





The *Euscaphis japonica* genome and the evolution of malvids

Wei-Hong Sun^{1,2,3} , Zhen Li^{4,5} , Shuang Xiang^{1,2,3}, Lin Ni², Diyang Zhang³ , De-Qiang Chen^{1,2,3}, Meng-Yuan Qiu^{1,2,3}, Qi-Gong Zhang^{1,2}, Lin Xiao^{1,2}, Le Din^{1,2,3}, Yifan Li^{1,2}, Xing-Yu Liao³, Xue-Die Liu³, Yu-Ting Jiang³, Pei-Lan Zhang^{1,2,3}, Hui Ni^{1,2}, Yifan Wang^{1,2}, Yi-Xun Yue^{1,2}, Xi Wu^{1,2}, Xiang-Qing Din^{2,3}, Wei Huang^{1,2}, Zhi-Wen Wang⁶, Xiaokai Ma^{1,3}, Bobin Liu^{1,2}, Xiao-Xing Zou^{1,2}, Yves Van de Peer^{4,5,7,8,*}, Zhong-Jian Liu^{3,*}  and Shuang-Quan Zou^{1,2,3,*}

¹College of Forestry, Fujian Agriculture and Forestry University, Fuzhou 350002, China,

²Fujian Colleges and Universities Engineering Research Institute of Conservation and Utilization of Natural Bioresources, Fujian Agriculture and Forestry University, Fuzhou 350002, China,

³Key Laboratory of National Forestry and Grassland Administration for Orchid Conservation and Utilization at College of Landscape Architecture, College of Landscape Architecture, Fujian Agriculture and Forestry University, Fuzhou 350002, China,

⁴Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent 9052, Belgium,

⁵VIB Center for Plant Systems Biology, Ghent 9052, Belgium,

⁶PubBio-Tech, Wuhan 430070, China,

⁷Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria, South Africa, and

⁸Academy for Advanced Interdisciplinary Studies and College of Horticulture, Nanjing Agricultural University, Nanjing 210095, China

Received 14 June 2021; accepted 28 September 2021; published online 29 September 2021.

*For correspondence (e-mail: zjliu@fafu.edu.cn [Z.-J.L.]; yves.vandeppeer@psb.vib-ugent.be [Y.V.d.P.]; zou@fafu.edu [S.-Q.Z.]).

SUMMARY

Malvids is one of the largest clades of rosids, includes 58 families and exhibits remarkable morphological and ecological diversity. Here, we report a high-quality chromosome-level genome assembly for *Euscaphis japonica*, an early-diverging species within malvids. Genome-based phylogenetic analysis suggests that the unstable phylogenetic position of *E. japonica* may result from incomplete lineage sorting and hybridization event during the diversification of the ancestral population of malvids. *Euscaphis japonica* experienced two polyploidization events: the ancient whole genome triplication event shared with most eudicots (commonly known as the γ event) and a more recent whole genome duplication event, unique to *E. japonica*. By resequencing 101 samples from 11 populations, we speculate that the temperature has led to the differentiation of the evergreen and deciduous of *E. japonica* and the completely different population histories of these two groups. In total, 1012 candidate positively selected genes in the evergreen were detected, some of which are involved in flower and fruit development. We found that reddening and dehiscence of the *E. japonica* pericarp and long fruit-hanging time promoted the reproduction of *E. japonica* populations, and revealed the expression patterns of genes related to fruit reddening, dehiscence and abscission. The key genes involved in pentacyclic triterpene synthesis in *E. japonica* were identified, and different expression patterns of these genes may contribute to pentacyclic triterpene diversification. Our work sheds light on the evolution of *E. japonica* and malvids, particularly on the diversification of *E. japonica* and the genetic basis for their fruit dehiscence and abscission.

Keywords: *Euscaphis japonica*, malvids, genome, population history.

INTRODUCTION

Flowering plants (angiosperms) form by far the largest and most structurally and functionally diverse plant group on Earth. Among angiosperms, Mesangiospermae account for approximately 99.95% of extant species and include five lineages: eudicots, monocots, magnoliids, Chloranthales, and Ceratophyllales (Cantino et al., 2007). Within

Mesangiospermae, eudicots form the largest and most diverse clade, representing approximately 75% of angiosperm species (Yang et al., 2020). Superrosids and superasterids constitute the core eudicots, and among superrosids, the clade rosids mainly comprises fabids and malvids (The Angiosperm Phylogeny Group, 2016). Malvids comprise 58 families, exhibiting remarkable morphological and

ecological diversity (The Angiosperm Phylogeny Group, 2009, 2016). However, the phylogenetic position of malvids based on organelle or nuclear genes is still contested (Maia et al., 2014; Zhao et al., 2016). Although to date, the genomes of many malvid species have been sequenced, there is no species in the early-diverging group of malvids whose genome has been completely sequenced. The genome sequence of such species would provide valuable information for assessing the relationships of malvids.

Malvids consisting of Geraniales, Myrtales, Crossosomatales, Picramniales, Sapindales, Huerteales, Malvales and Brassicales, and Crossosomatales are a sister group of the other extant malvids (The Angiosperm Phylogeny Group, 2016). The family Staphyleaceae belongs to the order Crossosomatales, and comprises the genera *Staphylea*, *Turpinia*, and *Euscaphis*, with over 50 species of evergreen (ET) or deciduous (DT) trees and shrubs, mainly distributed in tropical and subtropical regions (Li et al., 2008; The Angiosperm Phylogeny Group, 2016). The monotypic genus *Euscaphis* includes only one species, *E. japonica*, which is found in the valleys of southeastern China, the semi-island of Korea, and Japan (Huang et al., 2015). *Euscaphis japonica* is widely planted as a street and ornamental tree because its fruits crack open into a butterfly shape when they mature, revealing the red endocarp and black seed (Liang et al., 2018, 2019; Yuan et al., 2018). Currently, natural populations of *E. japonica* are fragmented and threatened, owing to human interference, poor seed germination, and slow growth (Sun et al., 2019). Here, we present a high-quality genome sequence of *E. japonica*. As shown, the genomic information of *E. japonica* can help us understand and clarify the evolution of malvids. The availability of the *E. japonica* genome is essential for resolving fundamental questions regarding its diversification and for providing new insights into its evolutionary history, with important implications for future conservation efforts.

RESULTS

Genome assembly and annotation

Karyotype analysis showed that *E. japonica* contains 24 chromosomes ($2n = 2x = 24$) (Supplementary Note 1 and Figure S1). K-mer ($k = 17$) analysis indicated that its genome size is about 1.39 Gb, and genome has a heterozygosity of 0.5% (Figure S2). For *de novo* whole-genome sequencing of *E. japonica*, we obtained a total of 132.71 Gb of clean reads by using PacBio Technology (Table S1). The final assembled genome was 1.20 Gb, with a contig N50 value of 11.65 Mb (Table S1). The quality of the assembly was evaluated using Benchmarking Universal Single-Copy Orthologs (BUSCO) (Simao et al., 2015) showed that the completeness of the gene set of the assembled genome was 97.01%, and core eukaryotic genes mapping approach (CEGMA) analysis resulted in gene

completeness of 98.39% (Table S2). These results indicated that the *E. japonica* genome assembly was very successful, complete, and of high quality. To assemble the contigs into pseudochromosomes, a high-throughput chromosome conformation capture (Hi-C) library was constructed and sequenced, yielding 188.76 Gb of clean data (Table S1). We anchored 99.05% of the assembly genome (1.19 Gb) on to 12 pseudochromosomes with the aid of the Hi-C sequence data using the hierarchical clustering strategy (Figure 1a and Table S3). The length of the pseudochromosomes ranged from 67.41 to 146.22 Mb, with a scaffold N50 value of 98.65 Mb (Tables S1, S3). In the chromosome interval interaction heat map, the diagonal interaction intensity was high, whereas the interaction intensity outside the diagonal line was low, indicating that the Hi-C assembly quality of *E. japonica* was very high (Figure S3).

In total, 32 950 protein-coding genes in the *E. japonica* genome were predicted (Figure S4 and Table S4), of which 30 873 (93.67%) were predicted protein-coding genes that could be annotated using functional databases (Table S5). We also identified 349 microRNAs, 626 transfer RNAs, 759 ribosomal RNAs, and 3940 small nuclear RNAs (Table S6). In addition, the proteome was estimated to be 98.03% complete based on BUSCO (Table S7). The repetitive sequences accounted for 72.74% (872.82 Mb) of the genome of *E. japonica* (Table S8), which was higher than those in the completely sequenced genomes of *Eucalyptus grandis* (34%) (Myburg et al., 2014), *Dimocarpus longan* (52.87%) (Lin et al., 2017), and *Bombax ceiba* (60.30%) (Gao et al., 2018). Among the repetitive sequences of *E. japonica*, long terminal repeat (LTR) elements accounted for 43.28% (Copia: 15.69%; Gypsy: 19.50%) and LINES accounted for 1.89% (L2/CR1/Rex, 0.8%; L1/CIN4, 1.09%), which were much larger values than those of *D. longan* (LTR, 36.54%; LINES, 0.04%) (Lin et al., 2017). However, the proportion of SINEs in the repetitive sequences of *D. longan* (2.43%) was much greater than that in the genomes of *E. japonica* (<0.01%) (Lin et al., 2017). We analyzed the LTR insertion time of *E. japonica* and found that the amplification of LTR-RTs occurred largely between 0.1 and 0.25 Mya (Figure 1b). Hence, it suggests that LTRs contribute most to the expansion of the *E. japonica* genome.

Gene family evolutionary analysis

By comparing the genomes of 17 plant species, we found that 737 gene families and 2775 genes were unique to the *E. japonica* genome (see Experimental procedures; Table S9). The Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway can be enriched for these genes involved in "glutathione metabolism," "arachidonic acid metabolism," "cyanoamino acid metabolism," and "taurine and hypotaurine metabolism" (Table S10). In addition, analysis of gene families revealed that 10 species in malvids share 9786 gene families (Figure S5). Gene Ontology (GO) and

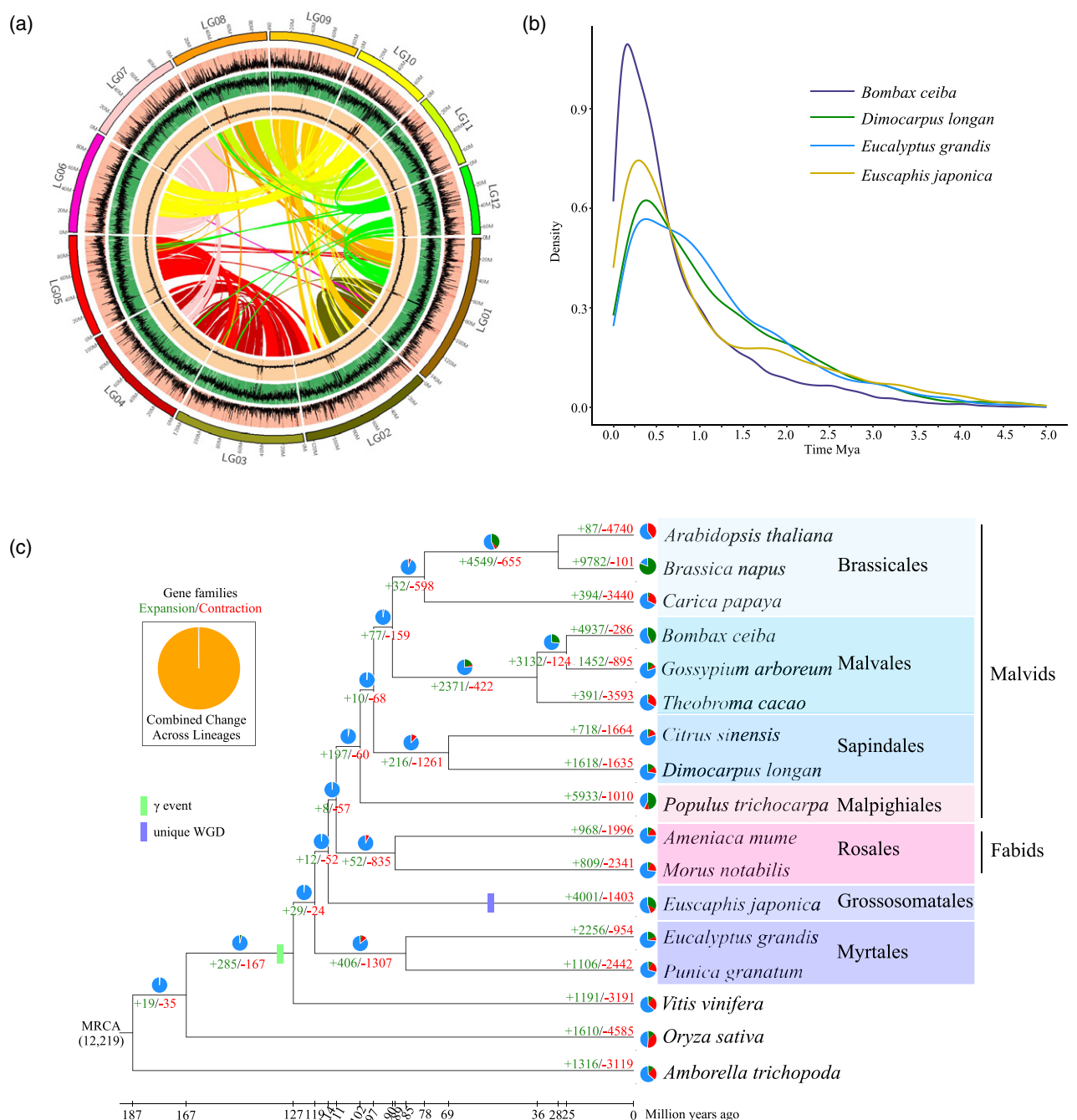


Figure 1. Comparative genomic analyses of *Euscaphis japonica* with other plants. (a) Genomic structure of *E. japonica*. Tracks from outside to inside are as follows: (i) 12 pseudochromosomes, (ii) the gene density, (iii) TE distribution, (iv) GC content, and (vi) synonymy of *E. japonica* genome. Circle figure was generated using Circos (<http://circoos.ca/>). (b) Insertion time distribution of the long terminal repeat in *E. japonica*, *Bombax ceiba*, and *Eucalyptus grandis*. (c) Bayesian tree showing divergence times and the evolution of gene family. Green and red numbers are the numbers of gene families that have expanded and contracted, respectively. In the pie chart, the blue portions represent the gene families with a constant copy number, and the orange portions represent the 12 219 gene families found in the most recent common ancestor (MRCA). WGD, whole genome duplication event.

KEGG enrichment analyses revealed that these gene families are particularly enriched in GO terms of “primary metabolic process,” “cellular metabolic process,” and “biosynthetic

process” and in the KEGG pathways of “metabolic pathways,” “plant hormone signal transduction,” and “biosynthesis of secondary metabolites” (Table S11).

Expansion and contraction analysis of orthologous gene families uncovered that a total of 4001 gene families in *E. japonica* expanded and 1403 gene families contracted (Figure 1c). GO and KEGG enrichment analyses found that the significantly expanded gene families are particularly enriched in GO terms of “catalytic activity,” “hydrolase activity,” and “threonine kinase activity” and in the KEGG pathways of “flavonoid biosynthesis” and “glutathione metabolism” (Table S12). GO enrichment analyses found that the significantly contracted gene families are particularly enriched in the GO terms of “ADP binding,” “anion binding,” and “heme binding” (Table S13). These expanding or contracting gene families may provide important clues to the evolution of the *E. japonica* genome.

Phylogeny of malvids

In APG IV, the rosids include malvids and fabids, and Myrtales belong to malvids and *Populus trichocarpa* belong to fabids (The Angiosperm Phylogeny Group, 2016). The phylogenetic tree constructed from chloroplast genomes from 17 species also supports this result and *E. japonica* belongs to malvids (see Experimental procedures; Figure S6). However, phylogenetic analysis based on genomes of rosids species supported Myrtales as a sister group to the rosids but not to the malvids clade within the eurosids, and *P. trichocarpa* was within malvids rather than in fabids (Argout et al., 2011; D’Hont et al., 2012; Myburg et al., 2014; Qin et al., 2017; Shulaev et al., 2011). We also constructed the Bayesian tree based on the single-copy genes derived from 17 species (see Experimental procedures; Table S9), and showed that the Myrtales clade form a sister group to the rosids, and *P. trichocarpa* is sister to malvids (Figure 1c). Interestingly, *E. japonica*, which represents the Crossosomatales clade formed a sister group with rosids. Phylogenetic analysis based on the chloroplast genome and single-copy genes showed that the phylogenetic position within rosids was unstable, suggesting the existence of ancient hybridization in rosids.

To determine further the relationship of *E. japonica*, fabids, and malvids, the concatenated and ASTRAL trees were constructed based on nucleotide and amino acid sequences, and their topological structures are the same as that of the Bayesian tree (Figure S7). However, the support rate of *E. japonica* and rosids order is weaker (Figure S7). Therefore, we used the Q value in ASTRAL to evaluate the discordance of gene trees in the single-copy gene data set. The branching order for *E. japonica*, fabids, and malvids displayed a high level of discordance among gene trees, with two nearly equally supported (and one slightly less supported) topologies in both nucleotide and amino acid sequence-based analyses (Figure S7b,d). Incomplete lineage sorting may play a role in confounding the resolution of diverging branches within angiosperms (Mirarab et al., 2014; Reichelt et al., 2021). Therefore, we

speculate that the incomplete lineage sorting during the diversification of the ancestral population may have led to the unstable phylogenetic position of *E. japonica*.

Whole-genome duplication

We build distributions of synonymous substitutions per synonymous sites (K_s) for *E. japonica* paralogs to infer more precisely the timing of polyploidization events that have occurred during the evolutionary past of *E. japonica*. Distributions of K_s for paralogs in the genome of *E. japonica* showed two clear peaks, one at $K_{s1} \approx 0.34$ and the other at $K_{s2} \approx 1.29$ (Figure S8), which indicates that the genome of *E. japonica* bears the traces of two polyploidization events. We further analyzed the K_s distribution of orthologs between *E. japonica* and *Vitis vinifera*. The results showed that the K_s differentiation peak of *E. japonica* – *V. vinifera* ($K_s \approx 0.74$) was between the two K_s peaks of *E. japonica* and was greater than that of *V. vinifera* ($K_s \approx 1.14$) (Figure 2a). This suggests that the ancestors of *E. japonica* and *V. vinifera* shared an ancient polyploidization event before their differentiation, that is, the polyploidization event corresponding to K_{s2} of *E. japonica* is likely the same polyploidization event as experienced by *V. vinifera*. The ancestor of *V. vinifera* only experienced the ancient polyploidization event shared by most eudicots, referred to as the γ event, which is an ancient whole genome triplication event (Wu et al., 2019). The collinearity dot plot diagram of *E. japonica* supports a γ event and a more recent whole genome duplication event (WGD) in *E. japonica* (Figure 2b). We estimated that the divergence time of *E. japonica* from other malvid species was 108 million years ago (Mya) (Figure 1c) and that the recent WGD event of *E. japonica* thus occurred approximately 53 Mya (Supplementary Note 2, Figure S9), suggesting that the recent WGD event of *E. japonica* is not shared with other malvid species. If the dating of the WGD is accurate and trustworthy, the timing of the recent WGD event might have coincided with a brief period of extreme global warming, referred to as the Paleocene-Eocene Thermal Maximum, which occurred approximately 55.5 Mya (Handley et al., 2011). Possibly, the WGD might have helped *E. japonica* to survive this high-temperature period. Our study unveiled two polyploidization events in *E. japonica*, the ancient whole genome triplication event shared with most eudicots (γ event), and the younger WGD unique to *E. japonica*.

Population genetic structure and demographic history

E. japonica, a monotypic species, exhibits high levels of morphological diversity at different altitudes, and its trees can be either DT or ET, which reflects their diverse genetic backgrounds (Li et al., 2008; Sun et al., 2019). We sequenced 101 samples from 11 natural populations with an average depth of 81.38% genome coverage, generating 214.94 Gb of raw data (Tables S14, Table S15). Using the

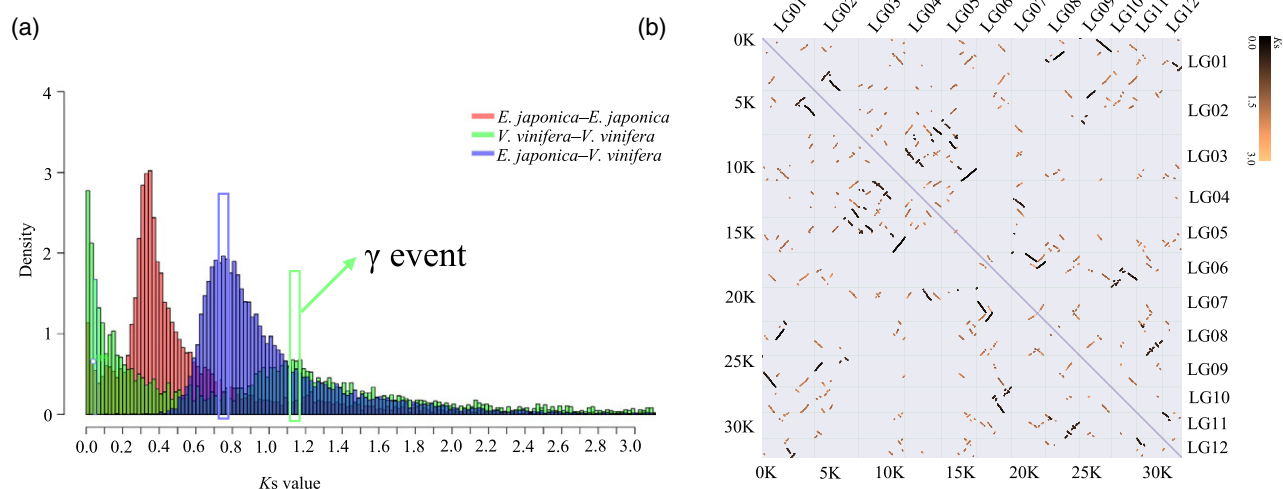


Figure 2. Whole-genome polyploidization event in *Euscaphis japonica*. (a) K_s distribution in *E. japonica* and *Vitis vinifera*. Peaks of intraspecies K_s distributions indicate ancient whole genome polyploidization events, and peaks of interspecies K_s distributions indicate speciation events. (b) Genome collinearity dot plot. Darker (black) collinear blocks represent the recent whole genome duplication event, and the lighter (orange) blocks represent the γ event. It can be clearly found that one dark collinearity block and four light collinearity blocks can be found in a region of the genome, which represent the recent whole genome duplication event and the γ event, respectively.

chromosome-level *E. japonica* as the reference genome, we obtained 13 254 963 single nucleotide polymorphisms (SNPs; Table S15). The neighbor-joining tree constructed based on SNP-calling showed that all samples were divided into DT and ET *E. japonica*, and the WYS and GYC populations are the existing DT and ET ancient populations, respectively (Figure 3a). ADMIXTURE analysis (Alexander et al., 2009) also confirmed the optimal classification of *E. japonica* populations ($K = 2$), and indicated that there was no genetic exchange between the DT and ET populations (Figure 3b, Figure S10). The fixation index (F_{ST}) values among DT and ET *E. japonica* populations fluctuated between 0.6996 and 0.7579 (Figure S11), indicating high genetic differentiation between DT and ET populations. In addition, the values of the Watterson's estimator (θ_w) and the average pairwise diversity within populations (θ_π) of the DT populations were higher than those of the ET populations (Table S16), indicating that the DT populations have rich genetic diversity. The values of Tajima's D in the DT populations (Tajima's $D > 1.3$) were also higher than those in the ET populations ($0 < \text{Tajima's } D < 1.0$) (Table S16). This is perhaps because of the population bottleneck effect, population structure, and unbalanced selection.

Multiple factors such as mating, selection, genetic drift, and effective population size affect the linkage disequilibrium (LD) patterns of natural populations (Ennis, 2007; Geng et al., 2021). We used `POPLDDECAY` to conduct LD analysis and found that LD of the DT populations decayed faster than the ET populations with increasing physical distance in DT populations, and the LD in DT populations were

lower than that of the ET populations (Figure 3c). Compared with self-pollinated species, cross-pollinated species have a faster LD decline because the latter has lower recombination efficiency (Campoy et al., 2016). Owing to the positive selection effect, the LD value of natural selection or domesticated populations will be higher (Suzuki, 2010). In addition, the low LD can be accounted for by the large effective population sizes (Neale and Savolainen, 2004). *Euscaphis japonica* is a long-lived plant with a mixed mating system. The distribution range of the DT is wider than that of the ET. Therefore, the low LD of the DT may be accounted for by the large effective population size, and the high LD of the ET may be accounted for by natural selection.

The natural populations of *E. japonica*, mainly distributed in Southeast Asia, are small and scattered. To determine whether their extinction is underway, we applied pairwise sequential Markovian coalescent analysis (Schiffels and Durbin, 2014) to assess demographic history (Figure 3d). A population bottleneck appeared in the ancestor of the *E. japonica* population at late Miocene (5–4 Mya), after which the ancestral population started to expand until the differentiation of the DT and ET populations in the late Pliocene (2.5–2.0 Mya). Since then, these two lineages have experienced completely different demographic histories. The increasingly colder and arid climate following the mid-Miocene Climatic Optimum not only led to the construction and fragmentation of numerous Tertiary relict trees in East Asia (Pound et al., 2012), but also promoted their speciation and lineage diversification (Zhao et al., 2019). In mid-Pliocene (3.3–3.0 Mya), the concentration of atmospheric

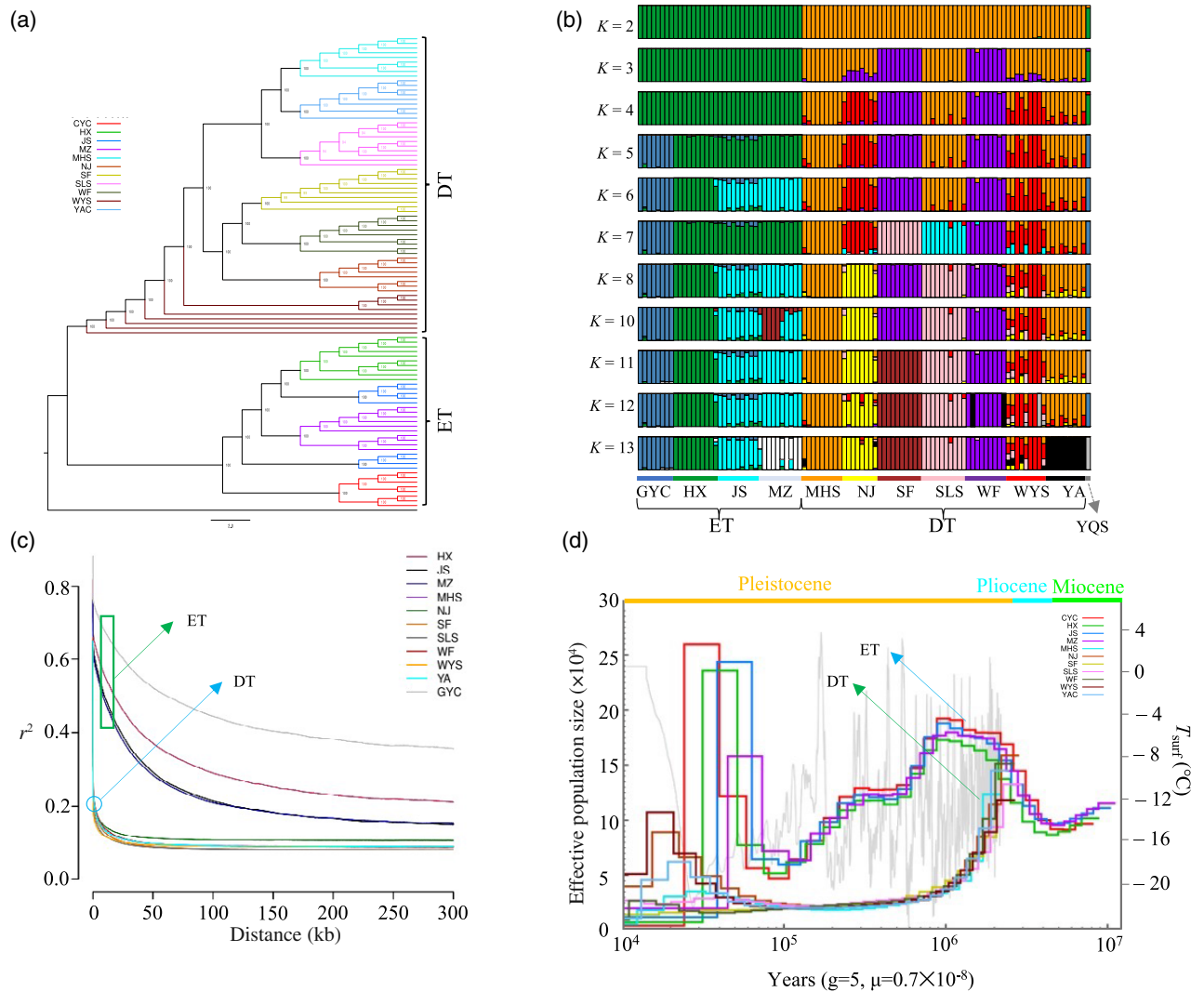


Figure 3. Population genetic structure and demographic history of *Euscaphis japonica* populations. (a) Neighbor-joining phylogenetic tree of all samples constructed using whole genome single nucleotide polymorphism data based on pairwise identity-by-state genetic distances. DT, deciduous; ET, evergreen. (b) Model-based population assignment by ADMIXTURE analysis for ancestral clusters ($K = 2-13$). The x-axis shows populations, and the y-axis quantifies the proportion of inferred ancestral lineages. When $K = 3$, the cross-validation error is the lowest (Figure S10), which means that the best grouping of *E. japonica* populations is three, which are the ET *E. japonica*, the DT *E. japonica* and an outer group *Tapiscia sinensis* (YQE). (c) Linkage disequilibrium patterns in different populations. x-axis: physical distances between two single nucleotide polymorphisms marked in kb; y-axis: r^2 used to measure linkage disequilibrium. (d) Effective physical size inferred by pairwise sequential Markovian coalescent analysis.

CO₂ was higher than or comparable with that in the present, and the climate was warmer (Burke et al., 2018; Fordham et al., 2020). The shrinkage and differentiation of *E. japonica* ancestor population may be related to the temperature change during this period.

After the ancestral DT population differentiated, its population declined sharply at early Pleistocene (2.0–1.0 Mya), and then remained stable until it rose 0.1 Mya years ago. The gradual colder climate may have caused the DT population size rapid reductions (Figure 3d). The ET population declined at about 1.0 Mya, the first bottleneck occurred at

0.7–0.4 Mya, and the second bottleneck occurred at about 0.15–0.1 Mya, at approximately the same time as the two largest Pleistocene glaciations in China, the Naynayxungla Glaciation (0.78–0.50 Mya), and the Penultimate Glaciation (0.30–0.13 Mya) (Zheng et al., 2002). The second population expansion of the DT and ET occurred after the Penultimate Glaciation retreated, and the DT and ET populations reached their pinnacle 30 000–10 000 and 80 000–30 000 years ago, respectively. The warm weather during the Greatest Lake Period (40 000–30 000 years ago) could have contributed to the population expansion (Hu and Wei, 2004).

Candidate positively selected genes

Selective sweep analysis combined with F_{ST} and $\theta\pi$ ratio has been an effective method to detect natural selection signals related to the living environment (Geng et al., 2021). Based on the high ET/DT p log ratio and the extreme divergence SNPs, we detected 1012 positively selected genes (PSGs) for the ET and 388 PSGs for the DT (Figure 4 and Table S17, Table S18). However, only the ET PSGs are significantly enriched GO terms, including, “phosphorelay sensor kinase activity,” “protein binding,” and “ATP binding” (Table S19). Long-term field observations have found that the ET has more flowers and fruits than the DT ones, and the fruit color is redder than that of the DT ones are. Interestingly, some genes involved in flower and fruit development were found among the ET PSGs (Table S17). Such as MIKC* (*Ej17442*) is involved in the development of pollen, and ANR (*Ej04020*) and AGL6 (*Ej30461*) are involved in the regulation of flowering time and flower development, respectively (Liu et al., 2013). NST1/3 (*Ej17793*) mediates fruit dehiscent by regulating secondary wall formation and lignification within the endocarp layers (Dardick and Callahan, 2014). The UDP glucose flavonoid 3-O-glucosyl transforms (*UFGT*, *Ej27609*) converts the colored anthocyanidins into anthocyanins, a more stable water-soluble plant pigment (Enoki et al., 2017). Highly expressed UFGT promotes the accumulation of anthocyanins and improves fruit coloring (Enoki et al., 2017).

Flower and fruit development

MADS-box genes and flower origin. The MADS-box gene family is mainly involved in flowering and flower development (Bai et al., 2019; Causier et al., 2002). The *E. japonica* inflorescence is a terminal panicle with small yellowish green flowers, five sepals, and five petals, which are 4–6 mm in diameter (Li et al., 2008). To study the flowering model of *E. japonica*, we identified 75 MADS-box genes in the *E. japonica* genome (Table S20), which were classified into Type I and Type II genes based on phylogenetic analysis (Figure S12). The Type II MADS genes of *E. japonica* were classified as MIKC* and MIKCC (Figure 5a). MIKC* genes retained a conserved role in the gametophyte during land plant evolution, because they equip the gametophyte with the physiological ability to cope with challenges inherent to life on land (Adamczyk and Fernandez, 2009; Kwantes et al., 2012; Liu et al., 2013; Verelst et al., 2007). The two MIKC* genes, *EjIMP.1* (*Ej15843*) and *EjIMP.2* (*Ej17442*) in *E. japonica* also perform the same function, because they were only expressed in the stamens (Figure 5b), thus enabling them to withstand environmental stresses and respond quickly under favorable conditions. Interestingly, the female gametes (immature seeds) develop in a closed carpel that are not directly exposed to the environment until the seeds mature, in which the MIKC* genes show no expression. The most well-known

MIKCC-type genes function as homeotic selector genes in the specification of floral organ identity (Chen et al., 2017). We identified 44 MIKCC-type genes, including homologs of the genes for the ABCDE model of floral organ identities, *A* (*FUL/AP*), *B* (*Bs/AP3/Pl*), and *SEP* subfamilies have been duplicated (Figure 5a). According to the ABCDE model of flower development, the four floral organ types in a typical flower are specified by five classes of floral organ identity genes: *A + E*, sepals; *A + B + E*, petals; *B + C + E*, stamens; and *C + E*, carpels; and *C + D + E*, ovules (Theissen et al., 2016). The expression analysis revealed that the flowering pattern of *E. japonica* was consistent with the ABCDE model of angiosperm flowering (Figure 5).

Fruit development

Plant evolution is largely driven by adaptations in seed protection and dispersal strategies that allow diversification into new niches; fruit color, fruit dehiscence, and fruit or seed abscission play important roles in seed development, protection, and dispersal (Dardick and Callahan, 2014; Estornell et al., 2013). As the fruit of *E. japonica* develops, the pericarps gradually turn red during August to September, and then crack to exposing mature black seeds (Figure S13). The seeds are tightly attached to the endocarp and begin to fall along with the pericarp during November, and it is not until March of the following year that all the fruits fall from the tree (Figure S13). We carried out field observations of three natural populations for a period of 1 year (Supplementary Note 3, Figure S14), and observed that the feeding of fruit by some birds and rodents promotes seed dispersal.

Fruit reddening. A previous study found that the change from green to red fruit color in *E. japonica* is caused by the accumulation of anthocyanins and the degradation of chlorophyll and carotenoids (Yuan et al., 2018). We identified 43 genes related to anthocyanin synthesis in *E. japonica* genome (Table S21), 16 of which were upregulated as the fruit matured (Figure S15). The expression level of *UFGT* (*Ej27609*), which converts anthocyan into anthocyanins, was the highest in the red pericarp stage (Fr_III) (Figure S15). The *TT19*-like gene (*Ej03746*) and *TT12*-like gene (*Ej13801*) are transport genes that carry anthocyanins from endoplasmic reticulum to the vacuole (Debeaujion et al., 2001; Kitamura et al., 2004; Zhao and Dixon, 2009), and their expression level is highest in the red pericarp stage (Fr_III), suggested that the accumulation of anthocyanins during the red pericarp stage is high. The four highly expressed regulatory genes, transparent testa glabra 1 (*TTG1*, *Ej15950*), transparent testa 8 (*TT8*, *Ej14142*), enhancer of glabra 3 (*EGL3*, *Ej16685*), and MYB domain protein 90 (*MYB90*, *Ej02336*), in *E. japonica* may regulate anthocyanin biosynthesis through the formation of the MYB-bHLH-WD40 protein complex

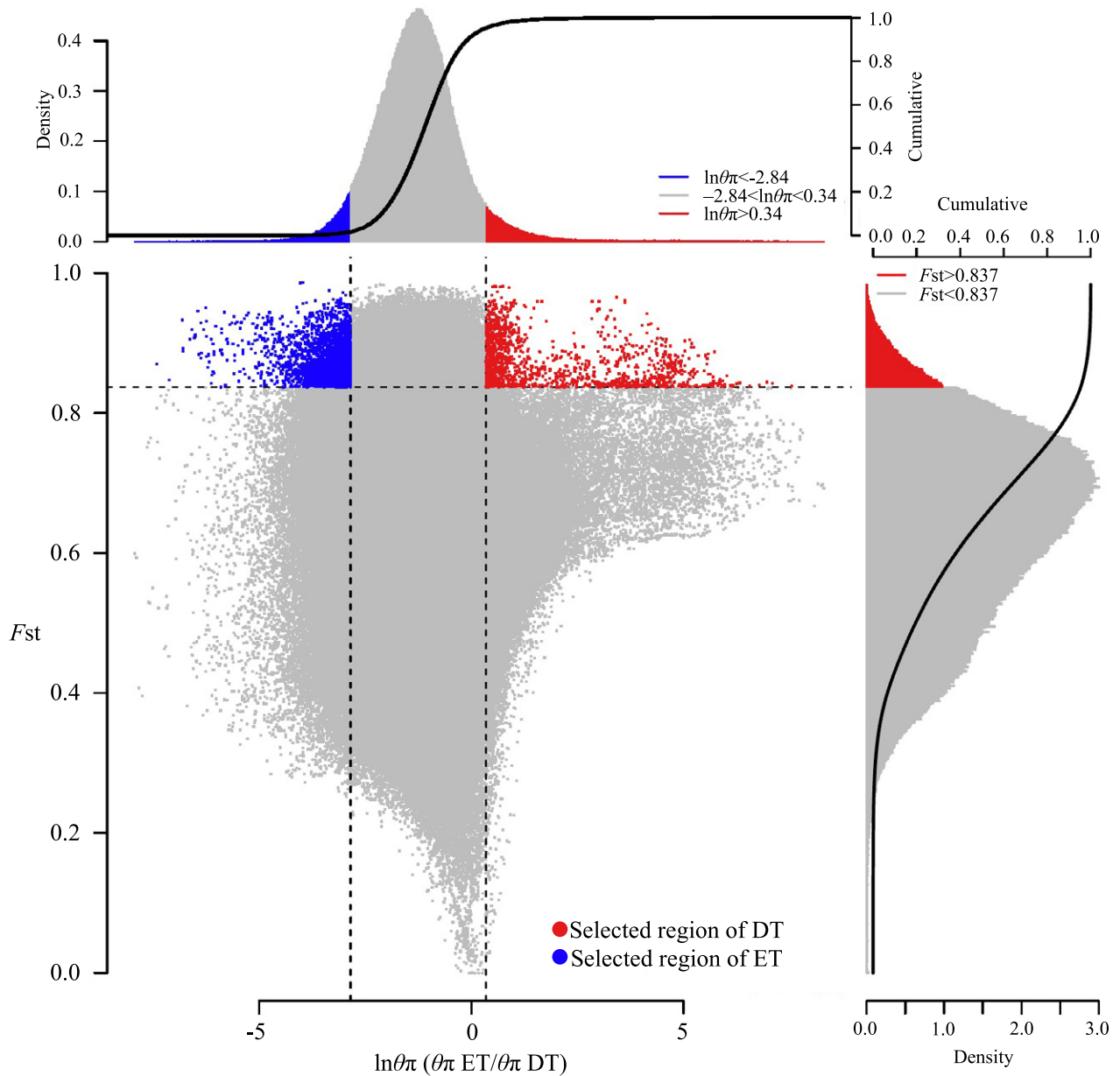


Figure 4. Distribution of the F_{st} and $\theta\pi$ values in the evergreen (ET) and deciduous (DT) *Euscaphis japonica*. Vertical and horizontal dashed lines correspond to the 3% right tails of the F_{st} and $\theta\pi$ value distribution, respectively.

(An et al., 2012; Baudry et al., 2004) (Figure S15). We also identified the genes related to carotenoid synthesis (16 members) and chlorophyll degradation (12 members) (Table S21), and found that most genes related to carotenoid synthesis were highly expressed in the green pericarp stage (Fr_I) and that most genes related to chlorophyll degradation were highly expressed in the fruit discoloration (Fr_II) and red pericarp stages (Fr_III) (Figure S16). These results indicated that the different expression patterns of these genes influenced the change in pericarp color of *E. japonica* to red.

Fruit dehiscence. The fruit structure of *E. japonica* is mainly composed of valves, valve margins, and a replum. The valves are fused by a valve margin and replum, and the valve margins delimit the borders between the valves and the replum and consist of a separation layer and a layer of lignified cells (Figure S17). We identified 29 genes related to fruit dehiscence, including *FRUITFULL* (*FUL*), *REPLUMLESS* (*RPL*), *SH1/2*, *INDEHISCENT* (*IND*), *GIBBERELLIN 3-OXIDASE 1* (*GA3OX1*), *ALCATRAZ* (*ALC*), and *DELLA* (Table S22). Among them, the copy number of the *DELLA* gene in *E. japonica* was higher than that in other

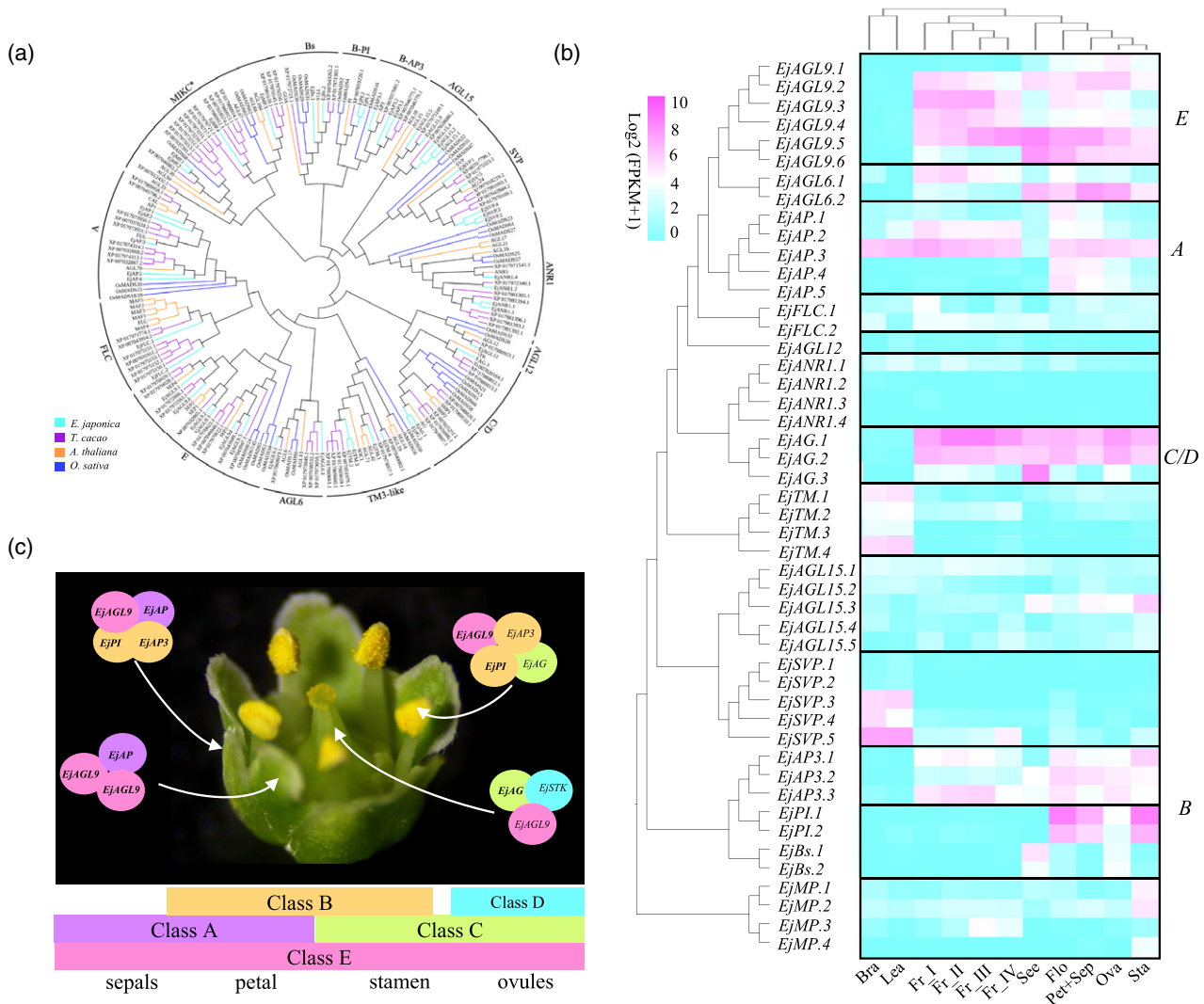


Figure 5. Flower model of *Euscaphis japonica*.

(a) Phylogenetic tree of MADS Type II genes from *E. japonica*, *Theobroma cacao*, *Arabidopsis thaliana*, and *Oryza sativa*.

(b) Expression profile of MADS Type II genes in reproductive organs and vegetative. In the mixture of petals and sepals, five A class genes and six E class genes are expressed, and two B-PI class and three B-AP3 class genes are expressed. In the ovary, six E class genes, two C class genes (*EjAG.1* and *EjAG.2*), and one D class gene (*EjAG.3*) are expressed. Five E class genes, two C class genes (*EjAG.1* and *EjAG.2*), two B-PI class, three B-AP3 class genes are expressed in the stamens. Expression patterns of these genes indicate that the flowering model of *E. japonica* conforms to the ABCDE model.

(c) Flower model of *E. japonica*.

malvid species was (Figure S18). *FUL* and *RPL* are specifically expressed in valves and replums, respectively, and inhibit the specific expression of *SHP1/2* in the valve margins (Ferrándiz, 2002; Roeder et al., 2003). In turn, the expression of *SHP1/2* promotes the expression of *IND* in the lignified layer and *ALC* in the separation layer, resulting in the formation of the lignification and separation layers (Girin et al., 2011; Kay et al., 2013; Ortiz-Ramírez et al., 2018). The rigidity of the lignified layer aids in the separation of the valves from the replum (Van Gelderen et al., 2016).

The expression analysis showed that the *SHP1/2*-like genes (*Ej24289* and *Ej06992*) were highly expressed during

fruit discoloration (Fr_II) and fruit dehiscence stages (Fr_III), but the expression levels of *IND*-like genes (*Ej28219*, *Ej23752*, and *Ej11988*) and *ALC*-like genes (*Ej23734*) were low (Figure S17). This was because the highly expressed *FUL*-like gene (*Ej24289*) inhibited the expression of *IND*-like genes, which in turn inhibited the expression of five downstream *GA3OXI*-like genes and five *NST1/3*-like genes (Figure S17). In addition, a highly expressed *DELLA*-like gene (*Ej17994*) also inhibited the expression of *ALC*-like genes. Among these genes that regulate fruit dehiscence, the *FUL*-like gene (*Ej24289*) may be the main factor that activates the metabolic pathways of fruit dehiscence and indirectly activates the metabolic pathways of lignin in the

valve margins (Figure S17). In total, 32 key genes of lignin synthesis that are highly expressed during pericarp dehiscence were identified (Figure S17 and Table S23). Cinnamyl alcohol dehydrogenase (*CAD*) catalyzes the last step in this pathway, and only genes belonging to the bona fide *CADs* reduces hydroxycinnamyl aldehydes to the corresponding alcohols (Guo et al., 2010). Our phylogenetic tree showed that three genes in *E. japonica* belonged to bona fide *CADs* (Figure S19), but only *Ej06344* was highly expressed during fruit dehiscence (Figure S17). These results suggest that these lignin synthesis-related genes, which are highly expressed in the fruit dehiscence stage, promote the synthesis of lignin in the lignified layer of the valve margins, resulting in fruit dehiscence due to the rigidity caused by the presence of lignin.

Fruit abscission. Ethylene plays an important role as a positive regulator of fruit ripening and abscission (Nakano et al., 2014). The key genes for ethylene synthesis, *S*-adenosine-*L*-methionine synthetase gene (*SAMS*) and 1-aminocyclopropane-1-carboxylic acid oxidase gene (*ACO*), were highly expressed in the young fruit (Fr_I) and fruit discoloration stages (Fr_II), but their expression was significantly decreased in the fruit maturity stage (Fr_III; red pericarp; fruit about to dehisce) (Figure S20 and Table S24). This might have been because sufficient ethylene is needed to promote fruit maturity during the young fruit stage, and the demand for ethylene decreases after the fruit matures. Surprisingly, the expression levels of *SAMS*-like and *ACO*-like genes increased significantly after 70 days of fruit ripening (Figure S20). The ethylene receptors (*ETRs*) act as negative regulatory elements that sense ethylene signals, inhibit downstream constitutive triple response 1 (*CTR1*), activate the positive regulatory factor ethylene insensitive 2 (*EIN2*) in the cytoplasm, and then transmit the signal to ethylene insensitive 3 (*EIN3*) in the nucleus, thus promoting the expression of ethylene responsive transcription factors (*ERF*) (Hall and Bleecker, 2003; Ju et al., 2012; Roberts et al., 2002). Some *ERF* genes have specific functions in the regulation of abscission zone development (Chen et al., 2015; Cui et al., 2016; Liao et al., 2016). During the fruit abscission period (Fr_IV), *CTR1*-like gene expression was low, and *EIN2*-like and *EIN3*-like gene expression was high (Figure S20). Because of the inhibition of EIN3-binding F-box protein 1/2 (*EBF1/2*) genes and the promotion of mitogen-activated protein kinase 3/6 (*MPK3/6*) genes, the expression pattern of *EIN3*-like genes differed among the four different periods of pericarp development (Figure S20). In addition, 10 *ERF* genes were highly expressed in the fruit abscission stage, which may promote fruit abscission (Figure S20). In summary, our results show that ethylene may promote fruit maturity in the early stages of fruit development as well as fruit abscission in the later stages of fruit development.

Pentacyclic triterpene synthesis

Pentacyclic triterpenes are mainly divided into oleanane, ursane, lupane, and friedelane, which have anti-inflammatory, antitumour, antihyperlipidemic, anti-ulcer, and antimicrobial properties (Liu, 1995, 2005). These compounds are involved in the normal growth and development process of plants, particularly as chemical barriers in the plant defense mechanism against insects, fungi, nematodes, and weeds (Murata et al., 2008; Wang et al., 2010). The branches, pericarp, and leaves of *E. japonica* were found to be rich in triterpenoids, among which the main type was pentacyclic triterpenoids (Supplementary Note 4, Figure S21 and Table S25). Pentacyclic triterpenoids are synthesized via the mevalonic acid (MVA) and methylerythritol 4-phosphate pathways by the cyclization of 2,3-oxidosqualene, thereby producing triterpenoid skeletons (Kim et al., 2020). We identified 78 genes related to pentacyclic triterpenoid biosynthesis pathways in the *E. japonica* genome, including the MVA and methylerythritol 4-phosphate pathways (Tables S26, S27). Notably, the expansion of 3-hydroxy-3-methylglutaryl-CoA reductase (*HMGGR*) that functions upstream of the MVA was identified in *E. japonica* (15 members) (Figure S22 and Table S27). Overexpression of this gene can increase the production of triterpenoids in plants (Darabi et al., 2012; Harker et al., 2010; Hey et al., 2006). Despite the considerable level of gene duplication in *E. japonica*, only two *EjHMGGR.1* and *EjHMGGR.2* were highly expressed in branches, leaves, and fruits (Figure 6). *SQS* and *SQE* genes are key genes involved in the synthesis of the triterpenoid precursor 2,3-oxidosqualene (Rasbery et al., 2007; Vishwakarma, 2015). Seven *SQS*-like and seven *SES*-like genes in the *E. japonica* genome were identified, among which *EjSQS1*, *EjSQS2*, and *EjSQE1* were highly expressed in branches, pericarps, and leaves (Figure 6, Table S27).

Plants have a large OSC gene family, which includes genes such as lupeol synthase (*LUS*), β -amyrin synthase (*bAS*), dammarenediol synthase (*DDS*), and cycloartenol synthase (*CAS*), which can catalyze 2,3-oxidosqualene to form more than 20 different pentacyclic or tetracyclic triterpenoids (Thimmappa et al., 2014). *bAS* and *LUS* are precursors for the synthesis of various pentacyclic triterpenes, and *CAS* is the precursor of tetracyclic sterols (Delis et al., 2011; Liu et al., 2009). We identified 11 genes from the OSC gene family, including six *bAS*, four *CAS*, and one *LUS* (Table S27). Except that *EjbAS6* was not expressed, *EjbAS1* and *EjLUS* were highly expressed in branches, leaves, and pericarps, whereas *EjCAS1*, *EjCAS2*, *EjCAS3*, and *EjCAS4* were more expressed in seeds than in leaves and pericarps (Figure 6).

The triterpenoid skeletons are modified by various cytochrome P450, dehydrogenases, reductases, and other modification enzymes (Huang et al., 2012). The members of CYP716 subfamily from P450 gene family were found to participate in pentacyclic triterpene scaffold modification

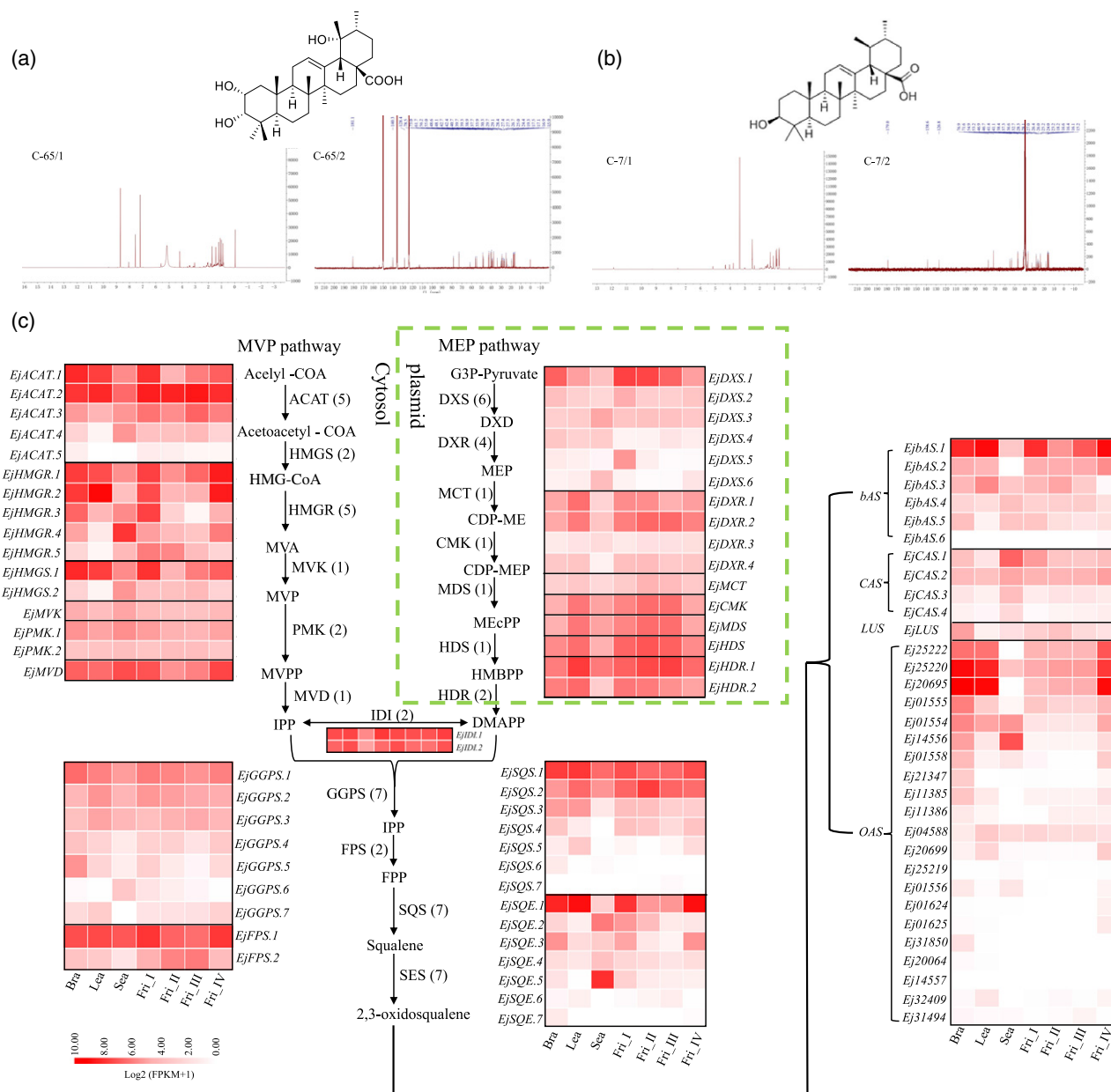


Figure 6. Pentacyclic triterpene biosynthesis in *Euscaphis japonica*.

(a,b) Two pentacyclic triterpenes that are abundant in *E. japonica*. (a) Euscaphic acid hydrogen spectrum (left) and carbon spectrum (right). (b) 19a-hydroxyursolic acid hydrogen spectrum (left) and carbon spectrum (right).

(c) Expansion expression patterns of pentacyclic triterpene biosynthesis-related genes in *E. japonica*. MEP, methylerythritol 4-phosphate; MVP, mevalonic acid. (See the full name of genes in Table S26).

(Ghosh, 2017). We constructed a phylogenetic tree of the *E. japonica* P450 gene family and found a large number of the oleanolic acid synthase gene (*OAS*) belonging to the CYP716A subfamily in the *E. japonica* genome (21 members) (Figure S23). The *OAS* gene is specialized in C-28 oxidation of the β -amyrin skeleton. Most of the *OAS*-like genes of *E. japonica* were highly expressed in branches, leaves, and fruits, and their expression patterns were similar to those of *bAS*-like genes (Figure 6). In conclusion, the

expansion and different expression patterns of genes related to the synthesis of pentacyclic triterpenes in the *E. japonica* genome might be contributing to pentacyclic triterpene diversification.

CONCLUSION

We assembled a high-quality chromosome level genome of *E. japonica*, an early-diverging species within malvids. The phylogenetic trees constructed based on complete

plastid genomes and single-copy genes showed that the phylogenetic position within malvids branches was inconsistent, implying the hybridization event of the malvids ancestor population. Although the topological structures of concatenated and ASTRAL trees were consistent, the support rates of *E. japonica* and malvids order are weaker, indicating that LTR may be the reason for the inconsistent position of *E. japonica* phylogeny. WGD analysis indicates that *E. japonica* experienced a unique WGD event approximately 53 Mya. Based on 13 254 963 SNPs, it was revealed that the ancestors of *E. japonica* differentiated into the ET and DT populations at 2–2.5 Mya, and temperature is the main reason for the completely different population histories of these two groups. We identified PSGs in the ET and DT, and found some genes related to flower and fruit development in ET PSGs. The redness, dehiscence, and abscission of *E. japonica* fruits promoted the spread of seeds, which is conducive to population reproduction. Therefore, we further analyzed the molecular regulation mechanism of flower and fruit development. In addition, *E. japonica* is rich in pentacyclic triterpenes. We analyzed the key genes involved in pentacyclic triterpene synthesis, and found that different expression patterns of these genes and the expansion of the terpene synthesis gene *HMGR* and the pentacyclic triterpene modified gene *OAS* may contribute to pentacyclic triterpene diversification of *E. japonica*. In conclusion, the present study has important scientific significance because it describes the phylogeny of the malvids, how the ancestor *E. japonica* differentiates into the ET and DT, and the evolutionary significance of the fruit.

EXPERIMENTAL PROCEDURES

Sample preparation and sequencing

All plant materials used for genome and transcriptome sequencing were collected from a wild mature *E. japonica* ET tree growing in Gaoyang Village, Quanzhou City, Fujian Province, China (27° 54' N, 117° 10' E). The Hi-C sequencing material was obtained from a two-year-old seedling cultivated from the seed of this tree.

For genome sequencing, a modified cetyltrimethylammonium bromide method was used to extract total genomic DNA from young leaves. We adopted the whole genome shotgun strategy and constructed a DNA library with 400 bp inserts. All obtained libraries were sequenced on Illumina HiSeq X ten platform. Additionally, SMRTbell template preparation involved DNA concentration, damage repair, end repair, ligation of hairpin adapters, and template purification, and these steps were carried out using AMPure PB Magnetic Beads (Pacific Biosciences, Menlo Park, CA, USA). We carried out 20 kb single-molecule real-time DNA sequencing using PacBio to sequence a DNA library on the PacBio Sequel platform. For Hi-C sequencing, a Hi-C library was prepared by Santa Cruz (Dovetail Genomics, Scotts Valley, CA, USA) and sequenced on an Illumina NovaSeq. For transcriptome sequencing, total RNA from the leaves, branches, flowers, fruits, and floral organs (stamens, ovaries, petals, and sepals) was extracted and purified. All transcriptome libraries were constructed using the

Illumina TruSeq library Stranded mRNA Prep Kit and sequenced on an Illumina HiSeq 2000 platform.

Genome assembly

We first evaluated the genome size of *E. japonica* based on *k*-mer analysis to determine the amount of sequencing data required to assemble the genome (Lander et al., 2001). The *E. japonica* genome was assembled based on the obtained Illumina data and Hi-C data. We filtered and removed organellar DNAs, reads of poor quality or short length, and chimeras in the raw data obtained by Illumina sequencing. FALCON assembler (<https://github.com/PacificBiosciences/FALCON/>) was used for self-correction and Smartdenovo assembler (<https://github.com/ruanjue/smartdenovo>) was used to preliminarily assemble. The draft assembly was polished using Arrow (<https://github.com/PacificBiosciences/GenomicConsensus>). To increase the consensus accuracy of the assembly, Illumina short reads were recruited for further polishing using Pilon v1.22 (<https://github.com/nanoporetech/ont-assembly-polish>) (Walker et al., 2014). The polished sequences were aligned with the NT library to remove the aligned bacterial and fungal sequences. Core Eukaryotic Genes Mapping Approach (CEGMA) (<http://korflab.ucdavis.edu/dataset/cegma/>) was used to determine the genome assembly. We also evaluated the integrity of the assembled genome by using BUSCO v3 (<https://busco.ezlab.org>) (Simao et al., 2015).

In addition, we used FASTQ (Chen et al., 2018a) to filter the raw data obtained by Hi-C sequencing and obtain high-quality clean reads, which were then mapped to the draft scaffolds using a fast and accurate short-read alignment with a Burrows–Wheeler transform (Li and Durbin, 2009), and the duplicated mapped reads and unmapped reads were removed using SAMtools (<https://github.com/samtools>) (Li et al., 2009). The separation of Hi-C read pairs mapped in draft scaffolds was analyzed using chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions (Burton et al., 2013) to produce a likelihood model for genomic distance between read pairs, and the model was used to identify putative misjoins. The greater the number of reads of interaction between two contigs, the greater the likelihood of a class. Contig clustering was performed according to the number of interaction reads. Then, Hi-C data were used to perform scaffolding using LACHESIS software, and finally, approximately 99.95% of sequences were grouped into 12 super scaffolds. The contigs were sorted according to the intensity of every two contig interactions and the mapping location of interaction reads.

Gene predictions and functional annotation

Prediction methods based on homology, *de novo*, and transcriptome data were used to predict the protein-coding genes in *E. japonica* genome. Homologous proteins from seven known whole genome sequences of *Amborella trichopoda*, *Arabidopsis thaliana*, *Brassica napus*, *D. longan*, *Gossypium arboreum*, *Oryza sativa*, and *V. vinifera* were downloaded from Phytozome 12 (<https://phytozome.jgi.doe.gov/pz/portal.html>). These homologous proteins were aligned to the *E. japonica* genome sequence using GEMOMA v1.3.1 (Keilwagen et al., 2016). AUGUSTUS v3.1 (<http://bioinf.uni-greifswald.de/augustus/>) (Stanke et al., 2006) with default parameters was applied for the *de novo* gene prediction. PASA v2.0.2 (Haas et al., 2003) was used to predict the unigenes sequence of the transcriptome assembly. The predicted gene structures from the above three methods were merged into a non-redundant gene model using EVM (Haas et al., 2008). The annotated results were further filtered using TransposonPSI (<http://transposonpsi.sourceforge.net>) to obtain the final gene structure

information. In addition, the completeness of the annotated genome was also assessed using BUSCO v3 (<https://busco.ezlab.org>) (Simao et al., 2015).

Gene functional annotation

The annotation results were aligned to five function databases by using BLAST v2.2.31 (Altschul et al., 1990) to obtain gene function information of *E. japonica*. The five protein databases include Swiss-Prot (<http://www.uniprot.org>) (Boeckmann et al., 2003), KEGG (<http://www.genome.jp/kegg/>) (Kanehisa and Goto, 2000), non-redundant protein sequence database (NR), Clusters of Orthologous Groups for eukaryotic complete genomes (KOG) (Koonin et al., 2004), and GO resource (Ashburner et al., 2000).

Identification of non-coding RNAs

Non-coding RNAs were further identified. Ribosomal RNAs were identified by aligning the ribosomal RNA template sequences from the Rfam database against the *E. japonica* genome using the BLASTN algorithm. Transfer RNAs (tRNA) were predicted using tRNAscan-SE 1.3.1 (<http://lowelab.ucsc.edu/tRNAscan-SE/>) (Lowe and Eddy, 1997). Other non-coding RNAs, including microRNAs and small nuclear RNAs, were predicted using INFERNAL (<http://infernal.janelia.org/>) (Nawrocki et al., 2009) to search the Rfam database (<http://infernal.janelia.org/>) (Griffiths-Jones et al., 2005).

Identification of repetitive elements

A repeat library of *E. japonica* was constructed using RepeatMasker v4.0.7 (www.repeatmasker.org), LTR_FINDER v1.06 (http://tlife.fudan.edu.cn/ltr_finder/) (Xu and Wang, 2007), and MITE-Hunter (Han and Wessler, 2010) based on the specific structure of repeated sequence. Then the database was merged with the Repbase database (Jurka et al., 2005) to form a final repeat sequence database. The tandem repeats across the *E. japonica* genome were predicted using Tandem Repeats Finder v4.09 (Gary, 1999) (<http://tandem.bu.edu/trf/trf.html>) based on the final database. Repeat sequences with $\geq 50\%$ identities were grouped into the same class. In addition, simple sequence repeat sequences were identified using MISA (<http://pgrc.ipk-gatersleben.de/misa/>) (Beier et al., 2017), and the results showed that there are 875 729 simple sequence repeat sequences in the *E. japonica* genome (Table S28).

Ortholog detection with OrthoMCL

The amino acid and nucleotide sequences of *E. japonica* and other 16 representative plant species were downloaded, including one basal angiosperm (*A. trichopoda*), one monocots (*O. sativa*), one rosids branch early differentiation plant (*V. vinifera*), and 11 rosids clade other plants (*A. thaliana*, *P. trichocarpa*, *Armeniaca mume*, *B. napus*, *D. longan*, *E. grandis*, *G. arboretum*, *Morus notabilis*, *B. ceiba*, *Carica papaya*, *Punica granatum*, *Theobroma cacao*, and *Citrus sinensis*). We constructed the protein datasets of these genomes and then used BLASTP (E-value of $1E-5$, similarity threshold of 30%, and coverage threshold of 50%) to align the protein datasets (Kent, 2002). OrthoMCL v1.4 (<http://orthomcl.org/orthomcl/>) (Li et al., 2003) was used to construct the orthologous groups. GO and KEGG enrichment analysis was performed on the identification of *E. japonica*-specific gene families. In addition, we conducted an enrichment analysis on the gene families shared by 10 malvids.

Genome evolution analysis

We filtered out the orthologous groups whose protein length was <200 bp to obtain a reliable single-copy gene family to construct

the phylogenetic tree. MUSCLE v3.8.31 (<http://www.drive5.com/muscle/>) (Edgar, 2004) was used to align the amino acid sequences of single-copy orthologous groups. We connected the nucleotide sequences of the single-copy orthologous groups into a supergene, which was used to construct a Bayesian phylogenetic tree. We further used RAxML with 500 bootstrap replicates to combine the single-copy orthologous set and constructed concatenated and ASTRAL phylogenetic trees based on nucleotide and amino acid sequences alignment, respectively. In addition, a phylogenetic tree was conducted based on complete plastid genomes of >17 species. These plastid genome data were downloaded from NCBI GenBank. The sequences were aligned by MAFFT v7.307 with 1000 bootstrap replicates (Katoh and Standley, 2013), and the phylogenetic tree constructed by RAxML (Stamatakis, 2014).

In Bayesian phylogenetic analysis, the divergence times and gene family expansion and contraction in each tree node were inferred using the MCMCTREE program (<http://abacus.gene.ucl.ac.uk/software/paml.html>) of the PAML package v4.7 (Yang, 2007) and CAFÉ 4.2 (<http://sourceforge.net/projects/cafehahnlab/>) (De Bie et al., 2006), respectively. Published *A. thaliana*-*A. trichopoda* (173–199 Mya), *A. thaliana*-*O. sativa* (115–308 Mya), *A. thaliana*-*B. ceiba* (84–95 Mya), and *A. thaliana*-*P. trichocarpa* (98–117 Mya) were used to calibrate divergence times. Functional enrichment analysis was performed on the unique gene families and the gene families that significantly contracted and expanded in the malvids and *E. japonica*.

Collinearity analysis and polyploidization event analysis

JCVI v0.9.14 (<https://pypi.org/project/jcvi/>) was used to analyze the protein sequences of *E. japonica* and *V. vinifera* to obtain the gene pairs in their collinear regions. The gene pairs in the collinear regions of *E. japonica*, *V. vinifera*, and *E. japonica*-*V. vinifera* were obtained, respectively. To estimate polyploidization event events, the distribution of synonymous substitutions per synonymous sites (K_s) in the genomes of *E. japonica* and *V. vinifera* were identified, respectively. The protein sequences of their genomes were self-aligned using Diamond and the mutual optimal alignment in the alignment results were extracted. The K_s values in the genome of *E. japonica* and *V. vinifera* were calculated using COMEML in the PAML package v4.4c (Yang, 2007). In addition, we calculated the time of WGD event of *E. japonica* based on the rate of K_s (λ) and the peak value in K_s differentiation; in brief, $\lambda = K_s/\text{time}$ (Supplementary Note 2).

Transcriptomic assembly and expression analysis

The raw reads obtained by Illumina HiSeq sequencing was further filtered using in-house Perl scripts to obtain clean reads. Clean reads that met the quality control requirements were assembled using Trinity (Grabherr et al., 2011), with the *E. japonica* genome as the reference genome. TransDecoder (<http://transdecoder.github.io>) was used to predict the protein sequences and coding sequences, and gene expression levels were estimated by using TopHat (Trapnell et al., 2012).

Gene family analysis

The HMM profiles of MADS (PF00319), P450 (PF00067), and AP2/ERF (PF00847) were obtained from Pfam (<http://pfam.xfam.org/>). The MADS-box, P450, and AP2/ERF candidate gene proteins were separately searched using HMMER 3.2.1 (with default parameters) and InterProScan (Zdobnov and Apweiler, 2001). The predicted genes were manually inspected, and their domains identified by SMART (Letunic et al., 2015). The candidate MADS-box, P450, and

AP2/ERF gene families were aligned using MEGA7 (Kumar et al., 2016), respectively. The phylogenetic trees were constructed on the CIPRES website (<https://www.phylo.org/portal2/>), and edited in iTOL (<https://itol.embl.de>).

Identification of fruit development-related and pentacyclic triterpene synthesis-related genes

We downloaded the fruit dehiscence, fruit abscission, anthocyanin synthesis, chlorophyll degradation, carotenoid synthesis, lignin synthesis, and pentacyclic triterpene synthesis-related genes of *Arabidopsis* from TAIR (<https://www.arabidopsis.org>) as the query sequences. TBLASTN (NCBI Blast v. 2.2.23) (Kent, 2002) was used to align the query sequences against *E. japonica* genome sequence for obtain candidate genes. Subsequently, the domains of candidate sequences were identified using SMART (<http://smart.embl-heidelberg.de/>) (Letunic et al., 2015), and putative shade candidate genes were further predicted using the NCBI BLAST (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

Demographic history

In total, 101 individuals from 11 natural *E. japonica* populations were selected to represent most of the known localities of *E. japonica* in China (Table S24). Genomic DNA extraction, library construction, and amplification followed standard protocols. All samples were sequenced using BGISEQ-500, and the raw reads were filtered using SOAPnuke v 1.5.4 (Chen et al., 2018b). The clean reads from all samples were aligned to *E. japonica* genome using BWA-MEM (<https://sourceforge.net/projects/bio-bwa/>). The obtained BAM files were sorted, and the mpileup + call in Bcftools v 1.10.2 (<https://github.com/samtools/bcftools>) were used to detect SNPs in a single sample. To obtain SNP calling, merge in Bcftools was used to obtain the vcf file of the SNP of the populations, which was filtered using VCFtools.

PLINK v 1.9 (Chang et al., 2015) was used to construct a bed file and a distance matrix for population structure analysis and phylogenetic tree construction. A neighbor-joining phylogenetic tree was constructed using MEGA v 7.0 (Kumar et al., 2016) based on the distance matrix. ADMIXTURE v 1.3.0 (Alexander et al., 2009) was used to infer the population structure and mixing of all samples through the input bed file; this was repeated five times, and cross-validation was $K = 2-12$.

We calculated the level of nucleotide diversity in each population, including θ_π and θ_w , the F_{ST} index to determine regions of differentiation between populations, and Tajima's D to evaluate any deviations from neutral evolution. These indexes were evaluated with VCFtools v 0.1.13. The LD was calculated using POPLDDECAY (Zhang et al., 2018). The history of population size was inferred by pairwise sequential Markovian coalescent analysis (Li and Durbin, 2011).

A combined method of F_{ST} and nucleotide heterozygosity θ_π was used to identify regions that were likely to be or have been under selection. We conducted enrichment analysis of candidate genes between DT and ET populations.

The R package ClusterProfiler performed GO enrichment (Yu et al., 2012), and KOBAS and BlastKOALA (Kanehisa and Goto, 2000; Xie et al., 2011) performed KEGG enrichment. The resulting *P* values were corrected for multiple comparisons using the method of Benjamini and Hochberg.

ACKNOWLEDGEMENTS

This research was jointly funded by Fund for Excellent Doctoral Dissertation of Fujian Agriculture and Forestry University, China

(no. 324-1122yb062), awarded to W.-H.S.; Fujian Provincial Department of Science *E. japonica* Evolution and Selection of Ornamental Medicinal Resources, China (no. 2020N5004), the Project of Forestry Peak Discipline at Fujian Agriculture and Forestry University, China (no. 712018007), and the Collection, Development and Utilization of *Euscaphis konlshli* Germplasm Resources, China (no. KSYLC004), awarded to S.-Q.Z.; YVdP acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (no. 833522) and from Ghent University (Methusalem funding, BOF.-MET.2021.0005.01).

AUTHOR CONTRIBUTIONS

S-QZ, Z-JL, and W-HS managed the project; S-QZ, Z-JL, YVdP, and W-HS planned and coordinated the project; S-QZ, Z-JL, YVdP, and W-HS wrote the manuscript; W-HS, X-KM, X-YL, X-DL, Y-TJ, P-LZ, Y-FW, XW, and X-QD collected plant material for genome, transcriptome, and population sequencing; W-HS, Z-WW, D-YZ, D-QC, and M-YQ performed the genome sequencing, assembly, annotation; W-HS, YVdP, Z-WW, and Y-XY participated in analyses of genome comparison and genome evolution; W-HS, ZL, D-YZ, Q-GZ, LX, Y-TJ, HL, and Y-FW contributed to the RNA-seq and corresponding analysis; W-HS, Q-XL, B-BL, X-XZ, and D-QC conducted flower and fruit of *E. japonica* analysis; LN, W-HS ZL, WH, and Q-XL performed the pentacyclic triterpene synthesis; W-HS, WH, Z-WW, D-YZ, and M-YQ were involved in the analyses of population history.

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

DATA AVAILABILITY STATEMENT

All sequences described in this manuscript have been submitted to the National Genomics Data Center (NGDC). The raw whole-genome data of *E. japonica* have been deposited in BioProject/GSA (<https://bigd.big.ac.cn/gsa>) under the accession codes PRJCA005268/CRA004271, and the assembly and annotation data have been deposited at BioProject/GWH (<https://bigd.big.ac.cn/gwh>) under the accession codes PRJCA005268/GWHBCHS00000000. The raw transcriptomes data of *E. japonica* have been deposited in BioProject/GSA (<https://bigd.big.ac.cn/gsa>) under the accession codes PRJCA005298/CRA004272.

SUPPORTING INFORMATION

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

Supplementary Note 1. Chromosome number assessment.

Supplementary Note 2. Whole-genome duplication identification and dating.

Supplementary Note 3. Observation of *E. japonica* seed dispersal.

Supplementary Note 4. Determination of pentacyclic triterpene substances.

Figure S1. Cytogenetic analysis of *E. japonica*.

Figure S2. Genome size and heterozygosity of *E. japonica* estimation using 17 *k*-mer distribution.

Figure S3. Interchromosomal of Hi-C chromosome contact map of *E. japonica* genome.

Figure S4. Gene structure prediction results of *E. japonica* and other species.

Figure S5. Venn diagram shows gene families of malvids.

Figure S6. Phylogenetic tree constructed by chloroplast genomes from 17 species.

Figure S7. Concatenated- and ASTRAL-based phylogenetic trees.

Figure S8. *Ks* distribution in *E. japonica*.

Figure S9. Distributions of synonymous substitutions per synonymous site (*Ks*) of one-to-one orthologs identified between *E. japonica* and *P. trichocarpa* and *V. vinifera*.

Figure S10. Population structure plot.

Figure S11. Fixation index (*F_{ST}*) heat map among *E. japonica* populations.

Figure S12. Phylogenetic analysis of MADS-box genes from *O. sativa*, *A. thaliana*, *E. japonica*, and *T. cacao*.

Figure S13. Observation the fruit development.

Figure S14. Animal seed dispersal.

Figure S15. Anthocyanin biosynthesis in *E. japonica* fruits.

Figure S16. Carotenoid accumulation and the chlorophyll degradation in *E. japonica* fruits.

Figure S17. Expression profile of fruit dehiscence-related genes.

Figure S18. Phylogenetic tree of *DELLA* genes obtained from six malvids species.

Figure S19. Phylogenetic tree of *CAD* genes obtained from seven malvids species.

Figure S20. Expression pattern of fruit abscission-related genes.

Figure S21. Structure of pentacyclic triterpene compounds separated from *Euscaphis*.

Figure S22. Phylogenetic tree of *HMGR* gene in plants.

Figure S23. Phylogenetic tree of P450s gene family obtained from *A. thaliana* and *E. japonica*.

Table S1. Assembled statistics of *E. japonica* genome.

Table S2. Evaluation of *E. japonica* genome assembly.

Table S3. Chromosome length of *E. japonica*.

Table S4. Prediction of gene structures of the *E. japonica* genome.

Table S5. Statistics on the function annotation of the *E. japonica* genome.

Table S6. Non-coding RNA annotation results of *E. japonica* genome.

Table S7. BUSCO assessment of the *E. japonica* annotated genome.

Table S8. Statistic of repeat sequence in *E. japonica* genome.

Table S9. Gene-clustering statistics for 17 species.

Table S10. KEGG enrichment result of unique genes families of *E. japonica*.

Table S11. Gene Ontology (GO) and KEGG enrichment result of significant shared by malvids species gene families.

Table S12. Gene Ontology (GO) and KEGG enrichment result of significant expansion of *E. japonica* gene families.

Table S13. Gene Ontology (GO) enrichment result of significant contraction of *E. japonica* gene families.

Table S14. Statistical sampling population information.

Table S15. Statistics population resequencing information.

Table S16. Statistical nucleotide polymorphisms in the populations.

Table S17. Candidate positive selection genes (PSGs) in the evergreen population.

Table S18. Candidate positive selection genes (PSGs) in the deciduous population.

Table S19. Gene Ontology (GO) enrichment result of significant PSGs in the evergreen population.

Table S20. List of MADS-box genes identified in *E. japonica*.

Table S21. Genes involved in anthocyanin biosynthesis, carotenoid biosynthesis, and chlorophyll degradation.

Table S22. Identification fruit dehiscence-related genes in *E. japonica*.

Table S23. Genes related to lignin synthesis that are highly expressed during pericarp dehiscence.

Table S24. Gene expression levels (FPKM) of fruit abscission-related genes in pericarp.

Table S25. Triterpene compounds separated from *Euscaphis*.

Table S26. Number of putative pentacyclic triterpene-related genes in the malvids species.

Table S27. Identified pentacyclic triterpene synthesis-related genes in *E. japonica* genome.

Table S28. Statistical simple sequence repeat.

REFERENCES

- Adamczyk, B.J. & Fernandez, D.E. (2009) MIK* MADS domain heterodimers are required for pollen maturation and tube growth in *Arabidopsis*. *Plant Physiology*, **149**, 1713–1723. <https://doi.org/10.1104/pp.109.135806>
- Alexander, D.H., Novembre, J. & Lange, K. (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, **19**, 1655–1664. <https://doi.org/10.1101/gr.094052.109>
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)
- An, X.H., Tian, Y., Chen, K.Q., Wang, X.F. & Hao, Y.J. (2012) The apple WD40 protein MdTTG1 interacts with bHLH but not MYB proteins to regulate anthocyanin accumulation. *Journal of Plant Physiology*, **169**, 710–717. <https://doi.org/10.1016/j.jplph.2012.01.015>
- Argout, X., Salse, J., Aury, J.-M., Guiltinan, M.J., Droc, G., Gouzy, J. et al. (2011) The genome of *Theobroma cacao*. *Nature Genetics*, **43**, 101–108. <https://doi.org/10.1038/ng.736>
- Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M. et al. (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, **25**, 25–29. <https://doi.org/10.1038/75556>
- Bai, G., Yang, D.H., Cao, P., Yao, H., Zhang, Y.H., Chen, X.J. et al. (2019) Genome-wide identification, gene structure and expression analysis of the MADS-box gene family indicate their function in the development of tobacco (*Nicotiana tabacum* L.). *International Journal of Molecular Sciences*, **20**, 5043. <https://doi.org/10.3390/ijms20205043>
- Baudry, A., Heim, M.A., Dubreucq, B., Caboche, M., Weisshaar, B. & Lepiniec, L. (2004) TT2, TT8, and TTG1 synergistically specify the expression of BANYULS and proanthocyanidin biosynthesis in *Arabidopsis thaliana*. *The Plant Journal*, **39**, 366–380. <https://doi.org/10.1111/j.1365-313X.2004.02138.x>
- Beier, S., Thiel, T., Münch, T., Scholz, U. & Mascher, M. (2017) MISA-web: a web server for microsatellite prediction. *Bioinformatics*, **33**, 2583–2585. <https://doi.org/10.1093/bioinformatics/btx198>
- Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E. et al. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research*, **31**, 365–370. <https://doi.org/10.1093/nar/gkg095>
- Burke, K.D., Williams, J.W., Chandler, M.A., Haywood, A.M., Lunt, D.J. & Otto-Bliesner, B.L. (2018) Pliocene and eocene provide best analogs for near-future climates. *Proceedings of the National Academy of Sciences of the United States of America*, **115**, 13288–13293. <https://doi.org/10.1073/pnas.1809600115>

- Burton, J.N., Adey, A., Patwardhan, R.P., Qiu, R., Kitzman, J.O. & Shendure, J. (2013) Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nature Biotechnology*, **31**, 1119–1125. <https://doi.org/10.1038/nbt.2727>
- Campoy, J., Lerigoleur-Balsemin, E., Christmann, H., Beauvieux, R., Girollet, N., Quero-García, J. *et al.* (2016) Genetic diversity, linkage disequilibrium, population structure and construction of a core collection of *Prunus avium* L. landraces and bred cultivars. *BMC Plant Biology*, **16**, 49. <https://doi.org/10.1186/s12870-016-0712-9>
- Cantino, P.D., Doyle, J.A., Graham, S.W., Judd, W.S., Olmstead, R.G., Soltis, D.E. *et al.* (2007) Towards a phylogenetic nomenclature of Tracheophyta. *Taxon*, **56**, 822–846.
- Causier, B., Kieffer, M. & Davies, B. (2002) MADS-box genes reach maturity. *Science*, **296**, 275–276. <https://doi.org/10.1126/science.1071401>
- Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S.S., Prucell, S.M. & Lee, J.J. (2015) Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, **4**, 7. <https://doi.org/10.1186/s13742-015-0047-8>
- Chen, F., Zhang, X., Liu, X. & Zhang, L. (2017) Evolutionary analysis of MIKCC-type MADS-box genes in gymnosperms and angiosperms. *Frontiers in Plant Science*, **8**, 895. <https://doi.org/10.3389/fpls.2017.00895>
- Chen, S.F., Zhou, Y., Chen, Y. & Gu, J. (2018a) Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **13**, 884–890. <https://doi.org/10.1093/bioinformatics/bty560>
- Chen, W.H., Li, P.F., Lee, Y.I. & Yang, C.H. (2015) FOREVER YOUNG FLOWER negatively regulates ethylene response DNA-Binding binding factors by activating an ethylene-responsive factor to control *Arabidopsis* floral organ senescence and abscission. *Plant Physiology*, **168**, 1666–1683. <https://doi.org/10.1104/pp.15.00433>
- Chen, Y., Chen, Y., Shi, C., Huang, Z., Zhang, Y., Li, S. *et al.* (2018b) SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *GigaScience*, **7**, 1–6. <https://doi.org/10.1093/gigascience/gix120>
- Cui, L., Feng, K., Wang, M.C., Wang, M., Deng, P.C., Song, W.N. *et al.* (2016) Genome-wide identification, phylogeny and expression analysis of AP2/ERF transcription factors family in *Brachypodium distachyon*. *BMC Genomics*, **17**, 636. <https://doi.org/10.1186/s12864-016-2968-8>
- D'Hont, A., Denoeud, F., Aury, J.-M., Baurens, F.-C., Carreel, F., Garsmeur, O. *et al.* (2012) The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature*, **488**, 213–217. <https://doi.org/10.1038/nature11241>
- Darabi, M. & Masoudi-Nejad, A. (2012) Bioinformatics study of the 3-hydroxy-3-methylglutaryl-coenzyme A reductase (HMGR) gene in Gramineae. *Molecular Biology Reports*, **39**, 8925–8935. <https://doi.org/10.1007/s11033-012-1761-2>
- Dardick, C. & Callahan, A.M. (2014) Evolution of the fruit endocarp: molecular mechanisms underlying adaptations in seed protection and dispersal strategies. *Frontiers in Plant Science*, **5**, 284. <https://doi.org/10.3389/fpls.2014.00284>
- De Bie, T., Cristianini, N., Demuth, J.P. & Hahn, M. (2006) CAFE: a computational tool for the study of gene family evolution. *Bioinformatics*, **22**, 1269–1271. <https://doi.org/10.1093/bioinformatics/btl097>
- Debeaujion, I., Peeters, A.J., Léon-Kloosterziel, K.M. & Koornneef, M. (2001) TRANSPARENT TESTA 12 gene of *Arabidopsis* encodes a multidrug secondary transporter-like protein required for flavonoid sequestration in vacuoles of the seed coat endothelium. *The Plant Cell*, **13**, 835–871. <https://doi.org/10.1105/tpc.13.4.853>
- Delis, C., Krokida, A., Georgiou, S., Pena-Rodriguez, L.M., Kavroulakis, N., Ioannou, E. *et al.* (2011) Role of lupeol synthase in *Lotus japonicus* nodule formation. *New Phytologist*, **189**, 335–346. <https://doi.org/10.1111/j.1469-8137.2010.03463.x>
- Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**, 1792–1797. <https://doi.org/10.1093/nar/gkh340>
- Ennis, S. (2007) Linkage disequilibrium as a tool for detecting signatures of natural selection. *Methods in Molecular Biology*, **376**, 59–70. https://doi.org/10.1007/978-1-59745-389-9_5
- Enoki, S., Hattori, T., Ishiai, S., Tanaka, S., Mikami, M., Arita, K. *et al.* (2017) Vanillylacetone up-regulates anthocyanin accumulation and expression of anthocyanin biosynthetic genes by inducing endogenous abscisic acid in grapevine tissues. *Journal of Plant Physiology*, **219**, 22–27. <https://doi.org/10.1016/j.jplph.2017.09.005>
- Estornell, L.H., Agustí, J., Merelo, P., Talón, M. & Tadeo, F.R. (2013) Elucidating mechanisms underlying organ abscission. *Plant Science*, **199–200**, 48–60. <https://doi.org/10.1016/j.plantsci.2012.10.008>
- Ferrándiz, C. (2002) Regulation of fruit dehiscence in *Arabidopsis*. *Journal of Experimental Botany*, **53**, 8. <https://doi.org/10.1093/jxb/erf082>
- Fordham, D.A., Jackson, S.T., Brown, S.C., Huntley, B., Brook, B.W., Dahl-Jensen, D. *et al.* (2020) Using paleo-archives to safeguard biodiversity under climate change. *Science*, **369**, 6507. <https://doi.org/10.1126/science.abc5654>
- Gao, Y., Wang, H., Liu, C., Chu, H., Dai, D., Song, S. *et al.* (2018) De novo genome assembly of the red silk cotton tree (*Bombax ceiba*). *GigaScience*, **7**, giy051. <https://doi.org/10.1093/gigascience/giy051>
- Gary, B. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*, **27**, 573–580. <https://doi.org/10.1093/nar/27.2.573>
- Geng, Y., Guan, Y., Qiong, L.A., Lu, S., An, M., Crabbe, M.J.C. *et al.* (2021) Genomic analysis of field pennycress (*Thlaspi arvense*) provides insights into mechanisms of adaptation to high elevation. *BMC Biology*, **19**, 143. <https://doi.org/10.1186/s12915-021-01079-0>
- Ghosh, S. (2017) Triterpene structural diversification by plant cytochrome P450 enzymes. *Frontiers in Plant Science*, **8**, 1886. <https://doi.org/10.3389/fpls.2017.01886>
- Girin, T., Paicu, T., Stephenson, P., Fuentes, S., Korner, E., O'Brien, M. *et al.* (2011) *INDEHISCENT* and *SPATULA* interact to specify carpel and valve margin tissue and thus promote seed dispersal in *Arabidopsis*. *The Plant Cell*, **23**, 3641–3714. <https://doi.org/10.1105/tpc.111.090944>
- Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I. *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, **29**, 644–652. <https://doi.org/10.1038/nbt.1883>
- Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S.R. & Bateman, A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Research*, **33**, D121–124. <https://doi.org/10.1093/nar/gki081>
- Guo, D.M., Ran, J.H. & Wang, X.Q. (2010) Evolution of the cinnamyl/sinapyl alcohol dehydrogenase (CAD/SAD) gene family: the emergence of real lignin is associated with the origin of bona fide CAD. *Journal of Molecular Evolution*, **71**, 202–218. <https://doi.org/10.1007/s00239-010-9378-3>
- Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith, R.K. Jr, Hannick, L. *et al.* (2003) Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research*, **31**, 5654–5666
- Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J. *et al.* (2008) Automated eukaryotic gene structure annotation using EVidence-Modeler and the program to assemble spliced alignments. *Genome Biology*, **9**, R7. <https://doi.org/10.1186/gb-2008-9-1-r7>
- Hall, A.E. & Blecker, A.B. (2003) Analysis of combinatorial loss-of-function mutants in the *Arabidopsis* ethylene receptors reveals that the ers1 etr1 double mutant has severe developmental defects that are EIN2 dependent. *The Plant Cell*, **15**, 2032–2041. <https://doi.org/10.1105/tpc.013060>
- Han, Y.J. & Wessler, S.R. (2010) MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Research*, **38**, e199. <https://doi.org/10.1093/nar/gkq862>
- Handley, L., Crouch, E.M. & Pancost, R.D. (2011) A New Zealand record of sea level rise and environmental change during the paleocene-eocene thermal maximum. *Palaeogeography, Palaeoclimatology, Palaeoecology*, **305**, 185–200. <https://doi.org/10.1016/j.palaeo.2011.03.001>
- Harker, M., Holmberg, N., Clayton, J.C., Gibbard, C., Wallace, A.D., Rawlins, S. *et al.* (2010) Enhancement of seed phytoesterol levels by expression of an N-terminal truncated Hevea brasiliensis (rubber tree) 3-hydroxy-3-methylglutaryl-CoA reductase. *Plant Biotechnology Journal*, **1**, 113–121. <https://doi.org/10.1046/j.1467-7652.2003.00011.x>
- Hey, S.J., Powers, S.J., Beale, M.H., Hawkins, N.D., Ward, J.L. & Halford, N.G. (2006) Enhanced seed phytoesterol accumulation through expression of a modified HMG-CoA reductase. *Plant Biotechnology Journal*, **4**, 219–229. <https://doi.org/10.1111/j.1467-7652.2005.00174.x>
- Hu, J. & Wei, F. (2004) Comparative ecology of giant pandas in the five mountain ranges of their distribution in China. In: *Giant Pandas: Biology and Conservation* (Lindburg, D. & Baragona, D., eds). London: University of California Press, pp. 137–148. <https://doi.org/10.1525/california/9780520238671.003.0015>

- Huang, L., Li, J., Ye, C., Li, C.F., Wang, H., Liu, B.Y. *et al.* (2012) Molecular characterization of the pentacyclic triterpenoid biosynthetic pathway in *Catharanthus roseus*. *Planta*, **236**, 1571–1581. <https://doi.org/10.1007/s00425-012-1712-0>
- Huang, Y.J., Liu, Y.S., Wen, J. & Quan, C. (2015) First fossil record of *Staphylea* L. (Staphyleaceae) from North America, and its biogeographic implications. *Plant Systematics and Evolution*, **301**, 2203–2218. <https://doi.org/10.1007/s00606-015-1224-z>
- Ju, C., Yoon, G.M., Shemansky, J.M., Lin, D.Y., Ying, Z.I., Chang, J. *et al.* (2012) *CTR1* phosphorylates the central regulator *EIN2* to control ethylene hormone signaling from the ER membrane to the nucleus in *Arabidopsis*. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 19486–19491. <https://doi.org/10.1073/pnas.1214848109>
- Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O. & Walichiewicz, J. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research*, **110**, 462–467. <https://doi.org/10.1159/000084979>
- Kanehisa, M. & Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, **28**, 27–30. <https://doi.org/10.1093/nar/28.1.27>
- Katoh, K. & Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, **30**, 772–780. <https://doi.org/10.1093/molbev/mst010>
- Kay, P., Groszmann, M., Ross, J.J., Parish, R.W. & Swain, S.M. (2013) Modifications of a conserved regulatory network involving *INDEHISCENT* controls multiple aspects of reproductive tissue development in *Arabidopsis*. *New Phytologist*, **197**, 73–87. <https://doi.org/10.1111/j.1469-8137.2012.04373.x>
- Keilwagen, J., Wenk, M., Erickson, J.L., Schattat, M.H., Grau, J. & Hartung, F. (2016) Using intron position conservation for homology-based gene prediction. *Nucleic Acids Research*, **44**, e89. <https://doi.org/10.1093/nar/gkw092-89>
- Kent, W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Research*, **12**, 656–664. <https://doi.org/10.1101/gr.229202>
- Kim, J., Kang, S.-H., Park, S.-G., Yang, T.-J., Lee, Y.I., Kim, O.T. *et al.* (2020) Whole-genome, transcriptome, and methylome analyses provide insights into the evolution of platycoside biosynthesis in *Platycodon grandiflorus*, a medicinal plant. *Horticulture Research*, **7**, 112. <https://doi.org/10.1038/s41438-020-0329-x>
- Kitamura, S., Shikazono, N. & Tanaka, A. (2004) *TRANSPARENT TESTA 19* is involved in the accumulation of both anthocyanins and proanthocyanidins in *Arabidopsis*. *The Plant Journal*, **37**, 104–114. <https://doi.org/10.1046/j.1365-3113x.2003.01943.x>
- Koonin, E.V., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Krylov, D.M., Makarova, K.S. *et al.* (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biology*, **5**, M5–M8R7. <https://doi.org/10.1186/gb-2004-5-2-r7>
- Kumar, S., Stecher, G. & Tamura, K. (2016) MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution*, **33**, 1870–1874. <https://doi.org/10.1093/molbev/msw054>
- Kwantes, M., Liebsch, D. & Verelst, W. (2012) How MIKC* MADS-box genes originated and evidence for their conserved function throughout the evolution of vascular plant gametophytes. *Molecular Biology and Evolution*, **29**, 293–302. <https://doi.org/10.1093/molbev/msr200>
- Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921. <https://doi.org/10.1038/35057062>
- Letunic, I., Doerks, T. & Bork, P. (2015) SMART: recent updates, new developments and status in 2015. *Nucleic Acids Research*, **43**, D257–D260. <https://doi.org/10.1093/nar/gku949>
- Li, D.Z., Cai, J. & Wen, J. (2008) Staphyleaceae. In: Wu, Z.Y., Raven, P.H. & Hong, D.Y., (Eds.) *Flora of China*. Vol. 11, Beijing (China): Science Press, pp. 498–504.
- Li, H. & Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>
- Li, H. & Durbin, R. (2011) Inference of human population history from individual whole-genome sequences. *Nature*, **475**, 493–496. <https://doi.org/10.1038/nature10231>
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N. *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
- Li, L., Stoekert, C.J. & Roos, D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research*, **13**, 2178–2189. <https://doi.org/10.1101/gr.1224503>
- Liang, W., Ni, L., Carballar-Lejarazú, R., Zou, X., Sun, W., Wu, L. *et al.* (2019) Comparative transcriptome among *Euscaphis konishii* Hayata tissues and analysis of genes involved in flavonoid biosynthesis and accumulation. *BMC Genomics*, **20**, 24. <https://doi.org/10.1186/s12864-018-5354-x>
- Liang, W., Zou, X., Carballar-Lejarazú, R., Wu, L., Sun, W., Yuan, X. *et al.* (2018) Selection and evaluation of reference genes for qRT-PCR analysis in *Euscaphis konishii* Hayata based on transcriptome data. *Plant Methods*, **14**, 42. <https://doi.org/10.1186/s13007-018-0311-x>
- Liao, W., Li, Y., Yang, Y., Wang, G. & Peng, M. (2016) Exposure to various abscission-promoting treatments suggests substantial ERF subfamily transcription factors involvement in the regulation of cassava leaf abscission. *BMC Genomics*, **17**, 538. <https://doi.org/10.1186/s12864-016-2845-5>
- Lin, Y., Min, J., Lai, R., Wu, Z., Chen, Y., Yu, L. *et al.* (2017) Genome-wide sequencing of longan (*Dimocarpus longan* Lour.) provides insights into molecular basis of its polyphenol-rich characteristics. *GigaScience*, **6**, 1–14. <https://doi.org/10.1093/gigascience/gix023>
- Liu, J. (1995) Pharmacology of oleanolic acid and ursolic acid. *Journal of Ethnopharmacology*, **49**, 57–68. [https://doi.org/10.1016/0378-8741\(95\)90032-2](https://doi.org/10.1016/0378-8741(95)90032-2)
- Liu, J. (2005) Oleanolic acid and ursolic acid: research perspectives. *Journal of Ethnopharmacology*, **100**, 92–94. <https://doi.org/10.1016/j.jep.2005.05.024>
- Liu, Y.L., Cai, Y.F., Zhao, Z.J., Wang, J.F., Xin, W., Xia, G.M. *et al.* (2009) Cloning and functional analysis of a β -Amyrin synthase gene associated with oleanolic acid biosynthesis in *Gentiana straminea* MAXIM. *Biological and Pharmaceutical Bulletin*, **32**, 818–824. <https://doi.org/10.1248/bpb.32.818>
- Liu, Y., Cui, S., Wu, F., Yan, S., Lin, X., Du, X. *et al.* (2013) Functional conservation of MIKC*-Type MADS box genes in *Arabidopsis* and rice pollen maturation. *The Plant Cell*, **25**, 1288–1303. <https://doi.org/10.1105/tpc.113.110049>
- Lowe, T.M. & Eddy, S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, **25**, 955–964. <https://doi.org/10.1093/nar/25.5.955>
- Maia, V.H., Gitzendanner, M.A., Soltis, P.S., Wong, K.S. & Soltis, D.E. (2014) Angiosperm phylogeny based on 18S/26S rDNA sequence data: constructing a large dataset using next-generation sequence data. *International Journal of Plant Sciences*, **175**, 613–650. <https://doi.org/10.1086/676675>
- Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., Swenson, M.S. & Warnow, T. (2014) ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, **30**, i541–i548
- Murata, J., Roepke, J., Gordon, H. & Luca, D.E. (2008) The leaf epidermome of *Catharanthus roseus* reveals its biochemical specialization. *The Plant Cell*, **20**, 524–542. <https://doi.org/10.1105/tpc.107.056630>
- Myburg, A.A., Grattapaglia, D., Tuskan, G.A., Hellsten, U., Hayes, R.D., Grimwood, J. *et al.* (2014) The genome of *Eucalyptus grandis*. *Nature*, **510**, 356–362. <https://doi.org/10.1038/nature13308>
- Nakano, T., Fujisawa, M., Shima, Y. & Ito, Y. (2014) The AP2/ERF transcription factor SIERF52 functions in flower pedicel abscission in tomato. *Journal of Experimental Botany*, **65**, 3111–3119. <https://doi.org/10.1093/jxb/eru154>
- Nawrocki, E.P., Kolbe, D.L. & Eddy, S.R. (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics*, **25**, 1335–1337. <https://doi.org/10.1093/bioinformatics/btp157>
- Neale, D.B. & Savolainen, O. (2004) Association genetics of complex traits in conifers. *Trends in Plant Science*, **9**, 325–330. <https://doi.org/10.1016/j.tplants.2004.05.006>
- Ortiz-Ramirez, C.I., Plata-Arboleda, S. & Pabón-Mora, N. (2018) Evolution of genes associated with gynoeceum patterning and fruit development in Solanaceae. *Annals of Botany*, **121**, 1211–1230. <https://doi.org/10.1093/aob/mcy007>
- Pound, M.J., Haywood, A.M., Salzmann, U. & Riding, J.B. (2012) Global vegetation dynamics and latitudinal temperature gradients during the Mid to Late Miocene (15.97–5.33 Ma). *Earth-Science Reviews*, **112**, 1–22. <https://doi.org/10.1016/j.earscirev.2012.02.005>

- Qin, G., Xu, C., Ming, R., Tang, H., Guyot, R., Kramer, E.M. *et al.* (2017) The pomegranate (*Punica granatum* L.) genome and the genomics of punicalagin biosynthesis. *The Plant Journal*, **91**, 1108–1128. <https://doi.org/10.1111/tpj.13625>
- Rasbery, J.M., Shan, H., Clair, R., Norman, M., Matsuda, S.P. & Bartel, B. (2007) *Arabidopsis thaliana* squalene epoxidase 1 is essential for root and seed development. *Journal of Biological Chemistry*, **282**, 17002–17013. <https://doi.org/10.1074/jbc.M611831200>
- Reichelt, N., Wen, J., Pätzold, C. & Appelhans, M.S. (2021) Target enrichment improves phylogenetic resolution in the *Zanthoxylum* (Rutaceae) and indicates both incomplete lineage sorting and hybridization events. *BioRxiv* [preprint]. <https://doi.org/10.1101/2021.04.12.439519>
- Roberts, J.A., Elliott, K.A. & Gonzalez-Carranza, Z.H. (2002) Abscission, dehiscence, and other cell separation processes. *Annual Review of Plant Biology*, **53**, 131–158. <https://doi.org/10.1146/annurev-arplant.53.092701.180236>
- Roeder, A.H., Ferrándiz, C. & Yanofsky, M.F. (2003) The role of the *REPLUMLESS* homeodomain protein in patterning the *Arabidopsis* fruit. *Current Biology*, **13**, 1630–1635. <https://doi.org/10.1016/j.cub.2003.08.027>
- Schiffels, S. & Durbin, R. (2014) Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, **46**, 919–925. <https://doi.org/10.1038/ng.3015>
- Shulaev, V., Sargent, D.J., Crowhurst, R.N., Mockler, T.C., Folkerts, O., Delcher, A.L. *et al.* (2011) The genome of woodland strawberry (*Fragaria vesca*). *Nature Genetics*, **43**, 109–116. <https://doi.org/10.1038/ng.740>
- Simao, F.A., Waterhouse, R.M., Panagiotis, I., Kriventseva, E.V. & Zdobnov, E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212. <https://doi.org/10.1093/bioinformatics/btv351>
- Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313. <https://doi.org/10.1093/bioinformatics/btu033>
- Stanke, M., Schöffmann, O., Morgenstern, B. & Waack, S. (2006) Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*, **7**, 62. <https://doi.org/10.1186/1471-2105-7-62>
- Sun, W.H., Yuan, X.Y., Liu, Z.J., Lan, S.R., Tsai, W.C. & Zou, S.Q. (2019) Multivariate analysis reveals phenotypic diversity of *Euscaphis japonica* population. *PLoS One*, **14**, e0219046. <https://doi.org/10.1371/journal.pone.0219046>
- Suzuki, Y. (2010) Statistical methods for detecting natural selection from genomic data. *Genes and Genetic Systems*, **6**, 359–376. <https://doi.org/10.1266/ggs.85.3.59>
- The Angiosperm Phylogeny Group (2009) An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG III. *Botanical Journal of the Linnean Society*, **161**, 105–121.
- The Angiosperm Phylogeny Group (2016) An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society*, **181**, 1–20. <https://doi.org/10.1111/boj.12385>
- Theissen, G., Melzer, R. & Rümpler, F. (2016) MADS-domain transcription factors and the floral quartet model of flower development: linking plant development and evolution. *Development*, **143**, 3259–3271. <https://doi.org/10.1242/dev.134080>
- Thimmappa, R., Geisler, K., Louveau, T., O'Maille, P. & Osbourn, A. (2014) Triterpene biosynthesis in plants. *Annual Review of Plant Biology*, **65**, 225–257. <https://doi.org/10.1146/annurev-arplant-050312-120229>
- Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R. *et al.* (2012) Differential gene and transcript expression analysis of RNA-seq. experiments with TopHat and Cufflinks. *Nature Protocols*, **7**, 562–578. <https://doi.org/10.1038/nprot.2012.016>
- Van Gelderen, K., van Rongen, M., Liu, A., Otten, A. & Offringa, R. (2016) An INDEHISCENT-controlled auxin response specifies the separation layer in early *Arabidopsis* fruit. *Molecular Plant*, **9**, 857–869. <https://doi.org/10.1016/j.molp.2016.03.005>
- Verelst, W., Saedler, H. & Münster, T. (2007) MIK* MADS-protein complexes bind motifs enriched in the proximal region of late pollen-specific *Arabidopsis* promoters. *Plant Physiology*, **143**, 447–460. <https://doi.org/10.1104/pp.106.089805>
- Vishwakarma, R.K., Patel, K., Sonawane, P., Kumari, U., Singh, S., Ruby, *et al.* (2015) Squalene synthase gene from medicinal herb *Bacopa monniera*: molecular characterization, differential expression, comparative modeling, and docking studies. *Plant Molecular Biology Reporter*, **33**, 1675–1685. <https://doi.org/10.1007/s11105-015-0864-z>
- Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S. *et al.* (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, **9**, e112963. <https://doi.org/10.1371/journal.pone.0112963>
- Wang, Z., Guhling, O., Yao, R., Li, F.L., Yeats, T.H., Rose, J.K. *et al.* (2010) Two oxidosqualene cyclases responsible for biosynthesis of tomato fruit cuticular triterpenoids. *Plant Physiology*, **155**, 540–552. <https://doi.org/10.1104/pp.110.162883>
- Wu, S.D., Han, B. & Jiao, Y. (2019) Genetic contribution of paleopolyploidy to adaptive evolution in angiosperms. *Molecular Plant*, **13**, 59–71. <https://doi.org/10.1016/j.molp.2019.10.012>
- Xie, C., Mao, X.Z., Huang, J.J., Ding, Y., Wu, J.M., Dong, S. *et al.* (2011) KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Research*, **39**, W316–W322. <https://doi.org/10.1093/nar/gkr483>
- Xu, Z. & Wang, H. (2007) LTR-FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research*, **35**, 265–268. <https://doi.org/10.1093/nar/gkm286>
- Yang, L., Su, D., Chang, X., Foster, C.S.P., Sun, L., Huang, C.-H. *et al.* (2020) Phylogenomic insights into deep phylogeny of angiosperms based on broad nuclear gene sampling. *Plant Communications*, **1**, 100027. <https://doi.org/10.1016/j.xplc.2020.100027>
- Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, **24**, 1586–1591.
- Yu, G.C., Wang, L.G., Han, Y.Y. & He, Q.Y. (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*, **16**, 284–287. <https://doi.org/10.1089/omi.2011.0118>
- Yuan, X.Y., Zou, X., Huang, W., Zhang, X., Chen, Z., Sun, W. *et al.* (2018) Study on the changes of contents of pigments of *Euscaphis konishii* Hayata fruit during fruit development. *Nonwood Forest Research*, **3**, 100–106.
- Zdobnov, E.M. & Apweiler, R. (2001) InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848. <https://doi.org/10.1093/bioinformatics/17.9.847>
- Zhang, C., Dong, S.S., Xu, J.Y., He, W.M. & Yang, T.L. (2018) PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics*, **10**, 10. <https://doi.org/10.1093/bioinformatics/bty875>
- Zhao, J. & Dixon, R.A. (2009) MATE transporters facilitate vacuolar uptake of epicatechin 3'-O-glucoside for proanthocyanidin biosynthesis in *Medicago truncatula* and *Arabidopsis*. *The Plant Cell*, **21**, 2323–2340. <https://doi.org/10.1105/tpc.109.067819>
- Zhao, L., Li, X., Zhang, N., Zhang, S.-D., Yi, T.-S., Ma, H. *et al.* (2016) (2016) Phylogenomic analyses of large-scale nuclear genes provide new insights into the evolutionary relationships within the Rosids. *Molecular Phylogenetics and Evolution*, **105**, 116–176. <https://doi.org/10.1016/j.ympev.2016.06.007>
- Zhao, Y.-P., Fan, G., Yin, P.-P., Sun, S., Li, N., Hong, X. *et al.* (2019) Resequencing 545 ginkgo genomes across the world reveals the evolutionary history of the living fossil. *Nature Communications*, **10**, 4201. <https://doi.org/10.1038/s41467-019-12133-5>
- Zheng, B., Xu, Q.Q. & Shen, Y.P. (2002) The relationship between climate change and quaternary glacial cycles on the Qinghai-Tibetan Plateau: review and speculation. *Quaternary International*, **97**, 93–101. [https://doi.org/10.1016/S1040-6182\(02\)00054-X](https://doi.org/10.1016/S1040-6182(02)00054-X)