This work was funded by European Research Council (DOUBLE-TROUBLE 833522).

Evolution of isoform-level gene expression patterns across tissues during lotus species divergence

Yue Zhang[1,2,3], Xingyu Yang[4], Yves Van de Peer[5,6,7], Jinming Chen[*,1,2], Kathleen Marchal[*,5,8], Tao Shi[*,1,2]

[1] CAS Key Laboratory of Aquatic Botany and Watershed Ecology, Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan 430074, China

[2] Center of Conservation Biology, Core Botanical Gardens, Chinese Academy of Sciences, Wuhan 430074, China

[3] University of Chinese Academy of Sciences, Beijing 100049, China

[4] Wuhan Institute of Landscape Architecture, Wuhan 430081, China

[5] Department of Plant Biotechnology and Bioinformatics, Ghent University, and VIB Center for Plant Systems Biology, Ghent 9052, Belgium.

[6] Centre for Microbial Ecology and Genomics, Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria 0028, South Africa.

[7] College of Horticulture, Academy for Advanced Interdisciplinary Studies, Nanjing Agricultural University, Nanjing 210095, China.

[8] Department of Information Technology, IDLab, IMEC, Ghent University, Ghent 9052, Belgium

***Corresponding authors: E-mail: jmchen@wbgcas.cn; Kathleen.Marchal@UGent.be; shitao323@wbgcas.cn**

Abstract

Both gene duplication and alternative splicing (AS) drive the functional diversity of gene products in plants, yet the relative contribution of the two key mechanisms to the evolution of gene function is largely unclear. Here, we studied AS in two closely related lotus plants, *Nelumbo lutea*, *N. nucifera*, and the outgroup *Arabidopsis thaliana*, for both single-copy and duplicated genes. We show that most splicing events evolved rapidly between orthologs, and that the origin of lineage-specific splice variants or isoforms contributed to gene functional changes during species divergence within *Nelumbo*. Single-copy genes contain more isoforms, have more AS events conserved across species, and show more complex tissue-dependent expression patterns than their duplicated counterparts. This suggests that expression divergence through isoforms is a mechanism to extend the expression breadth of genes with low copy numbers. As compared to isoforms from local, small-scale duplicated, isoforms of whole-genome duplicates are less conserved and display a less conserved tissue bias, pointing towards their contribution to subfunctionalization. Through comparative analysis of isoform expression networks, we identified orthologous genes of which the expression of at least some of their isoforms displays a conserved tissue bias across species, indicating a strong selection for maintaining a stable expression pattern of these isoforms. Overall, our study shows that both AS and gene duplication contribute to creating the diversity of gene function during the evolution of lotus.

Key words

Alternative splicing, Gene duplication, Co-expression network, Functional divergence, Nelumbo

Introduction

Alternative splicing (AS) is a post-transcriptional regulation mechanism in which multiple isoforms are generated from a single pre-mRNA, enriching RNA and protein diversity. The pre-mRNA splicing process is catalyzed by the spliceosome that recognizes genomic splice sites and hence triggers alternative splicing events (AS events). These isoforms are translated to proteins with sometimes different or even antagonistic functions (Lee and Rio, 2015; Wang, *et*

al., 2020). Hence, AS enriches an organism's functional diversity by generating multiple transcripts from the same limited gene pool (Syed, *et al.*, 2012; Kornblihtt, *et al.*, 2013; Reddy, *et al.*, 2013; Zhang *et al.*, 2017). Many tissue-specific isoforms have already been identified in plants (Wang *et al.*, 2016; DiMario *et al.*, 2017) and many AS events have been associated with abiotic and biotic stresses such as temperature, drought, salt stress, light, and pathogen infection (Jiang, *et al.*, 2017; Zhang *et al.*, 2017; Gu *et al.*, 2018; Rigo *et al.*, 2019; Tian *et al.*, 2019). Expression divergence of isoforms results in expression diversity that drives phenotypic differentiation between species (Wittkopp and Kalay, 2011; Yang and Wang, 2013; Zhou *et al.*, 2019).

Most previous studies on AS mainly emphasized the role of AS during tissue development within a single species (Thatcher *et al.*, 2016; Wang *et al.*, 2019). However, understanding how AS differentiation affects the functional divergence of genes is equally important (Yuan *et al.*, 2009; Thatcher *et al.*, 2014). Few studies analyzed how the expression of isoforms changes between orthologs during lineage divergence and/or contributed to the functional divergence between paralogs (duplicated genes). Previous studies, based on human ESTs described how AS events were conserved more frequently in single-copy than in duplicated genes in general (Su *et al.*, 2006). However, these studies did not account for the difference in origin of the duplicated genes. This difference in origin of duplicates, i.e., whether they result from a small-scale or large-scale (for instance whole-genome duplication, WGD) duplication event (Lan *et al.*, 2017; Van de Peer *et al.*, 2017) might impose differences in functional constraints and therefore might result in differential retention of AS events.

Isoform-level coexpression networks (Li *et al.*, 2014; Zhang *et al.*, 2020) that capture the degree to which isoforms from the same or different genes exhibit the same expression behavior are ideal to study the expression of isoforms across different conditions and/or species. Therefore, we performed a comparative analysis of tissue-specific isoform coexpression networks obtained from related plant taxa to unveil the role of AS in the evolution of gene function and to understand the relationship between AS and gene duplication.

We hereby focused on *Nelumbo*, a plant genus commonly known as lotus with two extant species, sacred lotus (*Nelumbo nucifera* Gaertn.) and American lotus (*Nelumbo lutea* Pers. or yellow-flower lotus). Both species share the same chromosome number (2n=16) with an

estimated divergence time of 1.50-11.81 million years (Xue *et al.*, 2012; Wu *et al.*, 2014; Cao *et al.*, 2016). *N. nucifera* is one of the few angiosperms that has undergone a single whole-genome duplication (WGD), making it a relatively simple model to investigate the evolution of AS following a WGD (Qiao *et al.*, 2019; Shi *et al.*, 2020). *N nucifera* also has a well-annotated genome and extensive gene isoform expression datasets (Zhang *et al.*, 2019; Shi *et al.*, 2020). To perform a reliable interspecific comparison of AS events and isoform expression between both *Nelumbo* species, we first improved gene and isoform annotations in *N. lutea* by complementing the already available transcriptome data with additional long (Nanopore transcriptome sequencing) and short-read sequencing (Illumina RNA-seq). Using *Arabidopsis thaliana* as an outgroup allowed further elucidation of which of the AS events, that were found to be conserved between both *Nelumbo* species, were also conserved over a longer evolutionary time.

Results

Full-length transcriptome sequencing and isoform identification in *N. lutea*

To obtain full-length transcriptome and isoform information for *N. lutea*, we used a combination of long- and short-read data (Li et al., 2017; Wang et al., 2017; Wang et al., 2018). A compendium of possible isoforms in *N. lutea* was obtained by subjecting a pool of high-quality RNAs obtained from 18 samples covering different developmental stages to long-read based nanopore sequencing. This resulted in 29,081,500 reads (3.1 billion bp) with an N50 of 1,202 bp (Table 1). After removing low-quality reads and rRNA sequences, a total of 28,827,854 reads were subdivided into 23,432,326 (81.28%) full-length non-chimeric (FLNC) reads and 5,395,528 (18.82%) non full-length (NFL) reads (Table 1). To perform error correction of the FLNCs, the NFL reads were combined with additional Illumina-based short-read sequence data obtained from the RNA of the same 18 tissue samples (see materials and methods). After error corrections and clustering (Fig. 1a), a total of 151,988 full-length consensus isoforms with an average length of 1211 bp and an N50 of 1482 bp were obtained from the long-read data (Fig. S1).

We also subjected 18 tissue samples to Illumina RNA-seq with two biological repeats per tissue sample (non-pooled), and the generated Illumina short reads were mapped onto the reference genome to obtain Illumina unique transcripts and for further expression analysis (Fig.

1a). After merging the Nanopore full-length consensus isoforms and Illumina unique transcripts, a total of 124,415 non-redundant transcripts (isoforms) were identified. Furthermore, non-redundant transcripts covered the full open reading frames of 38,175 gene loci distributed across the entire *N. lutea* genome (Fig. 1b). The average number of isoforms per gene was 3.25, with almost half of the genes (49.36%) containing more than two expressed isoforms (Fig. S2). For each consensus isoform, we quantified its expression in the samples of the 18 different developmental stages from different tissues using the Illumina reads.

Rapid evolution of splicing events within and across species

To study the evolution of alternative splicing, we compared isoforms between the closely related species *N. lutea*, *N. nucifera,* and the outgroup *Arabidopsis,* which has the greatest resource of transcriptomes and functional assay data among plants. Hereto, we relied on the well-annotated transcript isoforms in *N. lutea* obtained from this study and the isoforms in *N. nucifera* obtained from our previous study (Zhang *et al.*, 2019; Shi *et al.*, 2020). In addition, we improved the annotation of transcript isoforms in *Arabidopsis* using 66 published RNA-seq samples (Table S1) (Klepikova *et al.*, 2016). For each of the three species, we summarized the number of detected non-redundant isoforms, the average number of isoforms per gene, and the number of AS events (locations in the genome for which a discrepancy is found between the transcript and the genome sequence) giving rise to these isoforms (Table S2). Isoforms that resulted from more than two different alternative splicing events significantly outnumbered the isoforms that resulted from one AS event (chi-square test, p-value < 0.01) (Fig. 2a). This was valid for all three species, suggesting that most isoforms were derived from genes by undergoing multiple AS events (Fig. 2a). Seven major alternative types of AS events, classified as alternative 3' splice sites (A3), alternative 5' splice sites (A5), retained intron (RI), skipping exon (SE), mutually exclusive exons (MX), alternative first exons (AF), and alternative last exons (AL) (see Figure 2b for the splicing patterns) were identified in either of the species. The most common type of alternative splicing occurring in all three species was RI, followed by A3 and A5, consistent with what is observed in other plants (Mandadi and Scholthof, 2015; Wang *et al.*, 2018).

To study the extent to which AS events were conserved between *N. lutea, N. nucifera,* and *Arabidopsis*, we defined 10,245 orthologous gene groups (OGs) containing genes from all three species (see experimental procedures). For each OG we investigated whether isoforms of the orthologous genes shared conserved AS events. Hereto, in a pairwise fashion, we compared orthologous genes from different species belonging to the same OG and assessed whether they share the same AS events (see experimental procedures, Fig. S3). We identified 2420 interspecies conserved AS events between *N. lutea*, *N. nucifera*, and *Arabidopsis* (Table 2, Fig. S3). Remarkably, only 4 AS events are conserved among all three species. Expectedly, the two lotus species have more interspecies conserved AS events (97.6%) (Table 2). Again, most of the AS events that are conserved across species resulted from the following types: RI (37.43%) and A3 (28.51%) (Table 2). To validate the accuracy of our identified RNA-seq-based alternative splicing events, we randomly selected 20 interspecies conserved AS events from the same OG (amongst which 2 of the 4 were conserved in all three species) and validated the presence of conserved AS events in each of the relevant species using RT-PCR (Fig. 2C, Fig. S4, Table S3).

Besides the aforementioned interspecies conserved AS events, we also identified AS events shared between two paralogs. These were referred to as intraspecies conserved AS events (see experimental procedures). 517 (0.76%) intraspecies conserved AS events were identified for paralogous genes in *N. lutea*, 167 (0.39%) in *N. nucifera*, and 200 (0.21%) in *Arabidopsis* (Table S4). Calculating for each of the three species the percentage of conserved AS events (i.e. conserved at either interspecies or intraspecies level) on the total number of AS events occurring in that species (4.3% in *N. lutea*, 6.04% in *N. nucifera*, 0.28% in *Arabidopsis*) shows that only a relatively small fraction of AS events is conserved.

The tradeoff between duplication and alternative splicing as a source of functional divergence

To understand how copy number variation originating from gene duplication influences the diversity and number of AS events (Marshall *et al.*, 2018; Hurtig *et al.*, 2020), we classified genes based on their copy numbers for *N. lutea, N. nucifera*, and *Arabidopsis*, respectively, and compared how the number of AS events differ between genes with different copy numbers (Fig.

S5). To make the observed tendencies independent of the number of genes per species, we only considered genes with orthologs in each of the considered species. The number of AS events per gene decreased inversely with the gene's copy number. Significantly (Mann-Whitney *U* test, $p < 0.01$) fewer AS events were found in genes with 9 or more copies (Fig. S5), suggesting that genome duplications and splice variation tend to increase the functional diversity of a gene family in a mutually exclusive way. Our results also show that in all three species, single-copy genes underwent significantly more AS events than duplicated genes (chi-square test, p-value < 0.01, Fig. S6). In addition, single-copy genes display significantly (chi-square test, p-value < 0.01) more interspecies conserved AS events than duplicated genes, and this in both lotus species (Fig. S7). This is in line with previous findings in *Arabidopsis* (Su *et al.,* 2006; Mei *et al.*, 2017).

To investigate whether genes from different duplication origins (duplication types) (Sandve *et al.*, 2018; Shi *et al.*, 2020) differed in the number of splicing events they underwent, we classified genes according to their duplication origin and compared their number of AS events**.** From the non-orphan genes, the overall higher average number of splicing events in *Arabidopsis* than in the lotus species for the different classes of duplicate genes is at least partially due to the better annotation in *Arabidopsis* for which relatively more isoforms have been described (Fig. S8). As compared to genes resulting from WGD and to dispersed duplicates, local duplicates (including proximal and tandem duplicates) showed on average a lower number of AS events, an observation that holds in all three species (Fig. S8). To explore the level of intraspecies conserved AS events in duplicates of different origins, we classified paralogs according to their origin of duplication and within each class compared whether an AS event observed in a duplicated gene was also present in its closest paralog with the same origin of duplication (Fig. S9a). What is in common between the three species is that on average, proximal duplicate pairs tend to have more frequently conserved AS events than other duplicate pairs (chi-square test, p-value < 0.05) (Fig. S9b). Because of their generally younger age as compared to WGD or dispersed duplicates, relatively more proximal duplicates have survived pruning selection, and hence more redundant gene variants are retained. The duplicate orphan gene pairs contained no cases with at least one conserved AS event in either of the three species, suggesting that divergence of AS variation occurred relatively fast in orphan genes (Fig. S9b).

Lineage-specific isoforms as a source of divergence between closely related species

Because we used for each of the closely related lotus species a similar transcriptome sequencing strategy with a full-length RNA pool of similar composition (Zhang *et al.*, 2019), differences in isoforms observed between both lotuses can be attributed to lineage rather than tissue specificity. To identify the lineage-specific evolution of isoforms in both lotuses, we relied on the orthologous gene pairs (Fig. 3a) and identified conserved isoforms and lineage-specific isoforms (see experimental procedures). These lineage-specific isoforms were assumed to be newly generated during lineage evolution. To prevent a bias towards false-positive lineage-specific isoforms, we removed isoforms supported by less than two RNA-seq samples and encoding short CDS (see experimental procedures). This resulted in 11,592 lineage-specific isoforms covering 4,854 genes in *N. nucifera* and 14,221 lineage-specific isoforms covering 4,424 genes in *N. lutea*. Figure 3b shows a representative example of lineage-specific isoforms of respectively *N. lutea* and *N. nucifera*. Gene Ontology (GO) enrichment analysis on the set of genes containing *N. lutea* lineage-specific isoforms, showed enrichment in metabolic processes, related to “carbohydrate metabolism”, “lipid metabolism”, and “protein metabolism” (Fig. 3c). To exclude these lineage-specific isoforms were unique to our accessions, we verified the lineage-specificity of these isoforms on an independent dataset. Hereto we collected public RNA-seq data of the *N. nucifera* rhizome. More than 91.7% of the lineage-specific isoforms we identified in our *N. nucifera* cultivar could be detected in the data of this independent cultivar, suggesting that most of these lineage-specific isoforms were shared among individuals of the same species (Fig. S10). To investigate the expression pattern of lineage-specific isoforms between multiple tissues in *N. nucifera* and *N. lutea*, the tissue expression heatmap of lineage-specific isoforms shows that although most lineage-specific isoforms were expressed in multiple tissues, some display a high expression level in a specific tissue (see experimental procedures, Figs. 3d-e). Furthermore, as compared to interspecies conserved isoforms derived from the same genes, the lineage-specific isoforms showed a significantly (*t*-test, $p < 0.01$) higher tissue-specificity in both *N. lutea* and *N. nucifera* **(**Fig. 3f, Tau index). These findings suggest that during evolution, numerous lineage-specific isoforms originated in closely related lotus species, some of which display a highly tissue-specific expression.

Isoforms of the same gene often show expression divergence across tissues

To further study functional divergence between isoforms (Zhang *et al.*, 2020), we build for each species an isoform-resolved WGCNA co-expression network (Fig. S11, Table S6) which was clustered in coexpression modules. We subsequently selected for each species module sets that was representative of each of the profiled tissues (see experimental procedures) (Fig. 4a, Fig. S11). In this way, we could study tissue divergence of expression between isoforms of the same gene within and across species as a proxy of a gene's functional divergence. For each gene, we defined a 'polymorphism value' (PV) which measures the degree to which isoforms derived from the same gene in the same species are part of module sets representative of different tissues (Fig. 4a). A high PV indicates that the different isoforms of a gene belong to module sets representative of many different tissues and hence indicate a diversification of the gene's function in that species through alternative splicing. For most genes, at least two isoforms have a high PV (> 0.5), an observation that holds in each of the studied species (56.15% in *N. lutea*, 60.06% in *N. nucifera*, and 54.46% in *Arabidopsis*) with (Fig. S12). This indicates that during tissue development, the functional divergence of most genes was likely achieved by transcribing isoforms with different tissue-specific expression patterns.

Conservation of tissue-specific expression of isoforms from paralogous genes

We used the above-developed metrics (PV) to assess the degree to which the expression divergence of isoforms depends in case of paralogs on their duplication origin. Hereto we again grouped genes that underwent a duplication according to their origin of duplication and calculated a PV level for all these genes. Figs. 4d-d shows the distribution of these PVs, conditioned on their type of duplication. The average PV of isoforms of single-copy genes was higher than the PV of isoforms of duplicated genes in both Nelumbo species, being particularly significant for *N. nucifera* (Mann Whitney *U* test, $p < 0.01$, Figs 4b-d). Although this pattern was not observed in *Arabidopsis* (Fig. 4d). It is in line with conserved single-copy genes showing relatively large tissue expression diversity (De Smet *et al.*, 2013; Shi *et al.*, 2020). Except for tandem duplicated genes in *N. lutea*, the PV levels of isoforms of proximal and tandem duplications were lower than the PVs of isoforms of WGD/segmental and dispersed

duplicates in the three species (Figs 4b-d). This suggests that isoforms of local duplicates exhibit less expression divergence across tissues than isoforms originating from other types of duplication. This is consistent with the higher observed tissue specificity of gene expression in local duplicates (Shi *et al.*, 2020).

We also investigated whether the degree to which isoforms of paralogs show a bias towards the same tissues depends on their duplication origin. To perform this analysis we distinguished in each of the species between single-tissue and multiple-tissue bias genes, based on the number of tissue-specific module sets the isoforms of a gene belong to (see experimental procedures). The analyses were performed for paralogous genes that either both showed a single-tissue bias or both a multiple tissue bias. For paralogous genes that both show a single-tissue bias, we assessed whether all their isoforms belonged to the module set representative of the same tissue in a species (Fig. 4e). If so, the isoforms of these paralogs were said to have a conserved tissue bias in expression. We noticed that isoforms of paralogs originating from local duplications showed more frequently conserved tissue bias (chi-square test, $p < 0.05$) than isoforms of paralogs originating from WGD/segmental and dispersed duplication events (Fig. 4f-h). This did not hold true for isoforms of paralogs originating from tandem duplications in *N. lutea*. Conclusively, isoforms of paralogs originating from local duplications are more often relevant to the same single tissue.

For paralogous genes that both show a multiple-tissue bias, we calculated for each paralog the fraction of module sets to which isoforms of the query paralog and its closest relative were both assigned versus the total number of module sets to which the query paralog was assigned (see experimental procedures, Fig. 4i). The larger this fraction, the more the expression of the isoforms of both paralogs shows a conserved tissue bias. However, no significant (Mann-Whitney *U* test, $p> 0.05$) difference in this fraction between duplicates of different origins was observed. Overall, we noticed that in all three species the fraction of tissue-specific module sets to which the isoforms of both paralogs were commonly assigned versus the total number of module sets was low on average (36.35% in *Arabidopsis*, 24.24% in *N. nucifera*, and 21.63% in *N. lutea*), indicating that the occurrence of tissue expression divergence after gene duplication is at least partially mediated by differences in isoform expression (Figs 4j-l). This was further confirmed by the fact that the number of isoforms in genes displaying multiple-

tissue bias was significantly higher than in single-tissue bias genes, and this for each of the three species ((Mann-Whitney *U* test, $p<0.01$), Fig. S13).

Conservation of tissue-specific expression patterns of isoforms from orthologous genes

To assess whether isoforms of orthologs displayed conserved tissue-specific expression bias, we relied on the definition of matching tissues (see experimental procedures). Comparing the similarity between the expression patterns at the level of orthologous genes, hereby not yet considering isoforms (see experimental procedures). Our results suggest that the expression patterns of orthologs are significantly (Mann-Whiney *U* test, p-value < 0.05) more similar between conditions reflecting matching than between conditions reflecting non-matching tissues in all considered pairwise comparisons of the species investigated (Fig. S14). Therefore, we used the concept of matching tissues to further investigate whether isoforms from genes belonging to the same OG (orthologous group) exhibit the same tissue bias in expression across the different species: more specifically we assess whether isoforms of orthologous genes belong to the module sets in either species that are representative of the same matching tissues. Also here we distinguished two cases: a distinction was made between orthologous genes of which the isoforms display a bias towards a single tissue and orthologous genes of which the isoforms display a bias towards multiple tissues (Fig. S15a). From all pairwise comparisons between orthologs, we identified 10,394 sub-groups of OGs consisting of orthologs of which the isoforms displayed a conserved tissue bias (see experimental procedures, Fig. S15b) in at least two of the considered species. These include both cases that show conserved single- and multiple-tissue bias (Table S7). Figure S16 shows representative sub-groups containing isoforms of orthologous genes for which the expression shows conserved tissue bias towards the same matching tissue in three investigated species. Such cases were observed for both cases of orthologs with solely single tissue or solely multiple tissue bias.

Differentiation of tissue-specific expression patterns across orthologous isoforms associated with floral organ specification

To investigate whether differential tissue-specific expression between isoforms of orthologs contributed to changes in floral organs, we focused on the genes of the floral 'ABCE Model'

that regulate the formation of floral organs (Soltis *et al.*, 2007). In line with previous studies (Ó'Maoiléidigh *et al.*, 2014; Zhang *et al.*, 2019), we identified genes involved in the 'ABCE Model' using phylogenetic analysis of the MADS-box gene family across the three species (Fig. S17). As different combinations of 'ABCE Model' genes contribute to forming a particular floral tissue, we assumed that the expression of all isoforms of genes in a particular combination should be expressed in the same floral tissues and that this tissue-specific expression pattern should be conserved across the species investigated. However, analysis of the degree to which the isoforms of genes belonging to the 'ABCE Model' exhibit biases towards the same tissue (matching tissue representative modules) in each of the species shows that some isoforms of well-characterized genes show an expression bias toward tissue-specific module sets that are representative for floral tissues, other than the ones the genes have been characterized for (Fig. 5, Fig. S18). In *Arabidopsis* and *N. nucifera*, isoforms (three in *Arabidopsis* and one in *N. nucifera*) from A-class genes, known to be specifically expressed in sepal and petal also showed a tissue-specific expression in carpel and one isoform from the B-class genes which are supposed to be only expressed in petal and stamen shows a tissue-specific expression in sepal (Figs. 5, Fig. S18). These results suggest that alternative splicing and divergence in expression between isoforms of the genes in the 'ABCE Model' contributes to the formation of floral tissue. Interestingly, these unexpected tissue biases in expression of the isoforms of the ABCE model genes tend to be conserved between *N. nucifera* and *Arabidopsis*. In addition, in both lotuses but not in Arabidopsis, some of the isoforms of A-class genes showed a tissue-specific expression in the receptacle, suggesting that A-class isoforms might regulate the formation of lotus-specific metamorphosis receptacles (Fig. 5). Overall, our results show that subtle changes in the floral organ-specific expression of isoforms of the 'ABCE Model' genes might have implications for the specific characteristics of the floral organs in lotus.

Discussion

Here we studied how the conservation and expression of isoforms change between orthologs during lineage divergence and contributed to the functional divergence between paralogs (duplicated genes). We hereby assumed that isoforms of paralogs would evolve differently depending on the duplication origin of the paralogs. To perform this analysis we focused on

Lotus, as this system with its single ancient WGD (Shi *et al.*, 2020) is ideal to compare the evolution of alternative splicing, after either a WGD or after small-scale duplication events. To allow for a comprehensive study of the evolution of alternative splicing and its effect on gene expression in both the short (within-genus level) and longer-term (between a core eudicot and a basal eudicot) we included *Arabidopsis* as an outgroup in the analysis.

Systematically analyzing seven main AS types in the studied plant species, showed that most isoforms undergo more than two AS events. We identified 2,420 interspecies conserved AS events between orthologous gene pairs of different comparisons of *N. nucifera, N. lutea*, and *A. thaliana.* In line with previous studies, closely related species share more conserved AS events (Mei *et al.*, 2017), and only four AS events were conserved in all three species. This suggests that over time AS events are lost or are progressively being replaced by new AS variants in orthologous genes. When focusing on the faith of AS and isoforms after gene duplication, we found that the number of AS events is inversely related to the number of gene copies and that relatively more single-copy genes undergo AS events than genes with more copies. In addition**,** single-copy genes contain not only more isoforms than their duplicated counterparts, but also their AS events tend to be more often conserved across species. The fact that relatively fewer AS events occur in multicopy genes and that their AS events are less conserved might imply their functional decay or/and indicate that a balance exists between increasing genome complexity through either duplication versus alternative splicing. These observations are in line with the hypothesis that the number of AS events are gradually reduced after gene duplication, especially in large gene families to avoid producing functionally redundant isoforms and wasting resources in cells (Su *et al.*, 2006; Talavera *et al.*, 2007). Along the same lines, duplicates might inherit the splicing isoforms from their ancestral gene. If these isoforms had different functions, differential isoform retention can occur where isoforms with different biological functions are differentially retained between the paralogs. In such cases, the divergence between the isoforms of paralogs might be associated with the retention of the paralogs themselves through subfunctionalization (Santos *et al.*, 2011; Su and Gu, 2012). Although the absolute number of AS events depends on the species, we found that overall, local duplicates have a lower number of AS events than WGD or dispersed duplicates. In addition, the AS events of local duplicates tend to be more often conserved between paralogs than those

of WGD or dispersed duplicates. The latter observation is in line with what was observed in the human genome (Roux and Robinson-Rechavi, 2011). These results suggest that local duplicates, particularly proximal duplicates, inherit the splicing isoforms from their ancestral genes more strictly than WGD or dispersed duplicates, resulting in fewer, but more conserved AS events.

Previous studies showed how isoforms with species-specific expression resulted in functional divergence between proteins of the same family and hence drove divergence of phenotypic traits between species (Thatcher *et al.*, 2014; Zhang *et al.*, 2017; Smith *et al.*, 2018; Huang *et al.*, 2021; Smith *et al.*, 2021). Also, when comparing isoforms between the two *Nelumbo species*, we identified 4,854 *N. nucifera* genes and 4,424 *N. lutea* genes that gave rise to lineage-specific isoforms in each of their respective lineages. These *N. nucifera* lineage-specific isoforms were shown to also be present in related cultivars of the same lineage, further confirming their lineage specificity. The genes containing these lineage-specific isoforms were enriched in biological functions associated with phenotypic traits specific to the lotus. Herein our results show how lineage-specific isoforms can drive phenotypic differences between closely related lineages during evolution, in line with previous findings in other closely related species (Thatcher *et al.*, 2014; Zhang *et al*., 2017; Huang *et al.*, 2021).

Analyzing the tissue-specific isoform resolved coexpression networks, inferred for each of the three species, showed that for more than half of the genes their isoforms exhibit expression patterns biased towards different tissues (high PV value). This indicates that many genes extend their expression breath through their isoforms. Isoforms of single-copy genes in lotus tend to show a larger tissue divergence in expression than duplicated genes. This further suggests that expression divergence through isoforms is a mechanism to extend the expression breadth of genes with low copy numbers. We also observed that, in general, isoforms from local duplicates exhibit less expression divergence across tissues than isoforms resulting from WGD and tend to have the same tissue-specific expression bias. Overall, we noticed that irrespective of the number of isoforms, the expression of isoforms of duplicated genes rarely showed expression bias towards the same tissues (Figs 4j-l).

In contrast to previous isoform expression network analysis (Li *et al.*, 2014; Ma *et al.*, 2020), our analysis allows performing a comparison of the degree to which orthologous isoforms display the same pattern of tissue bias across species. We identified 10,394

orthologous sub-groups for which isoforms of the orthologous genes display a conserved tissue bias in at least two of the species analyzed. The presence of such orthologous isoforms with conserved tissue bias across species indicates a strong selection for maintaining conserved AS. When compared to isoforms of orthologous genes that show conserved multiple-tissue bias, we observed less conservation in expression behavior and tissue bias than for orthologs with conserved single-tissue bias. In addition, analysis of the genes and isoforms of the floral 'ABCE' gene model in *Arabidopsis*, *N. nucifera*, and *N. lutea* show that some isoforms show unexpected floral organ-specific expression. This indicates that the expression of the 'ABCE' gene model is more complex at the isoform level than what has been described at the gene level. The formation of the lotus-specific floral tissue receptacle might be the result of combinations of isoforms from A-class genes, expressed in a tissue-specific way that is unique to Lotus.

Conclusively, our results illustrate that a tradeoff exists between AS and gene duplication in driving the evolution of gene function. In general, single-copy genes tend to have more AS events, more conserved AS events, and the expression of their isoforms is more biased towards multiple tissues than that of other duplicates. This indicates that these isoforms tend to be generally important across tissues and conserved and drive the functional diversity of single-copy genes. From all duplicated genes, local duplicates have the lowest number of AS events and tend to have more often at least one isoform that is conserved between the paralogs. In addition, these isoforms display a less divergent expression bias (more biased towards a single tissue). This indicates that either more functional constraints are in place that keep the isoforms of both copies functionally similar or, alternatively that functional decay is not yet complete. Duplicates resulting from WGD display relatively fewer AS events and show isoforms of which the expression pattern and tissue bias are the least conserved, pointing towards subfunctionalization.

Experimental procedures

Plant materials

The seeds of *N. lutea* were collected from Pataula Creek, Fort Gaines, Georgia, USA (85°03'W, 31°45'N), and were cultivated in tanks in the Wuhan Botanical Garden Wuhan, China (114°30' E, 30°60' N). Leaf and petiole samples were collected before the flowering stage. Petal, sepal,

immature receptacle, immature stamen, and unpollinated carpel were collected before the blooming day (0 d post-anthesis, DPA), while the pollinated carpels were collected 12 hours after hand pollination. The mature stamen and mature receptacle were collected at 2 DPA, seed-coat at 12 DPA, embryo at 15 DPA, and cotyledon at 12 and 15 DPA. Root and rhizome (internode, node, and apical meristem) were collected after flowering. Overall, a total of 18 tissue samples from the different developmental stages (two biological repeats per tissue) were collected and frozen in liquid nitrogen.

RNA extraction and Illumina sequencing in *N. lutea*

For each sample, total RNA was extracted using the RNAprep pure Plant Kit (TIANGEN). After quality checking, the RNA of each sample was used to construct Illumina cDNA libraries following the recommendations of NEBNext Ultra^{TM} RNA Library Prep Kit for Illumina (NEB, USA). The sequencing of the cDNA libraries was performed using the Illumina HiSeq2000 with 150 bp paired-end reads.

Nanopore full-length transcript sequencing of *N. lutea* RNAs

The mixed RNA from each of the 18 samples was pooled in an equal quantity to obtain 1 ug RNA as required for the Oxford Nanopore library preparation. SuperScript IV First-Strand Synthesis System (Invitrogen) was used for full-length mRNA reverse transcription. According to the Oxford Nanopore recommended protocol, barcodes (Oxford Nanopore Technologies, ONT) were added to the RNA pool during cDNA amplification. The pooled cDNA was further end-repaired and dA-tailed using NEBNext Ultra End repair/dA-tailing module (NEB) and adaptor ligation using the T4 DNA ligase (NEB). Oxford Nanopore Technologies adapters were ligated to cDNA in a reaction containing adapter mix AMX1D (ONT) and Blunt/TA Ligase Master Mix (NEB). Libraries were purified by using Agencourt AMPure XP beads and eluted by Adapter Bead Binding buffer (ONT) and Elution Buffer (ONT). Libraries were then mixed with Fuel mix and Running buffer provided by ONT. Then, the final cDNA libraries were sequenced on one FLO-MIN109 flowcell and run on the PromethION platform.

Gene and isoform identification in *Nelumbo* species

The Illumina clean short reads were aligned to *the N. lutea* genome assembly (downloaded from http://nelumbo.biocloud.net) using HISAT2 (v2.1.0) with parameter -k 5 (Kim *et al.*, 2019). Mapped reads were clustered into transcripts using StringTie v1.3.5. The Illumina unique transcript annotations obtained from the different samples were merged using TACO (Niknafs *et al.*, 2017). For the Nanopore long-read data, short-length (< 500 bp) and low-quality (read quality score < 7) reads were first removed. To further reduce the transcriptomic noises, we removed rRNA from the ONT reads, using information obtained by mapping the reads to the rRNA dataset in Pfam (http://rfam.xfam.org/) using Minimap2 (Li 2018). The obtained clean Nanopore long reads were subsequently classified based on the presence of the Nanopore adapters at both ends of the reads into respectively full-length non-chimeric (FLNC) and non-full-length reads (NFL). The obtained FLNC reads were error corrected using both the NFL reads and the Illumina short-reads (a merged dataset) with Pilon (v1.22), similar to the error correction strategy employed in previous studies (Li *et al.*, 2017; Wang *et al.*, 2017; Wang *et al.*, 2018). Corrected FLNC reads were mapped to the reference genome using Minimap2 with parameters -p 1.0 -N 100, and the mapped results were further filtered out multi-loci mappings using Samtools with parameter -flag 2304 to obtain clusters of FLNC transcripts (Li, 2018). These parameters guarantee a very strict mapping setting so that for each transcript only the best mapping location is withheld. As the 5' end sequence of transcripts may be partial degradation during the process of sequencing and multiple FLNC transcripts can map on the same location, the clusters contained redundant reads. Subsequently, Nanopore consensus isoforms were obtained by removing redundant reads in each cluster using cDNA_Cupcake and pinfish with alignment coverage < 85% and alignment identity < 0.9.

Eventually, EVidenceModeler (EVM) was used to predict confident gene models using both the Illumina unique transcripts and the Nanopore consensus isoforms. A gene model is here defined as a gene locus together with its identified isoforms. TransDecoder v5.5.0 was applied to predict the open reading frame (ORF) of isoforms based on the merged transcriptome, and only isoforms with complete ORF and with a length over 30 amino acids were retained.

The gene together with their identified isoforms based on PacBio full-length transcript sequences and Illumina reads in *N. nucifera* genome from our previous study(Zhang *et al*., 2019) were downloaded from Nelumbo Genome Database (Li *et al.*, 2021).

Isoform identification in *Arabidopsis* genome

The gene annotation (Araport11) of the *Arabidopsis* genome was downloaded from https://www.arabidopsis.org. A total of 66 *Arabidopsis* RNA-seq samples were downloaded from a previous study (Klepikova, et al. 2016) and used to improve the *Arabidopsis* isoform annotation (Table S1). The *Arabidopsis* RNA-seq samples were mapped to the reference genome (TAIR10) using HISAT2 with parameter -k 5, and mapped reads with an average of low (5.76%) multiple-mapping ratio (Table S1) were clustered into transcripts using StringTie v1.3.5. The Illumina transcript annotations of each sample were merged using TACO (merged transcriptome). The TransDecoder v5.5.0 was applied to predict the open reading frame (ORF) of isoforms based on the *Arabidopsis* merged transcriptome. Only isoforms with a complete ORF in the Araport11 gene locus were retained.

Delineation of orthogroups and classification of genes by their duplication status

Orthogroups containing homologous genes from *N. lutea*, *N. nucifera,* and *Arabidopsis* were delineated using OrthoMCL (Li *et al.*, 2003) with an e-value < 1e-15 and an inflation parameter of 2.0. Only orthogroups that contained genes from all three species were retained for downstream analyses. In addition, the CDS of *N. nucifera* genes were mapped onto the *N. lutea* genome using Blat with min identity > 0.9. The reciprocal best-mapping hits were defined as the orthologs. In total, 17,832 orthologous gene pairs were obtained. Single-copy genes and genes belonging to dispersed duplications, proximal duplications, tandem duplications, or WGD/segmental duplications in each of the three species were identified by MCScanX, as described previously in *N. nucifera* (Wang *et al.*, 2012; Shi *et al.*, 2020). Given that orphan genes (genes that have no homology to other sequenced plants) are mostly transient and lineage-specific, they were considered as a separate class (Tautz and Domazet-Lošo, 2011). To distinguish orphan from non-orphan genes, we used the gene sequences of *N. lutea*, *N. nucifera,* and *Arabidopsis* in a BlastP search against the PLAZA 4.0 database with e-value < 1e-6 after excluding *N. nucifera, N. lutea*, and *Arabidopsis thaliana*, respectively.

Analysis of alternative splicing in *N. lutea*, *N. nucifera*, and *Arabidopsis*

SUPPA2 with parameter *-f ioe* was used to identify AS events (i.e. genomic locations in the gene locus giving rise to alternative splicing), respectively based on the transcript annotations in each of the genomes of *N. lutea*, *N. nucifera,* and *Arabidopsis* (Trincado *et al.*, 2018). Seven types of AS events were distinguished: 3' alternative splicing sites (A3), 5' alternative splicing sites (A5), retained introns (RI), skipped exons (SE), alternative first exons (AF), alternative last exons (AL), and mutually exclusive exons (MX). We identified for each orthologous group AS events that were conserved between *N. lutea*, *N. nucifera*, and *Arabidopsis.* We, hereby, only focused on the AS events that affected protein-coding sequence regions and we analyzed the AS events for each splicing type separately. We used the 'splice junction-based approach' of Mei et al. (Figure S3) (Mei *et al.*, 2017): the sequences 45 bp upstream and downstream of a splice junction site (referred to as SASJs or sequences around a splice junction) were extracted, concatenated and compared between gene loci of the same orthologous group using BlastN with e-value < 1e-5. Matches less than 100 bp in length were removed. Matching pairs for which the proportion of perfectly matching bases in the target was > 80% were identified as conserved AS events. Ideally, a conserved AS event involves two genes (1:1) that originate from the same orthologous group. However, as an orthologous group can contain both orthologs and paralogs (gene copies from the same species), a distinction is made between intraspecies and interspecies conserved AS events (Figure S9a, Figure S3). If the AS event is conserved between the paralogous gene and its closest relative (best hit of the result of BlastN with e-value < $1e^{-6}$) of the same species, it is considered an intraspecies conserved AS event. An AS event is considered to be conserved at the interspecific level, if the AS events occurring in an ortholog in one species are conserved in at least one of the orthologous genes of the second species.

Identification of conserved isoforms and Nelumbo lineage-specific isoforms

To identify conserved isoforms for orthologs in both *Nelumbo* species that belong to the same OG, isoform sequences of orthologs of either species were aligned using BlastN. Conserved isoforms were defined as follows: 1) The aligned regions covered over 99% of the CDS sequence of both compared genes with a p-value < $1e^{-6}$, 2) the average expression of the orthologs (Fragments per kilobase of exon model per million mapped fragments, FPKM) was

higher than 0.1 in the investigated tissue samples of either species. To identify lineage-specific isoforms (non-conserved isoforms), we searched in the same OG for isoforms that did not meet the criteria mentioned above. In addition, the following filtering criteria were used to reduce the number of spurious lineage-specific isoforms: we removed non-conserved isoforms 1) supported by less than two RNA-seq samples in either *N. lutea* or *N. nucifera*, 2) or non-conserved isoforms that did not encode complete protein sequences. In addition, to reduce spurious signals introduced by incomplete transcripts derived from short-read Illumina data, only isoforms of which the coding CDS length covered > 10% of the longest CDS derived from the corresponding gene were retained.

Expression analysis of isoforms

The short read RNA-seq data for the different tissue samples in *N. lutea* and *Arabidopsis* used for isoform identification were also used for expression quantification at gene and isoform levels. The expression matrix of genes and isoforms in *N. nucifera* was downloaded from http://nelumbo.biocloud.net (Li *et al.*, 2021). These isoform expression matrices of both *Nelumbo* species were used to visualize the expression of lineage-specific isoforms with a heatmap (using FPKM values). For each isoform, the $\log_2$ FPKM values plotted in the heatmap correspond to the average expression values (FPKM) observed in samples belonging to the same developmental stage of a tissue.

To investigate the degree to which an isoform exhibits tissue-specific expression in a species, we used the tau index (Kryuchkova-Mostacci and Robinson-Rechavi, 2017):

$$\text{tau} = \frac{\sum_{i=1}^{n}(1 - \hat{x}_i)}{n - 1} \; ; \; \hat{x}_i = \frac{x_i}{max_{1 \le i \le n}(x_i)} \; ;$$

Where xi = log (expression of isoform in sample i) and n = the number of samples in which the isoform is expressed. A high tissue-specific expression corresponds to a high tau value.

To compare the expression behavior of genes between species, we identified 11 matching tissues between the three species based on the tissue annotation of the samples in each of the species. Per species, samples that had the same tissue annotation were grouped.

To test whether orthologous genes have a similar expression pattern, we generated per species and per gene a vector containing for the gene its average expression observed across all

samples assigned to the tissue in that species. This results per species and for each gene in a gene expression vector that contains the degree of expression of the gene in each of the tissues in the species. We pairwisely compared these species-specific expression vectors of the orthologs using the Pearson correlation. However, because the number of gene copies in one orthologous group is different for the three species, the relationship between orthologs is not always a one-to-one relation, but could also be a one-to-many or many-to-many relation. To reduce this complexity, the tissue-specific expression values of all paralogs of the same species belonging to the considered orthologous group are collapsed in one value (sum of the expression values of the paralogs within a species) prior to calculating the correlation. A high correlation coefficient indicates that the orthologs of the compared species show a highly similar expression behavior across the matching tissues in each of the respective species. We compared whether the pairwise correlations were higher when the expression vectors of two orthologs were compared between matching tissues than when they were compared between matching tissue and non-matching tissue.

Construction of isoform co-expression networks in *N. lutea*, *N. nucifera*, and *Arabidopsis*

Silent and constitutively expressed isoforms with an average FPKM < 0.1 and coefficient of variation of FPKM < 1 in all of the three species were removed from their respective isoform expression matrices. Scale-free isoform co-expression networks were constructed for *N. lutea*, *N. nucifera*, and *Arabidopsis* using the "WGCNA" package based on default settings from the tutorial (Langfelder and Horvath, 2008). Correlation analysis between isoforms was used to build a correlation matrix. Based on the scale-free topology of the network represented by the correlation matrix, a soft threshold was derived that was used to build an adjacency matrix of the final coexpression network. We used the linkage hierarchical clustering coupled with the topological overlap dissimilarity and dynamic tree cut measure to build WGCNA modules.

The number and type of tissues for which the expression was profiled in the two Nelumbo species were similar. However, for *Arabidopsis*, more RNA-Seq samples with different development stages were available for each of the profiled tissues. Having a different number of samples for a particular tissue (biased tissue representation) might affect the degree to which the co-expression modules represent the different tissues in each of the species. To study

isoform expression divergence across species (proxied by the conservation of tissue specific expression), we assigned for each species, the modules to a particular tissue (defined as tissue-specific module sets) as follows: WGCNA assigns a significance of each isoform to belong to a module based on the degree to which the expression profile of the isoform associates with the trait (tissue) vector. To identify the degree to which a module associates with a tissue, the average significance of all isoforms in one module for that specific tissue represents the relevance of the module to the tissue. Modules that were associated significantly ($p < 0.01$) to the same tissue, were selected and used to represent a particular tissue (referred to as a tissue-specific module set).

Based on the degree to which isoforms of the same gene contribute to different tissue-specific module sets a polymorphism value (PV) is defined:

$$\mathrm{PV} = \frac{N_t}{N_i} \ (N_i > 1)$$

where N_t is the total number of different tissue module sets to which the different isoforms of a gene are assigned, and N_i is the number of isoforms of the gene. This PV value captures the degree to which the isoforms of the same gene are dispersed over different module sets. The larger the value, the larger the dispersion.

For both paralogous and orthologous genes, we compared whether the expression of their isoforms was biased towards the same (matching) tissue. To perform this comparison, genes were divided into two groups, according to the number (n) of tissue-specific module sets in which the isoforms of these genes were present: single-tissue bias genes **(**n=1 indicating that all isoforms of that gene belong to the module set representative of a single tissue) and multiple-tissue bias genes (n>1 or indicating that at least some of the isoforms of the gene are present in module sets representative of different tissues). We performed these comparisons only for orthologous and paralogous gene pairs for which either both genes in the pair showed a single tissue bias or both genes showed a multiple tissue bias. A paralog of a gene is here defined as its genetically closest relative of the same species that belongs to the same OG. For isoforms of paralogous genes derived from genes that display a single-tissue bias, the tissue bias was considered 'conserved' if all of the isoforms of the paralogs belong to the same module set that represents a particular tissue. For isoforms of paralogous genes derived from genes that display

multiple-tissue bias, we calculated the degree to which the tissue bias was conserved for the query gene and its paralog as the ratio of the number module sets representative of different tissues to which the isoforms of both paralogous genes belong versus the total number of different module sets to which the isoforms of the query gene belong.

For isoforms of pairs of orthologous genes that display a single-tissue bias in each of the species, the tissue bias was considered 'conserved' if all of the isoforms of the orthologs are in each of the species present in the module set that is representative of the same tissue (matching tissues between species). For isoforms of ortholog pairs derived from genes that display a multiple-tissue bias in each of the species, the tissue bias across species was considered conserved if the isoforms of the orthologous genes were present in module sets representing at least two matching tissues. Because the orthologous gene pairs for which a conserved tissue bias was observed constituted only part of one OG, we further referred to the collection of orthologous gene pairs with conserved tissue bias as an orthologous sub-group.

GO enrichment analysis

For genes containing the *N. lutea* lineage-specific isoforms, biological functions were assigned to *N. lutea* genes with BLAST2GO using the "non-redundant" database of plants and default settings (Conesa *et al.*, 2005). The TBtools was used to identify the significantly enriched GO terms (Chen *et al.*, 2020).

RT-PCR verification of selected AS events

The RNAs from samples were reverse transcribed into cDNA. Primers were designed to span the sequences that contained AS events (Table S4). The PCR fragments were tested in 1% agarose gel to verify the existence of AS events.

Data availability

The Oxford Nanopore full-length sequencing dataset generated for this work is accessible through NCBI Sequence Read Archive (SRA) under accession number PRJNA777451. The Illumina RNA-seq dataset is accessible through NCBI SRA accession number PRJNA705058.

Acknowledgements

This work was supported by the Strategic Priority Research Program of Chinese Academy of Sciences (No. XDB31000000), the National Natural Science Foundation of China (Nos 3217024, 31570220, 31870208), and Youth Innovation Promotion Association of Chinese Academy of Sciences (2019335), Flemish Fonds Wetenschappelijk Onderzoek-Vlaanderen (FWO) [3G046318, G.0371.06] and UGent BOF (BOF19/24J/062), European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (No. 833522), and Ghent University Methusalem funding (BOF.MET.2021.0005.01).

Conflict of interest

The authors declare no conflict of interest.

Author contributions

T.S. and J.C. conceived the idea. Y.Z. and X.Y. collected the tissues. Y.Z. analyzed the data and finished the confirmatory experiment. Y.Z., T.S., and M.K. wrote the manuscript. T.S., M.K., Y.V.P. and J.C. revised the manuscript.

Supporting information

Fig. S1. Density distribution of consensus full-length transcripts obtained by Nanopore sequencing.

Fig. S2. Distribution of the number of isoforms per gene in *N. lutea*, *N.nucifera* and *Arabidopsis*.

Fig. S3. Pipeline to identify conserved AS events in the same ortholog group.

Fig. S4. RT-PCR validation of interspecies conserved AS events.

Fig. S5. Box-plots showing the distribution of the number of AS events in per gene in genes with different copies in *N. lutea*, *N.nucifera* and *Arabidopsis*.

Fig. S6. Percentage of respectively single-copy or duplicated genes that undergo AS events.

Fig. S7. The bar graph shows the percentage of genes with at least one interspecies conserved AS event in single-copy genes or duplicated genes in *N. lutea* and *N. nucifera*.

Fig. S8. Bar chart showing the average number of AS events per gene.

Fig. S9. Intraspecies conserved AS events in paralogous gene pairs.

Fig. S10. Heatmap showing the presence/absence of the *N. nucifera* lineage specific isoforms in other *N. nucifera* cultivars as obtained from transcriptome analysis

Fig. S11. The heatmap of WGCNA module-tissue association in *N. lutea*, *N. nucifera*, and *Arabidopsis*.

Fig. S12. Distribution of the polymorphism value (PV) for the genes in *N. lutea*, *N. nucifera*, and *Arabidopsis*.

Fig. S13. Distribution of the number of isoforms in duplicated genes displaying respectively single- or multiple- tissue bias expression patterns plots.

Fig. S14. Verification of matching tissues between *N. lutea*, *N. nucifera*, and *Arabidopsis*.

Fig. S15. Schematic of genes with single- and multiple-tissue bias and the conserved tissue bias between orthologous genes.

Fig. S16. Examples of the conservation of tissue-specific expression for orthologous gene pairs that either show single-tissue or multiple-tissue biased expression.

Fig. S17. Phylogenetic analysis of MADS-box genes from the two Nelumbo species and *Arabidopsis*.

Fig. S18. Tissue-specific module networks of isoforms for the 'ABCE' module in *Arabidopsis thaliana*.

Table S1. Summary of the RNA-Seq samples in this study.

Table S2. Summary of the gene, isoforms and alternative splicing events in *N. lutea*, *N.nucifera* and *Arabidopsis*.

Table S3. Summary of the interspecies conserved AS events.

Table S4. Summary of the RT-PCR primer for validating the interspecies conserved AS events.

Table S5. Summary of the intraspecies conserved AS in *N. lutea* , *N. nucifera* , and *Arabidopsis.*

Table S6. Summary of the nodes of isoform co-expression network for each of *N. lutea*, *N. nucifera* and *Arabidopsis.*

Table S7. Summary of the Conserved tissue bias patterns in different comparisons between *N. lutea*, *N.nucifera* and *Arabidopsis.*

References

Cao YN, Comes HP, Sakaguchi S, Chen LY, Qiu YX. 2016. Evolution of East Asia's Arcto-

Tertiary relict Euptelea (Eupteleaceae) shaped by Late Neogene vicariance and Quaternary climate change. *BMC Evol Biol.* 16:66.

Chamala S, Feng G, Chavarro C, Barbazuk W. 2015. Genome-Wide Identification of Evolutionarily Conserved Alternative Splicing Events in Flowering Plants. *Front Bioeng Biotechnol.* 3:33.

Chen C, Chen H, Zhang Y, Thomas HR, Frank MH, He Y, Xia R. 2020. TBtools: An Integrative Toolkit Developed for Interactive Analyses of Big Biological Data. *Mol Plant.* 13:1194-1202.

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

Day IS, Golovkin M, Palusa SG, Link A, Ali GS, Thomas J, Richardson DN, Reddy AS. 2012. Interactions of SR45, an SR-like protein, with spliceosomal proteins and an intronic sequence: insights into regulated splicing. *Plant J.* 71:936-947.

De Smet R, Adams KL, Vandepoele K, Van Montagu MC, Maere S, Van de Peer Y. 2013. Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proc Natl Acad Sci U S A*. 110:2898-903.

DiMario RJ, Clayton H, Mukherjee A, Ludwig M, Moroney JV. 2017. Plant Carbonic Anhydrases: Structures, Locations, Evolution, and Physiological Roles. *Mol Plant.* 10:30-46.

Gu J, Xia Z, Luo Y, Jiang X, Qian B, Xie H, Zhu JK, Xiong L, Zhu J, Wang ZY. 2018. Spliceosomal protein U1A is involved in alternative splicing and salt stress tolerance in *Arabidopsis thaliana*. *Nucleic Acids Res*. 46:1777-1792.

Huang Y, Lack JB, Hoppel GT, Pool JE. 2021. Parallel and Population-specific Gene Regulatory Evolution in Cold-Adapted Fly Populations. *Genetics*. 218:iyab077

Hurtig JE, Kim M, Orlando-Coronel LJ, Ewan J, Foreman M, Notice LA, Steiger MA, van Hoof A. 2020. Origin, conservation, and loss of alternative splicing events that diversify the proteome in Saccharomycotina budding yeasts. *Rna*. 26:1464-1480.

Jiang J, Liu X, Liu C, Liu G, Li S, Wang L. 2017. Integrating Omics and Alternative Splicing Reveals Insights into Grape Response to High Temperature. *Plant Physiol.* 173:1502-1518.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol.* 37:907-915.

Klepikova AV, Kasianov AS, Gerasimov ES, Logacheva MD, Penin AA. 2016. A high resolution map of the *Arabidopsis thaliana* developmental transcriptome based on RNA-seq profiling. *Plant J*. 88:1058-1070.

Kornblihtt AR, Schor IE, Alló M, Dujardin G, Petrillo E, Muñoz MJ. 2013. Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nat Rev Mol Cell Biol*. 14:153-165.

Kryuchkova-Mostacci N, Robinson-Rechavi M. 2017. A benchmark of gene expression tissue-specificity metrics. *Brief Bioinform*. 18:205-214.

Lan T, Renner T, Ibarra-Laclette E, Farr KM, Chang TH, Cervantes-Pérez SA, Zheng C, Sankoff D, Tang H, Purbojati RW, et al. 2017. Long-read sequencing uncovers the adaptive topography of a carnivorous plant genome. *Proc Natl Acad Sci U S A*. 114:E4435-e4441.

Langfelder P, Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*. 9:559.

Lee Y, Rio DC. 2015. Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annu Rev Biochem*. 84:291-323.

Lev Maor G, Yearim A, Ast G. 2015. The alternative role of DNA methylation in splicing regulation. *Trends Genet*. 31:274-280.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 34:3094-3100.

Li H, Yang X, Zhang Y, Gao Z, Liang Y, Chen J, Shi T. 2021. Nelumbo genome database, an integrative resource for gene expression and variants of *Nelumbo nucifera*. *Sci Data*. 8:38.

Li L, Stoeckert CJ, Jr., Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*. 13:2178-2189.

Li W, Kang S, Liu CC, Zhang S, Shi Y, Liu Y, Zhou XJ. 2014. High-resolution functional annotation of human transcriptome: predicting isoform functions by a novel multiple instance-based label propagation method. *Nucleic Acids Res*. 42:e39.

Li YP, Dai C, Hu CG, Liu ZC, Kang CY. 2017. Global identification of alternative splicing via comparative analysis of SMAR- and Illumina-based RNA-seq in strawberry. *Plant J*.

90:164-176.

Ma J, Wang J, Ghoraie LS, Men X, Chen R, Dai P. 2020. Comprehensive expression-based isoform biomarkers predictive of drug responses based on isoform co-expression networks and clinical data. *Genomics*. 112:647-658.

Mandadi KK, Scholthof KB. 2015. Genome-wide analysis of alternative splicing landscapes modulated during plant-virus interactions in Brachypodium distachyon. *Plant Cell*. 27:71-85.

Marshall AN, Han J, Kim M, van Hoof A. 2018. Conservation of mRNA quality control factor Ski7 and its diversification through changes in alternative splicing and gene duplication. *Proc Natl Acad Sci U S A*. 115:E6808-e6816.

Mei W, Boatwright L, Feng G, Schnable JC, Barbazuk WB. 2017. Evolutionarily Conserved Alternative Splicing Across Monocots. *Genetics*. 207:465-480.

Niknafs YS, Pandian B, Iyer HK, Chinnaiyan AM, Iyer MK. 2017. TACO produces robust multisample transcriptome assemblies from RNA-seq. *Nat Methods*. 14:68-70.

Ó'Maoiléidigh DS, Graciet E, Wellmer F. 2014. Gene networks controlling Arabidopsis thaliana flower development. *New Phytol*. 201:16-30.

Qiao X, Li Q, Yin H, Qi K, Li L, Wang R, Zhang S, Paterson AH. 2019. Gene duplication and evolution in recurring polyploidization-diploidization cycles in plants. *Genome Biol*. 20:38.

Reddy AS, Marquez Y, Kalyna M, Barta A. 2013. Complexity of the alternative splicing landscape in plants. *Plant Cell*. 25:3657-3683.

Rigo R, Bazin JRM, Crespi M, Charon CL. 2019. Alternative Splicing in the Regulation of Plant-Microbe Interactions. *Plant Cell Physiol*. 60:1906-1916.

Roux J, Robinson-Rechavi M. 2011. Age-dependent gain of alternative splice forms and biased duplication explain the relation between splicing and duplication. *Genome Res*. 21:357-363.

Sandve SR, Rohlfs RV, Hvidsten TR. 2018. Subfunctionalization versus neofunctionalization after whole-genome duplication. *Nat Genet*. 50:908-909.

Santos ME, Athanasiadis A, Leitão AB, DuPasquier L, Sucena E. 2011. Alternative splicing and gene duplication in the evolution of the FoxP gene subfamily. *Mol Biol Evol*. 28:237-

247.

Shi T, Rahmani RS, Gugger PF, Wang M, Li H, Zhang Y, Li Z, Wang Q, Van de Peer Y, Marchal K, et al. 2020. Distinct Expression and Methylation Patterns for Genes with Different Fates following a Single Whole-Genome Duplication in Flowering Plants. *Mol Biol Evol*. 37:2394-2413.

Smith CCR, Rieseberg LH, Hulke BS, Kane NC. 2021. Aberrant RNA splicing due to genetic incompatibilities in sunflower hybrids. *Evolution*. 75:2747-2758.

Smith CCR, Tittes S, Mendieta JP, Collier-Zans E, Rowe HC, Rieseberg LH, Kane NC. 2018. Genetics of alternative splicing evolution during sunflower domestication. *Proc Natl Acad Sci U S A*. 115:6768-6773.

Soltis DE, Chanderbali AS, Kim S, Buzgo M, Soltis PS. 2007. The ABC model and its applicability to basal angiosperms. *Ann Bot*. 100:155-163.

Su Z, Gu X. 2012. Revisit on the evolutionary relationship between alternative splicing and gene duplication. *Gene*. 504:102-106.

Su Z, Wang J, Yu J, Huang X, Gu X. 2006. Evolution of alternative splicing after gene duplication. *Genome Res*. 16:182-189.

Syed NH, Kalyna M, Marquez Y, Barta A, Brown JW. 2012. Alternative splicing in plants--coming of age. *Trends Plant Sci*. 17:616-623.

Talavera D, Vogel C, Orozco M, Teichmann SA, de la Cruz X. 2007. The (in)dependence of alternative splicing and gene duplication. *PLoS Comput Biol*. 3:e33.

Tautz D, Domazet-Lošo T. 2011. The evolutionary origin of orphan genes. *Nat Rev Genet*. 12:692-702.

Thatcher SR, Danilevskaya ON, Meng X, Beatty M, Zastrow-Hayes G, Harris C, Van Allen B, Habben J, Li B. 2016. Genome-Wide Analysis of Alternative Splicing during Development and Drought Stress in Maize. *Plant Physiol*. 170:586-599.

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

Tian L, Zhao X, Liu H, Ku L, Wang S, Han Z, Wu L, Shi Y, Song X, Chen Y. 2019. Alternative splicing of ZmCCA1 mediates drought response in tropical maize. *PLoS One*.

14:e0211623.

Trincado JL, Entizne JC, Hysenaj G, Singh B, Skalic M, Elliott DJ, Eyras E. 2018. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol.* 19:40.

Van de Peer Y, Mizrachi E, Marchal K. 2017. The evolutionary significance of polyploidy. *Nat Rev Genet*. 18:411-424.

Wahl MC, Will CL, Lührmann R. 2009. The spliceosome: design principles of a dynamic RNP machine. *Cell*. 136:701-718.

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

Wang K, Wang D, Zheng X, Qin A, Zhou J, Guo B, Chen Y, Wen X, Ye W, Zhou Y, et al. 2019. Multi-strategic RNA-seq analysis reveals a high-resolution transcriptional landscape in cotton. *Nat Commun*. 10:4714.

Wang M, Wang P, Liang F, Ye Z, Li J, Shen C, Pei L, Wang F, Hu J, Tu L, et al. 2018. A global survey of alternative splicing in allopolyploid cotton: landscape, complexity and regulation. *New Phytol*. 217:163-178.

Wang W, Pu X, Yang S, Feng Y, Lin C, Li M, Li X, Li H, Meng C, Xie Q, et al. 2020. Alternative splicing of DSP1 enhances snRNA accumulation by promoting transcription termination and recycle of the processing complex. *Proc Natl Acad Sci U S A*. 117:20325-20333.

Wang TT, Wang HY, Cai DW, Gao YB, Zhang HX, Wang YS, Lin CT, Ma LY, Gu LF. 2017. Comprehensive profiling of rhizome-associated alternative splicing and alternative polyadenylation in moso bamboo (*Phyllostachys edulis*). *Plant J*. 91:684-699.

Wang X, Hu L, Wang X, Li N, Xu C, Gong L, Liu B. 2016. DNA Methylation Affects Gene Alternative Splicing in Plants: An Example from Rice. *Mol Plant*. 9:305-307.

Wang XM, Chen SY, Shi X, Liu DN, Zhao P, Lu YZ, Cheng YB, Liu ZS, Nie XJ, Song WN, et al. 2018. Hybrid sequencing reveals insight into heat sensing and signaling of bread wheat. *Plant J*. 98: 1015-1032.

Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, Lee TH, Jin H, Marler B, Guo H, et al. 2012. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and

collinearity. *Nucleic Acids Res.* 40:e49.

Will CL, Lührmann R. 2011. Spliceosome structure and function. *Cold Spring Harb Perspect Biol.* 3:a003707.

Wittkopp PJ, Kalay G. 2011. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat Rev Genet.* 13:59-69.

Wu Z, Gui S, Quan Z, Pan L, Wang S, Ke W, Liang D, Ding Y. 2014. A precise chloroplast genome of *Nelumbo nucifera* (Nelumbonaceae) evaluated with Sanger, Illumina MiSeq, and PacBio RS II sequencing platforms: insight into the plastid evolution of basal eudicots. *BMC Plant Biol.* 14:289.

Xue J, Dong Wa, Cheng T, Zhou Si. 2012. Nelumbonaceae: Systematic position and species diversification revealed by the complete chloroplast genome. *J. Syst. Evol.* 50:477-487.

Yang R, Wang X. 2013. Organ evolution in angiosperms driven by correlated divergences of gene sequences and expression patterns. *Plant Cell.* 25:71-82.

Yuan Y, Chung JD, Fu X, Johnson VE, Ranjan P, Booth SL, Harding SA, Tsai CJ. 2009. Alternative splicing and gene duplication differentially shaped the regulation of isochorismate synthase in Populus and Arabidopsis. *Proc Natl Acad Sci USA.* 106:22020-22025.

Zhang H, Lin C, Gu L. 2017. Light Regulation of Alternative Pre-mRNA Splicing in Plants. *Photochem Photobiol.* 93:159-165.

Zhang R, Calixto CPG, Marquez Y, Venhuizen P, Tzioutziou NA, Guo W, Spensley M, Entizne JC, Lewandowska D, Ten Have S, et al. 2017. A high quality *Arabidopsis* transcriptome for accurate transcript-level analysis of alternative splicing. *Nucleic Acids Res*. 45:5061-5073.

Zhang SJ, Wang C, Yan S, Fu A, Luan X, Li Y, Sunny Shen Q, Zhong X, Chen JY, Wang X, et al. 2017. Isoform Evolution in Primates through Independent Combination of Alternative RNA Processing Events. *Mol Biol Evol.* 34:2453-2468.

Zhang Y, Nyong AT, Shi T, Yang P. 2019. The complexity of alternative splicing and landscape of tissue-specific expression in lotus (*Nelumbo nucifera*) unveiled by Illumina- and single-molecule real-time-based RNA-sequencing. *DNA Res.* 26:301-311.

Zhang Y, Rahmani RS, Yang X, Chen J, Shi T. 2020. Integrative expression network analysis

of microRNA and gene isoforms in sacred lotus. *BMC Genomics.* 21:429.

Zhou P, Hirsch CN, Briggs SP, Springer NM. 2019. Dynamic Patterns of Gene Expression Additivity and Regulatory Variation throughout Maize Development. *Mol Plant.* 12:410-425.

Table 1. Summary of the full-length transcript sequencing of *N. lutea* on the Oxford Nanopore Technologies MinION platform.

Category	Nanopore full-length sequencing
Read Number	29,081,500
Base Number	31,673,918,600
N50 (bp)	1,202
Mean Length (bp)	1,089
Max Length (bp)	13,691
Mean Qscore	Q11
Number of clean reads (except rRNA)	28,827,854
Number of full-length reads	23,432,326
Full-Length Percentage (FL%)	81.28%

Table 2. Summary of interspecific conserved AS between different comparisons in *N. lutea* , *N. nucifera*, and *A. thaliana* .

	A3	A5	AF	AL	RI	SE	Total
A. thaliana vs *N. lutea*	12	1	0	0	26	0	39
A. thaliana vs *N. nucifera*	7	0	0	0	7	1	15
N. lutea vs *N. nucifera*	670	335	5	2	871	479	2362
A. thaliana vs *N. nucifera* vs *N. lutea*	1	0	0	0	2	1	4
Total	690	336	5	2	906	481	2420

Figures

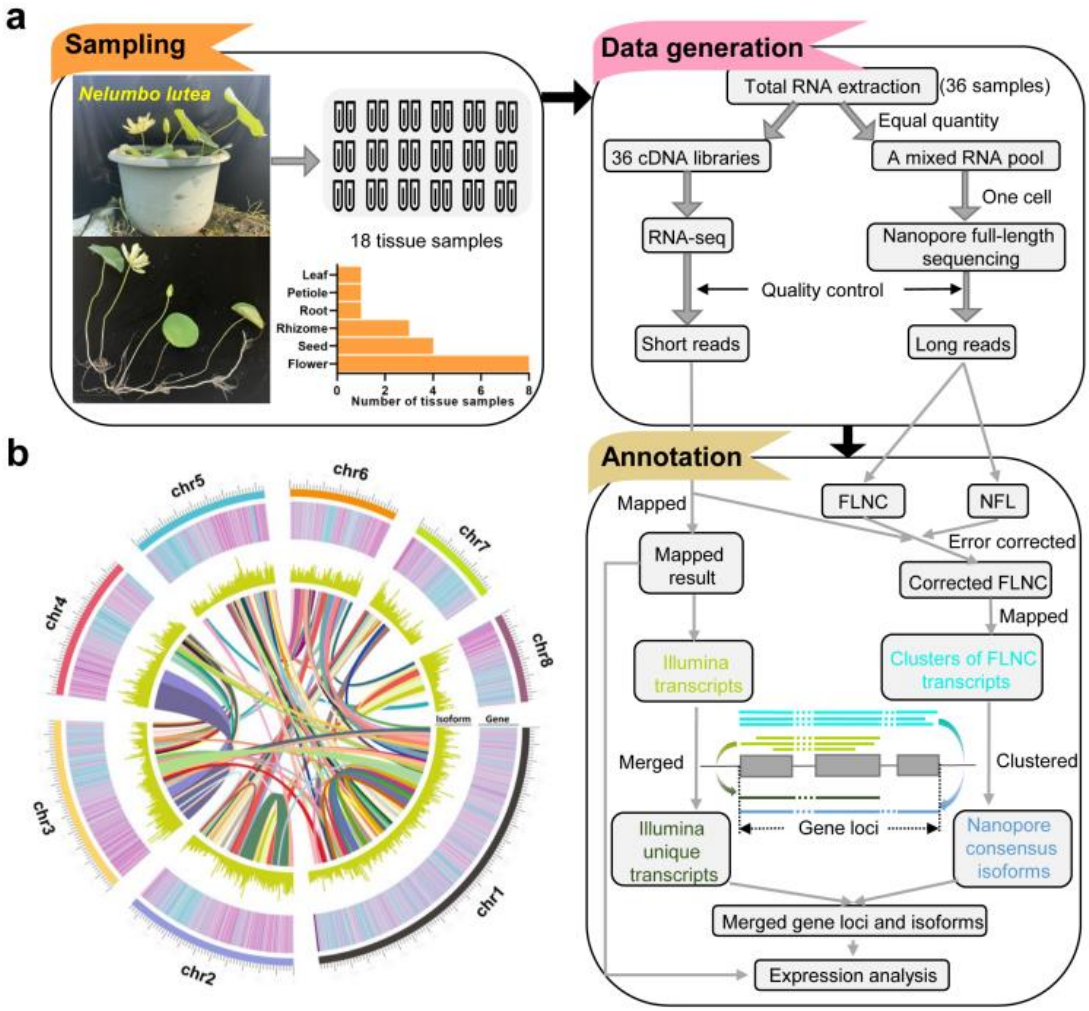

Figure 1. The annotation of genes and isoforms in the *N. lutea* genome.

(a) Flowchart for gene and isoform annotation by Nanopore full-length sequencing and Illumina RNA-seq in *N. lutea*. Sampling: the number of collected tissue samples in *N. lutea* organs was shown in the bar chart. Each of the 18 tissue samples from the different development stages (two biological repeats per tissue) was collected. Data generation: Total RNA for each of 36 samples was extracted, and the RNA of each sample was used for building a cDNA library and further sequenced by Illumina RNA-seq. In addition, an RNA pool obtained by mixing the equal quantity of RNAs extracted from 36 samples was subjected to Nanopore full-length sequencing. Annotation: The color schematic diagram shows examples of Illumina transcript merging and Nanopore full-length transcript clustering. The lines in different colors mean transcripts of the corresponding type. The gray boxes and lines were the exon and intron regions of merged gene loci. FLNC: full-length non-chimeric reads. NFL: non-full-length reads.

(b) CIRCOS plot showing the distribution of different annotations in the *N. lutea* genome. Four rings from outside to inside show the chromosomal positions (1st), gene density (the red means high density and blue means low density) (2nd), isoform density (3rd), and colored links show syntenic blocks (4th).

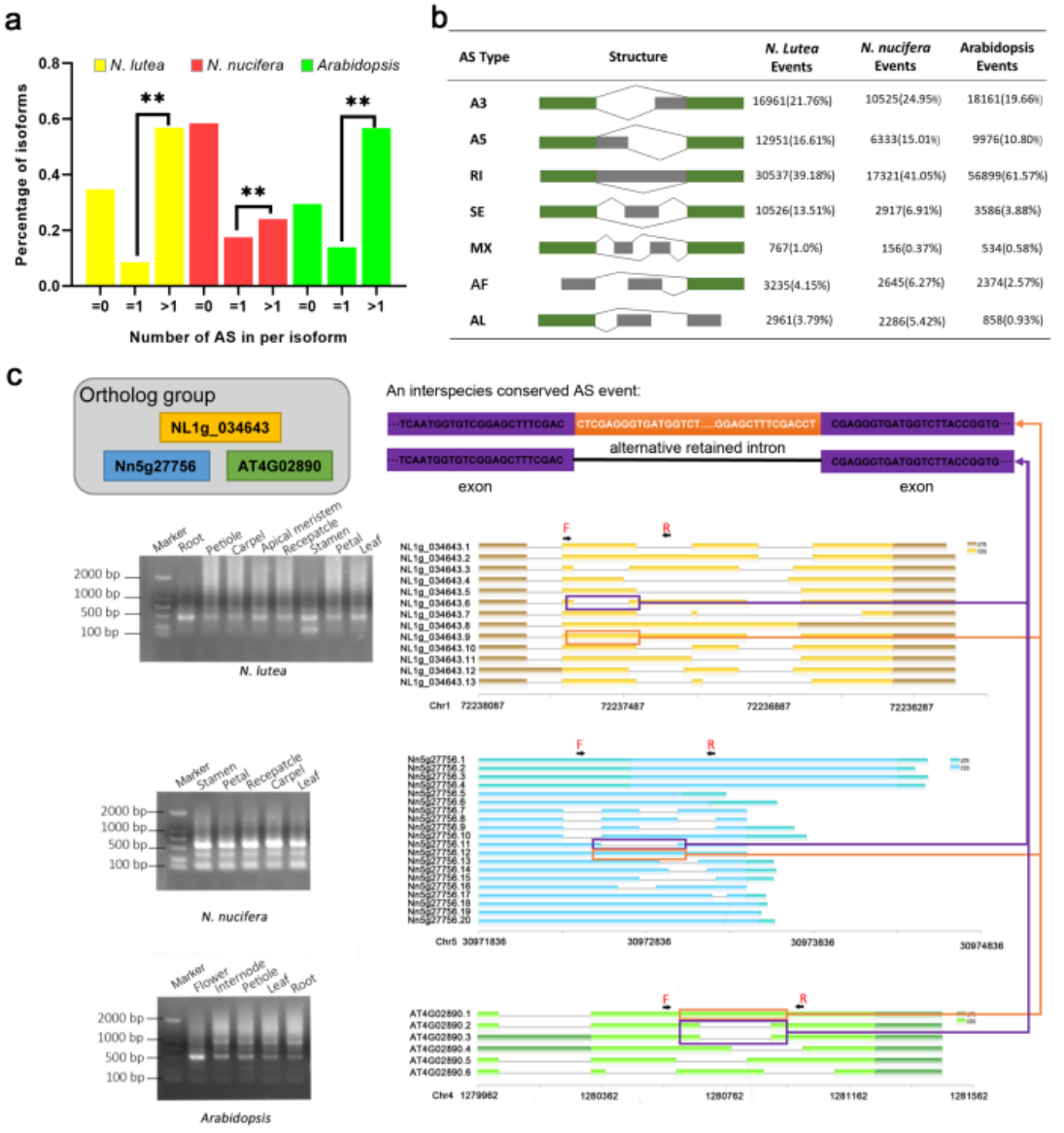

Figure 2. Characterization of alternative splicing (AS) events.

(a) Percentage of the number of AS events per isoform in *N. lutea*, *N. nucifera*, and *Arabidopsis*. Isoforms were classified into three categories: isoforms without AS (=0); isoforms with one AS event (=1); isoforms with more than one AS event ($\geqslant$1). The difference between isoforms with one AS event and more than one AS event was tested by the chi-square test (** means $p < 0.01$).

(b) Classification of AS events in *N. lutea*, *N. nucifera*, and *Arabidopsis*. Different types of AS events are shown: alternative 3' splice sites (A3), alternative 5' splice sites (A5), retained intron (RI), skipping exon (SE), mutually exclusive exons (MX), alternative first exons (AF), and alternative last exons (AL).

(c) RT-PCR validation of an interspecies conserved AS event. The orthologous genes of *N. lutea*

(*NL1g_034643*), *N. nucifera* (*Nn5g27756*), and *Arabidopsis* (*AT4G02890*) are classed into one ortholog group and share a conserved RI event. The orange sequence indicates the alternative retained intron, and the purple sequence indicates the exon region around this conserved RI event. The genomic locus of this interspecies conserved AS event in each of the three species can be found in Table S3. Yellow boxes (exons) and lines (introns) with arrows show the predicted structure of each isoform in *N. lutea*, blue in *N. nucifera*, and green in *Arabidopsis*. The arrows show the loci of the PCR primers (F, forward, and R, reverse) on the first isoform of each gene. The RT-PCR amplifications in each species among tissues are shown in the left panel.

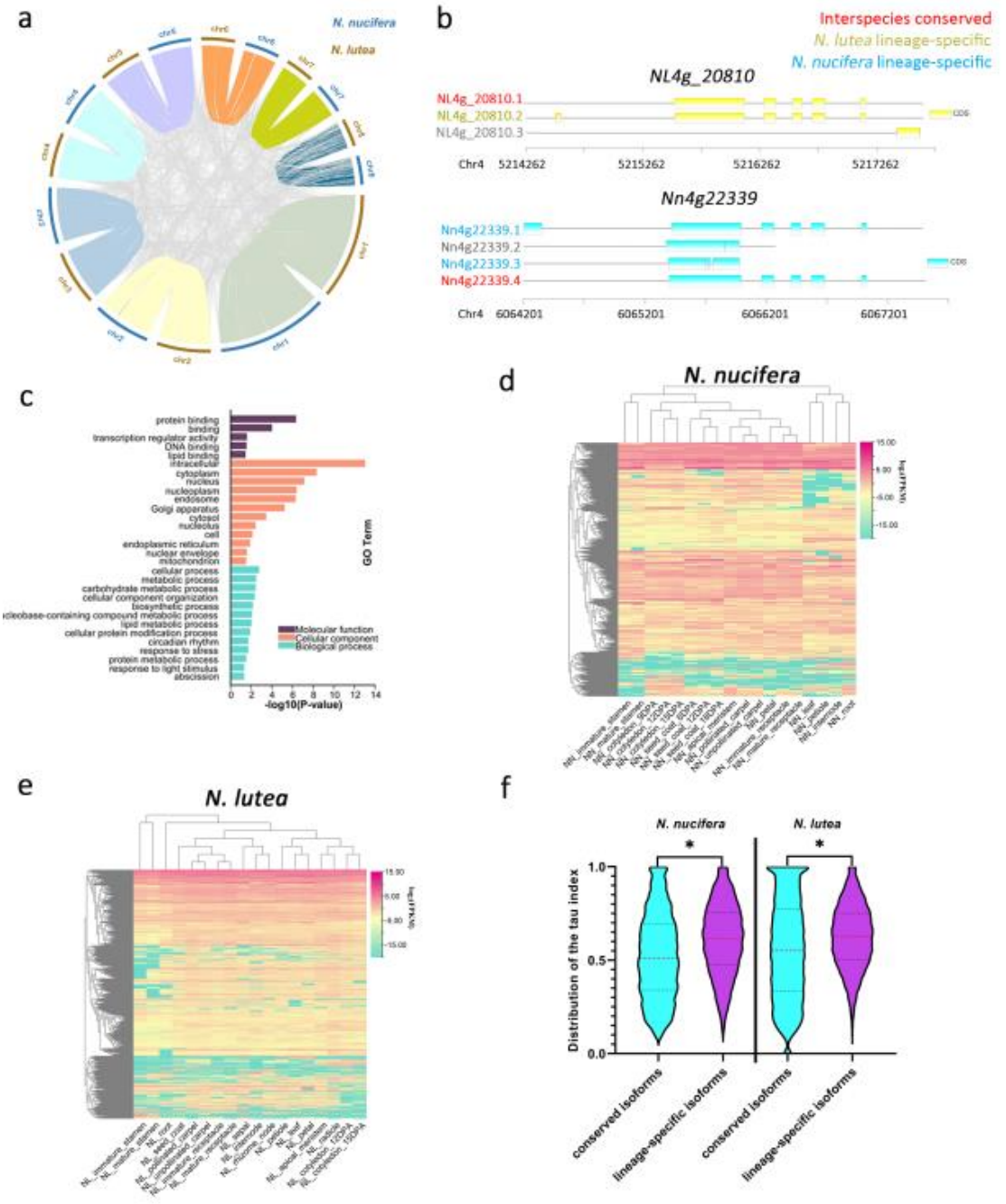

Figure 3. Identification and verification of lineage-specific isoforms in *N. lutea* and *N. nucifera*.

(a) CIRCOS plot showing the orthologous gene pairs between *N. lutea* and *N. nucifera* genomes. The outside brown rings are the chromosomes from the *N. lutea* genome, blue ones are from *N. nucifera*. The inner lines show the orthologous gene pairs.

(b) Validation of lineage-specific isoforms in *N. lutea* and *N. nucifera*. The CDS (coding sequence) and UTR (untranslated region) of isoforms of the *N. lutea* gene "*NL4g_20810*" are shown in yellow boxes and lines, and the corresponding CDS and UTRs of the isoforms of the

corresponding orthologous *N. nucifera* gene "*Nn4g22339*" are shown in blue. The names of the isoforms indicated in red are identified as interspecies conserved isoforms, whereas those indicated in yellow are *N. lutea* lineage-specific isoforms, and in blue *N. nucifera* lineage-specific isoforms.

(c) GO enrichment analysis of genes with *N. lutea* lineage-specific isoforms.

(d-e) Heatmap showing the expression profile (log2 FPKM) of respectively *N. lutea* and *N. nucifera* lineage-specific isoforms. Labels correspond to the investigated developmental stages per tissue profiled in respectively *N. lutea* and *N. nucifera*.

(f) Violin plot shows the distribution of the tau index for lineage-specific (cyan) and conserved (purple) isoforms obtained for respectively *N. lutea* and *N. nucifera.* The median is indicated with a red line. The difference in mean tau index between conserved isoforms and lineage-specific isoforms was tested with a *t*-test (two-tailed), * means p-value < 0.01.

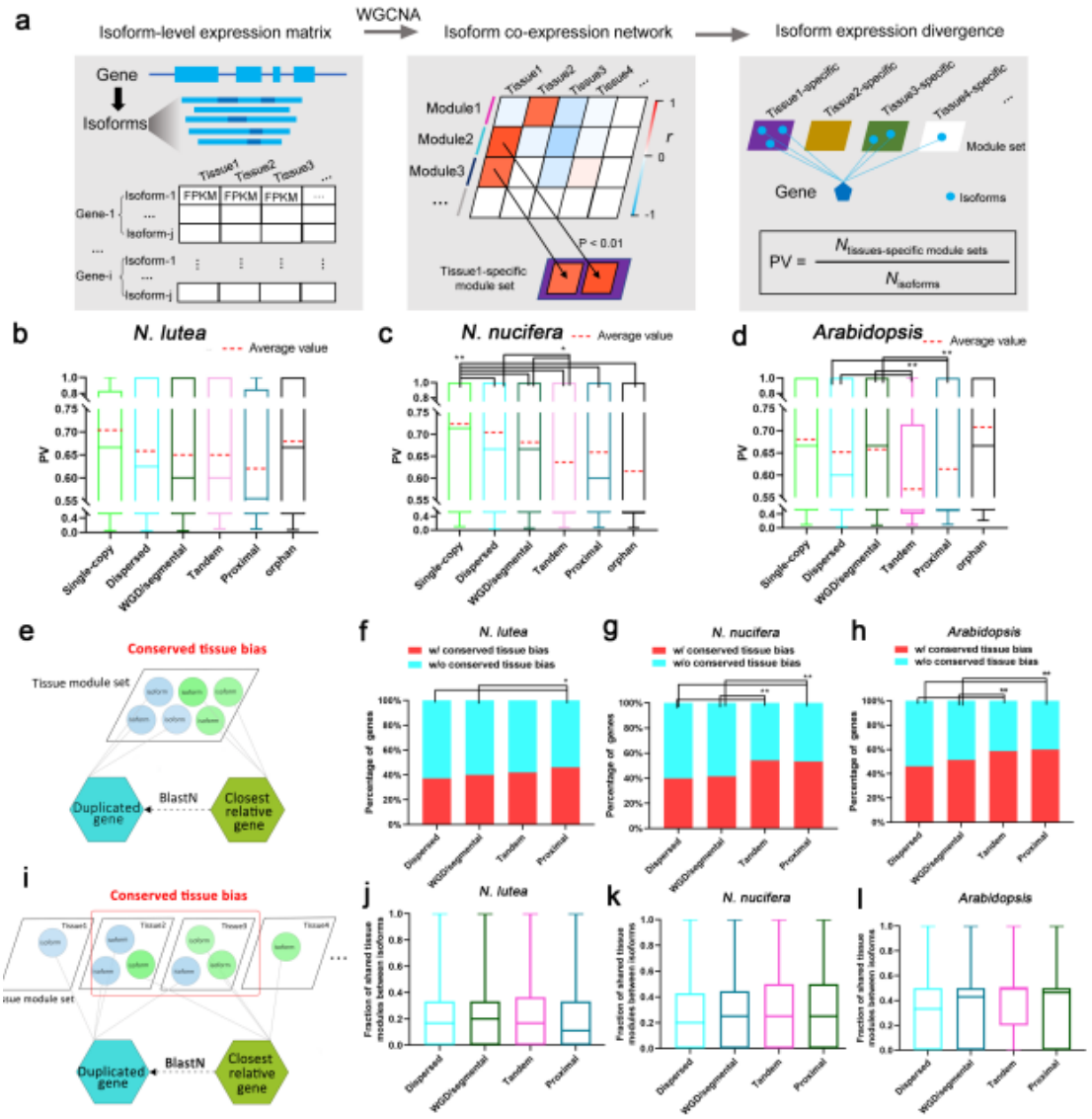

Figure 4. The expression divergence of different types of duplicated genes in *N. lutea*, *N. nucifera*, and *Arabidopsis*.

(a) Schematic overview of how isoform level coexpression networks and tissue-specific module sets were built and how the polymorphism value (PV) is used in combination with the tissue-specific module sets to identify a gene's divergence in isoform expression.

(b-d) box plots showing the PV levels of isoforms of duplicated genes, conditioned on their origin of duplication for respectively *N. lut*ea (b), *N. nucifera* (c), and *Arabidopsis* (d). The red dashed lines in the box plots indicate the average value of PV for different duplicated gene groups. Pairwise comparisons of the PV level between groups were performed using a Mann-Whitney *U* test, * means p-value < 0.05, and ** means p-value < 0.01.

(e) Assessing conserved tissue bias in expression for isoforms of paralogs of which the

expression of their isoforms is biased towards a single tissue (i.e both paralogs are single-tissue biased genes). Paralogs are represented by respectively the blue and green hexagons. The blue and green circles represent the isoforms of the corresponding paralogous genes. The parallelogram represents the module set representative of the tissue to which the isoforms of both paralogs were assigned. The isoforms of two paralogs are said to show a conserved tissue bias if all isoforms of both paralogs belong to the same tissue-specific module set (as illustrated in the figure).

(f-h) Percentage of single tissue bias genes with a duplication origin of which the isoforms show conserved tissue bias (red part of the bar). Genes are subdivided according to their origin of duplication. Results are shown for respectively *N. lutea* (f), *N. nucifera* (g), and *Arabidopsis* (h). The difference in the percentages of genes showing isoform-level tissue bias in the pairwise between the different duplication groups was tested by the chi-square test (* means $p < 0.05$; ** means $p < 0.01$).

(i) Assessing conserved tissue bias for isoforms of paralogs of which the expression is biased towards multiple-tissues. Paralogs are indicated by respectively the blue and green hexagons. The circles with different colors represent the isoforms of these corresponding paralogs. For this analysis, only the closest relative of the query gene was considered as a paralog. The parallelograms in the red rectangle represent the tissues to which the expression of the isoforms of the paralogs show a shared tissue bias. The isoform of a gene is said to display a bias towards a certain tissue if it belongs to a module set representative of that tissue.

(j-l) Violin plots illustrating the degree to which the tissue bias was conserved for isoforms of paralogs of which the expression is biased towards multiple tissues. Genes are subdivided according to their origin of duplication. This degree of conserved tissue bias is estimated as the ratio of the number of module sets representative of different tissues to which the isoforms of both paralogous genes belong versus the total number of module sets to which the query gene belongs. For this analysis, only the closest relative of the query gene was considered as a paralog. The ratios are shown for respectively *N. lutea* (j), *N. nucifera* (k), and *Arabidopsis* (l).

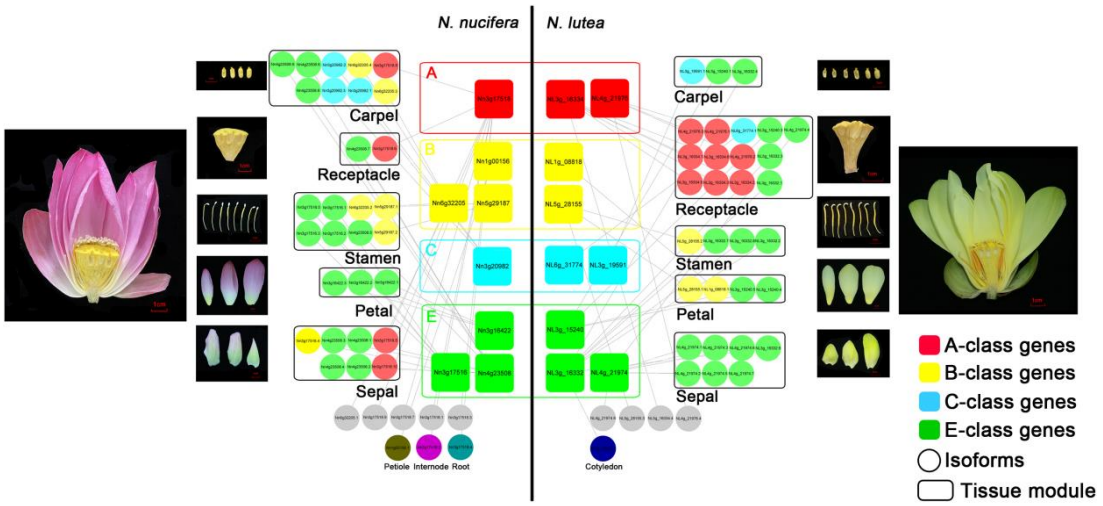

Figure 5. Tissue-specific module networks of isoforms for the 'ABCE' module in two Nelumbo species.

The filled squares in red (A-class), yellow (B-class), blue (C-class), and green (E-class) represent the 'ABCE' model genes. The colored circles (red, green, yellow, and blue) represent the isoforms of the genes belonging to distinct classes of the 'ABCE' model, and the grey circles represent isoforms not present in the tissue-specific expression modules. The circle's colors other than the ones indicated above are isoforms that are specifically expressed in non-floral tissue modules. The lines link genes and their corresponding isoforms. Different combinations of isoforms from the 'ABCE' model genes participate in the formation of different floral tissues.