR in poly(A) tail dynamics. On the other hand, short-tailed isoform in maize showing higher expression levels than long-tailed isoforms (Fig. 4g,h), while isoforms with different poly(A) tail lengths exhibited distinct functional trends in maize kernel and seedling. The similar correlations have also been observed in other studies (Chang *et al.*, 2014; Lima *et al.*, 2017; Liu *et al.*, 2019), suggesting that poly(A) tail length may play an important role in many aspects of transcript lifecycle.



Furthermore, since AS can be affected by *cis*-acting sequence polymorphisms, ASAS analysis helps the identification of splicing differences derived from two haplotypes within a hybrid individual (Demirdjian *et al.*, 2020) and facilitates the discovery of genes with breeding value. In this study, we found that *ARF28* in B73 × Ki11 expressed two different parent-derived isoforms with distinct Aux/IAA domains, and the same pattern was also observed in two other well-known maize hybrid combinations, namely B73 × Mo17 and Zheng58 × Chang7-2 (Fig. 5). Another ASAS gene, *Zm00001d020461* (Fig. S11), located within a grain weight and width related quantitative trait locus called qKW7 (Li *et al.*, 2016), was also found to express different isoforms in multiple maize inbred lines. Given that *ARF28* and *Zm00001d020461* are known to play an important role in growth and development processes, our findings highlight their promising potential value in breeding programs.

While iFLAS provides new insights into isoform identification and functional analysis from the perspective of full-length transcriptome, the integration of multiple tools may reduce its efficiency. Additionally, we did not consider the combined correlation between AS and APA in the PA site identification process, as well as the correlation between AS and allele-specific expression in the ASAS identification process, despite their correlation being extensively discussed (Park *et al.*, 2018; Blake & Lynch, 2021). Since iFLAS is still in development, we are dedicated to optimizing its analytical processes, improving its operational efficiency, and enhancing the comprehensiveness of its functional modules. For instance, RNA methylation analysis could be introduced to further explore potential correlations between AS and epigenetic modifications.

In summary, iFLAS offers a 'one-stop' solution for the identification and functional study of AS events in plants from a full-length perspective. We expect that it will serve as an efficient resource for plant AS research and further facilitate the exploration of functional genes and the improvement of important agronomic traits such as stress resistance, quality, and yield.

## Acknowledgements

This work was supported by the Beijing Natural Science Foundation (5232013), the National Key Research and Development Program of China (2023YFF1000100 and 2021YFD1201003), the Pinduoduo-China Agricultural University Research Fund (PC2023B01012), the Chinese Universities Scientific Fund (2023TC182), the STI 2030-Major Projects (2023ZD0407 501), the Key Science and Technology Project of Liaoning Province (2022JH1/10200001), the Hebei Provincial Science and Technology Plan Project Modern Breeding Industry Science and Technology Innovation Special Project (21326316D and 21326302D), and the 2115 Talent Development Program of China Agricultural University.




## Competing interests

None declared.

## Author contributions

FX, XW and JY conceived and designed the experiments. XC, XZ and JZ collected the maize transcriptome data. FX, SL, MS, QW, SJ, QC and JY conducted the data analysis and interpretation. FX and JY developed the ML models. FX and AZ developed the software. FX and JY wrote the manuscript, and all co-authors revised the paper.

## ORCID

Qian Wang  <https://orcid.org/0000-0002-0360-0859>  
Xiangfeng Wang  <https://orcid.org/0000-0002-6406-5597>  
Jun Yan  <https://orcid.org/0000-0002-3806-6457>

## Data availability

The MCP and MDRS datasets were obtained from publicly available data repositories. The MCP dataset can be accessed at ArrayExpress under accession numbers E-MTAB-7837 and E-MTAB-7394, while the MDRS dataset is available at NCBI BioProject under accession no. PRJNA64316. The MIP dataset has been deposited in NCBI BioProject with accession number PRJNA983493. The datasets for Arabidopsis, rice, wheat, and potato can be accessed with NCBI BioProject PRJNA382842 (Arabidopsis), PRJNA760839 (rice; Yu *et al.*, 2023), NGDC BioProject PRJCA007997 (potato; Bao *et al.*, 2022), and ENA BioProject PRJEB15048 (wheat; Clavijo *et al.*, 2017). All source codes in the iFLAS toolkit can be available at the GitHub repository: <https://github.com/CrazyHsu/iFLAS>.

## References

- Abdel-Ghany SE, Hamilton M, Jacobi JL, Ngam P, Devitt N, Schilkey F, Ben-Hur A, Reddy AS. 2016. A survey of the sorghum transcriptome using single-molecule long reads. *Nature Communications* 7: 1–11.
- Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. 2020. Opportunities and challenges in long-read sequencing data analysis. *Genome Biology* 21: 30.
- Bao Z, Li C, Li G, Wang P, Peng Z, Cheng L, Li H, Zhang Z, Li Y, Huang W. 2022. Genome architecture and tetrasomic inheritance of autotetraploid potato. *Molecular Plant* 15: 1211–1226.
- Bekker J, Davis J. 2020. Learning from positive and unlabeled data: a survey. *Machine Learning* 109: 719–760.
- Blake D, Lynch KW. 2021. The three as: alternative splicing, alternative polyadenylation and their impact on apoptosis in immune function. *Immunological Reviews* 304: 30–50.
- Chang H, Lim J, Ha M, Kim VN. 2014. TAIL-seq: genome-wide determination of poly(A) tail length and 3' end modifications. *Molecular Cell* 53: 1044–1052.
- Chaudhary S, Khokhar W, Jabre I, Reddy AS, Byrne LJ, Wilson CM, Syed NH. 2019. Alternative splicing and protein diversity: plants versus animals. *Frontiers in Plant Science* 10: 708.
- Chen Q, Han Y, Liu H, Wang X, Sun J, Zhao B, Li W, Tian J, Liang Y, Yan J *et al.* 2018. Genome-wide association analyses reveal the importance of alternative splicing in diversifying gene function and regulating phenotypic variation in maize. *Plant Cell* 30: 1404–1423.
- Chen S, Zhou Y, Chen Y, Gu J. 2018. FASTP: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34: i884–i890.
- Clavijo BJ, Venturini L, Schudoma C, Accinelli GG, Kaithakottil G, Wright J, Borrill P, Kettleborough G, Heavens D, Chapman H. 2017. An improved assembly and annotation of the allohexaploid wheat genome identifies



- complete families of agronomic genes and provides genomic evidence for chromosomal translocations. *Genome Research* 27: 885–896.
- Demirdjian L, Xu Y, Bahrami-Samani E, Pan Y, Stein S, Xie Z, Park E, Wu YN, Xing Y. 2020. Detecting allele-specific alternative splicing from population-scale RNA-seq data. *American Journal of Human Genetics* 107: 461–472.
- Feng Y-Y, Ramu A, Cotto KC, Skidmore ZL, Kunisaki J, Conrad DF, Lin Y, Chapman W, Uppaluri R, Govindan R. 2018. REGTOOLS: integrated analysis of genomic and transcriptomic data for discovery of splicing variants in cancer. *bioRxiv*. doi: [10.1101/436634](https://doi.org/10.1101/436634).
- de la Fuente L, Arzalluz-Luque A, Tardaguila M, Del Risco H, Marti C, Tarazona S, Salguero P, Scott R, Lerma A, Alastrue-Agudo A *et al.* 2020. TAPPAS: a comprehensive computational framework for the analysis of the functional impact of differential splicing. *Genome Biology* 21: 119.
- Guo W, Tzioutziou NA, Stephen G, Milne I, Calixto CP, Waugh R, Brown JW, Zhang R. 2021. 3D RNA-seq: a powerful and flexible tool for rapid and accurate differential expression and alternative splicing analysis of RNA-seq data for biologists. *RNA Biology* 18: 1574–1587.
- Hahne F, Ivanek R. 2016. Visualizing genomic data using Gviz and bioconductor. *Statistical Genomics: Methods and Protocols* 1418: 335–351.
- Hu Y, Fang L, Chen X, Zhong JF, Li M, Wang K. 2021. LIQA: long-read isoform quantification and analysis. *Genome Biology* 22: 182.
- Jacob AG, Smith CW. 2017. Intron retention as a component of regulated gene expression programs. *Human Genetics* 136: 1043–1057.
- Jia J, Long Y, Zhang H, Li Z, Liu Z, Zhao Y, Lu D, Jin X, Deng X, Xia R. 2020. Post-transcriptional splicing of nascent RNA contributes to widespread intron retention in plants. *Nature Plants* 6: 780–788.
- Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G. 2014. INTERPROSCAN 5: genome-scale protein function classification. *Bioinformatics* 30: 1236–1240.
- Keren H, Lev-Maor G, Ast G. 2010. Alternative splicing and evolution: diversification, exon definition and function. *Nature Reviews Genetics* 11: 345–355.
- Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology* 37: 907–915.
- Kolde R. 2012. PHEATMAP: pretty heatmaps. *R Package Version* 1: 726.
- Kolosov N, Daly MJ, Artomov M. 2021. Prioritization of disease genes from GWAS using ensemble-based positive-unlabeled learning. *European Journal of Human Genetics* 29: 1527–1535.
- Lan W, Li M, Zhao K, Liu J, Wu F-X, Pan Y, Wang J. 2017. LDAP: a web server for lncRNA-disease association prediction. *Bioinformatics* 33: 458–460.
- Lê S, Josse J, Husson F. 2008. FACTOMINE: an R package for multivariate analysis. *Journal of Statistical Software* 25: 1–18.
- Li H. 2018. MINIMAP2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34: 3094–3100.
- Li X, Li Y-x, Chen L, Wu X, Qin W, Song Y, Zhang D, Wang T, Li Y, Shi Y. 2016. Fine mapping of qKW7, a major QTL for kernel weight and kernel width in maize, confirmed by the combined analytic approaches of linkage and association analysis. *Euphytica* 210: 221–232.
- Liao Y, Smyth GK, Shi W. 2014. FEATURECOUNTS: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30: 923–930.
- Lienhard M, van den Beucken T, Timmermann B, Hochradel M, Boerno S, Caiment F, Vingron M, Herwig R. 2023. IsoTOOLS: a flexible workflow for long-read transcriptome sequencing analysis. *Bioinformatics* 39: btad364.
- Lima SA, Chipman LB, Nicholson AL, Chen YH, Yee BA, Yeo GW, Collier J, Pasquinelli AE. 2017. Short poly(A) tails are a conserved feature of highly expressed genes. *Nature Structural & Molecular Biology* 24: 1057–1063.
- Liu B, Liu J, Xiao Y, Chen Q, Wang K, Huang R, Li L. 2022. A new self-paced learning method for privilege-based positive and unlabeled learning. *Information Sciences* 609: 996–1009.
- Liu K, Jia S, Du Q, Zhang C. 2017. NanoAsPipe: a transcriptome analysis and alternative splicing detection pipeline for MinION long-read RNA-seq. In: *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Kansas City, MO, USA: IEEE, 1823–1826.
- Liu Y, Nie H, Liu H, Lu F. 2019. Poly (A) inclusive RNA isoform sequencing (PAIso-seq) reveals wide-spread non-adenosine residues within RNA poly (A) tails. *Nature Communications* 10: 1–13.
- Long Y, Jia J, Mo W, Jin X, Zhai J. 2021. FLEP-seq: simultaneous detection of RNA polymerase II position, splicing status, polyadenylation site and poly (A) tail length at genome-wide scale by single-molecule nascent RNA sequencing. *Nature Protocols* 16: 4355–4381.
- Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* 15: 1–21.
- Marasco LE, Kornblihtt AR. 2023. The physiology of alternative splicing. *Nature Reviews Molecular Cell Biology* 24: 242–254.
- Marquez Y, Brown JW, Simpson C, Barta A, Kalyna M. 2012. Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. *Genome Research* 22: 1184–1195.
- Mei W, Liu S, Schnable JC, Yeh C-T, Springer NM, Schnable PS, Barbazuk WB. 2017. A comprehensive analysis of alternative splicing in paleopolyploid maize. *Frontiers in Plant Science* 8: 694.
- Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics* 40: 1413–1415.
- Park E, Pan Z, Zhang Z, Lin L, Xing Y. 2018. The expanding landscape of alternative splicing variation in human populations. *The American Journal of Human Genetics* 102: 11–26.
- Passmore LA, Collier J. 2021. Roles of mRNA poly (A) tails in regulation of eukaryotic gene expression. *Nature Reviews Molecular Cell Biology* 23: 1–14.
- Pimentel H, Parra M, Gee SL, Mohandas N, Pachter L, Conboy JG. 2016. A dynamic intron retention program enriched in RNA processing genes regulates gene expression during terminal erythropoiesis. *Nucleic Acids Research* 44: 838–851.
- Proudfoot NJ. 2011. Ending the message: poly (A) signals then and now. *Genes & Development* 25: 1770–1782.
- Qi H, Guo X, Wang T, Zhang Z. 2022. ASTOOL: an easy-to-use tool to accurately identify alternative splicing events from plant RNA-Seq data. *International Journal of Molecular Sciences* 23: 4079.
- Rogers MF, Thomas J, Reddy AS, Ben-Hur A. 2012. SPLICEGRAPHER: detecting patterns of alternative splicing from RNA-Seq data in the context of gene models and EST data. *Genome Biology* 13: R4.
- Shang X, Cao Y, Ma L. 2017. Alternative splicing in plant genes: a means of regulating the environmental fitness of plants. *International Journal of Molecular Sciences* 18: 432.
- Shen S, Park JW, Lu Z-x, Lin L, Henry MD, Wu YN, Zhou Q, Xing Y. 2014. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proceedings of the National Academy of Sciences, USA* 111: E5593–E5601.
- Stark R, Grzelak M, Hadfield J. 2019. RNA sequencing: the teenage years. *Nature Reviews Genetics* 20: 631–656.
- Subtelny AO, Eichhorn SW, Chen GR, Sive H, Bartel DP. 2014. Poly(A)-tail profiling reveals an embryonic switch in translational control. *Nature* 508: 66–71.
- Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D. 2008. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* 321: 956–960.
- Tang AD, Soulette CM, van Baren MJ, Hart K, Hrabeta-Robinson E, Wu CJ, Brooks AN. 2020. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nature Communications* 11: 1438.
- Tapia J, Ha KC, Sterne-Weiler T, Gohr A, Braunschweig U, Hermoso-Pulido A, Quesnel-Vallières M, Permyanov J, Sodaei R, Marquez Y. 2017. An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. *Genome Research* 27: 1759–1768.
- Tardaguila M, De La Fuente L, Marti C, Pereira C, Pardo-Palacios FJ, Del Risco H, Ferrell M, Mellado M, Macchietto M, Verheggen K. 2018. SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Research* 28: 396–411.
- Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2013. Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics* 14: 178–192.



- Trincado JL, Entizne JC, Hysenaj G, Singh B, Skalic M, Elliott DJ, Eyraas E. 2018. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biology* 19: 40.
- Walley JW, Sartor RC, Shen Z, Schmitz RJ, Wu KJ, Ulrich MA, Nery JR, Smith LG, Schnable JC, Ecker JR. 2016. Integration of omic networks in a developmental atlas of maize. *Science* 353: 814–818.
- Wang B, Kumar V, Olson A, Ware D. 2019. Reviving the transcriptome studies: an insight into the emergence of single-molecule transcriptome sequencing. *Frontiers in Genetics* 10: 384.
- Wang B, Tseng E, Baybayan P, Eng K, Regulski M, Jiao Y, Wang L, Olson A, Chougule K, Van Buren P. 2020. Variant phasing and haplotypic expression from long-read sequencing in maize. *Communications Biology* 3: 1–11.
- Wang F, Zhang X, Zhang L, Li J, Yue J-X. 2022. NanoTrans: an integrated computational framework for comprehensive transcriptome analyses with Nanopore direct-RNA sequencing. *BioRxiv*. doi: 10.1101/2022.11.29.518309.
- Wang JR, Holt J, McMillan L, Jones CD. 2018. FMLRC: hybrid long read error correction using an FM-index. *BMC Bioinformatics* 19: 1–11.
- Workman RE, Tang AD, Tang PS, Jain M, Tyson JR, Razaghi R, Zuzarte PC, Gilpatrick T, Payne A, Quick J *et al.* 2019. Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nature Methods* 16: 1297–1305.
- Xing H, Pudake RN, Guo G, Xing G, Hu Z, Zhang Y, Sun Q, Ni Z. 2011. Genome-wide identification and expression profiling of auxin response factor (ARF) gene family in maize. *BMC Genomics* 12: 1–13.
- Yu G, Wang L-G, Han Y, He Q-Y. 2012. CLUSTERPROFILER: an R package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology* 16: 284–287.
- Yu Z, Chen Y, Zhou Y, Zhang Y, Li M, Ouyang Y, Chebotarov D, Mauleon R, Zhao H, Xie W. 2023. Rice Gene Index: a comprehensive pan-genome database for comparative and functional genomics of Asian rice. *Molecular Plant* 16: 798–801.
- Zheng Y, Peng H, Zhang X, Zhao Z, Gao X, Li J. 2019. DDI-PULearn: a positive-unlabeled learning method for large-scale prediction of drug-drug interactions. *BMC Bioinformatics* 20: 1–12.

## Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

**Fig. S1** Principles for junction refinement of long-read mapping results using short-read data.

**Fig. S2** Evidences for isoforms filtered by GB-PU, FLAIR and SQANTI3.

**Fig. S3** Comparison of expression levels for isoforms identified by FLAIR, SQANTI3, and GB-PU using RNA-Seq data.

**Fig. S4** Performance comparison of SQANTI3, FLAIR, and GB-PU.

**Fig. S5** Functional analysis of reliable novel isoforms identified in the MCP, MDRS, and MIP datasets.

**Fig. S6** Evidence for different types of alternative splicing event of maize.

**Fig. S7** GO enrichment analysis of differentially spliced genes in three comparison groups of the MCP dataset. Emb, embryo; Endo, endosperm.

**Fig. S8** Median length distribution of ORFs and 3' UTRs of inclusive and exclusive isoforms associated with SE, A3SS, and A5SS events.

**Fig. S9** Functional differences of isoforms with different poly(A) tail lengths in the MDRS dataset.

**Fig. S10** Variant annotation and GO enrichment analysis of the allele-specific alternative splicing genes.

**Fig. S11** Allele-specific alternative splicing pattern for gene *Zm00001d020461*.

**Fig. S12** Comparison of expression levels for isoforms identified in Arabidopsis, rice, wheat, and potato.

**Fig. S13** Evidence for different type of alternative RNA processing event of Arabidopsis, rice, wheat, and potato.

**Notes S1** Details of feature selection in GB-PU, the workflow used in iFLAS to identify allele-specific alternative splicing events, the performance comparison of GB-PU, SQANTI3, and FLAIR and the wide applicability of iFLAS in isoform and alternative splicing identification for Arabidopsis, rice, wheat, and potato.

**Table S1** Description of features used in SQANTI3 and GB-PU.

**Table S2** Sequencing statistics for the MCP, MIP, and MDRS datasets.

**Table S3** Comparison of four methods for identifying novel isoforms.

**Table S4** Information of all isoforms.

**Table S5** Differential spliced genes and differential expressed genes in different comparison groups.

**Table S6** Allele-specific alternative splicing genes detected in B73 × Ki11 data.

**Table S7** Functional description of analysis modules in integrated full-length alternative splicing analysis.

**Table S8** Sequencing statistics of Arabidopsis, rice, wheat, and potato leaf transcriptome datasets.

**Table S9** Statistics of identified isoform and alternative splicing events in Arabidopsis, rice, wheat, and potato.

Please note: Wiley is not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.