JIPB Journal of Integrative Plant Biology

New Resources
https://doi.org/10.1111/jipb.13318

Cultivated hawthorn (*Crataegus pinnatifida* var. major) genome sheds light on the evolution of Maleae (apple tribe)[oo]

Ticao Zhang[1†*], Qin Qiao[2†*], Xiao Du[3,4], Xiao Zhang[3], Yali Hou[3], Xin Wei[3], Chao Sun[3], Rengang Zhang[5], Quanzheng Yun[5], M. James C. Crabbe[6,7,8], Yves Van de Peer[9,10,11] and Wenxuan Dong[3*]

1. College of Chinese Material Medica, Yunnan University of Chinese Medicine, Kunming 650500, China
2. School of Agriculture, Yunnan University, Kunming 650091, China
3. College of Horticulture, Shenyang Agricultural University, Shenyang 110866, China
4. Key Laboratory of Beibu Gulf Environment Change and Resources Utilization of Ministry of Education, Nanning Normal University, Guangxi, Nanning 530001, China
5. Beijing Ori-Gene Science and Technology Co. Ltd, Beijing 102206, China
6. Wolfson College, Oxford University, Oxford, UK
7. Institute of Biomedical and Environmental Science & Technology, School of Life Sciences, University of Bedfordshire, Park Square, Luton, UK
8. School of Life Sciences, Shanxi University, Taiyuan 030006, China
9. Department of Plant Biotechnology and Bioinformatics, Center for Plant Systems Biology, Ghent University, VIB, 9052, Ghent, Belgium
10. Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria 0028, South Africa
11. College of Horticulture, Nanjing Agricultural University, Nanjing 210095, China

[†]These authors contributed equally to this work.

*Correspondences: Ticao Zhang (ticaozhang@126.com, Dr. Zhang is fully responsible for the distribution of the materials associated with this article); Qin Qiao (qiaoqin@ynu.edu.cn) and Wenxuan Dong (wxdong63@126.com)

Ticao Zhang

Wenxuan Dong

ABSTRACT

Cultivated hawthorn (*Crataegus pinnatifida* var. major) is an important medicinal and edible plant with a long history of use for health protection in China. Herein, we provide a *de novo* chromosome-level genome sequence of the hawthorn cultivar "Qiu Jinxing." We assembled an 823.41 Mb genome encoding 40 571 genes and further anchored the 779.24 Mb sequence into 17 pseudo-chromosomes, which account for 94.64% of the assembled genome. Phylogenomic analyses revealed that cultivated hawthorn diverged from other species within the Maleae (apple tribe) at approximately 35.4 Mya. Notably, genes involved in the flavonoid and triterpenoid biosynthetic pathways have been significantly amplified in the hawthorn genome. In addition, our results indicated that the Maleae share a unique ancient tetraploidization event; however, no recent independent whole-genome duplication event was specifically detected in hawthorn. The amplification of non-specific long terminal repeat retrotransposons contributed the most to the expansion of the hawthorn genome. Furthermore, we identified two paleo-sub-genomes in extant species of Maleae and found that these two sub-genomes showed different rearrangement mechanisms. We also reconstructed the ancestral chromosomes of Rosaceae and discussed two possible paleo-polyploid origin patterns (autopolyploidization or allopolyploidization) of Maleae. Overall, our study provides an improved context for understanding the evolution of Maleae species, and this new high-quality reference genome provides a useful resource for the horticultural improvement of hawthorn.

Keywords: ancestral chromosome reconstruction, hawthorn (*Crataegus* spp.), long terminal repeat retrotransposons (LTR-RTs), medicinal and edible plants, sub-genome

Zhang, T., Qiao, Q., Du, X., Zhang, X., Hou, Y., Wei, X., Sun, C., Zhang, R., Yun, Q., Crabbe, M. J. C., Van de Peer, Y., and Dong, W. (2022). Cultivated hawthorn (*Crataegus pinnatifida* var. major) genome sheds light on the evolution of Maleae (apple tribe). J. Integr. Plant Biol. **64:** 1487`–1501.

INTRODUCTION

Rosaceae, the rose family of plants, includes many species that are of economic importance as fruit trees, ornamentals and medicinal plants. With the rapid development of high-throughput sequencing technologies, whole-genome data of dozens of species in the Rosaceae family have been released and in most cases deposited in the Genome Database for Rosaceae (GDR; https://www.rosaceae.org) (Jung et al., 2019; Li et al., 2021), constituting a powerful genomic resource for the study of this family. Phylogenetic analysis based on plastid and nuclear genomes suggests that the Rosaceae can be divided into three subfamilies: Dryadoideae (Dryas subfamily), Rosoideae (Rosa subfamily) and Amygdaloideae (peach subfamily) (Xiang et al., 2017; Zhang et al., 2017). The Maleae (the apple tribe in the Amygdaloideae) includes commercially important fruit trees, including apple (*Malus* × *domestica*), pear (*Pyrus* spp.), loquat (*Eriobotrya japonica*) and hawthorn (*Crataegus* spp.). Among these four fruits, hawthorn is the only one for which the genome sequence has not been reported. Maleae tribe members have a basal haploid chromosome count of 17, instead of 7, 8 or 9 as in the other Rosaceae (Goldblatt, 1976). The evolutionary origin (autopolyploidization or allopolyploidization) of the Maleae has long been debated (Chevreau et al., 1985; Robertson et al., 1991; Raspé and Sloover, 1998; Evans and Campbell, 2002; Velasco et al., 2010; Verde et al., 2013; Xiang et al., 2017; Wang et al., 2019). The hawthorn genome data are an important resource for comparative analyses and could provide valuable clues to the genome evolution of Maleae.

Hawthorn (*Crataegus*) is a genus comprising small trees and shrubs found in the temperate zone of the Northern Hemisphere (i.e., Eurasia and North America) (Lo et al., 2009). The genus contains over 200 species (Phipps, 2015), among which ~20 species are widely distributed across China (Dong and Li, 2015). Our previous phylogenetic analyses based on specific locus amplified fragment sequencing (SLAF-seq) indicated there were two independent speciation events in Chinese hawthorn taxa. Species located in southwest China shared the gene pool with the European species, whereas species in northeastern China may have originated from the North American species (Du et al., 2019).

The cultivated hawthorn (*Crataegus pinnatifida* var. major) is an important medicinal and edible plant with a long history of traditional uses for health protection in China, particularly in facilitating digestion. In addition, a few other species are cultivated for their fruits in Europe (e.g., *Crataegus azarolus* and *Crataegus germanica*) and the Americas (e.g., *Crataegus mexicana* and *Crataegus opaca*) (Mehdi et al., 2015). The pulpy fruits of hawthorn have excellent flavor and attractive color, and are rich in nutrients. Hawthorn fruits contain abundant bioactive compounds, such as flavonoids, triterpenoids, organic acids, phenols, and procyanidins (Edwards et al., 2012; Yang and Liu, 2012; Xu et al., 2016). Furthermore, the potency of hawthorn in the treatment or prevention of cardiovascular diseases has been demonstrated in laboratory tests as well as in clinical trials (Edwards et al., 2012; Cloud et al., 2020).

Although there are transcriptomic and SLAF-seq studies on hawthorn (Xu et al., 2016; Guo et al., 2018; Du et al., 2019), hitherto no whole-genome sequence of hawthorn has been reported; this has limited the in-depth study of this species. Whole-genome level studies can provide insights into the evolution and entire compound metabolic pathways of plants (Wang et al., 2021; Zhou and Liu, 2022). In the present study, using Illumina Novaseq, Oxford Nanopore and chromosome conformation capture (Hi-C) sequencing technologies, we provide a *de novo* high-quality chromosome-scale genome sequence of the hawthorn cultivar "Qiu Jinxing" (*C. pinnatifida* var. major) ($2n = 2x = 34$). Based on the assembled genome, we conducted a genome comparison between hawthorn and related species and analyzed the genes responsible for the rich bioactive compounds in hawthorn. We mainly focused on two factors affecting the genome size of hawthorn: transposon amplification and whole-genome duplication (WGD) events. Furthermore, we reconstructed hypothetical ancestral chromosomes for Rosaceae and identified two paleo-sub-genomes in extant species of Maleae. Our study provides a high-quality genomic resource for further studies on the evolution of Maleae and the horticultural improvement of the hawthorn.

RESULTS AND DISCUSSION

Genome assembly and annotation of cultivated hawthorn

We sequenced the genome of the cultivated hawthorn cultivar "Qiu Jinxing" (Figure S1) *de novo* using the Oxford Nanopore platform with 87.0 Gb of generated reads (Table 1). Additional short-fragment libraries (400 bp) were constructed and 112.83 Gb of Illumina sequencing data were generated. We used these high-quality short reads generated by the Illumina platform for *k*-mer frequency analyses (Liu et al., 2013) to estimate the genome size (856.88 Mb, Table S1; Figure S2) and to correct for sequencing errors in the Nanopore assembly. The final assembled genome size was 823.41 Mb, with a high sequencing depth of 242.68 *X*. The contig N50 size of 1.74 Mb was not very high, partly due to the predicted high heterozygous (1.77%) and repetitive elements proportion (67.89%) (Tables 1, S1).

Based on 86.95 Gb of Hi-C data, the 779.24 Mb hawthorn genome sequence was anchored to 17 pseudo-chromosomes, which accounted for 94.64% of the final assembled sequence (Figure 1A; Tables 1, S2). Pseudo-chromosome lengths ranged from 36.02 to 55.84 Mb with scaffolds N50 = 44.94 Mb (Table S3). To assess assembly accuracy, we remapped Illumina short and Nanopore long sequencing reads to the assembled genome. With 99.70%

Table 1. Genome assembly and annotation of cultivated hawthorn (*Crataegus pinnatifida* var. major)

Assembly parameters	Results
Predicted genome size	856.9 Mb
Predicted heterozygous	1.77%
Illumina reads (400 bp)	112.83 Gb
Nanopore reads	87.0 Gb
Hi-C reads	86.95 Gb
Assembled genome size	823.41 Mb
Total contigs number	744
Length of contig N50	1.74 Mb
Number of contig N50	124
Total scaffolds number	666
Length of scaffold N50	44.94 Mb
Number of scaffold N50	8
Anchored chromosomes size	779.24 Mb
Anchored chromosomes (%)	94.64
Gene number	40,571
Repetitive elements	67.89%
Transposon elements	66.03%
Benchmarking Universal Single-Copy Orthologs assessment	1,342 (97.60%)

and 99.91% mapping rates of short and long sequencing reads, the reads covered (depth >1*X*) 91.43% and 99.95% of the whole genome, respectively. We integrated *ab initio* gene prediction programs, RNA-seq analysis and homology searches to annotate the genome of hawthorn (see Materials and Methods section). In total, 40,571 protein-coding genes were predicted, including 39,097 genes (96.37%) which could be functionally annotated in the Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) databases (Tables S4–S6; Figure S3). The completeness of the assembled dataset exceeded 97.60% when evaluated using the BUSCO method (Benchmarking Universal Single-Copy Orthologs; Table S7) (Simão et al., 2015). The above results confirm the high quality of our hawthorn genome assembly.

Comparative and evolutionary genomics of hawthorn and its relatives

Based on orthologous clustering by OrthoFinder2 software (Emms and Kelly, 2018), we identified 24,234 orthogroups (which can be roughly regarded as gene families), including 505,910 across the 18 species with 274 single-copy gene families (Figure S4). Among them, six gene families were hawthorn-specific, and these included 13 unique genes (Figure 1C). In addition, we identified 3,709 unclustered genes in the hawthorn genome. Thus, there were a total 3,722 (13 + 3,709) species-specific genes in the hawthorn genome. Functional annotation of these species-specific genes indicated they were primarily enriched in biosynthetic and metabolic processes, including arginine biosynthesis, alanine, aspartate and glutamate metabolism, the citrate acid (tricarboxylic acid (TCA)) cycle, starch and sucrose metabolism, fatty acid degradation, glycolysis/gluconeogenesis and biosynthesis of secondary metabolites (Table S8).

Using 1,194 orthogroups with a minimum of 88.9% of species having single-copy genes in any orthogroup, we constructed a phylogenetic tree for the 12 genera with representative species of Rosaceae the genomes of which have been sequenced to date, with six other species as outgroups (Figures 1D, S5). Consistent with previous results, the phylogenomic tree revealed that species in the Amygdaloideae and Rosoideae subfamilies formed two distinct clades (Xiang et al., 2017; Zhang et al., 2017). Hawthorn was placed in the Amygdaloideae clade and as a sister lineage to the combined clade of *Malus*, *Pyrus*, *Sorbus*, and *Eriobotrya*. Our results indicated that cultivated hawthorn separated from that combined clade at 35.4 (28.1–46.70) Mya, which overlapped with the previous estimated time of the split of hawthorn from other genera in Maleae (Xiang et al., 2017; Zhang et al., 2017).

We then examined the rates and direction of change in gene family size among taxa using CAFE (Han et al., 2013). The results showed that hawthorn exhibited larger numbers of expanded gene families than contracted ones (Figure 1D). Notably, the genes involved in many secondary metabolite biosynthetic pathways (e.g., flavone and flavonol biosynthesis, isoflavonoid biosynthesis, monoterpenoid biosynthesis, isoquinoline alkaloid biosynthesis, and triterpenoid biosynthesis) have been significantly amplified in the hawthorn genome (Figure 1E; Table S9). These rapidly expanded genes would be expected to be associated with the abundant biologically active compounds found in hawthorn. For example, the gene family encoding squalene epoxidase (SQE), which has been recognized as the common rate-limiting enzyme in the triterpene saponin and phytosterol biosynthetic pathways (Han et al., 2010), was significantly expanded in hawthorn (25 genes) as compared with four Maleae species from the sister clade: apple (20 genes), pear (19 genes), loquat (17 genes), and *Sorbus pohuashanensis* (16 genes). This may be related to the rich triterpenoid content of cultivated hawthorn (Edwards et al., 2012; Yang and Liu, 2012).

To detect the changes in gene expression of active compound production-related genes, we compared our coauthors' previously published transcriptomes data of two fruit developmental stages of hard- ("Qiu Jinxing") and soft-fleshed ("Ruanrou Shanlihong #3") hawthorn cultivars (Xu et al., 2016). We reanalyzed the transcriptome data based on the assembled reference genome in the present study (Figures S6, S7; Tables S10, S11). We found that most of the genes in the triterpenoid biosynthetic pathway were highly expressed (Figure S8). In addition to the triterpenoids, we also investigated the gene expressions of flavonoids biosynthetic pathway. We found 25 ($P = 0.0058$) and 28 ($P = 8.40E{-}06$) genes in hard- and soft-fleshed hawthorn cultivars, respectively, that were differentially expressed between the middle and late stages of fruit development and were significantly enriched in the flavonoid biosynthesis pathway (Figure 2; Tables S10, S11). Furthermore, the expression levels

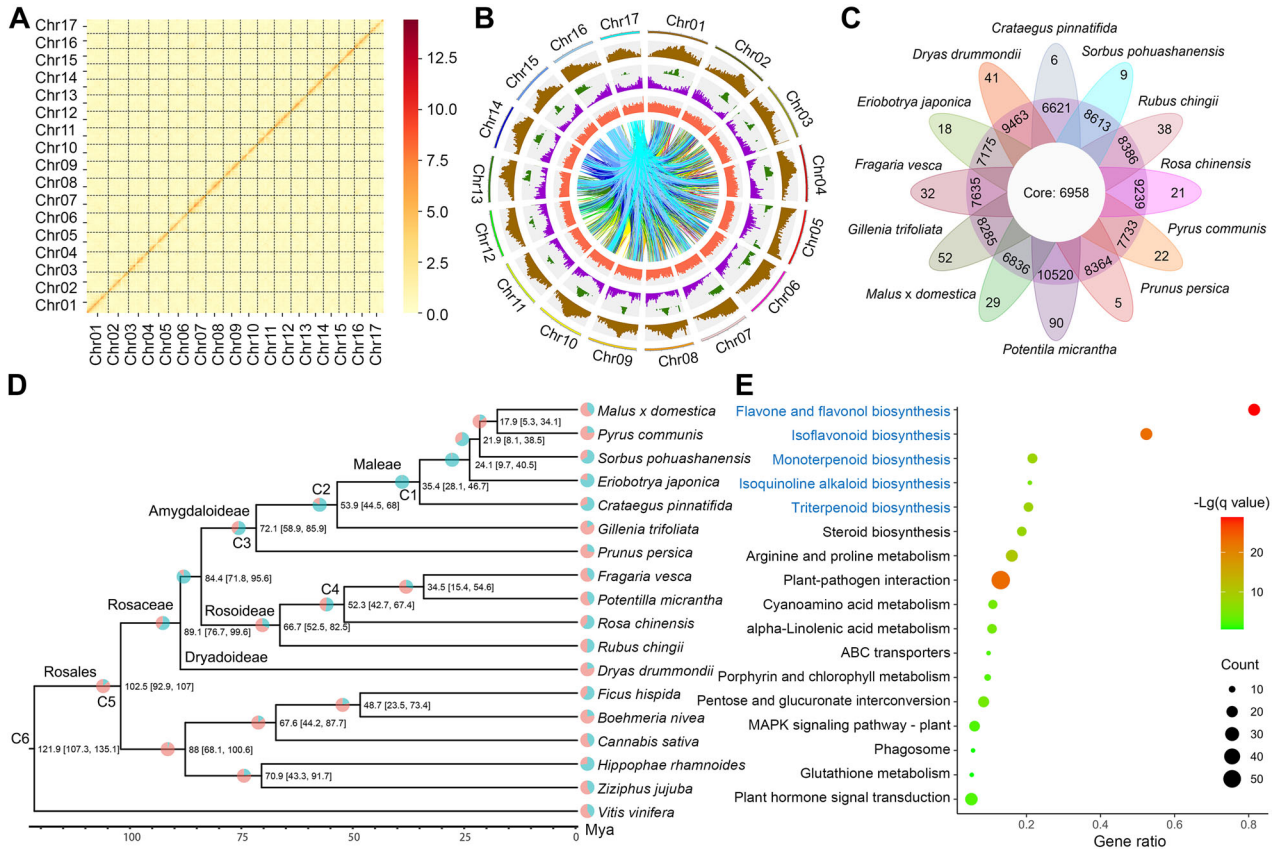

Figure 1. Genome assembly and comparative and evolutionary genomic analysis of cultivated hawthorn

(A) Hi-C interaction heatmap of cultivated hawthorn genome showing the interactions among 17 chromosomes (Chr1–17). **(B)** Genomic features of cultivated hawthorn. The outer colored square represents the 17 chromosomes of hawthorn, with tick marks every 10 Mb. Tracks show, from outside to inside, the densities of genes (red), DNA transposons (green), RNA transposons (blue), and other types of genome components (purple), respectively. The window size used in the circles was 500 kb. **(C)** Flower plot displaying the shared core orthogroups (in the center) and the species-specific orthogroups (in the petals) for the relatives. **(D)** Maximum likelihood phylogenetic tree and expanded and contracted gene families in Rosaceae. The numbers at the branch node in the tree indicate the divergence time and 95% confidence interval, and the pie charts show the relative sizes of expansion (lake blue) and contraction (red). C1–C6 mark calibration points used to estimate the divergence times. **(E)** Kyoto Encyclopedia of Genes and Genomes (KEGG) classification of gene function of genes that have rapidly expanded in cultivated hawthorn.

of most flavonoid biosynthesis-related genes, such as the chalcone synthase (*CHS*), chalcone isomerase (*CHI*), naringenin 3-dioxygenase (*F3H*), and dihydroflavonol 4-reductase (*DFR*) genes, decreased in the late stage of ripening in the hard-fleshed cultivar compared with the soft-fleshed cultivar. In contrast, the expression levels of some key lignin biosynthesis genes, such as the cinnamoyl-CoA reductase (*CCR*), caffeic acid O-methyltransferase (*COMT*), and cinnamyl alcohol dehydrogenase (*CAD*) genes, at the late stage were higher in the hard-fleshed fruit. Flavonoids and monolignols (the precursor of lignin) share a common biosynthetic origin, p-Coumaroyl-CoA. It was reported that when the carbon flow down the flavonoid pathway became limited, increased levels of monolignols were formed (Lunkenbein et al., 2006; Yeh et al., 2014). Moreover, the content of flavonoids changes dynamically during hawthorn fruit development (Zhang et al., 1994). According to the above results, we suggest collecting hawthorn fruits at the middle stage of fruit ripening in order to obtain greater amounts of flavonoid compounds.

Amplification of repetitive elements in the cultivated hawthorn genome

The assembled genome size of hawthorn (823.41 Mb) is larger than those of its closely related species apple (652–668 Mb) (Zhang et al., 2019; Sun et al., 2020), pear (498.27 Mb) (Linsmith et al., 2019), and loquat (760.1 Mb) (Jiang et al., 2020). Transposable element (TE) amplification and polyploidization (WGD) events are two main reasons for genome expansion (Bennetzen, 2002; Van de Peer et al., 2009). Therefore, we analyzed these two aspects separately. To reveal the causes of the large genome size in hawthorn, we first examined the evolution of TEs and their potential contribution to genome growth in hawthorn. We compared the TE content in hawthorn with that in three related Maleae species (apple, pear, and loquat) and discovered that hawthorn had the greatest TE content (514.55 Mb) among these four species. The most abundant TEs were long terminal repeat retrotransposons (LTR-RTs), with a total length of 426.31 Mb (53.99%) in the whole genome (Figure 3A;

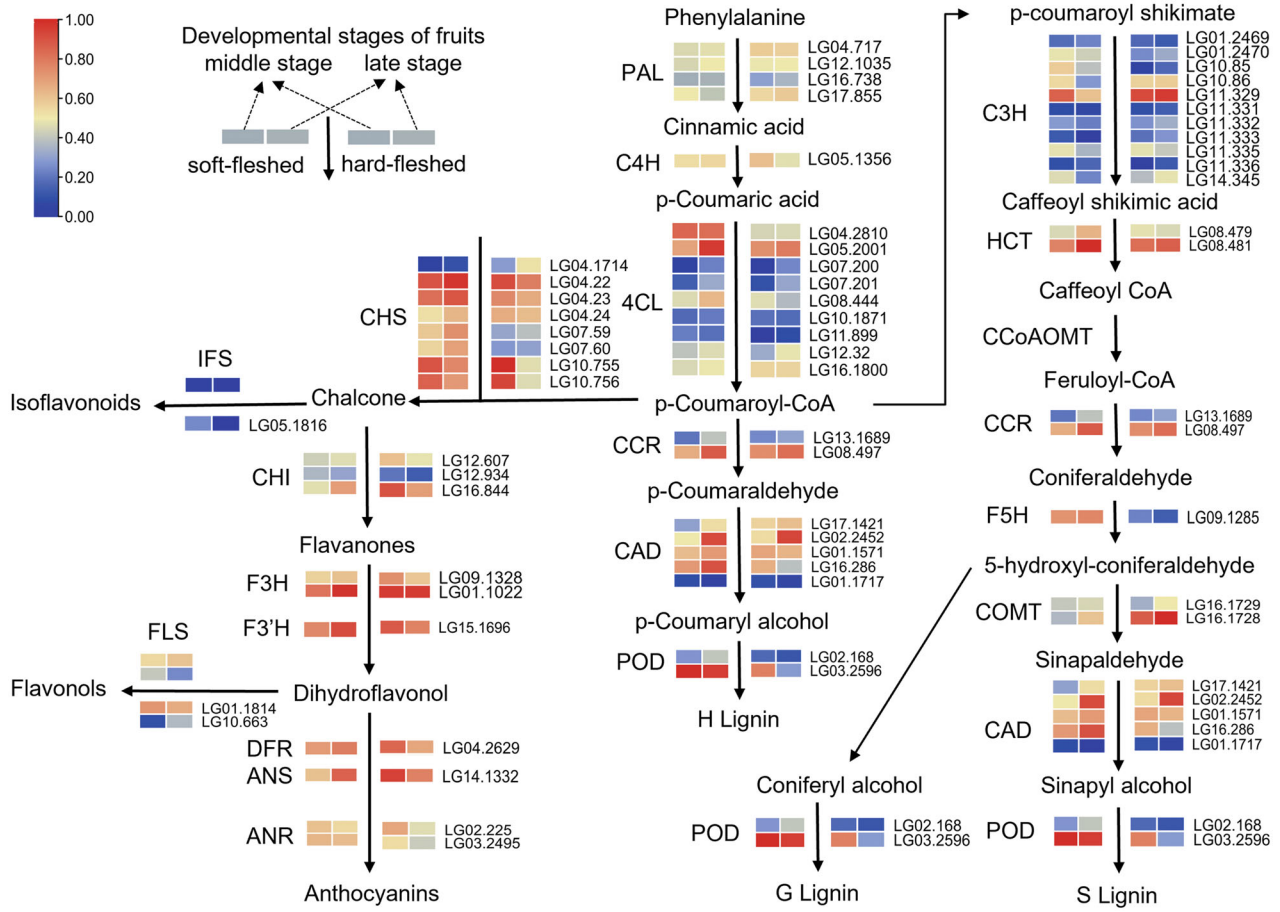

Figure 2. Biosynthetic pathway of flavonoids and related compounds, along with expression of the related genes in hard- and soft-fleshed hawthorn fruits

The gene expression profiles ($\log_{10}$[transcripts per million + 1]) at two developmental stages (middle stage at left, late stage at right) for the hard- and soft-fleshed fruits are presented in heatmaps alongside the gene names.

Table S12). To clarify the evolutionary history of LTR-RTs, we divided the genomes into non-overlapping categories, including non-TE, species-specific LTR-RTs and species-specific non-LTR-RTs, as well as non-specific LTR-RTs and non-specific non-LTR-RTs. Our results indicated that hawthorn genome contained the largest extent of non-specific LTR-RTs, indicating that the hawthorn genome is the least reduced. Both specific and non-specific LTR-RTs have contracted most in the pear genome (Figure 3A; Table S12). Therefore, the amplification and contraction of the different specific and non-specific LTR-RTs could be one reason for the differences in genome size in the Maleae tribe.

The insertion times of LTR-RTs differed among the four Maleae species, although all four had relatively recent insertions (Figure 3B). The proliferation of LTR-RTs in hawthorn peaked at ~0.56 Mya (Figure 3B), which was earlier than in apple (0.188 Mya) and pear (0.175 Mya) but later than in loquat (1.44 Mya). In the hawthorn genome, Ty3/gypsy families contributed 186.54 Mb (22.66%) of the genome and were 2.98-fold more abundant than Ty1/copia families (62.66 Mb, 7.61%) (Figures 3C, D, S12). Furthermore, the amplification of the Ty1/copia families Athila and Tekay (13.17 and 11.58 Mb, respectively), as well as that of the Ty3/gypsy family Ale (12.79 Mb), contributed most to the expansion of the hawthorn genome (Figure 3B). The proliferation of Ty1/copia, Ty3/gypsy, and unknown LTR-RTs peaked at 0.076, 0.499, and 0.51 Mya, respectively. These results suggest that most of the LTR-RTs in these four genomes were recently inserted, and these insertions occurred well after the divergence of these species (~20 Mya) (see Figure 1D).

Synteny analysis and sub-genome assignment in Maleae

The genome sequences in Rosaceae, which are becoming increasingly available (Jung et al., 2019; Li et al., 2021), together with the genome of hawthorn sequenced here, present an opportunity to elucidate the evolutionary history of extant genomes in Rosaceae, especially in Maleae. According to the four-fold degenerate synonymous sites of the third codon (4dTv) results, the Rosaceae overall did not experience a shared polyploidization event after the ancient hexaploidization of angiosperms ($K_s = 0.6–0.7$), whereas the

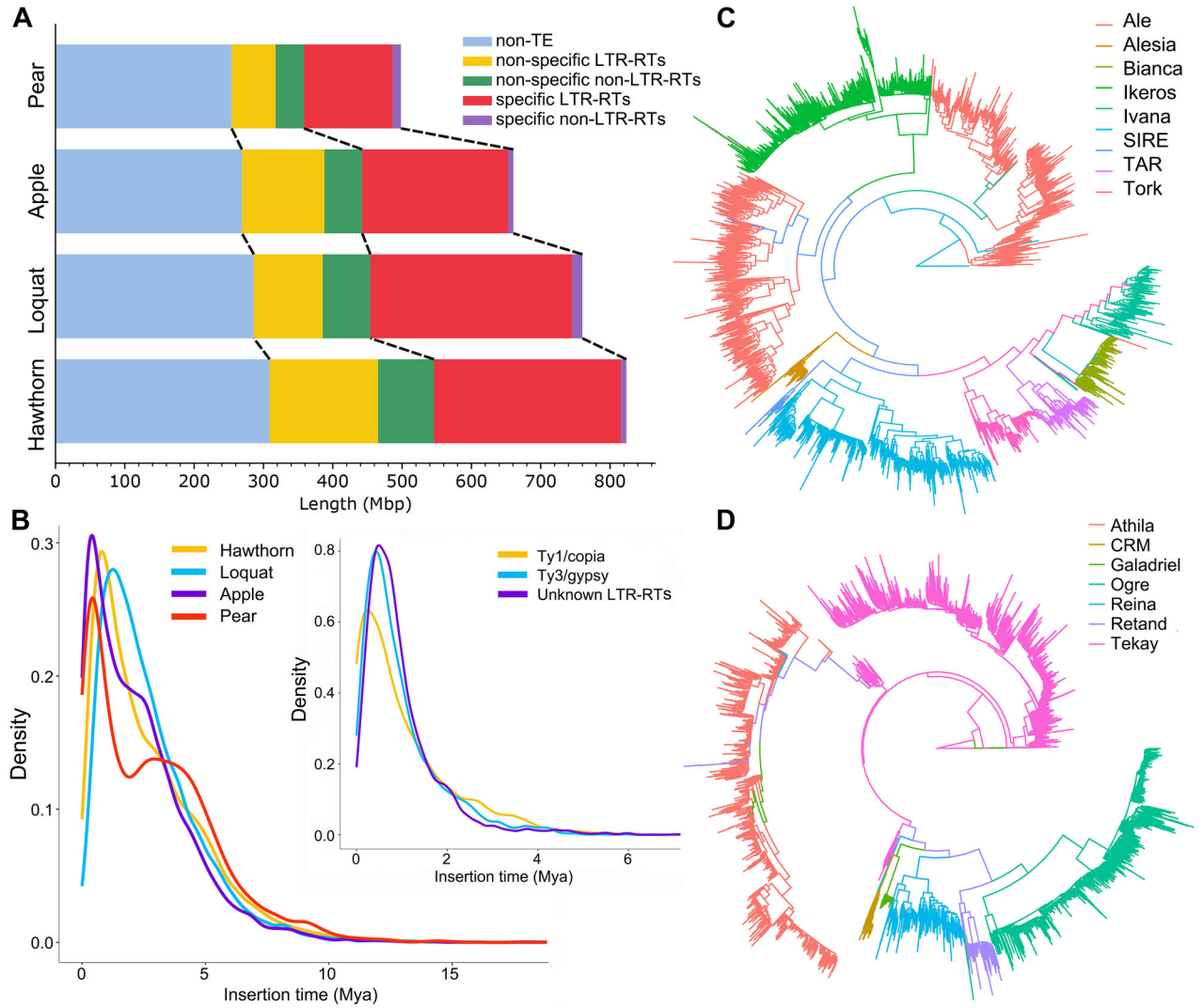

Figure 3. Evolutionary analysis of long terminal repeat retrotransposon (LTR-RT) families in the cultivated hawthorn genome
(A) Sequence size statistics for different types of transposable elements (TEs) in four species in the Maleae tribe. **(B)** Insertion times in the four species of LTR-RTs overall (left) as well as Ty1/copia, Ty3/gypsy, and an unknown LTR-RT (right). **(C, D)** Phylogenetic trees of Ty1/copia **(C)** and Ty3/gypsy **(D)** with the names of each gene family of LTR-RTs.

Maleae tribe exclusively shared a unique ancient tetraploidization event that occurred at K_s = ~0.15 (Figure 4A). This is consistent with previous studies (Shulaev et al., 2011; Shuang et al., 2020). Our results also indicate that no recent WGD event occurred specifically for hawthorn except for the tetraploidization in Maleae and that amplification of the LTR-RTs should have been a major cause of the hawthorn genome expansion. Consistent with this result, the genomic synteny analyses among the genomes of hawthorn, apple and loquat also indicated that these three species present a 1:1:1 syntenic depth ratio and have a strong genomic collinearity (Figure 4B – E).

Previous studies have assigned chromosomes of apple or pear into two sets of sub-genomes according to homologous relationships (Velasco et al., 2010; Jung et al., 2012; Li et al., 2019; Sun et al., 2020). Consistent with relationships among apple chromosomes detected in these reports, we found that apple chromosomes (Md3 vs. Md11, Md5 vs. Md10, Md9 vs. Md17, Md4/14/16 vs. Md6/12/13, and Md7/15 vs. Md1/2/8) had large syntenic blocks (Figures 4E, S9). With the well-assembled apple genome, we also found that Md13 not only has strong collinearity with Md16, but also has regions of collinearity with Md6 (Figure S9). Similarly, we also found two sets of chromosomes in each of hawthorn and loquat based on the genomic collinearity dot plot analysis. In the hawthorn genome, Cp2 versus Cp3, Cp5 versus Cp11, Cp13 versus Cp8, and Cp1/4/6/12/15 versus Cp7/9/10/14/16/17 had large syntenic blocks; and in the loquat genome, we detected large syntenic blocks between Ej1 versus Ej2, Ej11 versus Ej13, and Ej3/6/7/9/10/12 versus Ej4/5/8/14/15/16/17 (Figures 4C, D, S10). Similar to what was found in previous studies (Velasco et al., 2010; Jung et al., 2012), we ultimately assigned each (group) pair of

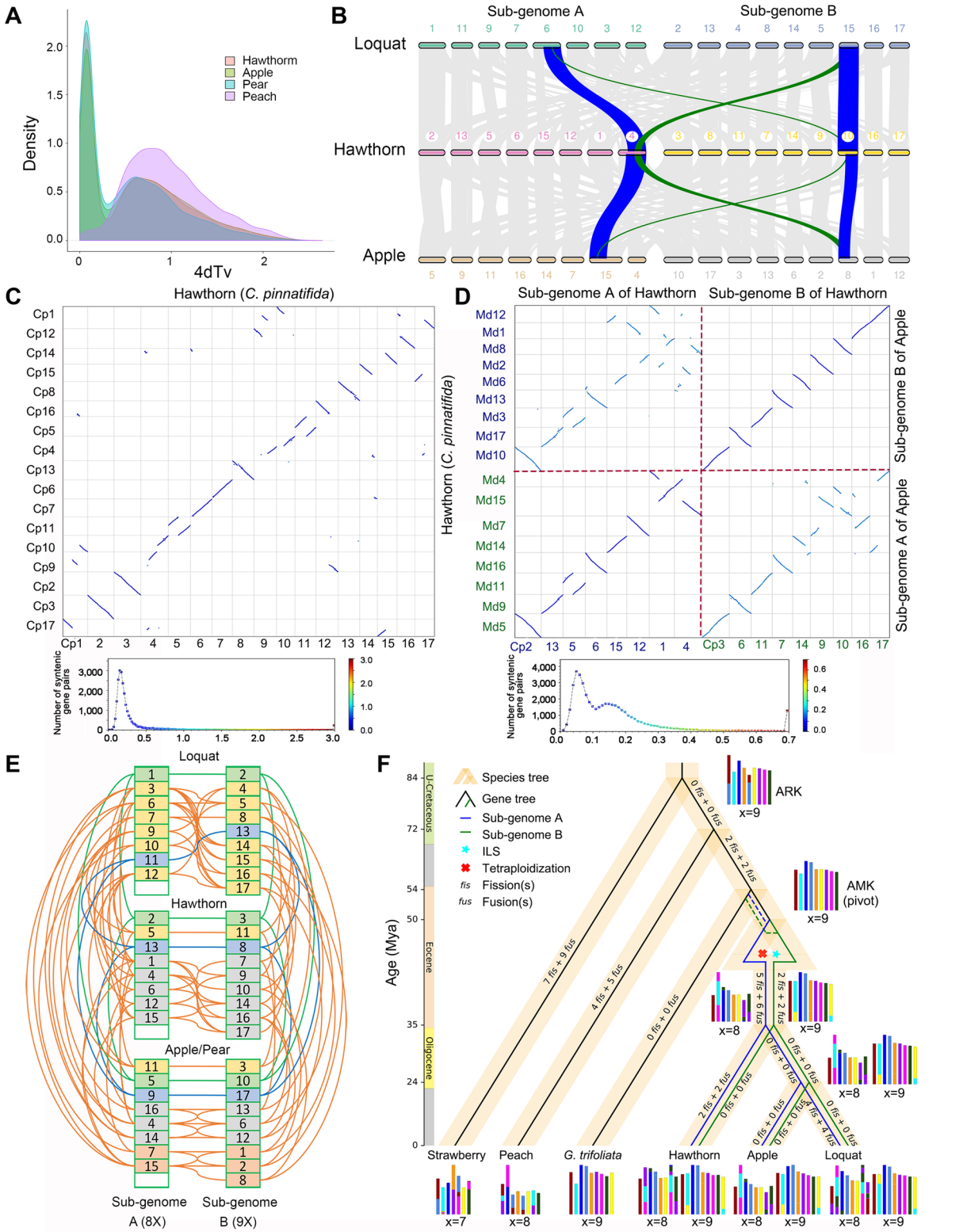

Figure 4. Syntenic analysis and ancestral chromosome reconstruction of the Rosaceae

(A) Age distribution of transversion substitutions at four-fold degenerate sites (4DTv) distance in hawthorn and three related species. **(B)** Synteny analyses among the genomes of hawthorn, apple, and loquat. Synteny blocks between paired chromosomes are connected by gray lines; one representative orthologous block (blue lines) and one out-paralogous block (green lines) are noted. (**C**, **D**) Syntenic dot plot and K_s distribution within the hawthorn genome (Cp; **C**) and between two sub-genomes of hawthorn and apple (Md; **D**). **(E)** Homologous collinear blocks identified among species in Maleae. The squares show the chromosomes and the different colors represent different homologous groups. The connecting lines denote homologous collinear blocks between chromosomes. **(F)** Reconstruction of Rosaceae ancestral chromosomes and evolutionary scenario of the chromosome changes that occurred in Maleae, with the gene tree embedded in the true species tree.

homologous chromosomes into two sub-genomes according to the homologous relationship of each chromosome among the three species (Figures 4D, E, S9, S10).

Sub-genome assignment could make the genomic relationships clear and help us to investigate the evolutionary pattern of the two sub-genomes. We classified collinear blocks into orthologs (in the same sub-genome between different species), in-paralogs (in a different sub-genome within the same species) and out-paralogs (in a different sub-genome between different species), corresponding to different evolutionary histories (species differentiation or WGDs) and divergence times (estimated using synonymous substitution rates, K_s). We observed that, among the different species, sub-genome A versus A or B versus B showed 1:1 orthologs (K_s = ~0.05, representing the speciation events). Within the same species, sub-genome A and B exhibited 1:1 in-paralogs when K_s peaked at around 0.15, corresponding to the shared recent tetraploidization event (Figures 4D, E, S9, S10; Table S13), but none of the homologous blocks were detected within the same sub-genome. Interestingly, although syntenic blocks were detected in sub-genome A versus B among different species, they were out-paralogs (K_s = ~0.15, also representing a shared recent tetraploidization event), but not orthologs, according to the distribution of K_s, suggesting that very few unbalanced rearrangements (e.g., fission and fusion) occurred between these two sub-genomes after speciation. The stability of sub-genomes may be attributed to low rates of homoeologous exchanges. It has been reported that although homoeologous exchanges can generate novel gene combinations and phenotypes, they may also lead to aberrant meiotic behavior in polyploids (Gaeta and Chris Pires, 2010). This phenomenon was also observed in allopolyploid *Cucumis hytivus* (cucumber) *and Eragrostis tef* (cereal teff), where no large-scale chromosomal rearrangements were identified between sub-genomes (Wang et al., 2017; VanBuren et al., 2020).

Furthermore, sub-genome A in Maleae exhibited more complex rearrangement events (e.g., chromosome inversion and translocation) than sub-genome B, which showed few rearrangements (Figures 4D, S10). This suggests that these two sub-genomes had different rearrangement mechanisms and that sub-genome B has an underlying mechanism for chromosome structural stability, for example, specific DNA maintenance or repair pathways. No obvious major inversions or genome rearrangements occurring between sub-genomes might also explain why we could divide genomes of Maleae into two sub-genomes. We consider the plausible explanation for this phenomenon to be that these two sub-genomes derive from two different ancestors, which experienced distinct rearrangement mechanisms. To test this hypothesis, we reconstructed ancestral chromosomes and investigated the chromosomal rearrangement history of Rosaceae.

Ancestral chromosome reconstruction and paleo-polyploid origin

We used the MGRA (Multiple Genome Rearrangements and Ancestors) method (Avdeyev et al., 2016) to reconstruct a hypothetical ancestral genome for Rosaceae. Our results identified an ancestral Rosaceae genome containing nine protochromosomes with 12,543 protogenes based on conserved gene colocations among the investigated genomes (Figure 4F), which was roughly consistent with previous reports, although our results contained more protogenes (Raymond et al., 2018; Groppi et al., 2021). The strawberry and peach chromosomes differentiated from the reconstructed Rosaceae ancestor (ARK) by seven fissions/nine fusions and four fissions/five fusions, respectively, which resulted in extensive chromosome changes in these two species (Figure 4F). However, fewer chromosomal rearrangements (two fissions/two fusions) were detected along the path from the ARK to the ancestral node (ancestral Maleae kindred, AMK, $n = 9$) of Maleae. To illustrate the evolutionary route of Maleae, we took two sub-genomes as different gene components and drew the gene trees embedded in the true species tree with AMK as a pivot. Consistent with the results of dot plot analysis, sub-genome A showed more rearrangement events than sub-genome B from AMK to extant species of Maleae (Figure 4F). The reconstructed progenitors and chromosomal rearrangement history further our understanding of Maleae evolution.

The origin of the ancestor of Maleae remains a subject of dispute. It has been proposed that members of the Maleae derived from a relative of the ancestor of *Gillenia trifoliata* ($x = 9$) via a within-species WGD (autopolyploidization) event (Velasco et al., 2010; Verde et al., 2013; Xiang et al., 2017; Wang et al., 2019). However, earlier studies suggested that a hybridization event (allopolyploidization) occurred between ancient sister species in Maleae, which possess chromosome numbers of 9 and 8, respectively (Chevreau et al., 1985; Robertson et al., 1991; Raspé and Sloover, 1998; Evans and Campbell, 2002). To provide a better resolution to this dispute, we conducted a phylogenetic analysis with single-copy genes from the two sub-genomes of apple (MdA, MdB) and hawthorn (CpA and CpB), as well as diploid *G. trifoliata*, the putative Maleae ancestor (Velasco et al., 2010). We identified 5,414 single-copy (1:1:1) orthologs and then constructed corresponding gene trees based on each protein sequence. Our results indicated that three major topologies (Types I, II, III) accounted for 87.57% (4,741/5,414) of the total phylogenetic trees (Figure 5A). The number of gene trees supporting the three major topologies were approximately 1:1:1 (1,668 vs. 1,562 vs. 1,511), with Type I being slightly more prevalent than the other two topologies (Figure 5A, B). We then mapped all the single-copy genes supporting the three topologies on the nine chromosomes of *G. trifoliata* and found that these genes were evenly distributed on the chromosomes (Figure 5C). This incongruence phenomenon among single-gene trees indicated that incomplete lineage sorting (ILS) and/or hybridization events may have occurred during early speciation in Maleae. Considering that Type II and Type III, which conflicted with the most likely species tree (Figure 1D), had equal proportions of supporting gene trees, we assumed that ILS should play an important role in the gene tree incongruence pattern. Then, we used the ILS test model (Degnan and

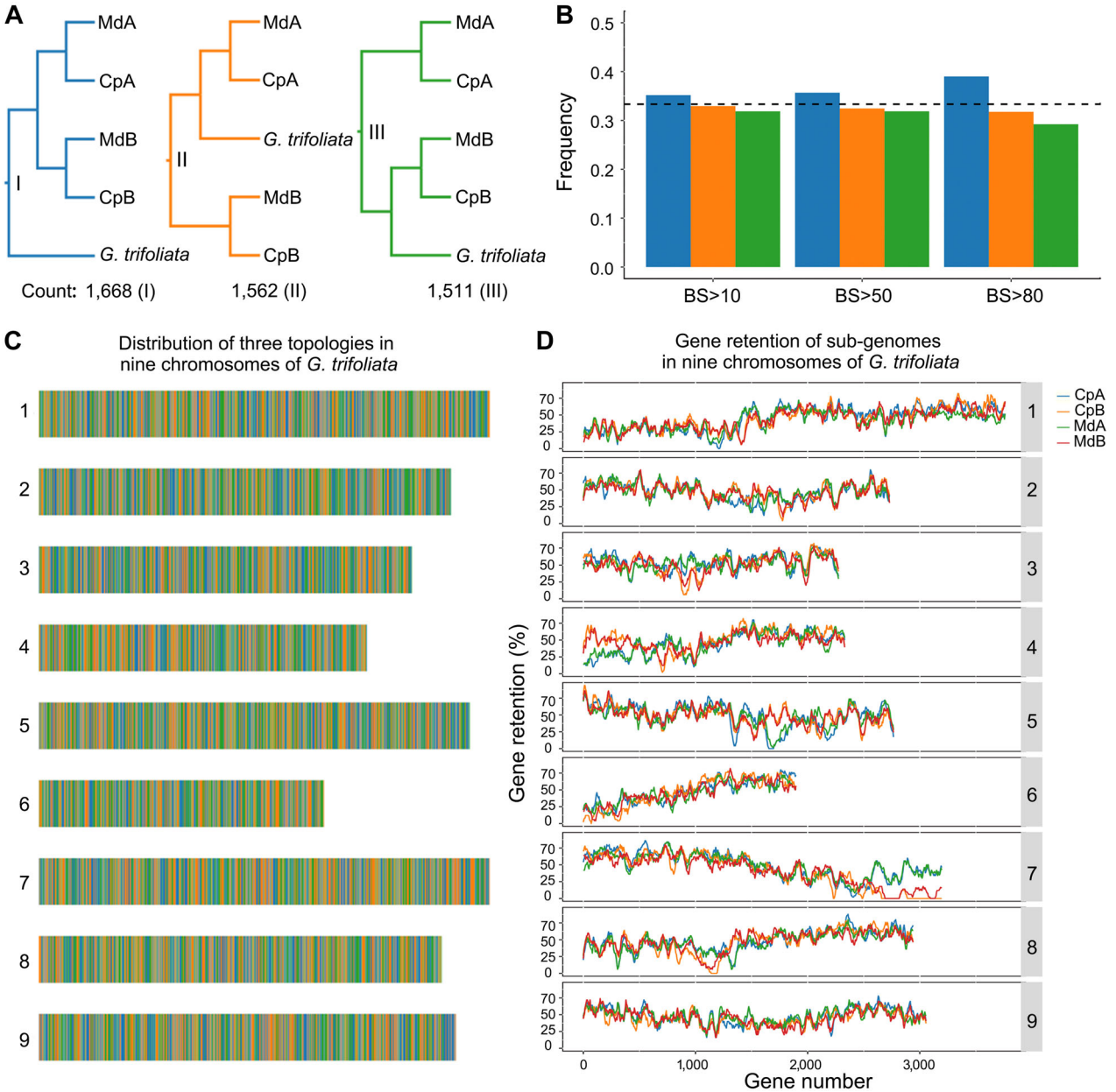

Figure 5. Gene tree discordances among sub-genomes of hawthorn, apple, and diploid *Gillenia trifoliata*

(A) Three major topologies supported by gene trees (count: number of gene trees). **(B)** The proportion of trees supporting the three topologies at different bootstrap (BS) levels. The column colors correspond to the colors of the three tree types in **(A)**. **(C)** Distribution of the single-copy genes in the three tree types on the nine chromosomes of *G. trifoliata*.**(D)** Gene retention pattern in the sub-genomes of apple and hawthorn corresponding to nine chromosomes of *G. trifoliata*. The *Y* axis indicates the percentage of gene retention in sub-genomes of hawthorn and apple corresponding to *G. trifoliata* chromosomes. It was calculated with 500-gene sliding windows along each *G. trifoliata* chromosome.

Rosenberg, 2009) to compare observed gene tree frequencies with expected ones under ILS for each topology. The χ^2 statistics results indicated that there was no significant difference between them (P-value = 0.045), suggesting that conflicting phylogenetic signals were likely to have arisen under ILS.

If evolutionary origins of polyploidy in Maleae are inferred from interspecific hybridization, then one sub-genome should retain significantly more genes than the other "submissive" sub-genomes as reported in previous studies (Schnable et al., 2011; Bird et al., 2018). Therefore, in order to detect if there is a bias in gene retention of sub-genomes, we calculated the gene retention ratios of the two sub-genomes corresponding to the nine chromosomes of *G. trifoliata*. The distribution results indicated that most locations showed no obvious gene retention bias between sub-genomes A and B along each chromosome, while some regions (e.g., the start of Chr. 4 and end of Chr. 7) showed weak gene retention bias (Figure 5D). This is roughly consistent with the observation of

unbiased fractionation between the two sub-genomes in pear (Li et al., 2019).

Thus, the above results suggest that the extant species in Maleae are likely to have derived from an ancestor species via a within-species WGD (autopolyploidization) event as reported in previous studies (Li et al., 2019). However, the different rearrangement mechanisms and weak gene retention bias in several genome regions between the two sub-genomes observed in our study, as well as a biased evolution pattern found in the singletons and homeologs within each sub-genome in pear (Li et al., 2019), could not be perfectly explained. Our results therefore could not rule out a hybrid origin of Maleae between two very close ancestral relatives of *Gillenia* leading to allopolyploidization. In reality, an increasing number of studies have shown that complex evolutionary processes affect the genome in polyploids and that there should be a continuum between two theoretical extremes (strict allopolyploidy and autopolyploidy) (Ramsey and Schemske, 2002; Bomblies, 2020), which might be the case with Maleae. Future studies on the other possible ancestor of Maleae could provide new clues for the origin of Maleae.

MATERIALS AND METHODS

Sample collection and sequencing

Healthy young leaves of the hard-fleshed hawthorn (*Crataegus pinnatifida* var. major) cultivar "Qiu Jinxing" were collected from the National Hawthorn Germplasm Repository of China (https://www.cgris.net/query/croplist.php, identification number: SZP016) at Shenyang Agricultural University (Figure S1) for high-quality genomic DNA extraction. For Nanopore sequencing, a size of 30–80 kb genomic DNA was selected with BluePippin (Sage Science) and processed according to the Ligation Sequencing Kit 1D protocol (SQK-LSK109). The final library was sequenced on R9.4 flow cells using the PromethION DNA sequencer (Oxford Nanopore Technologies, NY, USA). Base-calling was completed on the PromethION instrument using MinKnow ver. 2.2.

For Illumina sequencing, paired-end (PE) libraries with insert sizes of 400 bp were constructed and sequenced on the Illumina HiSeq X Ten platform. These short reads were used for genome information estimation, genome assembly correction and evaluation. Based on the Illumina reads, genome size and heterozygosity of the hawthorn were estimated using *k*-mer statistics (Liu et al., 2013). Previously released transcriptome data were used for genome annotation (Xu et al., 2016; Guo et al., 2018). For Hi-C sequencing, the library preparation procedure was conducted as previously described (Lieberman-Aiden et al., 2009). The Hi-C libraries were controlled for quality and sequenced on the Illumina HiSeq platform.

Genome assembly and annotation

For *de novo* genome assembly, Nanopore reads with an average quality score higher than seven were retained and further corrected with NextDenovo (https://github.com/Nextomics/NextDenovo). These reads were assembled into contigs by wtdbg v2.4 (Ruan and Li, 2020). The draft genome assembly was polished with Pilon v1.22 (Walker et al., 2014) using the Illumina short reads. To evaluate the accuracy of assembly, BWA (burrows wheel aligner) software (Li and Durbin, 2009) was used to map all the Illumina paired-end reads to the assembled genome and SAMtools v0.1.1855 (Li and Hua, 2009) was used to evaluate the mapping rate and genome coverage of sequencing reads. Next, the Hi-C paired-end clean reads were aligned to the assembled contigs with BWA-mem (Li and Durbin, 2009) and then clustered onto 17 chromosomes with LACHESIS software (http://shendurelab.github.io/LACHESIS/). The integrity of the genome assembly was evaluated using the BUSCO method (Simão et al., 2015).

Three gene prediction methods (*de novo*-based, RNA-seq-based, and homolog-based) were used in combination to identify the protein-coding genes. For *de novo*-based prediction, Augustus v2.4 (Stanke and Waack, 2003; Stanke et al., 2008) and GlimmerHMM v3.0.4 (Majoros et al., 2004) with default parameters were used for gene prediction. For the RNA-seq-based prediction, GeneMark-ST v5.1 (Tang et al., 2015) and PASA v2.0.2 (Mount et al., 2006) were used. For homology-based predictions, protein sequences of 10 species, *Arabidopsis thaliana*, *Oryza sativa*, *Pyrus communis*, *Prunus persica*, *Fragaria vesca*, *Fragaria iinumae*, *Rosa chinensis*, *Rubus occidentalis*, *Potentilla micrantha* and *Malus* × *domestica*, were used as references. Finally, EVM v1.1.1 (Haas et al., 2008) was used to integrate the results of the three methods. All genes were annotated by aligning with the NR database, Swiss-Prot, KEGG database (release 84.0). Then, the predicted genes were annotated using the InterPro database and InterProScan (Quevillon et al., 2005) software package.

Gene family expansion and contraction

To investigate the evolutionary position of cultivated hawthorn (*Crataegus pinnatifida* var. major), gene family clustering analysis was performed using OrthoFinder2 (Emms and Kelly, 2018) on the protein-coding genes of hawthorn and 16 additional sequenced Rosales species (*Dryas drummondii*, *Gillenia trifoliata*, *Eriobotrya japonica*, *Malus* × *domestica*, *Pyrus communis*, *Prunus persica*, *Fragaria vesca*, *Rubus chingii*, *Rosa chinensis*, *Potentilla micrantha*, *Morus notabilis*, *Cannabis sativa*, *Ziziphus jujuba*, *Ficus hispida*, *Hippophae rhamnoides*, and *Boehmeria nivea*) with *Vitis vinifera* as an outgroup. Expansions and contractions of orthologous gene families were detected using CAFE v3.0 (Han et al., 2013). The significantly expanded and contracted gene families were functionally annotated with GO and KEGG enrichment levels.

Phylogenetic tree construction

In total, 1,194 orthogroups with at least 88.9% of species having single-copy genes in any orthogroup were selected and aligned by MAFFT (Katoh and Standley, 2013). All the

aligned protein sequences were merged. A phylogenetic tree was constructed using IQ-TREE (Nguyen et al., 2014) with the JTT + F + R3 model and 1,000 bootstraps. Finally, the divergence times were estimated based on one-to-one orthologs using a Bayesian method implemented in MCMCtree of the PAML 4.9 package (Yang, 2007) with the options "independent rates" and "GTR" model. Using a burn-in of 1,000 iterations, a Markov Chain Monte Carlo analysis was run for 10,000 generations. Five fossil records were used as time-calibrated points, including the Crown Rosales (C5: 106.5–90.0 Mya), Stem Prunus (C3: >55 Mya), Stem Maleae (C2: >47.8 Mya), Stem Crataegus (C1: >33.9 Mya) and Stem Rosa (C4: >47.8 Mya) (Xiang et al., 2017; Zhang et al., 2017; Silvestro et al., 2021). Due to the lack of fossils at the root of our phylogenetic tree, we used the estimated time (C6: 107–135 Mya) in Timetree (Kumar et al., 2017) for secondary calibration.

Transcriptomic analysis of cultivated hawthorn fruits

Our co-authors previously published transcriptomic sequencing data for hard- ("Qiu Jinxing," NCBI SRA number: SAMN05607047, SAMN05607049, SAMN05607041, SAMN05607052) and soft-fleshed ("Ruanrou Shanlihong #3," NCBI SRA number: SAMN05607044, SAMN05607114, SAMN05607090, SAMN05607054) hawthorn cultivars (Xu et al., 2016). We reanalyzed these transcriptome data based on our assembled reference genome in this study. The mapped fragments for each gene were counted and genes with averaged transcripts per million (TPM) ≥1 were considered to be expressed. Hisat v2 (Kim et al., 2015) was used to compare the sequences with the reference genome, and Stringtie (Pertea et al., 2015) was used to quantify the expression of genes and transcripts. Hierarchical clustering and heatmaps of expressed genes among fruit ripening stages in the two fleshed fruit hawthorn types were generated using the Pheatmap package in R.

Repetitive elements identification

Transposable element annotation was conducted using the Extensive *de novo* TE Annotator (EDTA) pipeline (Ou et al., 2019). The --step option was set to "all" to run the entire annotation pipeline of the software. The --sensitive option was set to "1" to detect additional TEs using the EDTA RepeatModeler tool (Tarailo-Graovac and Chen, 2009). The –anno option was set to "1" to conduct whole-genome annotation of the TEs. For species-specific LTR-RT identification, classification and insert age estimation, the package SubPhaser (Jia et al., 2022) was used ($-q = 100$). The species-specific *k*-mers were identified first, then an LTR-RT library was constructed through scanning the assembled genome by calling LTR harvest (Ellinghaus et al., 2008) and LTR Finder (Xu and Wang, 2007). The clade level of LTR-RTs was classified by TEsorter (https://github.com/zhangrengang/TEsorter) (Zhang et al., 2022) and the LTR-RT protein domains (INT, RH and RT domains) were identified and extracted. LTR-RT protein sequences were aligned with Mafft (Katoh and Standley, 2013) and then merged into one sequence. The phylogenetic tree was built with FastTree (Price, 2009) and visualized with ggtree (Yu et al., 2017). The genetic distance between 5′ and 3′ LTR-RTs was estimated using the JC69 model and the insertion time was estimated according to: $T = d/2\mu$, where T is time, d is the genetic distance, and μ is the substitution rate (1.3E−8 per site per year).

Whole-genome synteny and duplication analysis

To identify syntenic regions within and between genomes, protein sequences of three species were aligned against themselves and each other using BLASTP (v2.2.31). Then, credible collinear blocks within these genomes were detected using MCScanX (Wang et al., 2012). The amino acid alignments were reverse-translated to the corresponding codon-based nucleotide alignments using PAL2NAL (Suyama et al., 2006), and K_aK_s_Caculator 2.0 (Wang et al., 2010) was used to calculate the nonsynonymous (K_a) and synonymous (K_s) substitution rates. Then, the 4DTv distance was calculated. The values of all gene pairs were plotted to determine putative WGD events and divergence between each two species.

Ancestral chromosome reconstruction of Rosaceae

Although increasing numbers of genome sequences for Rosaceae are available, only some of these are of sufficient quality for large-scale synteny analyses in ancestral genome reconstruction. Therefore, the dot plot method was used to select well-assembled genomes with fewer assembly errors and better chromosome continuity in Rosaceae for ancestral genome reconstruction. The species were *C. pinnatifida*, *E. japonica*, *M. domestica*, *G. trifoliata*, *P. persica*, and *F. vesca*. Only the genome of *M. domestica* was used to represent the clade of *Malus* and *Pyrus* because these two genomes showed a high degree of collinearity. To simplify the homologous collinearity between species, only one gene of a tandem repeat cluster was kept and the others were discarded (i.e., the tandem gene cluster was considered a gene). The SubPhaser pipeline (Jia et al., 2022) was also used for partition and phase sub-genomes using repetitive *K*-mers as the "differential signatures."

Then, MRGA2 (Avdeyev et al., 2016) was used to infer the ancestral chromosome, which supported the gene gain or loss. This method is restricted to genomes with equal gene content, that is, to cases in which each gene is present in every genome in exactly one copy. Since species in Maleae shared a polyploidization event, this suggests that only a small number of single-copy genes could be used (Jung et al., 2012). Therefore, we analyzed the two sub-genomes separately in order to identify more orthologs in each sub-genome for ancestral genome reconstruction and investigated the evolutionary pattern of the two sub-genomes.

Data deposition

The raw genomic reads generated in this study have been deposited in the National Center for Biotechnology

Information Sequence Read Archive (BioProject PRJNA823924). The genome assembly and annotation files are available at the Genome Database for Rosaceae (https://www.rosaceae.org/).

ACKNOWLEDGEMENTS

This work is supported by grants from National Natural Science Foundation of China (32060237 to T.Z. and 32060085 to Q.Q.). 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).

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

T.Z., Q.Q., and W.D. conceived and designed the study; X.D., X.Z., Y.H., X.W., C.S., and W.D. prepared the materials; Q.Q., T.Z., X.D., R.Z., Q.Y., and Y.V.P. conducted the experiments, analyzed data, and prepared the results; T.Z., Q.Q., and M.J. C.C. wrote and improved the manuscript. All authors approved the final manuscript.

Edited by: Xuehui Huang, Shanghai Normal University, China

Received Apr. 18, 2022; **Accepted** Jun. 13, 2022; **Published** Jun. 24, 2022

OO: OnlineOpen

REFERENCES

Avdeyev, P., Jiang, S., Aganezov, S., Hu, F., and Alekseyev, M.A. (2016). Reconstruction of ancestral genomes in presence of gene gain and loss. J. Comput. Biol. **23:** 150–164.

Bennetzen, J.L. (2002). Mechanisms and rates of genome expansion and contraction in flowering plants. Genetica **115:** 29–36.

Bird, K.A., Vanburen, R., Puzey, J.R., and Edger, P.P. (2018). The causes and consequences of subgenome dominance in hybrids and recent polyploids. New Phytol. **220:** 87–93.

Bomblies, K. (2020). When everything changes at once: Finding a new normal after genome duplication. Proc. Biol. Sci. **287:** 20202154.

Chevreau, E., Lespinasse, Y., and Gallet, M. (1985). Inheritance of pollen enzymes and polyplod origin of apple (*Malus x domestica* Borkh.). Theor. Appl. Genet. **71:** 268–277.

Cloud, A.M.E., Vilcins, D., and Mcewen, B.J. (2020). The effect of hawthorn (*Crataegus* spp.) on blood pressure: A systematic review. Adv. Integr. Med. **7:** 167–175.

Degnan, J.H., and Rosenberg, N.A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends. Ecol. Evol. **24:** 332–340.

Dong, W., Li, Z.X. (2015). *The Science and Practice of Chinese Fruit Tree: Hawthorn*. Science Press, Shanxi.

Du, X., Zhang, X., Bu, H., Zhang, T., Lao, Y., and Dong, W. (2019). Molecular analysis of evolution and origins of cultivated hawthorn (*Crataegus* spp.) and related species in China. Front. Plant Sci. **10:** 443.

Edwards, J.E., Brown, P.N., Talent, N., Dickinson, T.A., and Shipley, P. R. (2012). A review of the chemistry of the genus *Crataegus*. Phytochemistry **79:** 5–26.

Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. BMC Bioinform. **9:** 18.

Emms, D.M., and Kelly, S. (2018). OrthoFinder2: Fast and accurate phylogenomic orthology analysis from gene sequences. BioRxiv.

Evans, R.C., and Campbell, C.S. (2002). The origin of the apple subfamily (Maloideae; Rosaceae) is clarified by DNA sequence data from duplicated GBSSI genes. Am. J. Bot. **89:** 1478–1484.

Gaeta, R.T., and Chris Pires, J. (2010). Homoeologous recombination in allopolyploids: The polyploid ratchet. New Phytol. **186:** 18–28.

Goldblatt, P. (1976). Cytotaxonomic studies in the tribe quillajeae (Rosaceae). Ann. Mo. Bot. Gard. **63:** 200–206.

Groppi, A., Liu, S., Cornille, A., Decroocq, S., Bui, Q.T., Tricon, D., Cruaud, C., Arribat, S., Belser, C., Marande, W., Salse, J., Huneau, C., Rodde, N., Rhalloussi, W., Cauet, S., Istace, B., Denis, E., Carrère, S., Audergon, J.M., Roch, G., Lambert, P., Zhebentyayeva, T., Liu, W.S., Bouchez, O., Lopez-Roques, C., Serre, R.F., Debuchy, R., Tran, J., Wincker, P., Chen, X., Pétriacq, P., Barre, A., Nikolski, M., Aury, J.M., Abbott, A.G., Giraud, T., and Decroocq, V. (2021). Population genomics of apricots unravels domestication history and adaptive events. Nat. Commun. **12:** 3956

Guo, W., Guo, N., Li, W., and Dai, H. (2018). Transcriptome analysis reveals the hawthorn response to the infection of apple chlorotic leaf spot virus. Sci. Hortic **239:** 171–180.

Haas, B.J., Salzberg, S.L., Wei, Z., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R., and Wortman, J.R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. **9:** R7.

Han, J.Y., In, J.G., Kwon, Y.S., and Choi, Y.E. (2010). Regulation of ginsenoside and phytosterol biosynthesis by RNA interferences of squalene epoxidase gene in *Panax ginseng*. Phytochemistry **71:** 36–46.

Han, M.V., Thomas, G.W.C., Lugomartinez, J., and Hahn, M.W. (2013). Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. Mol. Biol. Evol. **30:** 1987–1997.

Jia, K.H., Wang, Z.X., Wang, L., Li, G.Y., Zhang, W., Wang, X.L., Xu, F.J., Jiao, S.Q., Zhou, S.S., Liu, H., Ma, Y., Bi, G., Zhao, W., ElKassaby, Y.A., Porth, I., Li, G., Zhang, R.G., and Mao, J.F. (2022). SubPhaser: A robust allopolyploid subgenome phasing method based on subgenome-specific k-mers. New Phytol. **235:** 801–809.

Jiang, S., An, H., Xu, F., and Zhang, X. (2020). Chromosome-level genome assembly and annotation of the loquat (*Eriobotrya japonica*) genome. GigaScience **9:** giaa0 15.

Jung, S., Cestaro, A., Troggio, M., Main, D., Zheng, P., Cho, I., Folta, K. M., Sosinski, B., Abbott, A., Celton, J.M., Arús, P., Shulaev, V., Verde, I., Morgante, M., Rokhsar, D., Velasco, R., and Sargent, D.J. (2012). Whole genome comparisons of *Fragaria*, *Prunus* and *Malus* reveal different modes of evolution between Rosaceous subfamilies. BMC Genomics **13:** 129.

Jung, S., Lee, T., Cheng, C.-H., Buble, K., Zheng, P., Yu, J., Humann, J., Ficklin, S.P., Gasic, K., Scott, K., Frank, M., Ru, S., Hough, H., Evans, K., Peace, C., Olmstead, M., DeVetter, L.W., McFerson, J., Coe, M., Wegrzyn, J.L., Staton, M.E., Abbott, A.G., and Main, D. (2019). 15 years of GDR: New data and functionality in the Genome Database for Rosaceae. Nucleic Acids Res. **47:** D1137–D1145.

Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol. Biol. Evol. **30:** 772–780.

Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: A fast spliced aligner with low memory requirements. Nat. Methods **12:** 357–360.

Kumar, S., Stecher, G., Suleski, M., and Hedges, S.B. (2017). TimeTree: A resource for timelines, timetrees, and divergence times. Mol. Biol. Evol. **34:** 1812–1819.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25:** 1754–1760.

Li, M., Xiao, Y., Mount, S., and Liu, Z. (2021). An atlas of genomic resources for studying Rosaceae fruits and ornamentals. Front. Plant Sci. **12:** 644881.

Li, Q., Qiao, X., Yin, H., Zhou, Y., Dong, H., Qi, K., Li, L., and Zhang, S. (2019). Unbiased subgenome evolution following a recent whole-genome duplication in pear (*Pyrus bretschneideri* Rehd.). Hortic. Res. **6:** 34.

Li, Z., and Hua, L. (2009). The sequence alignment-map format and SAMtools. Bioinformatics **25:** 2078–2079.

Lieberman-Aiden, E., Berkum, N.V., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., and Dorschner, M.O. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science **326:** 289.

Linsmith, G., Rombauts, S., Montanari, S., Deng, C.H., Celton, J.M., Guérif, P., Liu, C., Lohaus, R., Zurn, J.D., Cestaro, A., Bassil, N.V., Bakker, L.V., Schijlen, E., Gardiner, S.E., Lespinasse, Y., Durel, C. E., Velasco, R., Neale, D.B., Chagné, D., Van de Peer, Y., Troggio, M., and Bianco, L. (2019). Pseudo-chromosome-length genome assembly of a double haploid "Bartlett" pear (*Pyrus communis* L.). Gigascience **8:** giz138.

Liu, B., Shi, Y., Yuan, J., Hu, X., Zhang, H., Li, N., Li, Z., Chen, Y., Mu, D., and Fan, W. (2013). Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. Quant. Biol. **35:** 62–67.

Lo, E.Y.Y., Stefanović, S., Christensen, K.I., and Dickinson, T.A. (2009). Evidence for genetic association between East Asian and western North American *Crataegus* L. (Rosaceae) and rapid divergence of the eastern North American lineages based on multiple DNA sequences. Mol. Biol. Evol. **51:** 157–168.

Lunkenbein, S., Coiner, H., Vos, C., Schaart, J.G., and Salentijn, E. (2006). Molecular characterization of a stable antisense chalcone synthase phenotype in strawberry (*Fragaria x ananassa*). J. Agric. Food Chem. **54:** 2145–2153.

Majoros, W.H., Pertea, M., and Salzberg, S.L. (2004). TigrScan and GlimmerHMM: Two open source ab initio eukaryotic gene-finders. Bioinformatics **20:** 2878–2879.

Mehdi, Z., Nadia, T., Maria, K., Jeanette, L., Jensen, L., Shipley, P.R., Saša, S., and Dickinson, T.A. (2015). DNA barcodes from four loci provide poor resolution of taxonomic groups in the genus *Crataegus*. AoB Plants **7:** plv0 45.

Mount, S.M., Hamilton, J.P., Haas, B.J., Campbell, M.A., and Robin, B.C. (2006). Comprehensive analysis of alternative splicing in rice and comparative analyses with *Arabidopsis*. BMC Genomics **7:** 327.

Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2014). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. **32:** 268–274.

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J.R.A., Hellinga, A.J., Lugo, C.S.B., Elliott, T.A., Ware, D., Peterson, T., Jiang, N., Hirsch, C.N., and Hufford, M.B. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. **20:** 275.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. **33:** 290–295.

Phipps, J.B. (2015). *Crataegus*. Flora of North America North of Mexico . Oxford University Press, New York and Oxford. pp. 491–643.

Price, M.N. (2009). FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. Mol. Biol. Evol. **26:** 1641–1650.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., and Lopez, R. (2005). InterProScan: Protein domains identifier. Nucleic Acids Res. **33:** W116–W120.

Ramsey, J., and Schemske, D.W. (2002). Neopolyploidy in flowering plants. Annu. Rev. Ecol. Syst. **33:** 589–639.

Raspé, O., and Sloover, J. (1998). Isozymes in *Sorbus aucuparia* (Rosaceae: Maloideae): Genetic analysis and evolutionary significance of zymograms. Int. J. Plant Sci. **159:** 627–636.

Raymond, O., Gouzy, J., Just, J., Badouin, H., Verdenaud, M., Lemainque, A., Vergne, P., Moja, S., Choisne, N., Pont, C., Carrère, S., Caissard, J.C., Couloux, A., Cottret, L., Aury, J.M., Szécsi, J., Latrasse, D., Madoui, M.A., François, L., Fu, X., Yang, S.H., Dubois, A., Piola, F., Larrieu, A., Perez, M., Labadie, K., Perrier, L., Govetto, B., Labrousse, Y., Villand, P., Bardoux, C., Boltz, V., Lopez-Roques, C., Heitzler, P., Vernoux, T., Vandenbussche, M., Quesneville, H., Boualem, A., Benhamed, M., Wincker, P., and Bendahmane, M. (2018). The Rosa genome provides new insights into the domestication of modern roses. Nat. Genet. **50:** 772–777.

Robertson, K.R., Phipps, J.B., Rohrer, J.R., and Smith, P.G. (1991). A synopsis of genera in Maloideae (Rosaceae). Syst. Bot. **16:** 376–394.

Ruan, J., and Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. Nat. Methods **17:** 155–158.

Schnable, J.C., Springer, N.M., and Freeling, M. (2011). Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc. Natl. Acad. Sci. U.S.A. **108**: 4069–4074.

Shuang, J., Haishan, A., Fangjie, X., and Xueying, Z. (2020). Chromosome-level genome assembly and annotation of the loquat (*Eriobotrya japonica*) genome. Gigaence **9:** 1–9.

Shulaev, V., Sargent, D.J., Crowhurst, R.N., Mockler, T.C., Folkerts, O., Delcher, A.L., Jaiswal, P., Mockaitis, K., Liston, A., and Mane, S.P. (2011). The genome of woodland strawberry (*Fragaria vesca*). Nat. Genet. **43:** 109–116.

Silvestro, D., Bacon, C.D., Ding, W., Zhang, Q., Donoghue, P.C.J., Antonelli, A., and Xing, Y. (2021). Fossil data support a pre-Cretaceous origin of flowering plants. Nat. Ecol. Evol. **5:** 449–457.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics **31:** 3210–3212.

Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. Bioinformatics **24:** 637–644.

Stanke, M., and Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics **19:** 215–225.

Sun, X., Jiao, C., Schwaninger, H., Chao, C.T., Ma, Y., Duan, N., Khan, A., Ban, S., Xu, K., Cheng, L., Zhong, G.Y., and Fei, Z. (2020). Phased diploid genome assemblies and pan-genomes provide insights into the genetic history of apple domestication. Nat. Genet. **52:** 1423–1432.

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. **34:** W609–W612.

Tang, S., Lomsadze, A., and Borodovsky, M. (2015). Identification of protein coding regions in RNA transcripts. Nucleic Acids Res. **43:** e78.

Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. Curr. Protocol. Bioinformat. **Chapter 4**: 4.10.11–14.10.14.

Van de Peer, Y., Maere, S., and Meyer, A. (2009). The evolutionary significance of ancient genome duplications. Nat. Rev. Genet. **10:** 725–732.

VanBuren, R., Man Wai, C., Wang, X., Pardo, J., Yocca, A.E., Wang, H., Chaluvadi, S.R., Han, G., Bryant, D., Edger, P.P., Messing, J., Sorrells, M.E., Mockler, T.C., Bennetzen, J.L., and Michael, T.P. (2020). Exceptional subgenome stability and functional divergence in the allotetraploid Ethiopian cereal teff. Nat. Commun. **11:** 884.

Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., Fontana, P., Bhatnagar, S.K., Troggio, M. (2010). The genome of the domesticated apple (*Malus* × *domestica* Borkh.). Nat. Genet. **42:** 833–839.

Verde, I., Abbott, A.G., Scalabrin, S., Jung, S., Shu, S., Marroni, F., Zhebentyayeva, T., Dettori, M.T., Grimwood, J. (2013). The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. Nat. Genet. **45:** 487–494.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., and Young, S.K. (2014). Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS ONE **9:** e112963.

Wang, D., Zhang, Y., Zhang, Z., Zhu, J., and Yu, J. (2010). KaKs_Calculator 2.0: A toolkit incorporating gamma-series methods and sliding window strategies. Genom. Proteom. Bioinform. **8:** 77–80.

Wang, J., Qin, J., Sun, P., Ma, X., and Wang, X. (2019). Polyploidy index and its implications for the evolution of polyploids. Front. Genet. **10:** 807.

Wang, L., Lei, T., Han, G., Yue, J., Zhang, X., Yang, Q., Ruan, H., Gu, C., Zhang, Q., Qian, T., Zhang, N., Qian, W., Wang, Q., Pang, X., Shu, Y., Gao, L., and Wang, Y. (2021). The chromosome-scale reference genome of *Rubus chingii* Hu provides insight into the biosynthetic pathway of hydrolyzable tannins. Plant J. **107:** 1466–1477.

Wang, Y., Tang, H., Debarry, J.D., Tan, X., Li, J., Wang, X., Lee, T.H., Jin, H., Marler, B., and Guo, H. (2012). MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. **40:** e49.

Wang, Y., Zhao, Q., Qin, X., Yang, S., Li, Z., Li, J., Lou, Q., and Chen, J. (2017). Identification of all homoeologous chromosomes of newly synthetic allotetraploid *Cucumis* × *hytivus* and its wild parent reveals stable subgenome structure. Chromosoma **126:** 713–728.

Xiang, Y., Huang, C.-H., Hu, Y., Wen, J., Li, S., Yi, T., Chen, H., Xiang, J., and Ma, H. (2017). Evolution of Rosaceae fruit types based on nuclear phylogeny in the context of geological times and genome duplication. Mol. Biol. Evol. **34:** 262–281.

Xu, J., Zhao, Y., Zhang, X., Zhang, L., Hou, Y., and Dong, W. (2016). Transcriptome analysis and ultrastructure observation reveal that hawthorn fruit softening is due to cellulose/hemicellulose degradation. Front. Plant Sci. **7:** 1524.

Xu, Z., and Wang, H. (2007). LTR_FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. **35:** W265–W268.

Yang, B., and Liu, P. (2012). Composition and health effects of phenolic compounds in hawthorn (*Crataegus* spp.) of different origins. J. Sci. Food Agric. **92:** 1578–1590.

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. **24:** 1586–1591.

Yeh, S.Y., Huang, F.C., Hoffmann, T., Mayershofer, M., and Schwab, W. (2014). FaPOD27 functions in the metabolism of polyphenols in strawberry fruit (*Fragaria* sp.). Front. Plant Sci. **5:** 518.

Yu, G., Smith, D.K., Zhu, H., Guan, Y., and Lam, T.T.Y. (2017). ggtree: An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol. Evol. **8:** 28–36.

Zhang, H., Cui, S.X., and Zhang, W. (1994). The dynamic changes of flavonols in hawthorn fruits during growing and developing periods. J. GanSu Sci. **23:** 43–46 (in Chinese).

Zhang, L., Hu, J., Han, X., Li, J., Gao, Y., Richards, C.M., Zhang, C., Tian, Y., Liu, G., Gul, H., Wang, D., Tian, Y., Yang, C., Meng, M., Yuan, G., Kang, G., Wu, Y., Wang, K., Zhang, H., Wang, D., and Cong, P. (2019). A high-quality apple genome assembly reveals the association of a retrotransposon and red fruit colour. Nat. Commun. **10:** 1494.

Zhang, R.-G., Li, G.Y., Wang, X.L., Dainat, J., Wang, Z.X., Ou, S., and Ma, Y. (2022). TEsorter: An accurate and fast method to classify LTR-retrotransposons in plant genomes. Hortic. Res. **9:** uhac017.

Zhang, S., Jin, J., Chen, S., Chase, M.W., Soltis, D.E., Li, H., Yang, J., Li, D., and Yi, T. (2017). Diversification of rosaceae since the late cretaceous based on plastid phylogenomics. New Phytol. **214:** 1355–1367.

Zhou, X., and Liu, Z. (2022). Unlocking plant metabolic diversity: A (pan)-genomic view. Plant Commun. **3:** 100300.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article: http://onlinelibrary.wiley.com/doi/10.1111/jipb.13318/suppinfo

Figure S1. Information of collected sample
Figure S2. Frequency distribution of depth of 17-mer (upper) and *K*-mer (below) in genome survey of cultivated hawthorn
Figure S3. The genome annotation of hawthorn
Figure S4. Classification statistics of cultivated hawthorn genes
Figure S5. The maximum likelihood phylogenetic tree of Rosaceae with bootstraps
Figure S6. Changes in messenger RNA (mRNA) expression in hard-fleshed hawthorn "Qiujinxing"
Figure S7. Changes in messenger RNA (mRNA) expression in soft-fleshed hawthorn "Ruanrou Shanlihong #3"
Figure S8. Triterpene biosynthesis pathway in *Crataegus pinnatifida*
Figure S9. Syntenic dot plot and K_s distribution within the apple genome
Figure S10. Syntenic dot plot and K_s distribution between two sub-genomes of hawthorn and loquat
(A), apple and loquat **(B)**. Syntenic dot plot and K_s distribution between *Gillenia trifoliata* and sub-genome A of hawthorn **(C)** and sub-genome B of hawthorn **(D)**.
Table S1. Statistics of genome survey data
Table S2. Statistics of paired-end reads based on Hi-C technology
Table S3. Statistics of the lengths of 17 pseudo-chromosomes in the cultivated hawthorn genome
Table S4. Predicted genes and gene features of the cultivated hawthorn
Table S5. Gene functional annotation of the cultivated hawthorn
Table S6. Predicted RNA features of the cultivated hawthorn
Table S7. Conserved genes using the BUSCO (Benchmarking Universal Single-Copy Orthologs) method
Table S8. The enriched Kyoto Encyclopedia of Genes and Genomes (KEGG) functional categories of species-specific genes (*P*-value <0.05) in the cultivated hawthorn
Table S9. The enriched Kyoto Encyclopedia of Genes and Genomes (KEGG) functional categories of significantly (*P*-value <0.05) expanded genes in the cultivated hawthorn
Table S10. The enriched Kyoto Encyclopedia of Genes and Genomes (KEGG) functional categories of significantly (*P*-value <0.05) differentially

expressed genes between two fruit developmental stages in the hard-fleshed ("Qiu Jinxing") hawthorn cultivar

Table S11. The enriched Kyoto Encyclopedia of Genes and Genomes (KEGG) functional categories of significantly (*P*-value <0.05) differentially expressed genes between two fruit developmental stages in the soft-fleshed ("Ruanrou Shanlihong #3") hawthorn cultivar

Table S12. Statistics of repeat sequences, including transposable elements (TEs) in hawthorn, loquat, apple and pear genomes

Table S13. Statistics of orthologs between hawthorn, apple and loquat genomes. The color of the table corresponds to the squares of chromosomes in Figure 4E in the paper