nature communications

Article

https://doi.org/10.1038/s41467-025-66816-3

Phylogenomic profile of exon-intron organization across angiosperms, their relationships with protein domains, and functional implications

Received: 19 August 2024

Accepted: 12 November 2025

Published online: 07 January 2026

Check for updates

Taikui Zhang[1,2], Lin Zhang[3], Hong Ma[2] ✉ & Jun Wang[4] ✉

Eukaryotic genes usually encode proteins and contain exons in different size, including micro and small exon categories. However, genome-wide gene exon-intron organizations and their conservation across angiosperms have not been investigated. Among exons of protein-coding genes from 46 angiosperms and four gymnosperms, ~35% are micro and small exons affecting most genes and the remaining exons are defined here as medium, large, or super exons. Exon-intron structural comparison using our new bioinformatic tool, ExonEvo, reveals that, ~69% of conserved exons in homologous genes are found in two or more major angiosperm subgroups, including ~47% as the seed plant-wide conserved exons that are present in most genes in individual species. An examination of domain-exon correspondence revealed that most conserved protein domains match two or more conserved exon families and genes encoding the same domain family members can have different domain-exon correspondence. Moreover, most flower-preferential genes contain conserved exons and most exons with exon skipping during splicing are shared by seed plants, suggesting that alternative splicing of the exons could be a conserved mechanism for regulating protein structure. Our phylogenomic landscape of exon structure and domain-exon organization supports a model for conservation of exon/domain structure in relation to post-transcriptional regulation of gene/protein function.

Most eukaryotic genes have exons and introns and can have diverse exon-intron organization and functions. The introns are removed via splicing to generate mature mRNAs; RNA splicing is essential for proper expression of intron-containing genes and requires conserved spliceosome and splicing factors[1–3]. Eukaryotic proteins generally have one or more functional domains, with lengths from ~35 to ~250 amino acids[4]. Domains have diverse activities, such as ligand binding, catalysis, and macro-molecular interaction and often define conserved protein families across animals, fungi, and plants[5]. Gene family members are thought to be derived from a common ancestral gene and

[1]Co-Innovation Center for Sustainable Forestry in Southern China, State Key Laboratory of Tree Genetics and Breeding, College of Life Sciences, Nanjing Forestry University, Nanjing, Jiangsu, China. [2]Department of Biology, the Eberly College of Science, and the Huck Institutes of the Life Sciences, the Pennsylvania State University, University Park, State College, PA, USA. [3]Chongqing Key Laboratory of Plant Resource Conservation and Germplasm Innovation, School of Life Sciences, Southwest University, Chongqing, China. [4]School of Life Sciences, State Key Laboratory of Genetics and Development of Complex Phenotypes, Department of Biochemistry and Biophysics, Institute of Plant Biology, Fudan University, Shanghai, China. ✉e-mail: hxm16@psu.edu; junwang90@fudan.edu.cn

share one or more conserved domains, often with similar domain organizations and similar known or predicted functions. Analyses of genes in human and eight other eukaryotic genomes revealed that exon boundaries often correspond closely to those of protein domains[6]. In addition, transcripts with different combinations of exons and/or different exon boundaries can be generated from the same gene via alternative splicing, resulting in the production of distinct protein isoforms with diverse domain structures[7], including the presence/absence of all/part of domains. Therefore, splicing is a crucial step for regulating gene expression and protein function. Furthermore, during evolution exon shuffling alters the number of exons (and introns) of genes, thereby changing protein sequences and function[8]. Thus, understanding the patterns of conservation and divergence of exon sizes and their relationships to protein domains is an important aspect of evolutionary analyses of gene structure and function.

Flowering plants (angiosperms) are the most species-rich group of land plants, with over 370,000 species, and are the sister group to gymnosperms (~1000 non-flowering seed plants, such as pine and ginkgo)[9–11]. Seed plants are also highly complex in structure and function[12,13] and are extremely important for the human society, providing foods, clothing, shelter, medicine and many other uses; they are also crucial for the health of diverse ecosystems, supporting many animals, fungi and other organisms[14,15]. In seed plants, exon lengths range from one bp [e.g., in the *Arabidopsis thaliana* (*Arabidopsis* hereafter for brevity) *Anaphase Promoting Complex subunit 11* gene[16]] to >10 kb, with average exon lengths in specific plants spanning from 201 to 319 bp[17]. The smallest group of exons with fewer than 51 bp has been referred to as micro exon, found in both plants and animals[18–21]; ~8000 micro exons were detected in >6000 genes with introns in *Arabidopsis* and ~13,000 micro exons were detected in human genes[18,22]. More recently, the next shortest exons (with 51~100 bp) have been called small exons and 39,025 small exons were detected in 12,525 *Arabidopsis* genes[22]. Micro exons are found in *Arabidopsis* floral homeotic genes (also called the ABCE genes), such as *AG*, *AP1*, *AP2*, and *AP3*, and encode portions of DNA-binding or protein-interaction domains[22]. These floral homeotic genes belong to the MADS-box and AP2 gene families that have significantly higher frequencies of micro-exon-containing genes[23]. Are the patterns of micro and small exons conserved across flowering plants for floral homeotic genes and other genes? To address the question, genome-wide patterns of the number and size of exons need to be investigated in angiosperms.

Among angiosperms, alternative splicing of specific genes has been described for several species, including *Arabidopsis*, rice, tomato, soybean, cotton, *Populus*, and bamboo[24–30]. A form of alternative splicing, exon skipping, results in the lack of a specific exon in the mature mRNA and leads to the absence or shortening of functional domains or altered protein reading frame[31]. For example, the cotton *Ga05g01029* transcript, including a 45-bp micro exon encoding part of the DNA-binding AP2 domain, was detected in both vegetative and reproductive organs; however, its isoform excluding this micro exon was detected in vegetative organs (cotyledon, stem, root, leaf) but not in floral organs, such as the sepal, petal, stigma, and anther (stamen)[32]. DNA-binding proteins often contain both DNA-binding and protein-interaction domains. In some DNA-binding protein families, some paralogues are truncated and lack some domain(s) and have been reported to modulate the function of other full-length paralogues[33–35]. It is possible that isoforms lacking (part of) a domain due to alternative splicing can affect the protein activities. For example, a SEP3 protein (an E-function floral homeotic protein) lacking a portion of the K domain, which is required for the formation of the quartet protein complex, due to the skipping of a micro exon, was reported to affect flower development[36–38]. In addition, the *Arabidopsis* GRP20 protein was shown to be required for normal flower development and for proper retention of micro exons in transcripts of several floral homeotic genes, such as *AG*, *AP1*, *AP2*, *AP3*, and *SEP3*[22]. Also, the skip of micro exons of these genes was found to cause floral phenotype changes (e.g., *SEP3* in ref. 36), showing the importance of micro exons in function of these developmental regulators. Although the micro exons in these floral homeotic genes and other members of the MADS-box and AP2 gene families are conserved in representative angiosperms[22,23], genome-wide exploration of conservation pattern of micro exons and exons of other size groups across angiosperms is needed to provide further understanding the evolution of gene structures and their relation to functional domain.

Among protein-coding plant genes, ~62 % contain named protein domains; ~65 % of these domains belong to protein families that are conserved across land plants and green algae from genomic analyses of seven angiosperms and seven other green plants[39]. Phylogenetic analyses of specific gene families using plant genomes support general conservation of structures (e.g., refs. 22,40–47). For example, phylogenetic analysis of 101 genes encoding ICK/KRP kinases from 38 angiosperms and two gymnosperms supports three clades named as A, B, and C groups; all three groups share two exons encoding the conserved CDK/cyclin interacting/inhibiting domain, whereas groups A and B share two additional exons, but group C genes have a distinct long exon[45]. Also, most members of the cytochrome P450 (CYP) superfamily have similar exon-intron structures[40], but in pomegranate some *CYP75* paralogues lack an intron and have different expression patterns during fruit development from other paralogues[40]. However, the variations in exon/intron structures for most angiosperm genes, including those containing micro/small exons, are largely unknown. Furthermore, regarding the relationship between exon-intron structure and their encoded protein domains, a comparison of seven angiosperms and three other land plants identified 35 conserved exon structures, including smaller micro exons of <=15 bp and their flanking exon sequences, and presented information about the corresponding domains such as AP2, peptidase, and gelsolin[48]. A genome-wide investigation of correspondence between exons and protein domains across angiosperms can further facilitate the understanding of functional similarities and differences among genes and through evolutionary history and is now feasible with the publication of high-quality sequenced angiosperm genomes, including >10 Telomere-to-Telomere (T2T) genomes with assembled chromosomes (e.g., refs. 49–58).

RNA expression of genes provides valuable clues to gene functions, including differential expression patterns related to organs and environmental conditions. Specifically, expression patterns of genes containing micro and small exons might suggest function in various plant organs and can be influenced by various environmental cues. In addition, alternative RNA splicing could affect micro and small exons and support hypotheses about regulation of expression and protein diversity for these genes[22]. One of major alternative splicing types commonly found in plants, humans and other animals is exon skipping. For example, the skipping of a micro exon in the *SEP3* transcript can lead to defects in flower development[36,38]. Also, a conserved 9-base micro exon in the *Arabidopsis WRINKLED1* (*WRI1*) gene belonging to the *AP2* family is important for fatty acid synthesis[23,59]. In particular, because micro and small exons encode relatively short peptide sequences, protein isoforms related to the presence/absence of micro/small exon-encoded peptide regions might have relatively subtle functional differences, in comparison to those involving larger exons/domains. However, the landscapes of expression patterns of genes containing micro/small exons vs exons of other sizes and alternative splicing involving exons of different sizes are largely unknown in different plant organs.

To address the questions regarding genome-wide conservation and patterns of angiosperm exon sizes, their relationships to protein domains, expression of genes with exons of various sizes and their alternative splicing, we selected genomes from 46 representative angiosperms and four representative gymnosperms and profiled the

exon sizes and numbers in protein-coding genes. In addition, we performed phylogenomic analyses to estimate the conserved exons shared by two or more homologous genes, particularly those across seed plants or among angiosperms. Among these conserved exons, their matched domains and the domain-exon correspondence were analyzed. Then, we analyzed public RNA-Seq datasets to examine expression patterns of genes containing conserved exons and to detect alternative splicing of conserved exons. These results present a comprehensive and global angiosperm evolutionary landscape of exon structure and relationships to domain organization, with additional information on expression patterns of gene containing exons of different sizes and provide valuable resources for diverse investigations of gene structure and function.

Results

Landscape of the number and size of exons among angiosperms

Relatively recent high-quality assembly of sequenced plant genomes has provided accurate gene structure annotation for genome-wide comparative examination of exons across angiosperms. To survey the exon size distribution, here we selected 46 angiosperm genomes, including 13 T2T genomes and 33 other high-quality genomes with chromosomal assemblies and four gymnosperm chromosomal genomes for comparison as outgroups (see methods and Supplementary Data 1). We examined a total of 9,188,769 exons of protein-coding genes from these 50 genomes and summarized the number of exons per genome in range from 79,276 to 607,568 (median: 165,147) (Fig. 1a–c, Supplementary Fig. 1; Supplementary Data 2), providing a dataset for large-scale in-depth exon comparison.

To compare the exon length distribution patterns among these genomes, exons were divided into five categories according to their lengths (Fig. 1d and Supplementary Figs. 1 and 2). In addition to the previously defined micro and small exons, for convenience, we referred to the exons with length ranging from 101 to 200 bp as 'medium exons', those from 201 to 500 bp as 'large exons', and those larger than 500 bp as 'super exons'. Among 126,433 exons from 24,040 genes that are annotated in the T2T *Arabidopsis* genome[60], 6.2% (7865) were micro exons, 28.3% (35,818) were small exons, 33.3% (42,050) were medium exons, 22.2% (28,015) were large exons, and 10% (12,685) were super exons (Supplementary Fig. 2). In the rice (*Oryza sativa* cv. *japonica*) T2T genome[54], we detected 15,014 micro exons (6.5% of 230,786 exons), 53,135 (23%) small exons, 65,246 (28.3%) medium exons, 56,350 (24.4%) large exons, and 41,041 (17.8%) super exons (Fig. 1d, Supplementary Fig. 2). Patterns of exon distribution similar to those in *Arabidopsis* and rice were also detected in other angiosperms, with medium exons being the largest size category (median of ~32%) among the five sizes (Fig. 1d). Across the 50 genomes examined here, genes containing medium exon(s) account for ~58.5% of the protein-coding genes in a genome (Supplementary Fig. 3a, b). Medium exons encode peptides with 33–66 amino acids and preliminary analysis suggested that they might frequently match the size range for angiosperm protein functional units (domains). We examined all annotated domains encoded by conserved exons in 1,528,298 genes from 50 genomes, and found that medium exons represent 32.9% (median value) of exons matching part(s) of and/or whole domains in one of the genomes (Supplementary Fig. 4a, b). Medium exons account for 18.5% of these exons matching >50% of a domain (Supplementary Fig. 4c). Moreover, for 17 to 692 domain families (depending on different angiosperm species), medium exons are enriched (p-value < 0.05 under two-sided hypergeometric test; see methods) among all exons encoding all/part of domains; such domain families include the MYB DNA-binding domain (PF13921), RRM 1 (PF00076), and WD40 (PF12894) (Supplementary Fig. 4d, e). For example, GO enrichment analyses of the *Arabidopsis* genes containing medium exon(s) suggested that these genes are associated with GO terms for growth (e.g., seed growth), response to stresses (e.g., cellular response to heat), regulation of gene expression (e.g., RNA endonuclease activity), and others (Supplementary Data 3). The results imply that the medium exons have a broad range of functions in plant growth and adaptation.

Among the five categories, the micro and small exons together account for ~35% of all exons per genome (Fig. 1d). In particular, the number of small exons per genome (median: 35,024; range from 22,970 to 160,327) is more than 3 times that of micro exons (median: 9757; range from 5174 to 49,288) (Fig. 1e and Supplementary Fig. 2; see the previous paragraph for numbers of exons of different size categories in *Arabidopsis* and rice). Genes containing micro and/or small exons represent ~53.5% of genes per genome (Fig. 1f), implying that the micro/small exons might be important structure elements of angiosperm genes. GO analyses of the *Arabidopsis* genes showed that GO terms significantly enriched only in the micro exon containing genes are related to growth (e.g., vegetative growth), regulation of gene expression (e.g., cis-regulatory region sequence-specific DNA binding), post-translational modification (e.g., proteasome binding), transport (e.g., sucrose transport) (Supplementary Data 3). GO terms significantly enriched only in the small exon-containing genes are related to development (e.g., petal development, corolla development, endosperm development), cellular process (e.g., meiosis II cell cycle process), protein modification (e.g., protein methylation), and others (Supplementary Data 3). GO terms significantly enriched in above two gene sets involve nucleosome assembly, protein refolding, vernalization response, histone deacetylase activity, and others (Supplementary Data 3). The GO enrichment results further suggest that micro/small exon-containing genes have diverse and important functions.

Large and super exons account for ~23% of exons per genome (Fig. 1d) and correspond to most (97.4% = 35.7% + 61.7%) genes without introns (Supplementary Fig. 3c). For the 46 angiosperms analyses here, the genes with a single exon (lacking an intron) account for 4.97% ~40.33% (median value: 21.21%) of all genes in each genome (Supplementary Fig. 3d; Supplementary Data 2). For example, 5482 *Arabidopsis* genes lack an intron, including 1570 with lengths of 200-500 bp (large exons) and 3592 in the length range of super exons, accounting for ~94% of these single exon genes (Supplementary Fig. 1 and Supplementary Data 2). In rice (*Oryza sativa* ssp. *japonica* cv. Nipponbare, T2T genome), 15,136 (26.39%) of 57,359 genes contain a single exon[54], with 6233 being large exons and 8176 super exons, accounting for 95% of the rice single-exon genes, indicating that most intronless genes fall in the large and super exon categories. In addition, the Chinese pine (*Pinus tabuliformis*) genome, known as the largest seed-plant sequenced genome to date[17], includes 38,784 singe-exon genes (with 156-6807 bp of exon length) that account for 50.87% of 76,234 genes (Supplementary Fig. 3d; Supplementary Data 2). Of the 38,784 single exons, 37,370 (96%) are either large (19,933) or super (17,437) exons (Supplementary Fig. 1 and Supplementary Data 2). Our results suggest that these long single-exon genes are an important part of the genome. GO terms significantly enriched in genes with a single large/super exon include phloem development, mRNA transcription, signaling receptor binding, ubiquitin-protein transferase activity, response to decreased oxygen levels, and others (Supplementary Data 3). Their lack of introns indicates that they are probably not regulated at the level of splicing, further suggesting relatively uniform functions among organs and growth conditions.

In intron-containing genes (median of ~78.79% of all genes in an angiosperm genome), the exons have a median length of 127 bp (ranging from 110 to 151 bp among angiosperms) (Supplementary Data 2). For example, the eudicot carrot (*Daucus carota*; T2T genome[52]) has 28,606 (78.87%) intron-containing genes with a median exon length of 129 bp, including 19,380 genes with micro and/or small exons (Supplementary Data 2). In the triploid banana (*Musa acuminata*; a monocot) T2T genome[61], 24,807 intron-containing genes from the subgenome BXJ2 have a median exon length of 132 bp and 65% of these

https://doi.org/10.1038/s41467-025-66816-3

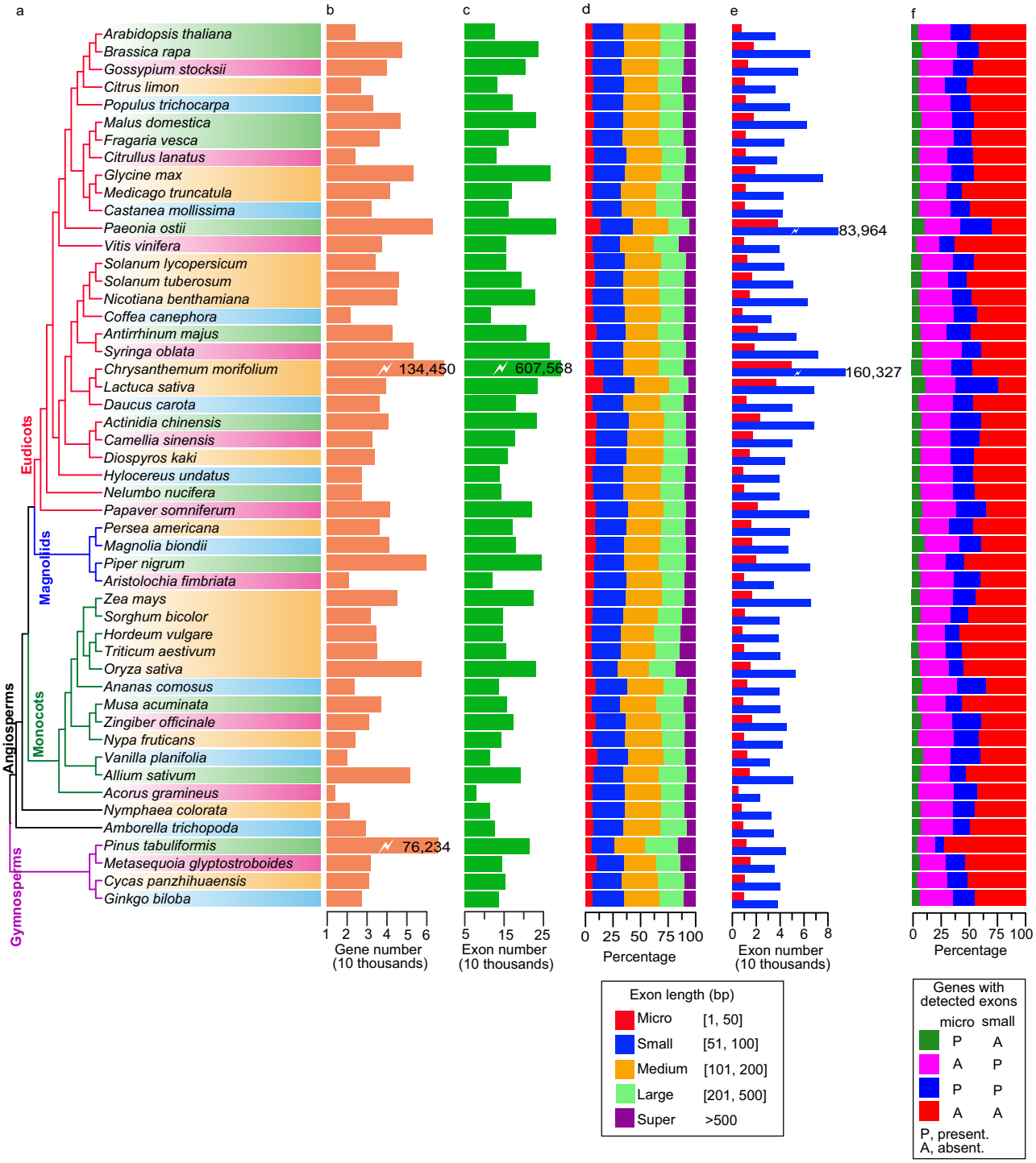


Fig. 1 | Distribution of exon numbers in 50 species. a A phylogeny of 46 angiosperm species, with 4 gymnosperms as outgroups. Branches with different colors represent different groups as the group name shown in left. To the right of terminal branches are species names, with two or more species from a family (see taxonomic levels in Supplementary Data 1) showing in the same background color. Each species corresponds to the information in its right following parts. **b** Bar-plots of number of the protein-coding genes in each genome suggesting genomic complexity across these species. See Supplementary Data 2 for number of genes from each genome. **c** Bar-plots of number of exons in protein-coding genes in each genome showing a pattern similar to that in part b. The number of exons in each genome can be found in Supplementary Fig. 2. **d** Bar-plots of percentage of exons with different length ranges in each genome. Five colors represent five exon length ranges as shown below. Red, blue, orange, green, and purple bars represent micro, small, medium, large, and super exons, respectively. See Supplementary Fig. 2 for the number of exons in different size-categories. The distribution of number of exons in each length category in each genome can be found in Supplementary Fig. 1. **e** Bar-plots of number of micro (red bars) and small (blue bars) exons in each genome indicating ~10 thousands of micro exons in most genomes and the small exons are more than 3 times of micro exons. **f** Bar-plots of percentage of number of genes with or without micro/small exons. Four colors represent four gene types as shown below. Green bars represent genes including micro exons and excluding small exons. Pink bars represent genes including small exons and excluding micro exons. Blue bars represent genes with micro and small exons. Red bars represent genes without micro and small exons. Source data are provided in a Source Data file.

genes carry micro and/or small exons (Supplementary Data 2). Magnoliids have a pattern similar to eudicots and monocots (Fig. 1b); for instance, 70% of 27,730 avocado (*Persea americana*) genes[62] include micro/small exons (Supplementary Data 2). The above results indicate that most of the protein-coding genes have introns, which depend on proper splicing for function and can be regulated at the splicing level. The possibility of alternative splicing can also increase the protein structural and functional diversity of these genes. To test this, we examined published alternative splicing datasets in *Arabidopsis* to estimate the number of affected exons in different length categories. Among 1531 exons exhibiting exon-skipping under 72-h heat stress relative to control in *Arabidopsis*[63], 323 (21.1%) are micro exons, 780 (50.95%) small exons, 325 (21.23%) medium exons, 91 (5.94%) large exons, and 12 (0.78%) super exons (Supplementary Data 4). The percentage (50.95%) of small exons among skipped exons is greater than genome-wide percentage (28.32%) of small exons (two-sided hypergeometric test, p-value = 1.11E-77; Supplementary Data 4). In another study of poplar root transcriptome under lead stress relative to control[24], among 113 skipped exons, 19 (16.81%), 52 (46.02%), 28 (24.78%), 12 (10.62%), and 2 (1.77%) are micro, small, medium, large, and super exons, respectively, suggesting that alternative splicing, including those involving micro and small exons, contributed to transcript/protein diversity in stress response. These cases further suggest that micro, small, and medium exons likely represent frequently skipped exons for increasing functional diversity.

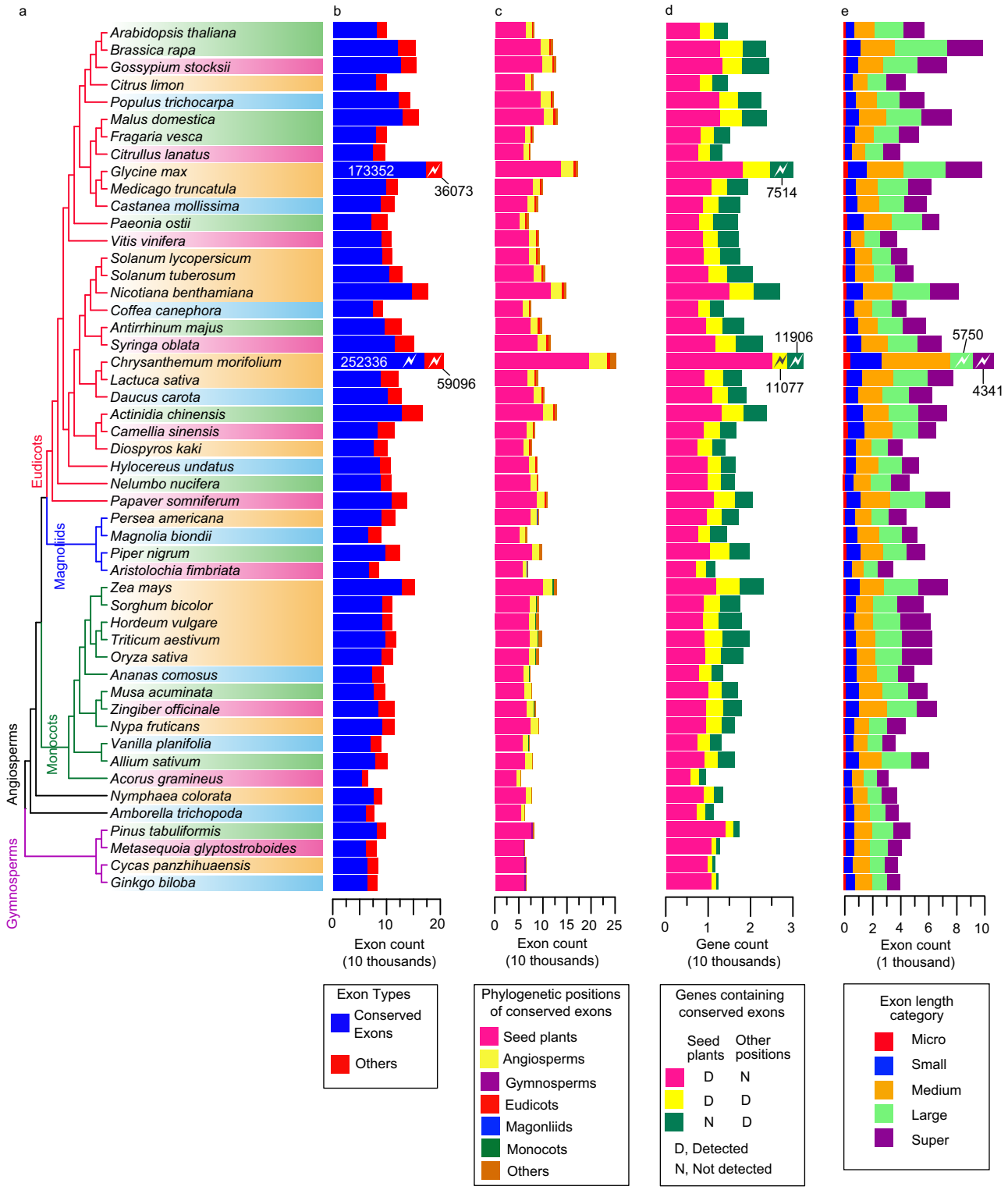
Phylogenomic analyses of conserved exons among angiosperm lineages

The above results of similar distributions among diverse angiosperms of both exon sizes and genes containing either micro/small exons and those containing a single large/super exon suggest possible evolutionary conservation of exon organizations among homologous genes, although these genes could differ in exon numbers and sizes (Fig. 1). To further detect the conservation of individual exons shared by genes of different angiosperm subgroups, we compared the exons among genes in an orthogroup with members related by gene ancestry (Fig. 2). Conserved exons were previously examined by manually comparing exon positions in multiple sequence alignments of the MADS-box[22] and *CYP75*[40] gene families; in addition, conserved short micro exons (≤15 bp) were determined along with the conservation of their upstream and downstream exon sequences[48]. However, an algorithm for genome-wide detection of conserved exons among multiple species is unavailable. To identify conserved exons, we performed phylogenomic analyses for orthogroups with >5 genes from >2 species (including ≥1 angiosperm), generating 16,754 gene trees (see methods). The genes in these gene trees contain a total of 6,347,314 exons, including 443,716 micro exons, 1,822,032 small exons, 2,123,073 medium exons, 1,345,267 large exons, and 613,226 super exons (Supplementary Fig. 3). We define conserved exons between two or more genes in the same orthogroup in three types: (1) highly conserved exons as having at least 50% amino acid sequence similarity; (2) moderately conserved micro exons with 30%-50% amino acid sequence similarity and moderately conserved exons in four other length categories showing at least 20% sequence similarity; and (3) candidate conserved exons as exons if their flanking exons are conserved as defined following criteria 1 or 2 (see methods). For statistics of conserved exons across angiosperms (rather than a species), each conserved exon family containing members from genes of an orthogroup (referred to as "conserved exon" hereafter for brevity, unless otherwise noted) is counted once regardless of the number of genes containing the exon. Following these criteria, we developed a new bioinformatic tool, ExonEvo, to identify conserved exons in annotated genomes using phylogenomics (see methods). We then used ExonEvo to compare genes in the above-mentioned gene trees and identified 141,229 conserved exons (including 122,059 highly conserved exons, 14,548 moderately conserved exons, and 4622 candidate conserved exons; Supplementary Fig. 5), each of which is present in two or more genes from the 50 genomes. In individual species, the conserved exons represent 69.8%-84.9% (median 79.9%) of total annotated exons (Fig. 2a, b). Among these conserved exons, 14,107 are micro exons (Supplementary Data 5), 44,336 small exons (Supplementary Data 6), 45,961 medium exons (Supplementary Data 7), 27,364 large exons (Supplementary Data 8), and 9461 super exons (Supplementary Data 9).

Among the 6347 *Arabidopsis* micro exons included in the phylogenomic analyses, 70.8% (4496) are conserved; phylogenomic results also indicate 71.1% (7007) of 9861 micro exons in maize (*Zea mays*) are conserved. Among the plant micro exons examined here by phylogenomics, the conserved micro exons account for 31.6%-74.1% (median: 60.55%) of the detected micro exons per species (Supplementary Fig. 6a, b), suggesting that different lineages might have retained different numbers of micro exons during evolution (see below), although it is possible that some species might also have gained some micro exons. In addition, conserved small exons account for 67.8%-91.4% (median: 85.7%) of the detected small exons per species (Supplementary Fig. 6c). For instance, 89.1% (30,069) of 33,747 detected small exons in tomato (*Solanum lycopersicum*) are conserved and 91.3% (30,294) of 33,183 detected small exons in barley (*Hordeum vulgare*) are conserved. The higher percentage of conserved small exons than that of the micro exons suggests that the small exons might have been under greater selection possibly due to their greater structural and functional contribution to the corresponding genes, whereas the higher variability of micro exons among species might allow differential splicing diversity in a species/lineage-dependent manner. Furthermore, the median percentage of the conserved medium, large, and super exons per species was estimated as 82.3%, 70.2%, and 44.9%, respectively (Supplementary Fig. 6d–f). The variability of the percentage of conserved exons per species among five categories suggests evolutionary plasticity of gene structures of different gene orthogroups among angiosperm species.

Among these conserved exons, 14,107 conserved micro exons were detected in 5762 orthogroups (containing a total of 182,094 genes, when counted separately for genes of individual species) from across 50 seed plants included here (Supplementary Fig. 7). Among these 5762 orthogroups, 80.9% (4659) have genes from at least one angiosperm and one gymnosperm and are considered conserved among seed plants. These 4659 seed-plant orthogroups contain 81.5% (11,502) of all conserved micro exons (Supplementary Fig. 7), implying that most conserved micro exons are derived from the most recent common ancestor (MRCA) of seed plants. In addition, conserved small exons (44,336) were detected in 10,345 orthogroups (containing 494,140 genes; Supplementary Fig. 7). A comparison of these 10,345 orthogroups indicated that 76.9% (7953) of these orthogroups have genes from at least one angiosperm and one gymnosperm and that these 7953 orthogroups include 36,399 conserved small exons, accounting for 82.1% of 44,336 conserved small exons, again suggesting that most conserved small exons are derived from the MRCA of angiosperms and gymnosperms. Moreover, conserved medium exons were found in 12,119 orthogroups (containing 613,886 genes); conserved large exons were identified in 12,395 orthogroups (with 553,124 genes); and conserved super exons in 6453 orthogroups (243,260 genes; Supplementary Fig. 7). A comparison of orthogroups including conserved medium, large, and super exons retrieved patterns similar to those for micro and small exons: 9074 seed-plant orthogroups contain 80.2% (36,854) of all conserved medium exons; 8802 seed-plant orthogroups include 74.1% (20,283) of all conserved large exons; and 4,102 seed-plant orthogroups contain 65.3% (6175) of all conserved super exons. Hence, most of the conserved exons are likely derived from the MRCA of seed plants.

https://doi.org/10.1038/s41467-025-66816-3



To further investigate the conservation patterns of exons, including those that are not found in both angiosperm and gymnosperm genes (as described above), we used ExonEvo to infer the MRCA of the genes containing conserved exons in gene trees and to reconcile the MRCA to the species tree (Supplementary Figs. 8–13). Among the retrieved patterns, the largest group (median of 78.7% among species) of the conserved exons are shared by seed plants (Fig. 2c), supporting the above results of orthogroup analyses. Among the conserved micro exons, 3411 exons were mapped to the MRCA of seed plants, 2353 to the MRCA of angiosperms, 1292 to that of eudicots, 274 to that of monocots, and 10,188 to other phylogenetic positions (Supplementary Fig. 9). The pattern of conserved micro exons shared by seed plants and various other groups implies that the retained exons have experienced diverse retention and loss events across angiosperms since they diverged from gymnosperms. In addition, our ExonEvo analyses of phylogenetic positions of the small exons showed that 22,120 exons are shared by seed plants, 7596 by angiosperms (but not gymnosperms), 3002 by eudicots, 1393 by rosids (124 + 146 + 318) or asterids (73 + 485 + 247) (both largest eudicot subgroups), 5614 by other groups represented by two or more species here, and 4515 by

Fig. 2 | Summary of conserved exons in 50 species. a A phylogeny of angiosperm species, with tree information same as that in Fig. 1a. **b** Bar-plot of the number of conserved exons and others in orthogroups. Each bar represents the exon count per species. Blue bars represent the number of conserved exons detected by our ExonEvo. Red bars represent the number of non-conserved exons and possibly failed exons for detecting conservation via ExonEvo. Summaries of conserved exons in each length category in each genome can be found in Supplementary Fig. 6. See micro, small, medium, large, and super conserved exons in Supplementary Data 5–9, respectively. **c** Bar-plot of number of the conserved exons with different phylogenetic distribution. Seven colors of bars represent seven different phylogenetic position categories of species shared an exon, as shown below. A detailed distribution of the number of conserved exons mapped at different phylogenetic positions of species-tree can be found in Supplementary Figs. 8–13. The largest part of bars in all species shows exons shared by seed-plants, suggesting most exons seed-plants are conserved. **d** Bar-plot of number of genes including or excluding conserved exons in each genome. Three colors of bars represent three categories of genes including or excluding the seed-plant specific or other conserved exons, as shown below. Red part represents the genes including the conserved exons exclusively shared by seed-plants. Yellow part represents the genes including the conserved exons mapped at seed-plant ancestor and that mapped at other phylogenetic positions of species tree. The green part represents the genes with conserved exons, excluding those mapped at the seed-plant ancestor. Genes including the seed-plant specific conserved exons account for the large part of all genes including conserved exons in each genome, suggesting the seed-plant specific conserved exons affecting most genes. **e** Bar-plot of number of the exons (part of the red exons in part b) that are merged from two or more conserved exons. Five colors represent five exon length ranges as shown below. A detailed example of the exon merge can be found in Supplementary Fig. 15. Source data are provided in a Source Data file.

paralogues within a single species (Supplementary Fig. 10). The diverse conservation landscapes of the small exons across angiosperms suggest that gene structures have changed involving small exons during angiosperm evolution. ExonEvo analyses also suggested that 53.6% (24,634) of all conserved medium exons, 45.8% (12,545) of all conserved large exons, and 38.8% (3675) of all conserved super exons are mapped to the ancestor of angiosperms and gymnosperms (Supplementary Figs. 11–14), implying different retention patterns of these three exon-size categories after angiosperms diverged from gymnosperms. Overall, 66,385 conserved exons were mapped to the MRCA of seed plants, 26,454 to angiosperm ancestor, 1919 to gymnosperm ancestor, 10,160 to eudicots, 2108 to monocots, 271 to magnoliids, 14,118 to different lineages, and 19,814 to other phylogenetic positions (Fig. 3a and Supplementary Figs. 9–13). It seems possible that the ancestral gene structure genes might be differentially retained during evolution, allowing for changes in splicing and protein function.

Among the 50 genomes, 913,316 genes have conserved exons; 7773-36,354 (median 12,989) genes in specific genomes contain seed plant-wide conserved exons, representing 65.2%-95.6% (median 76.7%) of the annotated genes per genome (Fig. 2d). The seed-plant conserved exons affect most genes and, therefore, likely have the greatest importance in gene evolution. For example, 11,355 *Arabidopsis* genes contain at least one seed-plant conserved exon; among these 11,355 genes, 8090 genes only contain conserved exons that are shared by seed plants, while the remaining 3265 genes also contain conserved exons shared by other groups, such as angiosperms. In rice, 13,021 genes contain at least one seed-plant conserved exon; among these 13,021 genes, 9361 genes only contain seed-plant conserved exons, while 3660 genes also contain conserved exons of other phylogenetic positions. Similar conservation patterns of exons were found in other genomes, with most genes containing conserved exons that are shared by seed plants (Fig. 2d). The results suggest that the ancestral exon structures have been retained in most genes, with likely conservation of possible splicing regulation and protein isoforms.

In addition to the conserved exons, we also surveyed the exons that are merged from two or more conserved exons [due to loss of intron(s)] and retrieved 3236-17,770 (median 5786) merged exons in individual species (Fig. 2e). For example, genes encoding ubiquitin-conjugating enzyme 19 (UBC19) have two conserved micro exons, two conserved small exons, one conserved medium exon, and one conserved large exon (Supplementary Fig. 15). Among these six exons shared by seed plants, five conserved exons were merged into a large exon in the grass *UBC19* genes (Supplementary Fig. 15), implying that grasses might have lost the functional plasticity of ancestral *UBC19* genes related to transcript/protein isoforms in comparison to other seed plants.

As the seed-plant conserved exons are the largest group, accounting for 47% of all conserved exons (Fig. 3a), we further examined their retention in major groups of angiosperms, including two major subclades of eudicots, rosids (such as *Arabidopsis* and soybean) and asterids (such as tomato), other eudicots, monocots, magnoliids, and other angiosperms. Among the seed-plant conserved exons, numbers of exons in each of the five size categories shared by various combination of the angiosperm groups are shown in Fig. 3b–e. Specifically, 99% (65,722) of the seed-plant exons were detected in at least one genome of core angiosperms (here including eudicots, monocots, and magnoliids), while 85.3% (56,630) were detected in at least one of the two basal angiosperm genomes (Fig. 3b). Small and medium exons together represent the majority (70.4%; Fig. 3c) of seed-plant conserved exons. Furthermore, 77.3% (51,309) of the seed-plant conserved exons are retained in all six angiosperm groups (Fig. 3d), including 2442 micro, 18,004 small, 19,478 medium, 9129 large, and 2256 super exons (Fig. 3e). Among 3411 micro exons, the largest portion (2442) represents the conserved exons shared by all six angiosperm groups (Fig. 3c–e). For instance, among the seed-plant homologs of the *Arabidopsis MKK6* gene (*AT5G56580*), the conserved micro exon (EXON0044515; Supplementary Data 5) is detected in rosids (such as *Arabidopsis* and apple) and asterids (such as tomato and kiwifruit), other eudicots (such as Papaver), monocots (such as rice and maize), magnoliids (such as avocado), and other angiosperms (such as *Amborella trichopoda*). Among the remaining 969 (=3411-2442) micro exons, 287 exons are detected in all but one group. For example, 127 exons are present in core eudicots and other angiosperms but absent in early diverging eudicots (*Nelumbo nucifera* and *Papaver somniferum*). Among the 127 exons, 41 were apparently merged with adjacent exon(s) with the loss of the intervening intron, resulting in longer exons in *N. nucifera* and *P. somniferum* and retention of the protein coding sequences. For another 42 exons that were absent in *N. nucifera* and *P. somniferum*, the entire genes were not detected in these plants. The remaining 44 (=127-41-42) conserved exons were part of members of orthogroups, but the genomic sequences of the corresponding genes in *N. nucifera* and *P. somniferum* lacked a detectable similar sequence to the conserved exons, suggesting that the exons are not required for the function of these early diverging eudicot genes. On the other hand, 300 exons are detected in only one angiosperm group; for example, 77 exons present in monocots but absent in other angiosperm groups include 44 exons that were merged into long exons in at least one other angiosperm groups, suggesting specific retention in one group and fusion or losses in others during the angiosperm history. Analyses of these 22,120 seed-plant conserved small exons showed that the largest portion (18,004) represents the small exons shared by all six angiosperm groups, again suggesting angiosperm-wide retention of these small exons. Among 4,116 other small exons, 1412 exons are absent in only one of the groups (including 370 ones merged into other exons in one corresponding group), whereas 1022 exons are present in only one group (543 of them fused into long exons in other groups), showing different retention and

https://doi.org/10.1038/s41467-025-66816-3

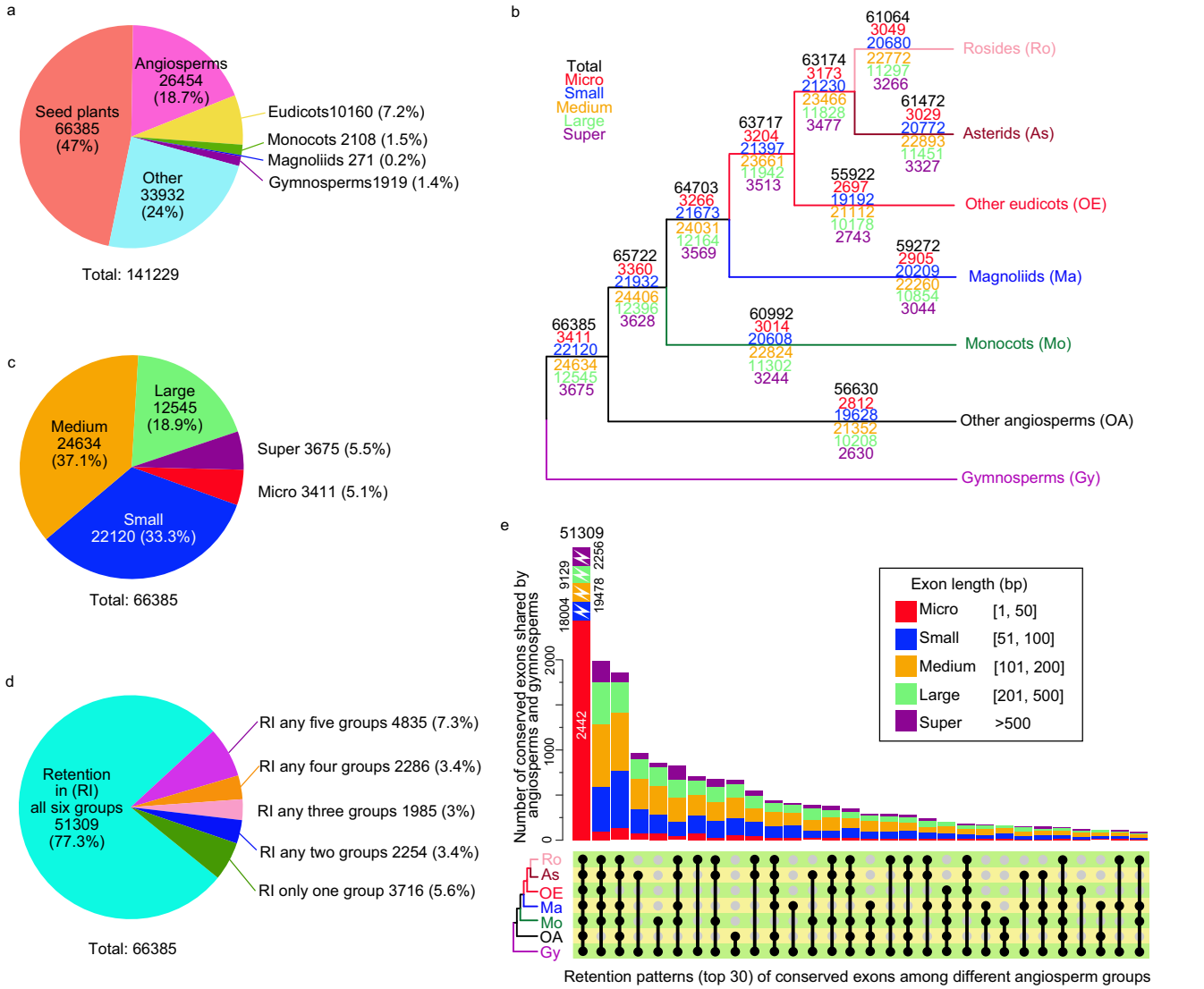


Fig. 3 | Comparison of the retention patterns of conserved exons among major groups. a A pie chart of the percentages of the conserved exons mapped at different phylogenetic positions of the species tree, showing that most of conserved exons are shared by seed plants. **b** Number of the 66,385 seed-plant conserved exons retained at the MRCA of different angiosperm groups. Numbers above/below a branch represent the number of conserved exons retained in at least one of the descendant groups derived from this branch. The black number represents the total number. Red, blue, orange, green, and purple numbers represent the number of micro, small, medium, large, and super exons, respectively. **c** A pie chart of the percentages of the seed-plant conserved exons in different exon length categories. **d** A pie chart of the percentages of the seed-plant conserved exons with different retention patterns among six angiosperm groups. See these angiosperm groups in part b. **e** An Up-Set plot of the number of seed-plant conserved exons retained in different plant groups. Points below bars represent retention of the exon in different groups of seed plants. Lines joining different points represent different retention patterns. The bar plot shows the number of exons with different retention patterns. The largest number of conserved exons mapped at the MRCA of seed plants represents these exons retained in all six angiosperm groups, indicating a high degree of conservation. Five colors of the bars represent five exon length categories as shown in the inset box. Source data are provided in a Source Data file.

fusion patterns of the ancestral exons among angiosperm groups. In addition, analyses of three other exon-size categories revealed that 19,478 medium exons, 9129 large exons, and 2256 super exons are shared by all six angiosperm groups, suggesting that all angiosperm groups retained these exons after their ancestor diverged from gymnosperms. Generally, 77.3% (51,309 = 2442 + 18,004 + 19,478 + 9129 + 2256) of the 66,385 seed-plant conserved exons are retained in all angiosperm groups (Fig. 3b, c). Among the remaining 15,076 exons, 4675 seed plant conserved exons are detected in nearly all groups but one group (1313 merged in one corresponding group), while 3716 exons are detected in only one of the groups (1867 merged in other groups). The patterns imply that most of the seed-plant conserved exons have been retained in all angiosperm groups, likely allowing the conservation of possible regulation at the splicing level during the diversification of the vast number of angiosperms.

Diverse patterns of conserved exons corresponding to protein domains

As both exons and protein domains have variable sizes, there could be several possible corresponding relationships between exons and their encoded protein domains. Sequences of two or more exons might encode a single domain, or a single (relatively large) exon could encode an entire domain. Conversely, a single exon can encode portions or all of multiple domains. When a domain is encoded by (parts of) multiple exons, it could have structurally different isoforms due to alternative splicing involving one or more of the corresponding exons, with potentially different functions. On the other hand, a domain

https://doi.org/10.1038/s41467-025-66816-3

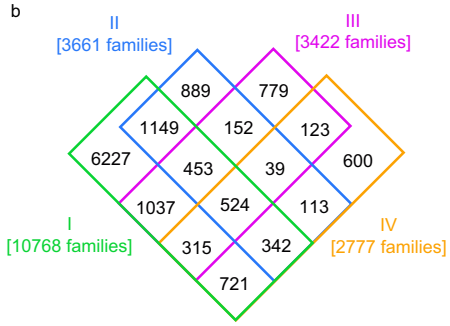
a

Summary of conserved exons matching domains

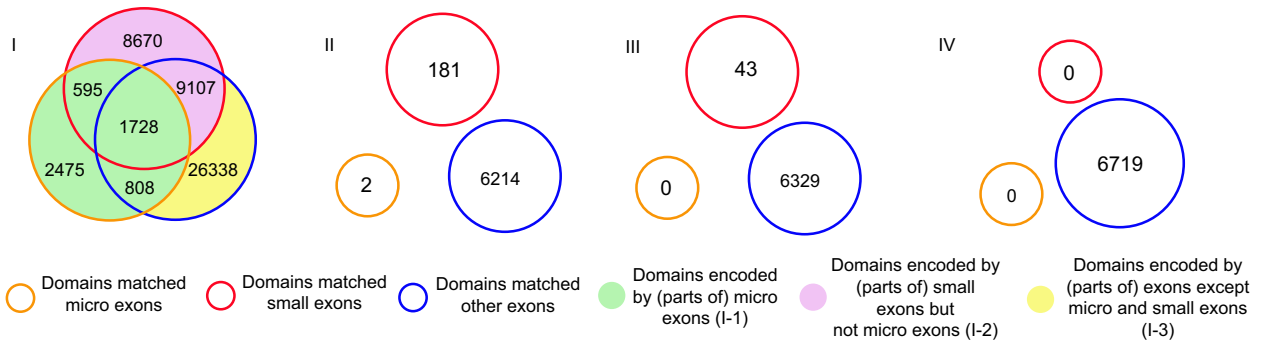
Category	Schematic correspondence between exons and domains	Number of conserved exons	Number of domains	Number of domain families
I	Domain / Exon 1 Exon ... Exon n	92868 (54195)	49721	10768
II	Domain / Exon	6400 (2196)	6400	3661
III	Domain 1 Domain ... Domain n / Exon	5087 (2637)	6372	3422
IV	Domain 1 Domain ... Domain n / Exon	2454 (919)	6719	2777

(#) number of exons shared by seed plants

b



c



d

Fig. 4 | Domains encoded by conserved exons. a Summary of conserved exons matching domains. Four categories of the correspondence are illustrated here. Category-I represents a domain matching two or more exons. Category-II represents a domain encoded by a single exon. Category-III represents an exon encoding at least one whole domain and matching part of other domains. Category-IV represents an exon encoding two or more whole domains. See examples of Category-I, II, III, and IV in Supplementary Figs. 16–19. To the right of these illustrations are the number of conserved exons, domains and domain families in each category. The number within parentheses represents the number of conserved exons shared by seed plants. See Supplementary Data 10 for the exon-domain correspondence in genes from 50 genomes. **b** Venn diagram of domain families among four categories suggesting different patterns of exon splicing. Families overlapped two or more categories implying that exon structures matching domains have evolved to be diverse in responsible for the possible different splicing roles (Supplementary Figs. 20–26). **c** Venn diagram of domains matching micro, small, and/or other conserved exons in each category of part a. Red, orange, and blue circles represent the domains matching micro, small, and other conserved exons, respectively. The green background of venn diagram represents the domains matching micro exons in category-I (I-1). The pink background of venn diagram represents the domains matching small exons but not micro exons in category-II (I-2). The yellow background of venn diagram represents the domains matching conserved exons excluding the micro and small exons (I-3). **d** Examples of domain-exon correspondences in I-1, I-2, and I-3. In each case, top orange box represents the focal domain that matches micro (red box), small (blue boxes), medium (green box), and/or large (purple box) conversed exons. Left: an I-1 case for the RNA recognition motif (RRM_1) domain in *ELF9* (*EARLY FLOWERING9*) gene. Medium: a case of I-2 is for the AP2 domain in *WRI5* gene, a homologous gene of AM-induced AP2 transcription factor. Right: an example of I-3 for the WRKY DNA-binding domain in *WRKY40* gene.

encoded by a single exon would have a uniform structure in all transcripts containing the exon with the same function. However, the genome-wide patterns of correspondence between exons and domains have not been reported in part due to the lack of analyses of the conservation of exons. In addition, different gene families can have similar sequences of a shared domain that is characterized by a PFAM profile. We defined a domain family as containing domains detected in one or more orthogroups using the same PFAM profile. The domain from each orthogroup is referred to as a member of the domain family. Whether different members of the same domain family can have different patterns of domain-exon organization is not clear on a genome level across many angiosperms, even though such knowledge can help us understand possible exon splicing patterns concerning domain functions. The above results on angiosperm exon conservation provide an opportunity to explore the evolutionary patterns of domains encoded by conserved exons in angiosperms. We found that 106,809 (=92,868 + 6400 + 5087 + 2454; ~76%) conserved exons have annotated domain information (Fig. 4a), whereas 34,420 (~24%) conserved exons with annotated open reading frames for protein-coding sequences (including 17,288 exons being the first or last exons in genes and 17,132 exons in other positions) do not match annotated domains (not even "functional unknown domains"); these conserved protein-coding regions could be added to future domain databases using updated gene annotation algorithms or supported by new functional studies.

According to the relative positions between exons and their corresponding domain, we assigned exons into four categories of "domain-exon organization" (Fig. 4a; see Supplementary Data 10 for the correspondence in individual conserved exons in 50 genomes). As an example, in seed plant homologs of the *MALE STERILITY 2* (*MS2*)[64], a region of ~912-bp encoding a NAD_binding_4 domain (an essential domain for *MS2* gene function[65]) corresponds to 6 conserved exons shared by seed plants (Supplementary Fig. 16). Such a pattern of multiple exons encoding a domain is referred to as category-I here. Among 106,809 conserved exons with annotated domain information,

92,868 exons (including 54,195 exons mapped at the MRCA of seed plants) belong to category-I and represent the largest fraction of all conserved exons that matched domains. These exons match 46,399 domains, which belong to 10,765 domain families according to sequence similarity and Pfam definitions (Fig. 4a). The functions of most domains in angiosperms are probably conserved based on their matched conserved exons. In addition, we defined the pattern of a single exon matching only one whole domain as category-II (Fig. 4a). For example, sequences (138 bp) of a single exon (in median length of 212 bp) shared by angiosperms encodes a MYB DNA-binding domain (PF00249; Supplementary Fig. 17). Category-II includes 6400 conserved exons encoding 6400 domains of 3661 domain families (Fig. 4a), representing the second largest portion of all conserved exons that matched domains. Furthermore, we also observed cases where a single exon matches multiple domains (Fig. 4a). For instance, in angiosperm *WRKY37* genes (Supplementary Fig. 18), the WRKY DNA-binding domain (PF03106) is encoded by a portion of a conserved medium exon (138 bp) shared by angiosperms and a portion of a 538-bp super exon, which also includes sequences encoding another domain (similar to Hid1: High-temperature-induced Dauer-formation1; PF12722) with a signaling function in worms. Such a pattern of an exon encoding at least one whole domain and part of another domain is defined here as category-III (Fig. 4a). To avoid overlapping definitions, domains entirely encoded by category-III exons are referred to as category-III domain(s), whereas the domains partially encoded by category-III exons and partially by additional exon(s) are assigned to category-I. There are 5087 category-III exons and they encode 11,455 domains of 3422 domain families (Fig. 4a). When a single exon encodes multiple whole domains and does not match any part of other domains, the correspondence was referred to as category-IV (Fig. 4a). An example of category-IV can be found in *AUXIN RESPONSE FACTOR* (*ARF*) genes (functioning in flora development, such as the ref. 66) with a super exon of 1101 bp conserved among seed plants; regions of this super exon with 306 bp and 252 bp encode, respectively, a B3 DNA-binding domain (PF02362) and an ARF domain (PF06507) (Supplementary Fig. 19). Category-IV has 2454 exons that encode 6692 domains of 2777 domain families (Fig. 4a). These four categories provide a general portrait of relationships between conserved exons and their encoded domains in seed plants.

Homologous genes encoding members of the same domain family can have different exon-intron structures; therefore, domains with different domain-exon categories (as described above) can be encoded by homologs and belong to the same domain family. To probe the evolutionary dynamics of exon-intron structure, we compared exon/domain categories among domain families and obtained the overlap between the four sets of domain families. Among 13,463 domain families recognized in the PFAM database and observed here (Fig. 4b), 8495 (~63.1%) have only one category of exon-domain correspondence (Categories-I, II, III, and IV with 6227, 889, 779, and 600 families, respectively); whereas only 524 (~3.9%) domain families have all four categories, with the remaining families having two or three categories (Fig. 4b). Thus, most of domain families have only one category of exon-domain correspondence, suggesting a high degree of evolutionary stability of exon-intron structure. Specifically, 6227 domain families are only found in category-I, such as the NAC domain (PF01849). The domains in these families match multiple exons and, hence, might have evolved to have diverse functions due to possible alternative splicing of exons. Also, 889 families have only category-II domains corresponding to a single exon, implying that these families might have more similar functions in comparison with the families with domains matching multiple exons (category-I). Further, 779 and 600 families were only detected in category-III and category-IV, respectively (Fig. 4b), suggesting that two or more domains encoded by the same exon are functionally related or coupled.

Although only ~37% of the domain families analyzed here include two or more categories of exon-domain correspondence, these domain families include homologous genes with different exon-intron structures, suggesting that different gene members might have differential regulation in splicing. Specifically, different members of the Homeodomain (PF00046) family were found for each of the four categories. In category-I, for example, the *SAWADEE HOMEODOMAIN HOMOLOG1* (*SHH1*) genes (Supplementary Fig. 20) contain a 46-bp micro exon and an 86-bp small exon encoding parts of the homeodomain and function upstream in the RNA-directed DNA methylation pathway in *Arabidopsis*[67]. Other members of this family, such as the angiosperm *WOX4* genes (Supplementary Fig. 21) belong to category-II, with a conserved large exon, part of which encodes the homeodomain of WOX4 proteins. In *Arabidopsis*, the *WOX4* gene acts to mediate the response of the cambium to auxin[68] and in *Populus WOX4* regulates cell division in the vascular cambium[69]. For category-III, the homeodomain is encoded by a conserved medium exon (Supplementary Fig. 22) in seed-plant homologs of the *LATE MERISTEM IDENTITY1* (*LMI1*) gene that modulate the organ shape, such as stipule proportions in *Arabidopsis* by regulating endoreduplication[70]. An example of a category-IV homeodomain is that encoded by a large exon conserved among angiosperms (Supplementary Fig. 23) in the *ZHD8* genes [the orthologues in the G13 clade of *ZHD* gene family tree[71]]. The rice *ZHD8* gene (*AGIS_Os04g031270* = *LOC_Os04g35500* = *Os04g03477*) expression was up-regulated under cold, drought, and ABA stresses[72], suggesting a role in abiotic stress response. In addition, the AP2 domain (PF00847) family includes members in categories-I, -II, -III, and -IV. For instance, in category-I, the AP2 domain corresponds to two micro exons and one small exon in the *AP2* genes (Supplementary Fig. 24); for category-II, the AP2 domain is encoded by part of a super exon in *ERF23* genes conserved in seed plants (Supplementary Fig. 25); for category-III, the AP2 domain is encoded by a conserved super exon (EXON0026954) in seed plants; in category-IV, an AP2 domain corresponds to a portion of an angiosperm super exon in the *RAV3* genes (Supplementary Fig. 26). We found 522 other domain families with members in each of the four exon-domain categories (Fig. 4b), such as the MYB-like (PF00249) and helix-loop-helix DNA-binding domains (PF00010), protein kinase domain (PF00069), RNA recognition motif (RRM_1; PF00076), and leucine rich repeat (PF13516). Similarly, other domains families with members in two or three categories include the cytochrome P450 domain (PF00067), the sugar (and other) transporter (PF00083), ubiquitin-conjugating enzyme (PF00179), and the universal stress protein family (PF00582). These results with families of homologous protein domains belonging to two to four exon-domain categories suggest that domains with similar molecular activities might be encoded by transcripts that could have different exon splicing patterns. It is possible that the variability of exon structures within the same domain family allow some members to have a stable protein structure, thus function, but facilitate other members to have structural/functional isoforms.

Protein isoforms that differ due to the presence or absence of relatively short peptides encoded by micro or small exons could differ in some functional aspects, but not losing all protein activities. To estimate the number of micro, small, and other exons encoding domains, we compared the exons among the four categories (Fig. 4c). In category-I, 1728 domains correspond to genic regions containing both micro and small exons, and exon(s) of other size(s); for example, the alternative splicing regulator domain (PF09750) in *SR16* genes corresponds to five seed-plant conserved exons (Supplementary Fig. 27), including one micro exon, one small exon, two medium exons, and one large exon. Including of other subcategories, 5606 (=2475 + 595 + 808 + 1728; Fig. 4c) domains correspond to genic regions with micro exons (for convenience, referred to as I-1) and

possibly exons of other sizes. A case of I-1 is the RRM_1 domain matching four (one micro, two small, and one medium) conserved exons in *ELF9* gene (Fig. 4d). Among those lacking micro exons, 17,777 (=8670 + 9107) domains match regions with small exons (called I-2; e.g., the AP2 domain in *WRI5* gene; Fig. 4d) and possible others; and genic regions for 26,338 domains lack micro and small exons (referred to as I-3; e.g., the WRKY domain in *WRKY40* gene; Fig. 4d). In short, 23,383 (=5606 + 17,777) domains correspond to regions with micro and/or small exons and represent 47% of the domains in category-I, making alternative splicing of these micro/small exons a possible regulatory mechanism of domain structures. In addition, medium, large, super exons contribute to the domains of category-IV and most of the domains in category-II and category-III (Fig. 4c).

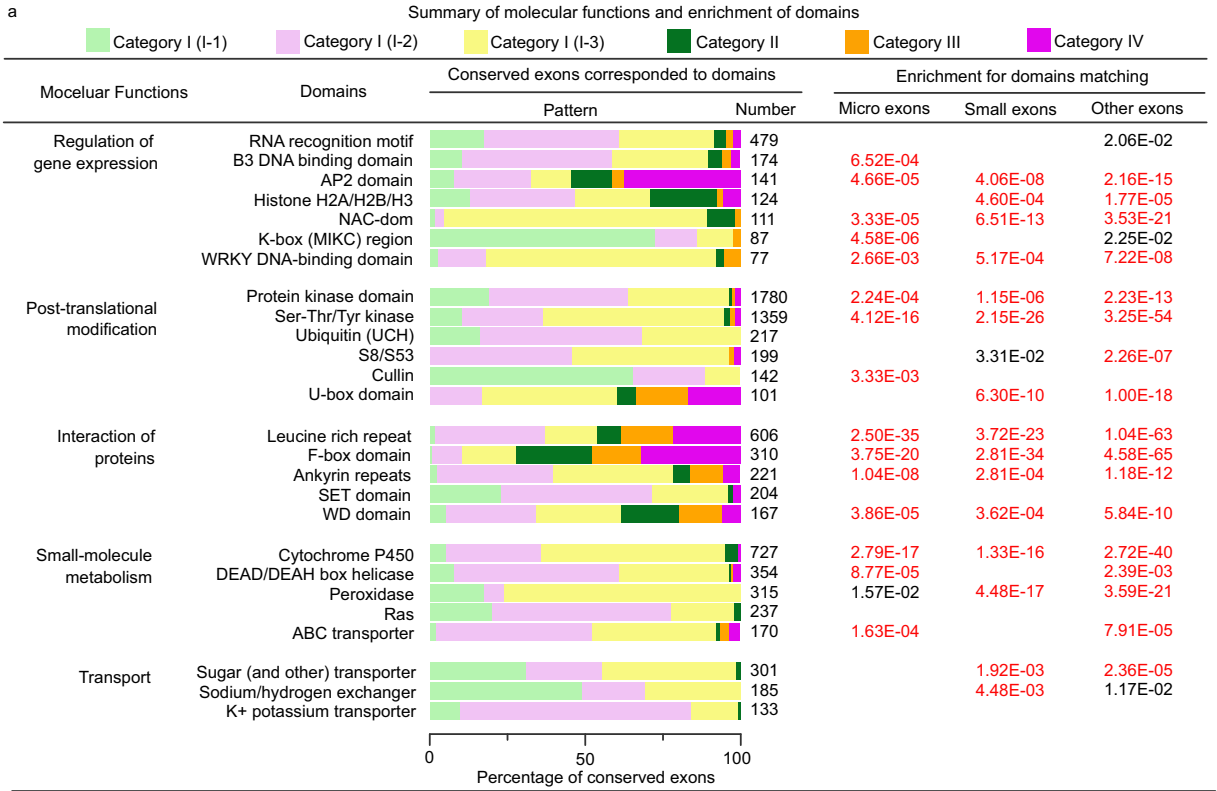
To obtain additional clues about the molecular functions of domains in different categories of exon organization, we examined various GO terms of these domains (methods) and present those associated with relatively large numbers of domains (Fig. 5a). Among these, 23 GO terms are enriched in domains matching gene regions containing micro exons, small exons, or other exons (as classified above), suggesting that domains with various categories of exon organization might be preferentially involved in specific functions (Fig. 5a). Among domains related to the regulation of gene expression, the RNA recognition motif (RRM_1) corresponds to a total of 479 conserved exons found in each of four categories, with 60.8% (291) being micro and small exons in category-I (Fig. 5a). For example, *ELF9* and its homologs contain an RRM_1 domain that corresponds to one micro, two small and one medium exons all shared by seed plants and belong to the hnRNP-like gene family with function in regulating splicing in *Arabidopsis*[73] (Supplementary Fig. 28). In addition to RRM1, the protein family of the RRM domain (CL0221) also includes the RRM_2 through RRM_10 domain types that are present in a variety of RNA-binding proteins with implications in regulation of alternative splicing[74]. In addition, the WRKY domain for DNA-binding is related to stress response and highly conserved in plants[75]; different members of the WRKY domain family belong to the domain-exon categories-I, II, III, but lack correspondence to micro exons and are not found in category-IV (Fig. 5a). Specifically, in seed-plant homologs of the rice *WRKY45/75* genes functioning in drought and salt resistances[76], the WRKY domain corresponds to two conserved exons (one medium and one large) (Supplementary Fig. 29); similarly, in seed-plant homologs of the *Arabidopsis WRKY18* gene related to pathogen defense[77], the region for the WRKY domain matches a large exon and a medium exon (Supplementary Fig. 30). Another domain related to the regulation of gene expression is the K domain (PF01486), which is a highly conserved domain encoded by MIKC-type MADS-box genes[78]. The K domain is enriched (p-value = 4.58E-6 < 0.01; two-sided hypergeometric test) in domains in category I-1 with micro exons and small exons (Fig. 5a), suggesting that alternative splicing of micro/small exons could contribute to diverse regulatory roles of genes containing this domain. For instance, the angiosperm *AGL17*-like genes[79] contain one micro exon and four small exons for the K domain (Supplementary Fig. 31), suggesting the K domain in the *AGL* genes might be regulated by splicing[80] in a way conserved among seed plants.

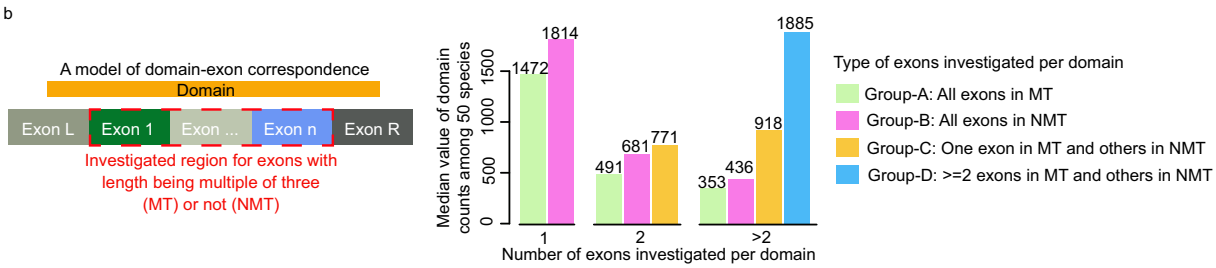
Among domains with GO terms for post-translational modification, protein kinase domain (Pkinase; PF00069) corresponds to 1780 conserved exons in all four categories, including 1136 exons from I-1 and I-2 (Fig. 5a). It is possible that splicing of micro and small exons contributes to the structure plasticity of Pkinase. For example, PHOTOTROPIN 2 (PHOT2) is a membrane-bound serine/threonine protein kinase that also functions as a blue light photoreceptor[81]. The *PHOT2* gene region for the Pkinase domain corresponds to 10 conserved exons (including one micro, four small, and five medium exons) shared by seed plants (Supplementary Fig. 32). NIMA-related serine/threonine kinases (NEKs) can regulate cell growth, microtubule organization, and stress responses[82]. The Pkinase domain of *NEK1/3/4* genes correspond to 13 (five micro, seven small, and one large) exons conserved in seed plants (Supplementary Fig. 33). The difference in domain-exon organization between *PHOT2* and *NEK1/3/4* genes imply that regulation involved micro and small exons (such as alternative splicing) might be relevant for functional diversity of protein kinases. Different organizations of micro/small exons encoding (part of) a domain are also observed for other domains, such as the serine carboxypeptidase (PF00450), the ubiquitin carboxyl-terminal hydrolase (UCH) domain, and the cullin domain (PF00888) (Fig. 5a).

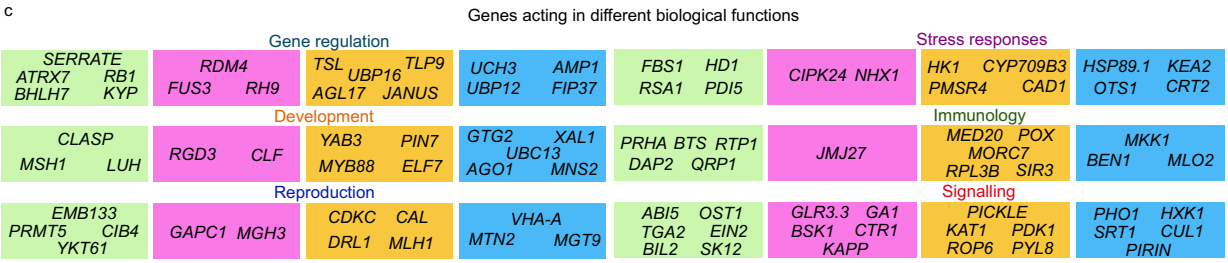
Among domains related to protein-protein interactions, the leucine rich repeat domain (LRR_8; PF13855) corresponds to 606 conserved exons found in each of four categories (Fig. 5a). In particular, nucleotide-binding leucine-rich-repeat (NLR) genes, the largest family of plant disease resistance genes, contain a C-terminal LRR_8 domain that can recognize pathogen-derived effectors[83–85]. Among NLR genes, LRR_8 domains are encoded by conserved exons of categories-III and IV, where the exons also correspond to a conserved NB-ARC domain (PF00931; see examples in Supplementary Data 11). The results suggest that exons encoding the LRR_8 domain are conserved specifically among NLR genes. The length variation of exons corresponding to the LRR_8 domains (Supplementary Data 11) may contribute to the diversification of these domains for recognizing different pathogen effectors[86]. In addition, the SET domain (PF00856) with an annotated activity of protein methylation corresponds to 204 conserved exons, including 47 exons belonging to I-1 (Fig. 5a), suggesting that splicing of micro and/or small exons might contribute to the structure plasticity of this domain family. Several SET domain proteins catalyze histone methylation[87], including the *Arabidopsis* ATXR5/ATXR6 for histone-3-lysine-27 mono-methylation, which affects transposon activity and genome stability[88]. In *ATXR5/ATXR6* genes, one large and two medium exons shared by seed-plants encode the SET domain (Supplementary Fig. 34). The angiosperm *ATXR5* and *ATXR6* gene clades likely resulted from a gene duplication of an angiosperm ancestral gene that probably had the same exon-intron structure corresponding to the SET domain (Supplementary Fig. 34). In addition, alternative splicing of *ATXR5* transcripts was detected with impact on the domain structure[89].

Another group of GO terms are for metabolism of small molecules, such as the cytochrome P450 domain (p450; PF00067), which corresponds to 727 conserved exons found in category-I, II, or IV (Fig. 5a). Exons in category-I account for 95.2% (692 including 261 from I-1 and I-2) of all conserved exons matching p450, suggesting that most of the p450 domains are encoded by multiple exons (including micro and/or small exons). Proteins with the p450 domain belong to the largest plant superfamily for enzymes and are involved in numerous reactions that are important for development and environmental responses[40]. The presence of micro/small exons in *CYP* genes suggests possible plasticity for *CYP* transcript and catalytic diversity of the encoded proteins. In addition, the DEAD/DEAH box helicase (DEAD; PF00270) family corresponds to 354 conserved domain-exon organizations in categories-I, II, and IV. Proteins with the DEAD domain include those regulating miRNA biogenesis in *Arabidopsis*, such as DEAD-BOX RNA HELICASE 27 (RH27) functioning in pri-miRNA processing[90] and RH6/RH8/RH12 regulating the formation of the Dicing body (D-body, for miRNAs generation)[91]. In seed-plant *RH27* genes (Supplementary Fig. 35), the DEAD domain corresponds to six exons conserved in seed plants, including one micro, two small, two medium, and one large exons. On the other hand, the DEAD domain in *RH6/RH8/RH12* (Supplementary Fig. 36) corresponds to three large exons conserved in seed plants, offering possible difference in the regulation of exon retention and protein activity. Furthermore, the peroxidase domain (PF00141) has been found in genes of bacteria, fungi, plants that play important roles in response to drought and other environmental stressess[92]. Our analyses retrieved the peroxidase domain corresponding to 315

https://doi.org/10.1038/s41467-025-66816-3

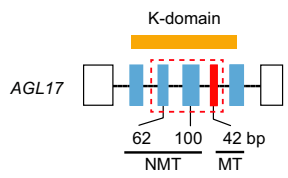


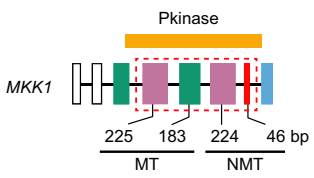




d

Types of exons in MT		Number of domains	
Micro	Small	One exon in MT and others in NMT	>=2 exons in MT and others in NMT
present	present/absent	229 (3.1E-3)	385 (3.7E-29)
absent	present	827 (1.2E-29)	1071 (2.29E-90)
absent	absent	634 (4E-37)	452 (6.29E-184)
example		see (e)	see (f)





conserved exons in categories-I, II, and III (Fig. 5a); 76.2% (240) of the matched exons are from the I-3 subcategory, implying that the relatively long exons can offer more stable peroxidase domain structure for diverse stress responses. The Ras family (Ras; PF00071) is part of the GTPase superfamily and also involved in response to diverse stresses. It matched 237 conserved exons in category-I and II (Fig. 5a), with 136 exons in I-2, implicating that the splicing of small exons could affect Ras transcript structural plasticity with possible links to diverse functions.

Examination of conserved exon lengths for multiple of three

If alternative splicing leads to skipping of an internal exon in a protein coding region of the transcript, whether the number of nucleotides in the skipped exon is multiple of three (MT) or not (NMT) can affect the

Fig. 5 | Functions of domains and the length of their corresponded exons. **a** Summary of molecular functions and enrichment of domains. Left are the major functional types. For each type, top 3-6 domain families ranked by the number of conserved exons are shown here. To the right of domain families are bar-plots of percentage of exons matching different categories. Six colors of bars represent conserved exons in three sub-categories of category-I and three other categories. The total number of conserved exons is shown to far right. For p-value of enrichment using the two-sided hypergeometric probability, we only present the value less than 0.05. Red numbers are p-value of <0.01. **b** Summary of exons with length being a multiple of three (MT) or not (NMT) investigated for domains of category-I. Left is a model of domain-exon correspondence. Inter exons are investigated to estimate the number of exons in MT. Domains are clustered into four types (from A to D) based on the investigated exons. Right is the bar-plot of number of domains. Colors of bars represent the domain types. These bars are compared according to the number of exons investigated per domain. **c** Biological functions of genes encoding domains in each type named in part **b**. Colors of boxes correspond to the colors of domain types in part **b**. Gene names are listed in boxes (see details in Supplementary Data 12). **d** Comparison of domains with at least one investigated exon in MT and other exons in NMT. For each of two domain types, the domains are grouped by their matched exons that are included or exclude micro/small exons. Numbers represent the number of domains. The number in round parentheses represents the p-value of <0.05 (two-sided hypergeometric test). **e** *AGL17* (*AGAMOUS-LIKE 17*) gene. **f** *MKK1* (*MITOGEN ACTIVATED PROTEIN KINASE KINASE1*) gene. Investigated exons in parts e and f are placed in a red dotted rectangle, with red, blue, green, and purple boxes representing micro, small, medium, and large exons (see their length below). Top orange box represents the focal domain. Source data are provided in a Source Data file.

reading frame of downstream exonic sequences and thus protein sequence. To examine the distribution of exon length in this aspect, we counted MT and NMT exons, with a null hypothesis of exons having 1/3 chance being MT. For domains encoded by multiple exons, the absence of an MT exon in a transcript would not affect the protein (domain) sequence encoded by downstream exons, but the skipping of an NMT exon would. In individual genomes, more than one third of exons in each size category are MT, although MT micro, small, and medium exons are fewer than half (Supplementary Fig. 37); for large and super exons, MT exons are greater than half, in part because many of them are for genes lacking introns. In addition, we examined exons being MT or NMT for different categories of the domain-exon organization (Fig. 5b). We subdivided category-I according to MT or NMT status of the intervening exon(s), resulting in 4 groups (Fig. 5b): Group A with all exons being MT (see green bars), group B with all exons being NMT, and groups C (one exon being MT, see orange bars) and D (two or more exon MT, see blue bar). A total of 5890 domain-exon types (=1472 + 491 + 353 + 771 + 918 + 1885; median value among 50 species shown in Fig. 5b) contain one or more MT exons, which could be skipped while retaining the reading frame of the downstream exons (see domain counts from individual species in Supplementary Fig. 38). On the other hand, group-B domain-exon structures (total 2931 = 1814 + 681 + 436 in median investigated among 50 species; the pink bars in Fig. 5b and Supplementary Fig. 38) contain all intervening exons being NMT, which, when skipped, would cause a frame shift in downstream exons. These four groups of domain-exon organizations are found in genes related to regulation of expression (Fig. 5c), such as the *ATRX7* genes encoding the SET domain (group-A), the *RH9* genes for the DEAD domain (group-B), the *JANUS* genes coding for the RRM_5 domain (group-C), and the *UBP12* genes for the UCH domain (group-D). The four groups are also found in genes for development/reproduction, stress responses, immunity, and signaling (Fig. 5c; see other examples in Supplementary Data 12), suggesting that they offer a potentially common modulation of domain-exon structures via splicing in a variety of plant genes. Furthermore, we compared the number of domains that correspond to intervening MT exon(s) (Fig. 5d–f). For domain-exon type I-1 (with micro exon, Fig. 4c), the fraction of domain-exon structures containing intervening MT micro exon(s) in group-C (13.6%; 229/1690; Fig. 5d) is higher than expected (p-value = 3.1E-3; FDR-BH). Similarly for domain-exon type I-2 (with small exon, but not micro exon, Fig. 4c), the fraction of MT small exon(s) (48.9%; 827/1690; Fig. 5d) is higher than expected (p-value = 1.2E-29; FDR-BH). In addition, for domain-exon organizations with only medium or large exons, the fraction of MT exons (37.5%; 634/1690; Fig. 5d) is less than expected (p-value = 4E-37; FDR-BH).

Expression of genes containing conserved exons among major organs

To gain clues about differential functions of genes containing conserved exons on a genome-wide scale, we analyzed transcriptomic data of the flower, fruit, and at least one of the root, stem and leaf from *Arabidopsis*, apple, watermelon, soybean, grape, tomato, *Papaver*, maize, wheat, pineapple, and *Amborella trichopoda* (see methods, also for analyses of alternative splicing as described in the next section). Specifically, for genes with likely functions in the flower, we identified preferentially expressed genes (PEGs) with a higher level of expression in the flower than one or more other tissues (see methods; Supplementary Data 13). For example, analyses of *Arabidopsis* transcriptomic data identified, respectively, 4960, 4944, and 3729 floral PE, in comparison with the root, leaf, and fruit (Supplementary Fig. 39). Then we counted the number of PEGs containing at least one conserved exon that were identified here and retrieved 3241 (65.3%), 3175 (64.2%), and 2248 (60.3%) floral PEGs that contain conserved exons and showed higher expression relative to root, to leaf, and to fruit, respectively (Supplementary Fig. 39). Among the conserved exons in these PEGs, >72% are conserved across seed plants; for example, 2363 (74.4%) of 3175 floral PEGs to leaf include the seed-plant conserved exons (Supplementary Fig. 39), suggesting that most of these PEGs could be regulated by conserved splicing processes. Similarly, in maize conserved exons are present in 3791 (67.7%) of 5600 floral PEGs to root, 2718 (63.6%) of 4273 floral PEGs to stem, 2133 (68.1%) of 3130 floral PEGs to leaf, and 5477 (68.5%) of 7995 floral PEGs to fruit (Supplementary Fig. 39). Among the maize PEGs with conserved exons, >70% have at least one exon shared by both angiosperm(s) and gymnosperm(s) (Supplementary Fig. 39). Patterns similar to those in *Arabidopsis* and maize were also found for the other nine species analyzed here, suggesting that most floral PEGs relative to the other tissues examined here are conserved at the splicing level across seed plants for the development and physiology of reproductive structures.

To gain clues about the possible molecular/biochemical functions of the PEGs with conserved exons, we examined domain families in these floral PEGs with conserved exons and detected the domains in PEGs from each dataset (Supplementary Fig. 40a). For example, we investigated 4390 domain members of 1963 domain families in 4944 *Arabidopsis* floral PEGs relative to leaf and found that 180 domain families are enriched, each with a significantly (two-sided hypergeometric test; p-value < 0.05; Supplementary Fig. 40a) greater percentage in these PEGs than the genome-wide percentage of the family (see methods). These enriched families include the domain families described in the previous section, such as LRR and K-domain. We then examined the domains in floral PEGs in all 11 species and retrieved ~3209 (median) domain members in floral PEGs relative to leaf, ~2718 (median) to stem, ~2832 (median) to fruit, and ~4261 (median) to root (see individual numbers in Supplementary Fig. 40a). Using enrichment in at least four species (including at least one eudicot and at least one monocot) as a cutoff, we obtained 147, 40, 164, and 135 enriched families, respectively, in floral PEGs relative to root, to stem, to leaf, or to fruit (Supplementary Data 14). Among these enriched protein domain families, 24 families are present in all four PEG patterns (Fig. 6a; Supplementary Data 14), such as K-domain (MADS) family,

https://doi.org/10.1038/s41467-025-66816-3

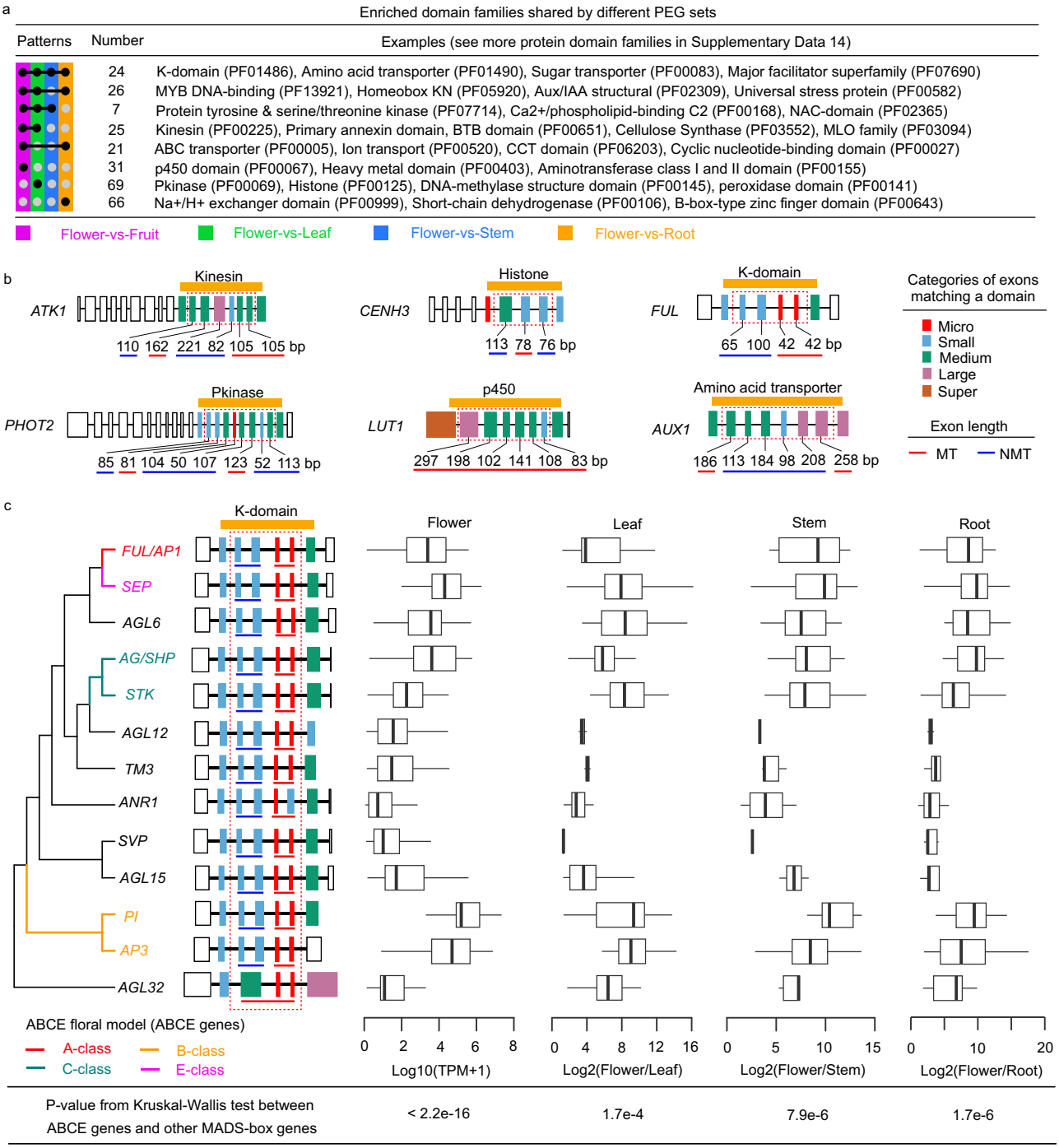


Fig. 6 | Protein domains encoded by flower-preferential genes in comparison at least one other organ. a Enriched domains encoded by preferentially expressed genes (PEGs) in the flower containing the conserved exon. Domain families enriched in >3 genomes are shown on the right, while their presence among specific sets of PEGs is indicated by black dots (see below for details) on the left with the background of green, blue, pink, orange bars, which respectively denote the corresponding PEGs in the flower relative to the leaf, stem, fruit, and root. **b** Domains in six specific PEGs. *ATK1, Arabidopsis Kinesin1. CENH3, CENTROMERIC HISTONE H3. FUL, FRUITFULL. PHOT2, PHOTOTROPIN 2. LUT1, LUTEIN DEFICIENT 1. AUX1, AUXIN RESISTANT 1.* Colored boxes represent focal exons, with their size ranges explained to the right. Exons with length being a multiple of three (MT) or not (NMT) are indicated with red or blue lines, respectively, below the exons. **c** Domain-exon correspondence patterns for K-domain in MADS-box genes related to flower development. The left tree illustrates the phylogenetic relationships among major MADS-box subfamilies, with the tip labels showing representative *Arabidopsis* names. Color branches represent the ABCE genes. To the right of phylogeny are the domain-exon correspondence in the ancestral gene model of each subfamily group. Colored exon boxes and red/blue lines below them have the same meaning as in part b. To the right of gene models are box-plots of gene expression levels [measured by log10(TPM + 1) in flowers] of genes in 11 angiosperms. Three additional sets of box-plots on the right are for the Log2 (fold changes in TPM) (LFC) of expression in the flower relative to leave, stem, and root, respectively. Box plots show the median (center line), interquartile range (box bounds), and whiskers extending to the smallest and largest values within 1.5 × IQR from the quartiles. Outliers beyond the whiskers are included in the source data but not displayed here. Kruskal-Wallis test between the ABCE genes and other MADS-box genes support greater changes of the expression of the ABCE genes than that of others (P-value < 2.2e-16, 1.7e-4, 7.9e-6, 1.7e-6, respectively). Source data are provided in a Source Data file.

with the known conserved functions of the K-domain-containing genes in flower development[93,94], and the sugar transporter domain for transport of sugar and other carbohydrates (e.g., proteins encoded by the *SWEET11* and *SWEET12* genes[95]), suggesting that these conserved functions (in four or more species) are floral preferential to each of the other four organs. The p450 (e.g., *RED ELONGATED 1*, *Brassinosteroid-6-Oxidase 2* genes) and 30 other domain families are enriched in floral PEGs relative to fruit but not to other tissues (Fig. 6a). These diverse functional domains with conserved exons suggest that their functions are likely conserved across angiosperms, potentially revealing new molecular mechanisms for development and physiology.

We further explored possible association of domain-exon structure and possible alternative splicing with differential expression, especially that involving genes encoding protein domains belonging to enriched families, using domain-exon types as described in the above section (Fig. 5b). We found that among those floral PEGs relative to only one of the four other tissues (Supplementary Fig. 41), >92% (median) of the corresponding domain-exon organizations are in category-I (Supplementary Fig. 40b), greater than the genome-wide percentage of this category (Fig. 4). Among these category-I domains encoded by floral PEGs, >66% match three or more exons, including at least one intervening exon in MT (Supplementary Fig. 40c). Specifically, a Kinesin domain is encoded by eight conserved exons of the *Arabidopsis ATK1* gene (with a crucial role in meiotic spindle morphogenesis[96]) and its seed-plant homologs (Fig. 6b; Supplementary Fig. 42). Among these six intervening matched exons for the domain, three medium exons are in MT (multiple of three nucleotides) and three other exons (one small, one medium, and one large) are in NMT (non-MT). Another example is CENTROMERIC HISTONE H3 (CENH3), which is an essential component of the specialized nucleosomes in the kinetochore and required for chromosome segregation during cell division, and is regulated by phospohrylation[97]. The CENH3 histone domain matches five conserved exons, including three intervening exons, one small MT and two NMT (Fig. 6b). The results suggest that alternative splicing skipping some of these MT-exons could shorten the corresponding protein domains without affect downstream sequences, providing a possible mechanism to regulate the activities of the domain.

Genes containing micro and small exons are known to be important for plant development and stress responses; some of these genes exhibit preferential expression patterns relevant to their functions[22]. One group of well-studied genes with micro exons is MIKC-type MADS-box genes including floral homeotic genes of the ABCE model; these genes encode proteins with the K-domain (for keratin-like) in angiosperm. For example, the A-class gene *APETALA1* (*AP1*) and its paralogue *FRUITFULL* (*FUL*) contain six conserved exons, including two intervening micro exons in MT and two intervening small exons in NMT (Fig. 6b; Supplementary Fig. 43). In *Arabidopsis FUL* gene affects many key genes for flower development[98]. The K-domain family is also enriched in ginkgo genes, such as homologs (*evm.model.chr1.1358* and *evm.model.chr1.1364*; Supplementary Fig. 44) of *Arabidopsis AGL2/4/9* [*SEPALLATA1/2/3* (*SEP1/2/3*); E class] genes required for all floral organ development[99]. Phylogenetic analyses of seed-plant *SEP* genes placed a duplication in the ancestor of angiosperms, likely as part of the angiosperm polyploidization event (Supplementary Fig. 44) and angiosperm *SEP* homologs are expressed in the floral meristem and multiple floral organs[100]. The K-domain corresponds to six conserved exons in *SEP* genes, where two intervening micro exons are in MT and one small exon in NMT, somewhat different from those in A-class genes (Fig. 6c). We noted that angiosperm ABCE genes exhibit significant higher expression levels in the flower than other MADS-box genes (p-value < 0.01 under Kruskal-Wallis Test; Fig. 6c). Similarly, among floral PEGs, ABCE genes also have a significantly greater increases in expression than other MADS-box genes (Fig. 6c). We compared the likely ancestral (commonly shared) gene models of different MADS-box subfamilies and found that K-domain in different MADS-box subfamilies matches 5-6 conserved exons (see the correspondence between conserved exons and K-domain in genes from 11 angiosperms in Supplementary Fig. 45). For example, C-class genes (*AG/SHP*) have six (two micro, three small, one median) exons and *AP3* homologs have (two micro, three small) exons (Fig. 6c). Similar gene structures were found in non-ABCE MADS-box genes, such as *TM3* (six exons) and *AGL12* (five exons). It is striking that all MADS-box genes described here contain an internal micro exon in MT (42 bp) encoding part of the K domain (Fig. 6c), suggesting a possible conserved splicing regulatory mechanism. Micro exons are unusually short and thus lack space to accommodate important cis-elements for splicing; therefore, they might require special regulation including specific splicing factors. This idea is supported by the recent discovery that the *Arabidopsis* GRP20 protein is required for the normal retention of a micro exon in several MADS-box genes (*AP1*, *SEP3*, *SEP4*, *AG*, *STK*, and *AP3*)[22]. The affected micro exons have 42 nucleotides, allowing the transcripts lacking the micro exon to be translated into proteins with a K-domain that is shortened by 14 amino acids. It is known that the K-domain is important for protein-protein interactions (such as dimerization between AP3 and PI, or higher order protein complex formation[101]), making it possible that a shortened K-domain might alter the pool of MADS-box protein complexes and thereby affecting flower development.

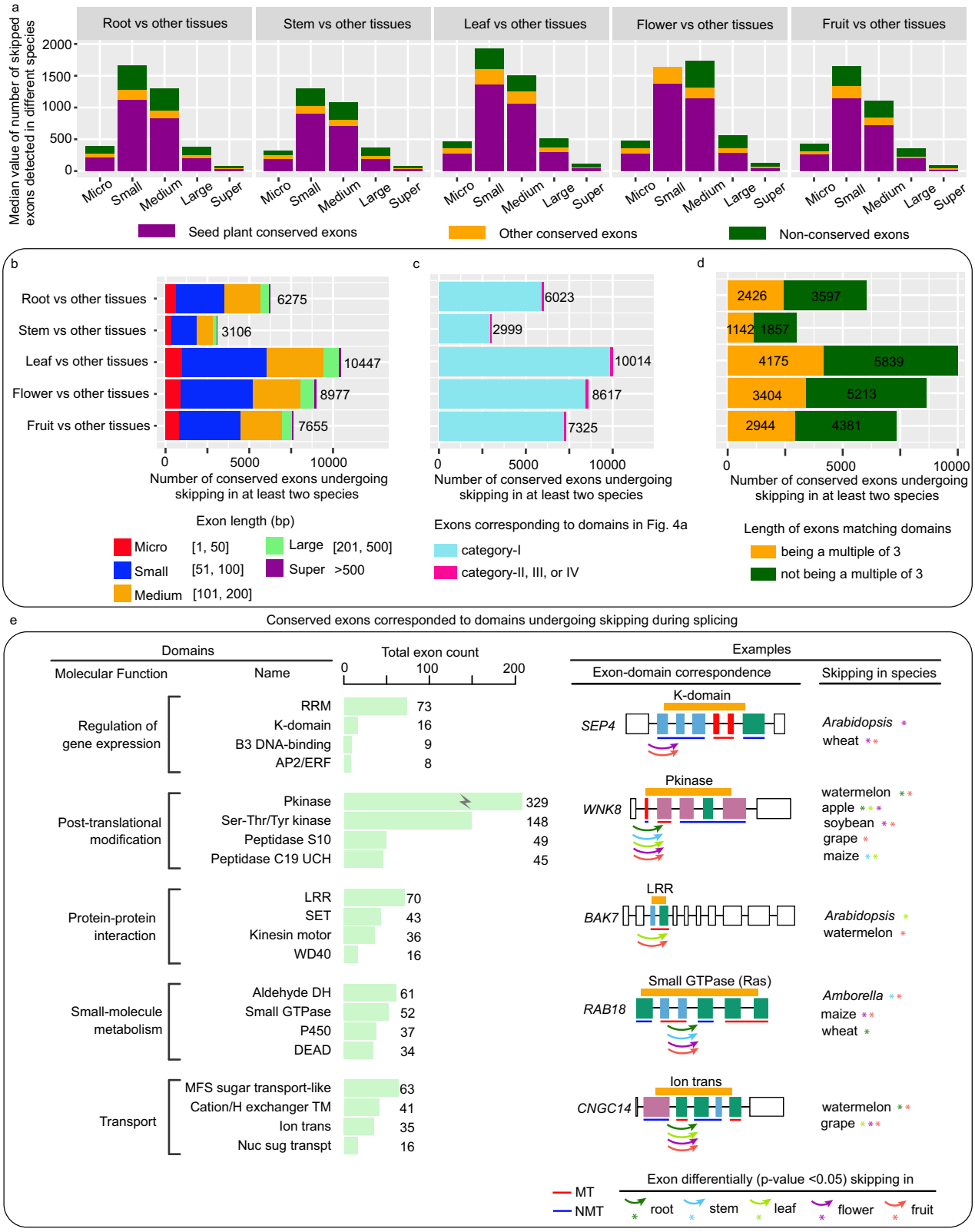
Auxin is an important plant hormone and regulate multiple developmental processes[102]. Among PEGs is the *AUXIN RESISTANT1* (*AUX1*) gene encoding an auxin influx transporter with a conserved Aa_trans domain (Fig. 6b; Supplementary Fig. 46). The Aa_trans domain matches six internal exons in *AUX1* homologs in most cases, but in Solanaceae and monocot species there is only one exon corresponding to the last two of the six exons in *AUX1* homologs of other species (Supplementary Fig. 46). Among the six exons, the first (3′) and last (5′) are in MT and others in NMT (Fig. 6b). Other PEGs related to development and metabolism with internal exons in MT encode the Pkinase domain, such as the *PHOT2* genes mentioned in above section (Fig. 6b; Supplementary Fig. 32) and the *Arabidopsis LUTEIN DEFICIENT1* (*LUT1*) gene encoding a heme-containing monooxygenase with a p450 domain[103] (Fig. 6b; Supplementary Fig. 47).

Conserved exons affected by exon skipping and functional implications in flower development

Alternative splicing is a crucial mechanism that increases the diversity of protein isoforms in eukaryotic organisms[7]. Genes containing conserved exons and with expression in multiple tissues and organs could have alternatively spliced transcripts, including those due to exon skipping, thereby increasing the functional diversity of their protein products. The results reported in the above sections indicate that most plant genes are conserved across angiosperms and encode protein domains matching multiple exons, suggesting that regulation by alternative splicing could affect their domain structure and function. However, whether alternative splicing affected conserved exons and patterns of alternative splicing that might affect specific conserved domain families are largely unknown.

To address these questions, we analyzed alternative splicing events affecting protein-coding exons using rMATS (see methods) in comparison with the annotated exon-intron structure, by analyses of RNA-seq datasets of 11 angiosperms (*Arabidopsis thaliana*, apple, soybean, watermelon, tomato, grape, *Papaver* [poppy], maize, wheat, pineapple, and *Amborella trichopoda*), which were selected to represent angiosperm diversity and for their available high-quality datasets. We found a total of 40,416 exon skipping events in root transcripts (10 species), 26,650 in stem transcripts (7 species), 58,224 in the leaf (11 species), 52,095 in the flower (11 species), and 47,003 in the fruit (11 species) (Supplementary Fig. 48). Among these skipped exons, small exons are the largest class and are enriched in each tissue of each

https://doi.org/10.1038/s41467-025-66816-3



species relative to the proportion of small exons among intron-containing genes from the respective species (Supplementary Fig. 48). Super exons are the smallest class among skipped exons in all tissues and large exons are the second-smallest class in most tissues (Fig. 7a, Supplementary Fig. 48); it is possible that the absence of a super exon can cause a severe functional defect of the protein, so such skipping is not beneficial to plant fitness. We also found other types of alternative splicing (A3SS and A5SS, Supplementary Fig. 49). However, intron retention results are prone to false positives due to incompletely spliced transcripts, thus not examined here.

We next investigated the extent of exon skipping that affected members of conserved exon families. In the floral transcriptome of a particular species the 11 angiosperms investigated here, 610-10,957 (median of 3850) events affecting conserved exons were detected,

Fig. 7 | Domains matching conserved exons undergoing skipping during splicing in different tissues. a Bar-plot of median value of number of seed-plant wide conserved exons (purple), other conserved exons (orange), and non-conserved exons (green) skipped in different species. See details of each species in Supplementary Fig. 48. **b** Bar-plot of number (see the right numbers) of conserved exons skipped in at least two species. Different colors of bars represent exon length categories as shown below. **c** Bar-plot of number (see the right numbers) of conserved exons in part b matching domains. Colors of bars represent different exon-domain correspondence in Fig. 4a. Most of these exons match the category-I (blue bars). **d** Bar-plot of number of conserved exons in part c in length being a multiple of three or not. Orange bars represent exons in length being a multiple of 3. Green bars represent exons in length being not a multiple of 3. **e** Major molecular function categories of domains corresponding to conserved exons with exon skipping signals. Left are five molecular function categories of domains. To the right of domain names are bar-plot of total count (see the right numbers) of exons. See more domain families in Supplementary Fig. 51. To the right of bars are domain examples showing the position of skipped exons within coding region for domains. See more examples in Supplementary Data 16. Source data are provided in a Source Data file.

with a combined total of 40,732 events (Supplementary Data 15). Similarly, thousands of exon skipping events were detected in the other tissues, with a total of 36,367 and 45,748 skipping events, respectively, detected in the fruit and the leaf and affecting conserved exons (Supplementary Data 15). The large fractions of conserved exons experiencing exon skipping suggest that alternative splicing (especially exon skipping) might be a conserved regulatory mechanism for these exons and that proteins encoded by transcripts lacking the exon might have a different function from proteins encoded by the full-length transcripts.

To identify shared exon skipping among the 11 species, we searched for the conserved exons that were skipped in a focal tissue in at least two species and found 6275 such skipped exons in root, 3106 in stem, 10,447 in leaf, 8977 in flower, and 7655 in fruit (Fig. 7b; see individual exons in Supplementary Data 16), suggesting that many of the exon skipping events are conserved to some extent. In individual species, about 50.3%-68.8% (median of 60.7%, in a species) of skipping events in the flower affected conserved exons in two or more species. In addition, among 36,367 skipping events affecting conserved exons detected in the fruit, 50.9%-64.6% (median of 57.2%, in a species) of them were shared by two or more species. Among 45,748 skipping events in the leaf affecting conserved exons, 55.8%-68.8% (median of 65.3%) of events detected in individual species were shared by two or more species. Specifically in the *Arabidopsis* flower, 675 skipping events were detected that affected individuals of conserved exons, including 235 events detected only in *Arabidopsis* and 440 (65.2%) events shared by *Arabidopsis* and other species. For example, a conserved exon (EXON0070818) in seed-plants homologs of *AT1G44760*, which is a proposed target gene of AP1 and AP3[104], was found to be skipped in the flower of *Arabidopsis* (Exon 2 in *AT1G44760.2*), soybean (Exon 2 in *GmW82.14G142700.1*), and wheat (Exon 2 in *TraesCS1B03G0773200.1*) (Table 1; see affected exons in Supplementary Data 16). Some exon skipping events were also shared by two or more the other 10 species analyzed here, such as the 409 events in tomato and other species (Table 1). For instance, members of the conserved exon family (EXON0023600) in angiosperm homologs of *ARF8* (*AT5G37020*), which has regulatory role in stamen elongation in *Arabidopsis*[105], were skipped in the flower of tomato (Exon 2 in *Solyc03g031970.4.1*) and grape (Exon 2 in both *Vitis12g00136.t01* and *Vitis10g01227.t01*) (Table 1 and Supplementary Data 16).

For exons skipped in the *Arabidopsis* flower, we found that 126 of 145 micro exons, 376 of 425 small exons, 150 of 188 medium exons, 21 of 29 large exons, and 2 of 4 super exons were conserved exons and that the seed-plant wide conserved exons account for 77.8% (98 of 126), 85.4% (321 of 376), 86% (129 of 150), 85.7% (18 of 21), and 100% (2 of 2) of conserved micro, small, medium, large, and super exons, respectively, that were skipped in the flower (Supplementary Fig. 48). Similar patterns were also found in other species. In general, seed-plant wide conserved exons are the majority (>52%) of micro, small, and medium exons affected by skipping in at least one tissue (Fig. 7a, Supplementary Fig. 48). In each tissue >56% of these exons are micro or small exons (Fig. 7b), implying that skipping particularly involving micro and small exons contributed substantially to transcript/protein diversity. In addition, a total of 17,032 conserved exons skipped in at least two species account for ~43% of these 39,721 conserved exons with skipping event detected in at least one of these 11 species (Supplementary Data 17), suggesting that skipping of conserved exons are conserved across angiosperms.

Splicing can be regulated via cis elements in introns[3,106]. To gain possible clues about introns, we also investigated the combined length of two introns flanking an exon affected by skipping and found that the median value of intron length increased from micro exons to longer ones in each of the 11 species, such as 684 bp, 813 bp, and 1068 bp, respectively, for median intron length corresponding to micro, small, and other conserved exons that were skipped in at least one tissue (Supplementary Fig. 50). In each of five exon size categories, the combined length of flanking introns for conserved exons that were skipped in at least one tissue is longer than that for all genome-wide exons (Supplementary Fig. 50), suggesting that exons regulated by exon skipping might be regulated by cis elements in flanking introns, similar to the cis-elements reported for regulating splicing[106]. In addition, the shorter intron length flanking conserved micro and small exons (Supplementary Fig. 50) suggest that they might be regulated via different mechanisms, such as cis elements found in exons and specific RNA-binding protein(s), as recently reported[22].

To assess possible impact of exon skipping on the molecular functions of protein domains, we examined the conserved exons affected by exon skipping and detected 6023 (96% of 6275) exons in root, 2999 (96.6% of 3106) in stem, 10,014 (95.9% of 10,447) in leaf, 8617 (96% of 8977) in flower, and 7325 (95.7% of 7655) in fruit matching domains (Fig. 7c), suggesting >95% of conserved exons skipped in two or more species affect domains. Indeed, the majority of those exons correspond to domains in category-I (Fig. 7c), implying that skipping micro and/or small exons could affect most of the domains encoded by multiple exons, at least one of which are affected by exon skipping. Investigation of the skipped exon in MT or NMT suggests that 38.1% ~41.7% of exons matching domains are in MT (Fig. 7d), implicating skipping these MT exons would not alter the downstream protein reading frame. In addition, the annotated functions of these domains include regulation of gene expression (e.g., K-domain, RRM), post-translational modification (e.g., Pkinase, Ser-Thr/Tyr kinase), protein-protein interactions (e.g., LRR, Kinesin), small-molecule metabolism (e.g., P450, Small GTPase), and transport (e.g., Sugar_tr) (Fig. 7e). Our results suggest that exon skipping probably affects domains with a variety of molecular functions in each tissue examined here. In particular, among conserved exons matching the K-domain in *SEP4*, an 82-bp exon experienced skipping in the flower and fruit, especially skipping differentially in flower in *Arabidopsis* and in flower and fruit in wheat (Fig. 7e). Another case is for conserved exons matching Pkinase in *WNK8* gene, where a 38-bp exon was excluded in transcripts of different tissues (Fig. 7e). Among conserved exons matching LRR in *BAK7* gene, a 72-bp exon underwent skipping in leaf and fruit (Fig. 7e). The number of skipped exons matching a domain family and enriched families against all members of the family in genome varies from different tissues (Supplementary Fig. 51), implying tissue-specific variation in proteins. Particularly, among those top 10 largest domain families, the k-domain is enriched in those conserved exons skipped in flower and fruit but not in other tissues (Supplementary Fig. 51),

https://doi.org/10.1038/s41467-025-66816-3

Table 1 | Summary of exon skipping (ES) events affecting conserved exons (CEs) in flower

Species	#Events	#EDIOS	Events detected in at least two species					
			#Events	Examples				
				CE ID	CE Type	Conservation	Detected ES events	Gene Functions
Arabidopsis	675	235	440	EXON0070818	Medium	Seed-plants	Exon 2 skipped in *AT1G44760.2*	*AT1G44760* highly expressed in flower, the flora organ-specific gene bound by AP1 and AP3 in *Arabidopsis*[104]
apple	2439	874	1565	EXON0010292	Small	Seed-plants	Exon3 skipped in *MdF0024132.1, MdF0008786.1, MdF0040511.1,* and *MdF0008786.1*	RBP45b functionally described as delaying flowering time in mutation in *Arabidopsis*[181]
watermelon	4106	1613	2493	EXON0047910	Small	Angiosperms	Exon 4 skipped in *Cl(G42_07g0087600.10*	UBC15 (interacted with CUL3A and LFH1) controlling SVP ubiquitination[182]
soybean	3287	1156	2131	EXON0013880	Micro	Angiosperms	Exon 10 skipped in *GmW82.13G350200.1*	Phosphofructokinase encoded genes upregulating in sepal development[183]
grape	10957	5450	5507	EXON0011187	Medium	Seed-plants	Exon 6 skipped in *Vitis01g00811.t01*	Homologs to RECEPTOR-LIKE PROTEIN KINASE 2 regulating the anther development in *Arabidopsis*[184]
tomato	610	201	409	EXON0023600	Medium	Angiosperms	Exon 2 skipped in *Solyc03g031970.4.1*	The *ARF8* orthogroup with *AT5G37020* exhibiting the regulatory role in stamen elongation in *Arabidopsis*[105]
poppy	1189	371	818	EXON0063870	Small	Angiosperms	Exon4 skipped in *PS0622810.1-mRNA-1*	AP3 (*AT3G54340*), one of class B genes for petal and stamen development[185]
wheat	4284	1757	2527	EXON0034035	Micro	Seed-plants	Exon 4 skipped in *TraesCS5B03G0939600.1*	In *Arabidopsis YAB5* (*AT2G26580*) functioning in early anther development[186]
maize	3850	1553	2297	EXON0099045	Large	Angiosperms	Exon 2 skipped in *Zm00014ba381110.1*	A homologs of *HUA1* (*AT3G12680*) acting in the C function in flower[187]
pineapple	4450	1885	2565	EXON0038400	Small	Seed-plants	Exon 4 skipped in *Aco007079.1.v3*	FERRETIN 1/3 functioning in iron homeostasis that is central to flower development and fertility[188]
Amborella	4885	2203	2682	EXON0028293	Micro	Seed-plants	Exon 6 skipped in *AmTr10g00432.1*	A homolog of *FUL* (*AT5G60910*) modulating the flower development[189]

EDIOS, Events detected in one species. See Supplementary Data 5–9 for specific exon members of the conserved exons in individual genes. See Supplementary Data 16 for exons skipped in different species.

supporting its key role in regulating flower and fruit development via exon skipping.

Discussion

Our phylogenomic analyses of exon landscapes across 46 angiosperms and 4 gymnosperms provide distribution profiles in five different length categories, including previously proposed micro and small exons, and present their conservation among seed plants, angiosperms, and angiosperm subgroups on gene structure and domain-exon correspondences. Among the five length categories, medium exons are the largest (32%) and contribute to ~25% (median) of genes; they are annotated to have crucial functions for plant development and response to diverse stresses (for example, *MPK6*, *ETR1*, *ABI1*, *TOC1*, *CHL3*, *IRT1*, *MYB75*, *DMT1*, and *JAR1*). Micro and small exons (accounting for ~8% and ~27% of all exons, respectively) are found in ~54% of genes and represent important structural elements of plant genes. Many of the micro/small exon-containing genes are important regulators of gene expression and other cellular processes (such as *FLC*, *FT*, *COP1*, *HY5*, *PIN1*, *ABI5*, *AGAMOUS*, *PIN2*, *WUS1*, and *PIF4*), with various biological functions. Large or super exons are present predominately in genes without introns (e.g., *RP1*, *BRI1*, *EIN3*, *SUN6*, *CBF3*, *GAI*, *RGA1*, *UBQ10*, *CBF1*, and *SHR*), suggesting that they are not affected by regulatory mechanisms at the splicing level. Most exons are shared by angiosperms and gymnosperms and these widely conserved exons contribute to ~76.7% of genes per genome. Analyses of protein domains encoded by conserved exons revealed that the vast majority of protein domains (category-I) match three or more exons. Among the exons with detected exon skipping, small exons are the largest fraction from previous studies and our analyses here, with additional ES events affecting micro and median exons, potentially allowing for different isoforms of protein domains to be translated from alternatively spliced transcripts. The information presented on conserved exons, their relationships to protein domains and their possible regulation by alternative splicing provide many hypotheses on regulation of gene expression and protein function, to be tested by further molecular genetic studies.

Our comparative analyses of exons across 50 seed plants revealed that a large majority of each of the four smaller size categories are conserved (with median values of 60.55%, 85.7%, 82.3%, and 70.2% for micro, small, medium, and large exons, respectively). However, only 44.9% of super exons are conserved; this relatively low proportion was puzzling at first and led to further examinations. One contributing factor is that some super exons correspond to two or more smaller exons in other genes of the same orthogroup and our definition of conserved exons would exclude such super exons from conserved exons when they only appear in one species. From 50 genomes we detected 823-4341 non-conserved exons per species (median of 1334), which correspond to two or more conserved exons in other genes of the same orthogroup, with a combined total of 76,762 non-conserved super exons for the 50 species. These non-conserved super exons probably resulted from fusion of two (or more) conserved exons via loss of intron(s) during angiosperm evolution, suggesting that potential alternative splicing of the ancestral (unfused) exons is not part of mechanisms regulating the transcript isoforms for the relevant exon(s). Analyses of all exon size categories revealed that 59,338 conserved exon families contain at least one member that was fused in one or more of the 50 species and also exhibit alternative splicing in one of the 11 species analyzed here (Supplementary Fig. 52a–c). In individual species, such fused exons account for ~65% (Supplementary Fig. 52d) of all exons likely fused from conserved exons (Fig. 2e), again suggesting changes in potential alternative splicing during angiosperm evolution for exons of different size.

A relatively large number (246,011 of 928,992 in total) of super exons correspond to intronless genes, which might be generated via reverse transcription and insertion into the genome and include

possible pseudogenes[107–109], which possess sequence features, such as poly-A tracks near the 3′ end and portions similar to repetitive elements. Among intronless genes, sequence conservation and detectable expression support functional genes[107,110], whereas pseudogenes might have accumulated mutations and sequence divergence. As the Chinese pine genome[17] has an unusually large number (17,437) of super intronless genes, we investigated these genes for the properties as possible evidence of pseudogenes and found that 13,975 (80.1% of 17,437) lacked characteristics of processed pseudogenes, including 913 conserved exons (Supplementary Data 18). The remaining 3,462 (19.9%) super intronless genes have features that could be associated with processed pseudogene, including 203 conserved exons/genes. The conservation of these 203 intronless genes suggests that they might be functional, similar to known functional intronless genes from genetic studies and supported by our observation that 182 of the 203 conserved intronless genes showed detectable expression in at least one of four organs (megagametophyte, leaf, stem, and needle; Supplementary Data 18). The remaining 21 (=203-182) conserved intronless genes did not have detectable expression in the four organs, accounting for ~1.9% of 1116 (=913 + 203) conserved super intronless genes, which were included in the subsequent analyses of conserved exons (Supplementary Data 18). Among 3259 (=3462–203) non-conserved intronless genes with pseudogene-like features, 2666 showed detectable expression in at least one organ, whereas the remaining 593 showed no detectable expression in any of the four organs and are candidate processed pseudogenes, corresponding to less than 3.4% of 17,437 super intronless genes in Chinese pine, and require further annotation using additional RNA-Seq datasets in future studies.

Our analyses of alternative splicing found that the conserved exon shared by multiple species can be commonly skipped in some of these species. For such conserved exons with skipping in one of five tissues, more than half are micro or small exons (Fig. 7b). Most of those conserved exons match domains that are encoded by multiple micro and/or other exons. It is likely that skipping such conserved exons is a common mechanism for protein diversity across species. In addition, other alternative splicing mechanisms, such as alternative 3′ splice site (A3SS) and alternative 5′ splice site (A5SS), were also found to alter exon size, including thousands of conserved exons that were shortened by A3SS, including 14,506 and 13,848 conserved exons in the flower and fruit, respectively, and shared by two or more species (Supplementary Fig. 49; also for A5SS). In short, 47,708 (~45%) and 29,955 (~28%) of 105,971 conserved exons present in at least one of the 11 angiosperms were affected by A3SS and A5SS, respectively (Supplementary Data 17). These results suggest that A3SS and A5SS events are also conserved alternative splicing mechanisms contributing to protein diversity across species. Further analyses of the conserved exons affected by A3SS/A5SS events indicate that nearly half of them have the altered length in MT in at least one species, implying that shortened exon size would not result into reading frame change in downstream domains (Supplementary Fig. 49). In general, our exon conservation investigation has a broad evolutionary implication in exon skipping and other alternative splicing mechanisms in plants. Alternative splicing contributes to some genes in a genome with two or more annotated transcripts; to estimate possible effect of alternative transcripts on the analyses of exon conservation, we examined the *Arabidopsis* genome with highly accurate annotations (Supplementary Data 19). Among the 18,197 *Arabidopsis* genes (in the 10,216 HOGs) included in our comparative analyses, 8015 genes have >=2 transcripts per gene and include 5762 genes with a total of 14,150 alternatively spliced exons[60]. In 3489 of these alternatively spliced genes, the longest transcript also has the longest coding region and most longer exons; however, for 2273 genes, the longest transcript has a shorter coding region with either fewer (1,110 genes) or shorter (1163 genes) exon(s). We examined 4433 exons from these 2273 genes, including 2672 exons with altered length ranges or intron retention and 1,761 additional exons that did not correspond to exons in the longest transcripts (see details in Supplementary Data 19). We assessed sequence similarity between these 4433 exons and the protein-coding sequences of other genes within their respective orthogroups (Supplementary Data 19), revealing that 2585 exons showed no similarity to other exons, whereas 952 exons matched individual conserved exons. These two types did not affect the overall analyses of conserved exons. In addition, 896 exons each matched two or more exons from different conserved exon families, suggesting that these 896 exons might have been the result of intronless in *Arabidopsis* (or a group including *Arabidopsis*). The analyses using the longest transcript did not include these 896 exons among the conserved exons, accounting for ~1.1% of 82,271 conserved individual exons in the *Arabidopsis* genome. Because our results on conserved exon families contain members from multiple genes, the annotation differences of this small number of genes do not affect the main conclusions of this study.

Our exon conservation, gene expression, and alternative splicing analyses also have implications in the evolution of gene families relative to the flower development, especially the MADS-box gene family that has been functionally verified in *Arabidopsis*, rice, and other plants (e.g., refs. 22,93,94,111,112). We analyzed the angiosperm MADS-box homologs by comparing their gene structures with reference to the gene phylogenetic positions and found that the MADS-domain matches the conserved medium exon across these homologs (Supplementary Fig. 45), representing a case of Category-II of the domain-exon correspondence illustrated in Fig. 4a. The K-domain corresponds to 5 or 6 conserved exons, with the first and last exons being typically small and medium exons, respectively (Supplementary Fig. 45). For most angiosperm ABCE genes and other related MADS-box genes, two of the intervening exons (for the K-domain) are small exons, but they are merged into a single medium exon in AGL32 homologs. Also, nearly all these MADS-box genes contain two intervening micro exons (namely, 'M1' and 'M2'; Supplementary Fig. 45), except in many *ANR1* homologs, where the last micro exon (M2) is replaced with a small exon. M1 and M2 in the ABCE genes are relatively frequently skipped in several species in the flower than other genes (Supplementary Fig. 45), implying that the skipping of micro exons have contributed to the protein diversity in relation to the flower development.

Our results provide genome-wide resources for gene structure annotation in genome projects by comparison between homologous genes in an orthologous group to optimize gene structures. For example, among homologs of the *MS2* gene, which is required for normal anther development, most seed plant genes have 9 exons, representing a likely ancestral gene structure (Supplementary Fig. 16). However, in coffee two annotated adjacent genes[113] on chromosome 6 are found here to be homologous to *MS2*: one of these (*GSCOCT00040587001*) has 4 exons corresponding to the four upstream ancestral *MS2* exons and the other (*GSCOCT00040586001*) has 5 exons corresponding to the five downstream ancestral exons. Our results support a likely gene model that combines the two coffee genes into a single gene, with 9 exons similar to other angiosperm *MS2* homologs. As new plant genome sequences are generated rapidly, our genome-wide landscape of gene structures across angiosperms supports in gene model prediction in plant genome projects. The detailed profiles of conserved exons and their encoded protein domains can serve to inform a variety of comparative and evolutionary analyses of gene families that perform diverse molecular functions. Future analyses at the genome/transcriptome scale in many more angiosperms can provide more insights into the evolution of gene structure and alternative splicing.

Methods

Genome data and source

We downloaded 13 Telomere-to-Telomere genomes and 37 genomes assembled in chromosome-scale from published genome databases

(see details in Supplementary Data 1). These 50 genomes represent 46 angiosperms (28 eudicots, 4 magnoliids, 12 monocots, 2 early-diverging groups) and 4 outgroups (gymnosperms)[17,51–57,60–62,113–152]. Particularly, to reduce the computation time, we selected chromosomes 1–14 and the Y chromosome of *Diospyros kaki*[124], the subgenome BXJ2 of *Musa acuminata*[61], and the subgenome B of *Triticum aestivum*[120]. According to the longest mRNA of each gene annotation in the genome annotation (GFF), protein-coding sequences and their translated proteins were used for phylogenomic analyses.

Orthogroup inference and phylogenetic analyses

We used diamond v2.1.8.162[153] and OrthoFinder v2.5.5[154] to analyze protein sequences of the 50 genomes to generate phylogenetic hierarchical orthogroups with the default MCL inflation value (I = 1.5). Specifically, a species-tree (Fig. 1) was selected from the resolved familial relationships among seed-plants from the OneKP project[9]. The species-tree was used as the reference tree for clustering orthogroups by OrthoFinder v2.5.5 with the MSA model, a more accurate way to detect orthogroups from gene trees inferred with default parameters[154]. To reduce computation cost, we also used PhyloMCL v2.0[155] to cluster each orthogroup including >200 genes into smaller orthogroups. Among the inferred orthogroups, to reduce the computation time, we selected orthogroups with >=6 genes from >=3 species (including >=1 angiosperm) for downstream gene tree analyses. In addition, protein sequences of each orthogroup were aligned by PASTA v1.8.6[156] with running in three iterations, each of which include tree inference by FastTree v2.1.11[157], sequence align by MAFFT v7.505[158], and alignment merge by MUSCLE v3.8.31[159]. Aligning coding DNA at the amino acid level shows strong phylogenetic signals because the phylogenetic signals from DNA sequences are more rapidly than that from their encoded protein sequences[160]. Therefore, we used custom script to backtranslate from the protein sequence alignment into nucleotide sequences[161]. Removal of poorly aligned or incomplete sequences compared with the rest of multiple sequence alignments can increase the accuracy of resulting gene trees[162]. Hence, we used trimAl v1.4.rev22[162] with the specific option (automated1) for Maximum Likelihood trees, the residue overlap threshold of 0.7 and the sequence overlap threshold of 75 to trim the nucleotide sequence alignments. The trimmed alignments were used to construct Maximum Likelihood gene tree by IQ-TREE v2.1.2[163] with 1000 ultrafast bootstrap replicates[164] and the substitution model of GTR + G.

In addition, each orthogroup was annotated based on genes from *Arabidopsis* (https://www.arabidopsis.org/index.jsp)[165], rice (http://www.ricesuperpir.com/web/nip/geneExpression)[54], and/or tomato (https://solgenomics.net/ftp/tomato_genome/annotation/ITAG4.1_release/)[57].

Exon Evolution

Exons were selected from the protein-coding regions in all 50 genomes because most published genomes did not include the UTR annotations. For the gene with coding strand, exons were ordered by their annotated positions from lowest to highest; otherwise, exons from genes with template strand were ordered by their positions from highest to lowest. We defined the conservation of two exons according to their amino acid sequence similarity, exon length and position. Specifically, the amino acid sequence similarity of any two exons was estimated by amino acid substitution matrices in BLOSUM-62[166]. The sequence similarity of exons A (from gene-1) and B (from gene-2) was defined as the ratio of the same or similar amino acids counts between the two exons to the number of amino acids in exon-A that were aligned to the amino acids in exon-B (excluding gaps) in the amino acid alignment among gene-1, gene-2, and other homologous genes in an orthogroup. We estimated the similarity between exon-A and exon-B, when there is a difference in length between the two exons less than 20% of the longer exon, and defined exons with >50% similarity as highly conserved exons. We also compared their flanking closest exons (if occurred) and defined such exons with flanking conserved exons as candidate conserved exons. For micro exons, we proposed exons in the similarity range from 30% to 50% as moderately conserved exons. For other exons, we proposed exons in the similarity range from 20% to 50% as moderately conserved exons. According to the defined conservation, we newly developed a software (ExonEvo) to detect the conserved micro/small exons between any two genes in gene tree of orthogroup. Phylogenetic positions of the conserved exons were estimated by the relationships between species (Fig. 2). Specifically, a seed-plant conserved exon was defined as at least one gene from angiosperms and one gene from gymnosperms. An angiosperm-pan conserved exon was defined as the exon shared by any two following groups: (1) eudicots and/or magnoliids; (2) monocots; (3) Nymphaeales; and (4) Amborellales. A gymnosperm-pan conserved exon was defined as the exon shared by genes mapped at the MRCA of Pinales, Cupressales, Cycadales and Ginkgoales. Exons shared by only one species were defined as lineage-specific exon. Finally, we estimated the number of conserved exons in each gene tree and their phylogenetic positions.

Gene ontology analyses

An orthogroup represents a set of homologous genes, with similar functions derived from their common ancestor, and GO terms of their consensus sequence could implicate possible functions of the orthogroup[167]. To compare GO annotations of orthogroups included genes with focal micro and/or small exons, we used the hmmemit tool in HMMER package v3.4[168] to infer the consensus sequences of orthogroups. Among sequences in an orthogroup, genes with possibly false or incomplete assembly and annotation could result into deletion or insertion regions in multiple sequence alignments and long terminal branches in gene tree. To reduce putatively negative effects of such noise sequence signals on the consensus sequence, we selected terminal branches ten times longer than their sister branches in gene tree and ranked them by their branch length from maximum to minimum. Among the descending sorted branches, we iteratively removed any branch two times longer than the average value (in 95% confidence interval) of the length of retained branches in gene tree until the average value did not change. Protein sequence alignments of retained genes were analyzed by using hmmemit with the simple majority-rule to generate a consensus sequence. In addition, we applied online InterPro (https://www.ebi.ac.uk/interpro/search/sequence/) to predict GO terms via searching 17,649 consensus sequences against default databases and mapping hits to InterPro2GO database[169–171], generating GO terms for 66.4% (11,724) of all consensus sequences. Among GO terms annotated by using known knowledge (basic-go database version: releases/2024-03-28; http://purl.obolibrary.org/obo/go/go-basic.obo)[172,173], we selected the ones belonging to the 'Molecular Function' category (corresponding to 8,188 orthogroups) for downstream analyses. Multiple GO terms derived from an ontology could be annotated on an orthogroup; to simplify the GO classification, we grouped the GO terms into regulation of gene expression, protein regulation and modification, metabolism, small-molecule metabolism, nucleic acid metabolism, interaction of proteins, transport, and other category. We used clusterProfiler v4.3.3 to perform enrichment analyses. For GO enrichment analyses, we retrieved the published *Arabidopsis* genome annotation package (org.At.tair.db, version 3.18.0; https://bioconductor.org/packages/release/data/annotation/src/contrib/org.At.tair.db_3.18.0.tar.gz).

Conserved domain analyses

To investigate the correspondence of conserved exons with protein domains, we utilized the hmmsearch tool in HMMER package v3.4 to align protein sequences against the Pfam database v37.0 (releases/2024-04-09; https://ftp.ebi.ac.uk/pub/databases/

Pfam/releases/Pfam37.0/Pfam-A.hmm.gz) with the E-value cutoff of 0.0001. To reduce computation time, we removed redundant domain names, when two or more domains were annotated for the same sequence region, by selecting the domain with the longest coverage. To include the domains frequently described in literatures and to reduce the number of predictions on a given sequence region, we compared such domains in same length, retrieved their literature records in the Europe PMC (https://europepmc.org/; retrieved on 2025-01-15), and preferred to select domains with higher records other than other domains. We also preferred to use the annotated domain other than the unannotated one (DUF). To reduce alternative PFAM annotations that are predicted on different exons matching a conserved exon, we retrieved a total of 22,702 InterPro accessions and their associated 22,950 PFAM accessions from the InterPro database (releases/2025-02-03; https://ftp.ebi.ac.uk/pub/databases/interpro/current_release/interpro.xml.gz) and used the unique InterPro accession to replace one or more homologs PFAM accessions. Specifically, 162 individuals feature at least two PFAMs per InterPro accession.

In addition, GO annotations of an InterPro group, defined as unique protein homologous genes matching the Pfam-based domains, were retrieved from the InterPro2GO database v99.0 (releases/2024-03-27; https://ftp.ebi.ac.uk/pub/databases/interpro/releases/99.0/interpro2go). GO terms were corresponded to each Pfam member according to the respective InterPro group (releases/2024-03-27; https://ftp.ebi.ac.uk/pub/databases/interpro/releases/99.0/interpro.xml.gz). GO terms belonging to the 'Molecular Function' category in the basic-go database (releases/2024-03-28) were used for further comparison.

Gene expression analyses

To infer functional implications of conserved exons, we analyzed RNA-Seq datasets of 11 species that have publicly available transcriptomic datasets from flower, fruit, and at least one of three other tissues (root, stem, and leaf), with each tissue having >=3 replicates released on the same SRA project in NCBI, totaling 178 transcriptomic datasets (see Supplementary Data 13 for accession numbers of transcriptomes in NCBI). We used fastq-dump v3.0.10 to retrieve RNA-Seq reads and remove technical reads (parameters: --skip-technical --readids --read-filter pass --dumpbase --split-3 -clip). The generated reads were processed by fastp v0.24.0[174] with default parameters to filter out low-quality reads. Resulted reads were analyzed by kallisto v0.50.1[175] to calculate transcript per million (TPM) values of genes. For each dataset, we used R-packages DESeq2 v1.42.1 and tximport v1.30.0 to estimate the number of differential expressed genes between two tissues (p-value < 0.05 under False Discovery Rate test). Among these differential expressed genes, we selected preferentially expressed genes (PEGs) that have a high fold change (>2) of expression in the flower relative to one of other tissues. These PEGs were further filtered out to contain at least one conserved exon in a PEG for downstream analyses.

To compare the domain-exon correspondence in the conserved exon-containing PEGs, we performed enrichment analyses for each species. Among PEGs for a pair of tissues with conserved exons, we compared the ratio of members of a domain family to all domains; the ratio was compared to that for all exon-containing genes to generate a p-value by a two-side hypergeometric test. A domain family with a p-value < 0.05 was selected as an enriched family. In addition, to compare the expression pattern between ABCE genes and other MADS-box genes across angiosperms, we performed the Kruskal-Wallis test in R using distinct genes within each subfamily as replicates.

Detection of alternative splicing events

To detect alternatively spliced exons in development of different tissues, we retrieved the RNA-Seq data that were used in the above section. For each species, those clean reads were mapped to the corresponding reference genome sequence by using HISAT2 v2.2.1[176] with default parameters. Alignments of reads against the genome sequence were transformed into Binary Alignment/Map (BAM) format files by using SAMtools v1.21[177]. To detect transcripts, we utilized StringTie v2.2.3[178] to analyze those BAM files under the reference genome guide and to merge those transcripts from different tissues into a non-redundant dataset of transcripts. To detect alternative splicing events between one tissue and one of other tissues, we used rMATS v4.3.0[179] with the parameters (--variable-read-length --novelSS) to compare those transcripts. We also filtered out the average inclusion and skipping junction number of <5 and required the PSI value of >0 and <1 for a focal tissue (see PSI in ref. 179). The alternative splicing events (including exon skipping, alternative 3′ splice site, and alternative 5′ splice site) of exon of a gene in corresponding tissue, we detected at least two kinds of transcripts with and without exons in this gene.

Detection of processed pseudogenes

Processed pseudogenes are characterized by the loss of introns, the presence of poly(A) sites near the 3′ end, and, frequently, the presence of repeat elements. To detect potential processed pseudogenes from genes lacking introns, we focused on the Chinese pine genome with publicly available repeat element annotations. We examined exons annotated with repeat elements and identified poly(A) signals that predicted in plant genomes[180] within a 200-bp region near the 3′ end. We employed Kallisto v0.50.1[175] to estimate the gene expression levels of the those single-exon genes (see Supplementary Data 13 for accession numbers of transcriptomes in NCBI). Single-exon genes featuring repeat elements, poly(A) motifs, a TPM value ≤ 1 were proposed as potential processed pseudogenes.

Reporting summary

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

Data availability

The accessions in Supplementary Data 1 and 13 are available in NCBI and other public databases. Specific databases used in our analyses include the *Arabidopsis* genome annotation package (https://bioconductor.org/packages/release/data/annotation/src/contrib/org.At.tair.db_3.18.0.tar.gz), and the basic-go database (releases/2024-03-28; http://purl.obolibrary.org/obo/go/go-basic.obo), the InterPro2GO database v99.0 (releases/2024-03-27; https://ftp.ebi.ac.uk/pub/databases/interpro/releases/99.0/interpro2go), and the respective InterPro group (releases/2025-02-03; https://ftp.ebi.ac.uk/pub/databases/interpro/current_release/interpro.xml.gz). Datasets including the sequence alignments, gene tree files of orthogroups, and the conserved exons and their matched domains in 50 genomes are available on FigShare (https://doi.org/10.6084/m9.figshare.28683905.v3). Source data are provided with this paper.

Code availability

The ExonEvo for estimating the conserved exons among homologous genes is available at https://github.com/TaikuiZhang/ExonEvo and code ocean (https://doi.org/10.24433/CO.4085198.v1).

References

1. Wan, R., Bai, R., Zhan, X. & Shi, Y. How is precursor messenger RNA spliced by the spliceosome? *Annu. Rev. Biochem.* **89**, 333–358 (2020).
2. Gehring, N. H. & Roignant, J.-Y. Anything but ordinary – emerging splicing mechanisms in eukaryotic gene regulation. *Trends Genet.* **37**, 355–372 (2021).
3. Choquet, K., Patop, I. L. & Churchman, L. S. The regulation and function of post-transcriptional RNA splicing. *Nat. Rev. Genet.* **26**, 378–394 (2025).

4. Jin, J. et al. Eukaryotic protein domains as functional units of cellular evolution. *Sci. Signal.* **2**, ra76-ra76 (2009).
5. Chothia, C., Gough, J., Vogel, C. & Teichmann, S. A. Evolution of the protein repertoire. *Science* **300**, 1701–1703 (2003).
6. Liu, M. & Grigoriev, A. Protein domains correlate strongly with exons in multiple eukaryotic genomes – evidence of exon shuffling? *Trends Genet.* **20**, 399–403 (2004).
7. Du, Y. et al. Differences in alternative splicing and their potential underlying factors between animals and plants. *J. Adv. Res.* **64**, 83–98 (2024).
8. Marsh, J. A. & Teichmann, S. A. How do proteins gain new domains? *Genome Biol.* **11**, 126 (2010).
9. Leebens-Mack, J. H. et al. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679–685 (2019).
10. Zhang, L. et al. Nuclear phylogenomics of angiosperms and evolutionary implications. *Diversity* **17**, 136 (2025).
11. Yang, Y. et al. Recent advances on phylogenomics of gymnosperms and a new classification. *Plant Diversity* **44**, 340–350 (2022).
12. Lange, B. M. The evolution of plant secretory structures and emergence of terpenoid chemical diversity. *Annu. Rev. Plant Biol.* **66**, 139–159 (2015).
13. Linkies, A., Graeber, K., Knight, C. & Leubner-Metzger, G. The evolution of seeds. *N. Phytol.* **186**, 817–831 (2010).
14. Leitch, A. R. & Leitch, I. J. Ecological and genetic factors linked to contrasting genome dynamics in seed plants. *N. Phytol.* **194**, 629–646 (2012).
15. Xie, D., Chen, T.-X., Du, H., Wu, H. & Ran, J.-H. Gymnosperm collecting in China: Past, present, and future. *J. Syst. Evol.* **63**, 479–491 (2025).
16. Guo, L. & Liu, C.-M. A single-nucleotide exon found in *Arabidopsis*. *Sci. Rep.* **5**, 18087 (2015).
17. Niu, S. et al. The Chinese pine genome and methylome unveil key features of conifer evolution. *Cell* **185**, 204–217 (2022).
18. Irimia, M. et al. A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* **159**, 1511–1523 (2014).
19. Black, D. L. Does steric interference between splice sites block the splicing of a short c-src neuron-specific exon in non-neuronal cells? *Genes Dev.* **5**, 389–402 (1991).
20. Carlo, T., Sierra, R. & Berget, S. M. A 5′ splice site-proximal enhancer binds SF1 and activates exon bridging of a microexon. *Mol. Cell. Biol.* **20**, 3988–3995 (2000).
21. Fairbrother, W. G., Yeh, R.-F., Sharp, P. A. & Burge, C. B. Predictive identification of exonic splicing enhancers in human genes. *Science* **297**, 1007–1013 (2002).
22. Wang, J. et al. Regulation of micro- and small-exon retention and other splicing processes by GRP20 for flower development. *Nat. Plants* **10**, 66–85 (2024).
23. Song, Q., Bari, A., Li, H. & Chen, L.-L. Identification and analysis of micro-exons in *AP2/ERF* and *MADS* gene families. *FEBS Open Bio* **10**, 2564–2577 (2020).
24. Zhu, F.-Y. et al. SWATH-MS-based proteogenomic analysis reveals the involvement of alternative splicing in poplar upon lead stress. *Genome Res.* **33**, 371–385 (2023).
25. Filichkin, S. A. et al. Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. *Genome Res.* **20**, 45–58 (2010).
26. Wang, B.-B. & Brendel, V. Genomewide comparative analysis of alternative splicing in plants. *Proc. Natl. Acad. Sci. USA* **103**, 7175–7180 (2006).
27. Keller, M. et al. Alternative splicing in tomato pollen in response to heat stress. *DNA Res.* **24**, 205–217 (2016).
28. Shen, Y. et al. Global dissection of alternative splicing in paleopolyploid soybean. *Plant Cell* **26**, 996–1008 (2014).
29. Wang, T. et al. Comprehensive profiling of rhizome-associated alternative splicing and alternative polyadenylation in moso bamboo (*Phyllostachys edulis*). *Plant J.* **91**, 684–699 (2017).
30. Wang, M. et al. A global survey of alternative splicing in allopolyploid cotton: landscape, complexity and regulation. *N. Phytol.* **217**, 163–178 (2018).
31. Kim, P., Yang, M., Yiya, K., Zhao, W. & Zhou, X. ExonSkipDB: functional annotation of exon skipping event in human. *Nucleic Acids Res.* **48**, D896–D907 (2019).
32. Wang, K. et al. Multi-strategic RNA-seq analysis reveals a high-resolution transcriptional landscape in cotton. *Nat. Commun.* **10**, 4714 (2019).
33. Zhao, C. & Beers, E. P. Alternative splicing of Myb-related genes *MYR1* and *MYR2* may modulate activities through changes in dimerization, localization, or protein folding. *Plant Signal. Behav.* **8**, e27325 (2013).
34. Szécsi, J. et al. *BIGPETALp*, a bHLH transcription factor is involved in the control of *Arabidopsis* petal size. *EMBO J.* **25**, 3912–3920 (2006).
35. Deng, Y. et al. Heat induces the splicing by *IRE1* of a mRNA encoding a transcription factor involved in the unfolded protein response in *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **108**, 7247–7252 (2011).
36. Hugouvieux, V. et al. Tetramerization of MADS family transcription factors SEPALLATA3 and AGAMOUS is required for floral meristem determinacy in *Arabidopsis*. *Nucleic Acids Res.* **46**, 4966–4977 (2018).
37. Puranik, S. et al. Structural basis for the oligomerization of the MADS domain transcription factor SEPALLATA3 in *Arabidopsis*. *Plant Cell* **26**, 3603–3615 (2014).
38. Conn, V. M. et al. A circRNA from *SEPALLATA3* regulates splicing of its cognate mRNA through R-loop formation. *Nat. Plants* **3**, 17053 (2017).
39. Zhang, X.-C. et al. Evolutionary dynamics of protein domain architecture in plants. *BMC Evol. Biol.* **12**, 6 (2012).
40. Zhang, T., Liu, C., Huang, X., Zhang, H. & Yuan, Z. Land-plant phylogenomic and pomegranate transcriptomic analyses reveal an evolutionary scenario of *CYP75* genes subsequent to whole genome duplications. *J. Plant Biol.* **62**, 48–60 (2019).
41. Feng, G., Burleigh, J. G., Braun, E. L., Mei, W. & Barbazuk, W. B. Evolution of the 3R-MYB gene family in plants. *Genome Biol. Evol.* **9**, 1013–1029 (2017).
42. Xu, G., Ma, H., Nei, M. & Kong, H. Evolution of F-box genes in plants: Different modes of sequence divergence and their relationships with functional diversification. *Proc. Natl Acad. Sci. USA* **106**, 835–840 (2009).
43. Finet, C., Berne-Dedieu, A., Scutt, C. P. & Marlétaz, F. Evolution of the *ARF* gene family in land plants: old domains, new tricks. *Mol. Biol. Evol.* **30**, 45–56 (2012).
44. Liu, J. et al. Genome-wide analysis and expression profile of the bZIP transcription factor gene family in grapevine (*Vitis vinifera*). *BMC Genomics* **15**, 281 (2014).
45. Torres Acosta, J. A., Fowke, L. C. & Wang, H. Analyses of phylogeny, evolution, conserved sequences and genome-wide expression of the *ICK/KRP* family of plant CDK inhibitors. *Ann. Bot.* **107**, 1141–1157 (2011).
46. Kofuji, R. et al. Evolution and divergence of the MADS-Box gene family based on genome-wide expression analyses. *Mol. Biol. Evol.* **20**, 1963–1977 (2003).
47. Mukherjee, K., Brocchieri, L. & Bürglin, T. R. A comprehensive classification and evolutionary analysis of plant homeobox genes. *Mol. Biol. Evol.* **26**, 2775–2794 (2009).
48. Yu, H. et al. Pervasive misannotation of microexons that are evolutionarily conserved and crucial for gene function in plants. *Nat. Commun.* **13**, 820 (2022).
49. Yuan, Z. et al. The pomegranate (*Punica granatum* L.) genome provides insights into fruit quality and ovule developmental biology. *Plant Biotechnol. J.* **16**, 1363–1374 (2018).

50. Arabidopsis Genome Initiative Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796 (2000).
51. Han, X. et al. Two haplotype-resolved, gap-free genome assemblies for *Actinidia latifolia* and *Actinidia chinensis* shed light on the regulatory mechanisms of vitamin C and sucrose metabolism in kiwifruit. *Mol. Plant* **16**, 452–470 (2023).
52. Wang, Y.-H. et al. Telomere-to-telomere carrot (*Daucus carota*) genome assembly reveals carotenoid characteristics. *Hortic. Res.* **10**, uhad103 (2023).
53. He, Z. et al. Evolution of coastal forests based on a full set of mangrove genomes. *Nat. Ecol. Evol.* **6**, 738–749 (2022).
54. Shang, L. et al. A complete assembly of the rice Nipponbare reference genome. *Mol. Plant* **16**, 1232–1236 (2023).
55. Ranawaka, B. et al. A multi-omic *Nicotiana benthamiana* resource for fundamental research and biotechnology. *Nat. Plants* **9**, 1558–1571 (2023).
56. Wang, M. et al. Genomic innovation and regulatory rewiring during evolution of the cotton genus *Gossypium*. *Nat. Genet.* **54**, 1959–1971 (2022).
57. Tomato Genome Consortium The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**, 635–641 (2012).
58. Albert, V. A. et al. The amborella genome and the evolution of flowering plants. *Science* **342**, 1241089 (2013).
59. Ma, W. et al. *WRINKLED1*, a ubiquitous regulator in oil accumulating tissues from *Arabidopsis* embryos to oil palm mesocarp. *PLoS ONE* **8**, e68887 (2013).
60. Wang, B. et al. High-quality *Arabidopsis thaliana* genome assembly with Nanopore and HiFi long reads. *Genomics Proteom. Bioinforma.* **20**, 4–13 (2022).
61. Huang, H.-R. et al. Telomere-to-telomere haplotype-resolved reference genome reveals subgenome divergence and disease resistance in triploid *Cavendish banana*. *Hortic. Res.* **10**, uhad153 (2023).
62. Nath, O. et al. A haplotype resolved chromosomal level avocado genome allows analysis of novel avocado genes. *Hortic. Res.* **9**, uhac157 (2022).
63. Isono, K. et al. *LHT1/MAC7* contributes to proper alternative splicing under long-term heat stress and mediates variation in the heat tolerance of *Arabidopsis*. *PNAS Nexus* **2**, pgad348 (2023).
64. Aarts, M. G. M. et al. The *Arabidopsis* MALE STERILITY 2 protein shares similarity with reductases in elongation/condensation complexes. *Plant J.* **12**, 615–623 (1997).
65. Chen, W. et al. Male sterile2 encodes a plastid-localized fatty acyl carrier protein reductase required for pollen exine development in *Arabidopsis*. *Plant Physiol.* **157**, 842–853 (2011).
66. Li, Q. et al. Interspecies transcriptomic comparisons reveal potential molecular genetic mechanisms underlying the evolutionary development of dimorphic flowers in *Viola*. *J. Syst. Evol.* **4**, 892–909 (2025).
67. Law, J. A. et al. Polymerase IV occupancy at RNA-directed DNA methylation sites requires SHH1. *Nature* **498**, 385–389 (2013).
68. Suer, S., Agusti, J., Sanchez, P., Schwarz, M. & Greb, T. *WOX4* imparts auxin responsiveness to cambium cells in *Arabidopsis*. *Plant Cell* **23**, 3247–3259 (2011).
69. Kucukoglu, M., Nilsson, J., Zheng, B., Chaabouni, S. & Nilsson, O. WUSCHEL-RELATED HOMEOBOX4 (*WOX4*)-like genes regulate cambial cell division activity and secondary growth in *Populus* trees. *N. Phytol.* **215**, 642–657 (2017).
70. Vuolo, F. et al. LMI1 homeodomain protein regulates organ proportions by spatial modulation of endoreduplication. *Genes Dev.* **32**, 1361–1366 (2018).
71. Jang, M.-J., Hong, W.-J., Park, Y.-S., Jung, K.-H. & Kim, S. Genomic basis of multiphase evolution driving divergent selection of zinc-finger homeodomain genes. *Nucleic Acids Res.* **51**, 7424–7437 (2023).
72. Figueiredo, D. D. et al. Seven zinc-finger transcription factors are novel regulators of the stress responsive gene *OsDREB1B*. *J. Exp. Bot.* **63**, 3643–3656 (2012).
73. Song, H.-R. et al. The RNA binding protein ELF9 directly reduces *SUPPRESSOR OF OVEREXPRESSION OF CO1* transcript levels in *Arabidopsis*, possibly via nonsense-mediated mRNA decay. *Plant Cell* **21**, 1195–1211 (2009).
74. Lorković, Z. J. & Barta, A. Genome analysis: RNA recognition motif (RRM) and K homology (KH) domain RNA-binding proteins from the flowering plant *Arabidopsis thaliana*. *Nucleic Acids Res.* **30**, 623–635 (2002).
75. Rushton, P. J., Somssich, I. E., Ringler, P. & Shen, Q. J. WRKY transcription factors. *Trends Plant Sci.* **15**, 247–258 (2010).
76. Tao, Z. et al. *OsWRKY45* alleles play different roles in abscisic acid signalling and salt stress tolerance but similar roles in drought and cold tolerance in rice. *J. Exp. Bot.* **62**, 4863–4874 (2011).
77. Xu, X., Chen, C., Fan, B. & Chen, Z. Physical and functional interactions between pathogen-induced *Arabidopsis* WRKY18, WRKY40, and WRKY60 transcription factors. *Plant Cell* **18**, 1310–1326 (2006).
78. Gramzow, L. & Theissen, G. A hitchhiker's guide to the MADS world of plants. *Genome Biol.* **11**, 214 (2010).
79. Becker, A. & Theißen, G. The major clades of MADS-box genes and their role in the development and evolution of flowering plants. *Mol. Phylogenet. Evol.* **29**, 464–489 (2003).
80. Severing, E. I. et al. Predicting the impact of alternative splicing on plant MADS domain protein function. *PLoS ONE* **7**, e30524 (2012).
81. Grunwald, Y., Gosa, S. C., Torne-Srivastava, T., Moran, N. & Moshelion, M. Out of the blue: Phototropins of the leaf vascular bundle sheath mediate the regulation of leaf hydraulic conductance by blue light. *Plant Cell* **34**, 2328–2342 (2022).
82. Vigneault, F. et al. Members of the plant NIMA-related kinases are involved in organ development and vascularization in poplar, *Arabidopsis* and rice. *Plant J.* **51**, 575–588 (2007).
83. Chia, K.-S. & Carella, P. Taking the lead: NLR immune receptor N-terminal domains execute plant immune responses. *N. Phytol.* **240**, 496–501 (2023).
84. Liu, Y. et al. An angiosperm NLR Atlas reveals that *NLR* gene reduction is associated with ecological specialization and signal transduction component deletion. *Mol. Plant* **14**, 2015–2031 (2021).
85. Li, S.-X., Liu, Y., Zhang, Y.-M., Chen, J.-Q. & Shao, Z.-Q. Convergent reduction of immune receptor repertoires during plant adaptation to diverse special lifestyles and habitats. *Nat. Plants* **11**, 248–262 (2025).
86. Godinho, D. P., Yanez, R. J. R. & Duque, P. Pathogen-responsive alternative splicing in plant immunity. *Trends Plant Sci.* **30**, 615–628 (2025).
87. Lei, L., Zhou, S.-L., Ma, H. & Zhang, L.-S. Expansion and diversification of the SET domain gene family following whole-genome duplications in *Populus trichocarpa*. *BMC Evol. Biol.* **12**, 51 (2012).
88. Ma, Z. et al. *Arabidopsis* serrate coordinates histone methyltransferases ATXR5/6 and RNA processing factor RDR6 to regulate transposon expression. *Dev. Cell* **45**, 769–784 (2018).
89. Ng, D. W. K. et al. Plant SET domain-containing proteins: structure, function and regulation. *Biochim. Biophys. Acta* **1769**, 316–329 (2007).
90. Hou, X.-L. et al. DEAD-BOX RNA HELICASE 27 regulates microRNA biogenesis, zygote division, and stem cell homeostasis. *Plant Cell* **33**, 66–84 (2020).
91. Li, Q. et al. DEAD-box helicases modulate dicing body formation in *Arabidopsis*. *Sci. Adv.* **7**, eabc6266 (2021).

92. Jouili, H., Bouazizi, H. & El Ferjani, E. Plant peroxidases: biomarkers of metallic stress. *Acta Physiol. Plant.* **33**, 2075–2082 (2011).
93. Ma, H., Yanofsky, M. F. & Meyerowitz, E. M. *AGL1-AGL6*, an *Arabidopsis* gene family with similarity to floral homeotic and transcription factor genes. *Genes Dev.* **5**, 484–495 (1991).
94. Soltis, D. E. et al. The floral genome: an evolutionary history of gene duplication and shifting patterns of gene expression. *Trends Plant Sci.* **12**, 358–367 (2007).
95. Fatima, U. et al. AtSWEET11 and AtSWEET12 transporters function in tandem to modulate sugar flux in plants. *Plant Direct* **7**, e481 (2023).
96. Chen, C. et al. The *Arabidopsis ATK1* gene is required for spindle morphogenesis in male meiosis. *Development* **129**, 2401–2409 (2002).
97. Demidov, D. et al. Deregulated phosphorylation of CENH3 at Ser65 affects the development of floral meristems in *Arabidopsis thaliana*. *Front. Plant Sci.* **10**, 928 (2019).
98. van Mourik, H. et al. Dual specificity and target gene selection by the MADS-domain protein FRUITFULL. *Nat. Plants* **9**, 473–485 (2023).
99. Pelaz, S., Ditta, G. S., Baumann, E., Wisman, E. & Yanofsky, M. F. B and C floral organ identity functions require *SEPALLATA MADS-box* genes. *Nature* **405**, 200–203 (2000).
100. Zahn, L. M. et al. The evolution of the *SEPALLATA* subfamily of MADS-Box genes: a preangiosperm origin with multiple duplications throughout angiosperm history. *Genetics* **169**, 2209–2223 (2005).
101. Yang, Y., Fanning, L. & Jack, T. The K domain mediates heterodimerization of the *Arabidopsis* floral organ identity proteins, APETALA3 and PISTILLATA. *Plant J.* **33**, 47–59 (2003).
102. Woodward, A. W. & Bartel, B. Auxin: regulation, action, and interaction. *Ann. Bot.* **95**, 707–735 (2005).
103. Kim, J. & DellaPenna, D. Defining the primary route for lutein synthesis in plants: The role of *Arabidopsis* carotenoid β-ring hydroxylase CYP97A3. *Proc. Natl Acad. Sci. USA* **103**, 3474–3479 (2006).
104. Chen, D., Yan, W., Fu, L.-Y. & Kaufmann, K. Architecture of gene regulatory networks controlling flower development in *Arabidopsis thaliana*. *Nat. Commun.* **9**, 4534 (2018).
105. Ghelli, R. et al. A newly identified flower-specific splice variant of *AUXIN RESPONSE FACTOR8* regulates stamen elongation and endothecium lignification in *Arabidopsis*. *Plant Cell* **30**, 620–637 (2018).
106. Carmel, L. & Chorev, M. The function of introns. *Front. Genet.* **3**, 55 (2012).
107. Kong, H. et al. Patterns of gene duplication in the plant *SKP1* gene family in angiosperms: evidence for multiple mechanisms of rapid gene birth. *Plant J.* **50**, 873–885 (2007).
108. Prade, V. M. et al. The pseudogenes of barley. *Plant J.* **93**, 502–514 (2018).
109. Drouin, G. & Dover, G. A. A plant processed pseudogene. *Nature* **328**, 557–558 (1987).
110. Zhao, D. et al. Members of the *Arabidopsis-SKP1-like* gene family exhibit a variety of expression patterns and may play diverse roles in *Arabidopsis*. *Plant Physiol.* **133**, 203–217 (2003).
111. Bowman, J. L. & Moyroud, E. Reflections on the ABC model of flower development. *Plant Cell* **36**, 1334–1357 (2024).
112. Li, L. et al. Interactions among proteins of floral MADS-box genes in *Nuphar pumila* (Nymphaeaceae) and the most recent common ancestor of extant angiosperms help understand the underlying mechanisms of the origin of the flower. *J. Syst. Evol.* **53**, 285–296 (2015).
113. Denoeud, F. et al. The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science* **345**, 1181–1184 (2014).
114. Ming, R. et al. The pineapple genome and the evolution of CAM photosynthesis. *Nat. Genet.* **47**, 1435–1442 (2015).
115. Yang, X. et al. The gap-free potato genome assembly reveals large tandem gene clusters of agronomical importance in highly repeated genomic regions. *Mol. Plant* **16**, 314–317 (2023).
116. Sun, X. et al. A chromosome-level genome assembly of garlic (*Allium sativum*) provides insights into genome evolution and allicin biosynthesis. *Mol. Plant* **13**, 1328–1339 (2020).
117. Wang, L. et al. A telomere-to-telomere gap-free assembly of soybean genome. *Mol. Plant* **16**, 1711–1714 (2023).
118. Mascher, M. et al. Long-read sequence assembly: a technical evaluation in barley. *Plant Cell* **33**, 1888–1906 (2021).
119. Cheng, S.-P. et al. Haplotype-resolved genome assembly and allele-specific gene expression in cultivated ginger. *Hortic. Res.* **8**, 188 (2021).
120. International Wheat Genome Sequencing Consortium Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* **361**, eaar7191 (2018).
121. Käfer, J. et al. A derived ZW chromosome system in *Amborella trichopoda*, representing the sister lineage to all other extant flowering plants. *N. Phytol.* **233**, 1636–1642 (2022).
122. Wang, J. et al. Construction of pseudomolecules for the Chinese chestnut (*Castanea mollissima*) genome. *G3 (Bethesda)* **10**, 3565–3574 (2020).
123. Bao, Y. et al. A gap-free and haplotype-resolved lemon genome provides insights into flavor synthesis and huanglongbing (HLB) tolerance. *Hortic. Res.* **10**, uhad020 (2023).
124. Horiuchi, A. et al. Ongoing rapid evolution of a post-Y region revealed by chromosome-scale genome assembly of a hexaploid monoecious persimmon (*Diospyros kaki*). *Mol. Biol. Evol.* **40**, msad151 (2023).
125. Tuskan, G. A. et al. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**, 1596–1604 (2006).
126. Fu, F. et al. The *Metasequoia* genome and evolutionary relationships among redwoods. *Plant Commun.* **4**, 100643 (2023).
127. Qin, L. et al. Insights into angiosperm evolution, floral development and chemical biosynthesis from the *Aristolochia fimbriata* genome. *Nat. Plants* **7**, 1239–1253 (2021).
128. Chen, J.-Y. et al. A chromosome-scale genome sequence of pitaya (*Hylocereus undatus*) provides novel insights into the genome evolution and regulation of betalain biosynthesis. *Hortic. Res.* **8**, 164 (2021).
129. Yuan, J. et al. Genomic basis of the giga-chromosomes and giga-genome of tree peony *Paeonia ostii*. *Nat. Commun.* **13**, 7328 (2022).
130. Chen, J. et al. A complete telomere-to-telomere assembly of the maize genome. *Nat. Genet.* **55**, 1221–1231 (2023).
131. Zhang, L. et al. A near-complete genome assembly of *Brassica rapa* provides new insights into the evolution of centromeres. *Plant Biotechnol. J.* **21**, 1022–1032 (2023).
132. Zhang, L. et al. The water lily genome and the early evolution of flowering plants. *Nature* **577**, 79–84 (2020).
133. Wang, Y. et al. A chromosome-level genome of *Syringa oblata* provides new insights into chromosome formation in Oleaceae and evolutionary history of lilacs. *Plant J.* **111**, 836–848 (2022).
134. Guo, L. et al. The opium poppy genome and morphinan production. *Science* **362**, 343–347 (2018).
135. Liu, Y. et al. The *Cycas* genome and the early evolution of seed plants. *Nat. Plants* **8**, 389–401 (2022).
136. Pecrix, Y. et al. Whole-genome landscape of *Medicago truncatula* symbiotic genes. *Nat. Plants* **4**, 1017–1025 (2018).
137. Zhang, W. et al. Genome assembly of wild tea tree DASZ reveals pedigree and selection history of tea varieties. *Nat. Commun.* **11**, 3719 (2020).

138. Zhou, Y. et al. The telomere-to-telomere genome of *Fragaria vesca* reveals the genomic evolution of *Fragaria* and the origin of cultivated octoploid strawberry. *Hortic. Res.* **10**, uhad027 (2023).
139. Qi, H., Yu, F., Deng, J., Zhang, L. & Yang, P. The high-quality genome of lotus reveals tandem duplicate genes involved in stress response and secondary metabolites biosynthesis. *Hortic. Res.* **10**, uhad040 (2023).
140. Shen, F. et al. Comparative genomics reveals a unique nitrogen-carbon balance system in Asteraceae. *Nat. Commun.* **14**, 4334 (2023).
141. Song, A. et al. Analyses of a chromosome-scale genome assembly reveal the origin and evolution of cultivated chrysanthemum. *Nat. Commun.* **14**, 2021 (2023).
142. Guo, X. et al. The genome of *Acorus* deciphers insights into early monocot evolution. *Nat. Commun.* **14**, 3662 (2023).
143. Shi, X. et al. The complete reference genome for grapevine (*Vitis vinifera* L.) genetics and breeding. *Hortic. Res.* **10**, uhad061 (2023).
144. Deng, Y. et al. A telomere-to-telomere gap-free reference genome of watermelon and its mutation library provide important resources for gene discovery and breeding. *Mol. Plant* **15**, 1268–1284 (2022).
145. Dong, S. et al. The genome of *Magnolia biondii* Pamp. provides insights into the evolution of Magnoliales and biosynthesis of terpenoids. *Hortic. Res.* **8**, 38 (2021).
146. Zhu, S. et al. The snapdragon genomes reveal the evolutionary dynamics of the *S-locus* supergene. *Mol. Biol. Evol.* **40**, msad080 (2023).
147. Liu, H. et al. The nearly complete genome of *Ginkgo biloba* illuminates gymnosperm evolution. *Nat. Plants* **7**, 748–756 (2021).
148. Piet, Q. et al. A chromosome-level, haplotype-phased *Vanilla planifolia* genome highlights the challenge of partial endoreplication for accurate whole-genome assembly. *Plant Commun.* **3**, 100330 (2022).
149. Paterson, A. H. et al. The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**, 551–556 (2009).
150. Hu, L. et al. The chromosome-scale reference genome of black pepper provides insight into piperine biosynthesis. *Nat. Commun.* **10**, 4702 (2019).
151. Li, W. et al. Near-gapless and haplotype-resolved apple genomes provide insights into the genetic basis of rootstock-induced dwarfing. *Nat. Genet.* **56**, 505–516 (2024).
152. Carey, S. B. et al. ZW sex chromosome structure in *Amborella trichopoda*. *Nat. Plants* **10**, 1944–1954 (2024).
153. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
154. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
155. Zhou, S., Chen, Y., Guo, C. & Qi, J. PhyloMCL: Accurate clustering of hierarchical orthogroups guided by phylogenetic relationship and inference of polyploidy events. *Methods Ecol. Evol.* **11**, 943–954 (2020).
156. Mirarab, S., Nguyen, N. & Warnow, T. PASTA: ultra-large multiple sequence alignment. *Res. Comput. Mol. Biol.* **8394**, 177–191 (2014).
157. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
158. Katoh, K. & Standley, D. M. A simple method to control over-alignment in the MAFFT multiple sequence alignment program. *Bioinformatics* **32**, 1933–1942 (2016).
159. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
160. Wernersson, R. & Pedersen, A. G. RevTrans: multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic Acids Res.* **31**, 3537–3539 (2003).
161. Li, W. et al. Pan-genome analysis reveals the evolution and diversity of *Malus*. *Nat. Genet.* **57**, 1274–1286 (2025).
162. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
163. Minh, B. Q. et al. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
164. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
165. Lamesch, P. et al. The *Arabidopsis* Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* **40**, D1202–D1210 (2011).
166. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **89**, 10915–10919 (1992).
167. Zhang, T. et al. Phylogenomic profiles of whole-genome duplications in Poaceae and landscape of differential duplicate retention and losses among major Poaceae lineages. *Nat. Commun.* **15**, 3305 (2024).
168. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
169. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
170. Burge, S. et al. Manual GO annotation of predictive protein signatures: the InterPro approach to GO curation. *Database* **2012**, bar068 (2012).
171. Blum, M. et al. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* **49**, D344–D354 (2020).
172. Ashburner, M. et al. Gene Ontology: Tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
173. The Gene Ontology Consortium The Gene Ontology resource: Enriching a GOld mine. *Nucleic Acids Res.* **49**, D325–D334 (2020).
174. Chen, S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *iMeta* **2**, e107 (2023).
175. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-Seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
176. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
177. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
178. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
179. Shen, S. et al. rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci. USA* **111**, 5593–5601 (2014).
180. Ji, G. et al. PASPA: a web server for mRNA poly(A) site predictions in plants and algae. *Bioinformatics* **31**, 1671–1673 (2015).
181. Muthuramalingam, M., Wang, Y., Li, Y. & Mahalingam, R. Interacting protein partners of *Arabidopsis* RNA-binding protein AtRBP45b. *Plant Bio.* **19**, 327–334 (2017).
182. Jin, S. et al. The CUL3A–LFH1–UBC15 ubiquitin ligase complex mediates SHORT VEGETATIVE PHASE degradation to accelerate flowering at high ambient temperature. *Plant Commun.* **5**, 100814 (2024).
183. Ma, X. et al. Post-pollination sepal longevity of female flower co-regulated by energy-associated multiple pathways in dioecious spinach. *Front. Plant Sci.* **13**, 1010149 (2022).
184. Mizuno, S. et al. Receptor-like protein kinase 2 (RPK 2) is a novel factor controlling anther development in *Arabidopsis thaliana*. *Plant J.* **50**, 751–766 (2007).

185. Krizek, B. A. & Meyerowitz, E. M. The *Arabidopsis* homeotic genes *APETALA3* and *PISTILLATA* are sufficient to provide the B class organ identity function. *Development* **122**, 11–22 (1996).
186. Wijeratne, A. J. et al. Differential gene expression in *Arabidopsis* wild-type and mutant anthers: insights into anther cell differentiation and regulatory networks. *Plant J.* **52**, 14–29 (2007).
187. Chen, X., Liu, J., Cheng, Y. & Jia, D. HEN1 functions pleiotropically in *Arabidopsis* development and acts in C function in the flower. *Development* **129**, 1085–1094 (2002).
188. Sudre, D. et al. Iron-dependent modifications of the flower transcriptome, proteome, metabolome, and hormonal content in an *Arabidopsis* ferritin mutant. *J. Exp. Bot.* **64**, 2665–2688 (2013).
189. Smaczniak, C. et al. Characterization of MADS-domain transcription factor complexes in *Arabidopsis* flower development. *Proc. Natl. Acad. Sci. USA* **109**, 1560–1565 (2012).

Acknowledgements

We thank Prof. Dawei Li for providing the kiwifruit genome resource; Profs. Ziwen He and Suhua Shi for providing the *Nypa fruticans* genome resource; Prof. Wei Li for providing the apple genome resource; and Dr. Hui Liu for providing the carrot genome resource. We thank Daisy Noe, João Vitor de S. Messeder, and Clancy Andrew McIntyre for valuable discussions. We thank Profs. Brad Barbazuk and Peipei Wang for very helpful comments on the manuscript. This work was supported by funds from Eberly College of Science and the Huck Institutes of the Life Sciences at the Pennsylvania State University (H.M.), the China Postdoctoral Science Foundation (2019M661344 to T.Z.), and the Research Startup Fund for High-Level and High-Educated Talents of Nanjing Forestry University (T.Z.). Computations for this research were performed on the Pennsylvania State University's Institute for Computational and Data Sciences' Roar supercomputer (USA), the National Supercomputing Center in Shenzhen (China), and the National Supercomputing Center in Wuzhen (China).

Author contributions

H.M., J.W., and T.Z. designed the research. H.M. supervised research and provided funds. T.Z. performed data analyses and wrote the draft manuscript. L.Z., J.W., and H.M. revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-66816-3.

Correspondence and requests for materials should be addressed to Hong Ma or Jun Wang.

Peer review information *Nature Communications* thanks Brad Barbazuk and Peipei Wang for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025