Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Chromosome-level genome assembly of Aralia cordata unveils evolution history and genomic landscape diversification in Aralia-Panax

Tae-Jin Yang

t.jyang@snu.ac.kr

Seoul National University https://orcid.org/0000-0002-9676-8801

Young Sang Park

Seoul National University

Padmanaban Mohanan

Sugentech Inc

Ki Jin Park

Gangwon State Agricultural Research and Extension Services

Yujin Lee

Seoul National University https://orcid.org/0000-0003-0013-4749

Jungyeon Kim

Institute of Food Industrialization, Institutes of Green Bioscience and Technology, Seoul National University

Young Hun Song

Seoul National University

Hyun Hee Kim

Sahmyook University https://orcid.org/0000-0002-2422-643X

Hong Thi Nguyen

Sahmyook University

Hyeji Lee

Seoul National University https://orcid.org/0009-0007-3245-8082

Jee Young Park

Seoul National University https://orcid.org/0009-0007-6557-0687

Article

Keywords: Aralia cordata, genome assembly, polyploidization, genome evolution, long terminal repeat, secondary metabolite

Posted Date: September 21st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10998138/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Abstract

Aralia and *Panax*, two neighboring medicinal plant genera, underwent diverse genomic evolution, including genome expansion and polyploidization. Their variable genome sizes, ploidy levels, and metabolite compositions make these genera valuable genomic resources for elucidating genome evolution. However, the lack of genomic data for the *Aralia* genus has hindered the full exploration of genomic features and evolution in this genus. Here, we assembled a high-quality chromosome-level genome of *Aralia cordata* with a length of 1.06 Gbp for 24 chromosomes. Candidate gene mining suggested that the gene loss and tandem gene duplication might be associated with the species-specific secondary metabolite compositions. Comprehensive genomic analyses with related species revealed their divergence history. Ancestral karyotype reconstruction suggested the modern chromosome evolution scenario of Araliaceae. Despite being tetraploid, *A. cordata* had a comparable genome size to the diploid *A. elata*, which has species-specific LTR expansions. We also inspected the distribution of the Araliaceae family-wide common 160 bp centromeric tandem repeat (TR160). The TR160s were identified in each chromosome of the *Aralia* and *Eleutherococcus* genera but not significantly in the *Panax* genus. The genomic analyses in this study provided valuable insight into chromosome evolution, polyploidization, and secondary metabolite diversification in the Araliaceae family.

INTRODUCTION

Aralia cordata is a plant species belonging to the Araliaceae family. This species is distributed in Northeast Asia, such as in China, Japan, and Korea. In Korea, the young shoots are used as vegetables for "dureup namul" (ttangdureup) and the roots are used as oriental medicine (Dok-hwal)[1, 2]. *A. cordata* had been reported to have diverse metabolites, such as diterpenes, saponins, and phenolic compounds, with analgesic, antiasthmatic, and anti-inflammatory effects[3, 4, 5]. However, the genomic approach to this genus remains limited, despite its diverse usage. Advances in sequencing technology have made it possible to construct chromosome-level genomes using hybrid methods that combine third-generation sequencing and second-generation sequencing, with Hi-C technology.

Chromosome-level genome assemblies of *A. elata* were published as this technology progressed[6, 7, 8]. Based on phylogenetic inference, the *Panax* genus was the most closely related genus to *Aralia* in the Araliaceae family, called the *Aralia-Panax* group[9, 10, 11]. Despite their close relationship, the characteristics of these species were quite distinct, such as chromosome numbers, ploidy levels, and genome sizes. It has been known that the Araliaceae have a basic chromosome number of n = 12[12, 13]. Additionally, the common Araliaceae whole genome duplication (WGD) event, Araliaceae-β, was widely recognized as occurring around 30 million years ago (MYA)[7, 14, 15]. Some species, including *Eleutherococcus senticosus* and *Panax ginseng*, subsequently underwent their own species-specific WGD events[14, 16]. Therefore, species that did not undergo WGD had 12 pairs of chromosomes, while species that did undergo WGD had 24 pairs of chromosomes. In the *Aralia* genus, *A. cordata* also possessed 24

pairs of chromosomes[12, 13]. However, the genome size of these species was not correlated with the WGD events. The genome sizes of the species having 12 pairs of chromosomes, including *A. elata* and *P. notoginseng*, were 1.05 Gbp and 2.3 Gbp, respectively[7, 17]. Meanwhile, the genome sizes of the species having 24 pairs of chromosomes, *E. senticosus* and *P. ginseng*, were 1.2 Gbp and 3.5 Gbp, respectively[14, 16, 18]. The larger genome size in *Panax* was contributed by long terminal repeat (LTR) expansion, especially *Gypsy del*-type[19, 20]. However, it was ambiguous which transposable elements (TEs) were affected in the genomic landscape of the *Aralia* genus compared to the *Panax* genus. Therefore, a comprehensive genomic analysis is required to clarify the evolutionary history of *Aralia* and *Panax*, particularly polyploidy, genome size expansion, and chromosome evolution.

In addition to their close evolutionary relationships, *Aralia* and *Panax* are utilized for medicinal purposes due to their beneficial secondary metabolites, which have distinct compositions. Herbaceous *Aralia* species, such as *A. cordata*, produce diterpenoids consisting of continentalic acid and kaurenoic acid, whereas woody *Aralia* species, including *A. elata*, mainly produce araloside and oleanane-type pentacyclic triterpenoids[1, 7]. In contrast, the *Panax* genus primarily produces dammarane-type tetracyclic triterpenoids, commonly called ginsenosides. Therefore, investigating genomic characteristics and candidate genes involved in metabolite biosynthesis is crucial to understanding and utilizing the unique metabolites in each species.

In this study, the chromosome-level genome of *A. cordata* was assembled with high quality. Comprehensive genomic comparisons revealed the genomic characteristics of this species and the evolution of the genomes of *Aralia* and *Panax*. We identified the candidate genes involved in the biosynthesis of secondary metabolites and proposed evolutionary scenarios for the genomic landscape diversification, including ancestral karyotype reconstruction, genome expansion, and tandem repeats. These findings provide valuable genomic information for understanding the evolution of the polyploid genomes, variations in genomic landscapes, and the biosynthesis of secondary metabolites.

RESULTS

Genome assembly and annotation of Aralia cordata

We assembled the genome for a cultivating variety, ‘Baek mi hyang’ of *A. cordata*, which was registered as a commercial variety in the Korea Seed & Variety Service (https://www.seed.go.kr/). The ‘Baek mi hyang’ was developed as a vegetable cultivar due to its aroma and tender texture compared to wild accessions (Registration Number: 4726). For the hybrid assembly, we produced 52 Gbp of Nanopore long reads (48.08X coverage), 113 Gbp of Illumina short reads (102.85X coverage), and 35 Gbp of Omni-C data (Supplementary Table 1). Using *k*-mer analysis with short reads, we estimated the genome size and ploidy level of *A. cordata* to be 1,097 Mbp and allotetraploid, respectively (Supplementary Fig. 1). Using this sequencing data, a total of 1,089 Mbp of the genome was assembled into 5,652 contigs with an N50 of 0.48 Mbp (Supplementary Table 2). Omni-C data were used to scaffold the contigs into chromosome-level scaffolds.

The twenty-four chromosome-level scaffolds presented 1,063 Mbp, accounting for 97.57% of the assembled genome (Supplementary Tables 2–3). The high-quality genome assembly was evaluated using various methods. The BUSCO assessment was 99.5%, and the LAI was 10.06 (Supplementary Table 3). High interaction in scaffolds was observed in the contact matrix (Supplementary Fig. 2). The mapping ratios of long reads and short reads were 97.35% and 99.74%, respectively. Long reads covered 98.92% of genomes with more than 5X coverage, while short reads covered 93.17% (Supplementary Table 4). Additionally, 95.7% of the mapped short reads were properly paired.

Repeat annotation revealed that 59.22% (629.46 Mbp) of the genome consists of TEs. LTRs were the most abundant element among the TEs, comprising 40.09% of the genome (Supplementary Table 5). *Gypsy*-type LTR retrotransposons were the most common LTR type, as observed in other previously published Araliaceae genomes (*A. elata*, *Panax* species, and *E. senticosus*)[7, 14, 15, 16, 17]. A total of 64,215 protein-coding genes were predicted using transcript evidence, protein homologs, and *ab initio* predictions. Of the 64,215 genes, 62,416 were functionally annotated from various sources with 99.0% of the BUSCO assessment (Supplementary Tables 6–8). The high BUSCO assessment in both the genome (99.5%) and gene model (99.0%), as well as the highly functionally annotated gene sets, implied that the genome assembly and annotation were done in high quality (Supplementary Table 8).

Whole genome duplication of Aralia cordata

As suggested by its tetraploid characteristics, the 24 chromosomes of *A. cordata* exhibited a 1:1 relationship with opposite density patterns between genes and TEs, which supports the high quality of this genome (Fig. 1). In order to understand the genome evolution of *A. cordata*, a phylogenetic tree was constructed with eight other species using 47 single-copy ortholog genes (Supplementary Fig. 3). This tree topology was employed for the divergence time estimation (Fig. 1). The estimated divergence time between *E. senticosus* and the *Aralia-Panax* group was 17.23 MYA, while it was 9.00 MYA between *Aralia* and *Panax*. The estimated divergence time between *A. cordata* and *A. elata* was 8.21 MYA, implying that the common ancestor of *Aralia* diverged after the divergence between *Aralia* and *Panax*. In *A. cordata*, 6,635 gene families were expanded, while 134 were contracted (Fig. 1 and Supplementary Fig. 4). Both GO and KEGG enrichment analyses showed that the expanded gene families were related to the secondary metabolism, such as phenylpropanoid and diterpenoid. Additionally, expanded genes related to plant hormone regulation or stress responses suggested the species-specific environmental adaptation.

The WGD relationships of *A. cordata* and its relatives were inferred by analyzing the *K*s relationships of synteny blocks (Supplementary Figs. 5–11). The collinear relationship with *V. vinifera* was 4:1. After core-eudicot triplication, *V. vinifera* did not undergo WGD events. Therefore, it can be inferred that *A. cordata* underwent WGD after triplication. The collinear relationships with *D. carota*, *E. senticosus*, and *P. ginseng* exhibited a 4:4 ratio, while those relationships were 2:1 with *P. notoginseng* and *A. elata*. Previous studies have identified a shared WGD event (Araliaceae-β) within the Araliaceae family, which occurred

independently from the Apiaceae family[7, 14, 15]. The block relationships indicated that the *A. cordata* underwent an additional WGD event, as was observed in other tetraploid species in the Araliaceae family.

The distribution of *K*s values exhibited a peak at 0.53 between *A. cordata* and *D. carota*, suggesting a divergence between the Apiaceae and Araliaceae (Supplementary Fig. 12). Subsequently, a peak at 0.32 was observed in all five species of the Araliaceae, indicating Araliaceae-β (Fig. 1). This result suggests that the WGD occurred distinctly in Araliaceae and Apiaceae, respectively. The peaks observed in *E. senticosus*, as well as between *A. cordata* and *E. senticosus*, were nearly 0.11. These peaks suggested that the divergence between *Aralia-Panax* and *E. senticosus*, as well as the divergence between the common ancestors of *E. senticosus*, occurred during a similar period, as the WGD in *E. senticosus* was a specific event. The observed peaks of *A. cordata* in relation to *P. ginseng*, *P. notoginseng*, and *A. elata* were 0.056, 0.058, and 0.045, respectively. Additionally, the self-comparison peaks of *A. cordata* and *P. ginseng* were identified at 0.03. These values indicated that the WGD of *A. cordata* and *P. ginseng* occurred species-specifically after the divergence of *Aralia* and *Panax*. In summary, *A. cordata* underwent two times of WGD events following core-eudicot triplication. The first WGD was Araliaceae-β, which commonly occurred in the Araliaceae family around 30 MYA, and the second one was specific to *A. cordata*, which emerged around 3 MYA.

Subgenome evolution of Aralia cordata

In the *k*-mer analysis, *A. cordata* was estimated as allotetraploid (Supplementary Fig. 1). We employed diverse methods, including TE, phylogenetic, and *K*s-based approaches, to identify each subgenome of *A. cordata* (Fig. 2 and Supplementary Fig. 13). SubPhaser was utilized to classify subgenomes according to their distinct TE compositions. Phylogenetic analysis was performed on each subgenome pair to validate TE-based classification (Supplementary Fig. 13). The phylogenetic locations of each set were identical due to the shared common ancestor, which was distinct from *A. elata*, as well as the absence of genome data for extant species. Therefore, *K*s analysis was conducted for the syntenic blocks between *A. cordata* and *A. elata* (Fig. 2). The *K*s values of shared syntenic blocks between each subgenome and *A. elata* were distinct. The subgenome with smaller *K*s values was designated as subA, while the larger one was designated as subB, which was consistent with the TE-based classification.

Gene retention, mutation rates, and expression levels were analyzed to confirm the subgenome dominance in this recent allotetraploid, *A. cordata* (Supplementary Figs. 14–17). When *A. elata* was utilized as a reference, the gene retention exhibited an unbiased pattern between subA and subB. The substitution rates and expression level comparison implied that subA may be dominant following a recent WGD. The overall *K*a and *K*s values were lower in subA than in subB. Furthermore, 19% of the genes were expressed higher in subA in all five investigated organs, despite 64% of genes exhibiting a neutral state. Despite the recent polyploidization, each subgenome of *A. cordata* has undergone different evolutionary forces, with subA being dominant.

The phylogenetic analysis and *K*s distribution suggested the evolutionary history and order of speciation in *Aralia-Panax* (Fig. 2 and Supplementary Fig. 18). The divergence history that was inferred from those

analyses occurred as follows: divergence between *Aralia* and *Panax*, divergence in the *Aralia* genus, divergence in the *Panax* genus, divergence between *A. cordata* subA and subB, and divergence between *P. ginseng* subB and *P. notoginseng*. The most recent *K*s peak of the subgenomes of *A. cordata* was observed at 0.03571, while it was observed at 0.03271 between *P. ginseng* subB and *P. notoginseng*. Those peaks indicated the divergence events of the common ancestor of *A. cordata* subA and subB, as well as the common ancestor of *P. ginseng* subB and *P. notoginseng*. This suggested that the polyploidization occurred after these peaks. Therefore, it could be inferred that the polyploidization events in *A. cordata* and *P. ginseng* occurred in a recent evolutionary period.

Secondary metabolites distinction in Aralia-Panax

The major triterpenoid compositions of *Aralia* and *Panax* differed based on their type. The oleanane-type triterpenoid was predominant in *Aralia*, while the dammarane-type was prevalent in *Panax*. Both types of triterpenoid shared 2,3-oxidosqualene as a precursor. At this point, 2,3-oxidosqualene was converted via β-amyrin synthase (bAS) into β-amyrin, which is a skeleton for oleanane-type triterpenoids[21, 22]. In contrast, dammarane-type triterpenoids were converted via dammarenediol-II synthase (DDS)[23]. Then, the P450 gene families, particularly CYP716, oxidized these skeletons to diverse structures[24, 25]. A previous study suggested that the exon deletion of the DDS gene in *A. elata* might have caused distinct metabolite compositions between *Aralia* and *Panax*[7]. Therefore, the oxidosqualene cyclase (OSC) gene family, including bAS and DDS, was analyzed to understand these distinctions in the *Aralia-Panax* group (Supplementary Fig. 19 and Supplementary Table 9).

In phylogenetic analysis, the putative DDS genes were divided into two groups with high bootstrap support (bb = 100) (Fig. 3 and Supplementary Fig. 20). One group comprised genes from *Panax* species, which had previously been reported as DDS genes. Another group included the amyrin synthase gene (AE10G00915.1), which had been reported in a previous study[7]. Furthermore, a comparison of the gene expression levels of putative bAS and putative DDS genes revealed that the putative bAS genes were expressed in all the species, while the putative DDS genes were solely expressed in *Panax* species. Therefore, the expressed putative DDS genes associated with the biosynthesis of dammarane-type triterpenoids were labeled as DDS, while the remaining were labeled as DDS-like. The gene blocks associated with the DDS and DDS-like genes were located in distinct blocks (Fig. 3B and Supplementary Fig. 21). The analysis revealed that DDS genes were solely present in the *Panax* species, while DDS-like genes were detected in the same gene blocks across all species.

In a previous study, exon residues of the DDS gene were identified in *A. elata*[7]. Similar to *A. elata*, *A. cordata* also exhibited the presence of DDS gene residues, while no such residues were detected in *E. senticosus*, *D. carota*, and *C. sativum* (Supplementary Table 10). The DDS gene residues in *Aralia* were situated in the DDS gene block and exhibited high homologies as well as phylogenetic support (Fig. 3 and Supplementary Figs. 21–22). The pairwise comparisons of the *K*s values revealed a clear distinction between DDS genes and DDS-like genes, suggesting that these genes underwent a distinct evolutionary history for each group (Supplementary Table 11). As previously reported, the deletion of DDS genes was

also observed in *A. cordata*[7]. The absence of DDS residue in other species, as well as the distinct *K*s distributions across DDS and DDS-like genes, suggested that DDS genes were activated in the common ancestor of the *Aralia-Panax* group. During the evolutionary history, speciation led to exon deletion within the *Aralia* genus, leading to a distinct triterpenoid composition between *Aralia* and *Panax*.

Triterpenoids were predominantly synthesized in woody *Aralia*, while diterpenoids were produced in herbal *Aralia*. *A. cordata*, which is an herbal species, primarily produced two diterpenoids, including kaurenoic acid and continentalic acid[1]. The ent-copalyl diphosphate was converted into the basic skeleton for the biosynthesis of both kaurenoic acid and continentalic acid by diTPS, especially TPS-e/f[1, 26, 27, 28]. The gene expansion for TPS-e/f in *A. cordata* was confirmed through gene mining (Fig. 3, Supplementary Figs. 23–24, and Supplementary Table 12). Furthermore, the expression levels and phylogenetic locations of TPS-e/f genes in *A. cordata* were distinct from those in other species. The micro-synteny block revealed that the presence of multiple copies of TPS-e/f genes in *A. cordata* Chr09 indicated that these genes underwent tandem gene duplication. In other species and on Chr21 in *A. cordata*, only one or two copies of TPS-e/f genes existed, whereas seven copies existed on Chr09 in *A. cordata* (Fig. 3 and Supplementary Fig. 25). The pairwise comparisons of *K*s values were distinct between *A. cordata* and other species (Supplementary Data 1). However, comparable *K*s values between Chr09 and Chr21 in *A. cordata* indicated that the tandem gene duplication of TPS-e/f in Chr09 occurred in a similar period with Ac-α.

The candidate genes associated with the biosynthesis of secondary metabolites, especially triterpenoids and diterpenoids, were identified through gene mining. With regard to chromosomal location, the genes that underwent gene loss or tandem duplication were not clustered together with other genes related to the biosynthesis pathway (Supplementary Figs. 26–27). The candidate genes were utilized to reconstruct the expression heatmap. Additionally, cluster analysis was utilized to infer the relationships between these genes (Supplementary Figs. 28–29). In triterpenoids, the DDS gene loss caused the absence of dammarane-type saponin biosynthesis, despite the presence of genes related to the following step. In diterpenoids, cluster analysis revealed a notable association among the genes related to diterpenoids, GGPPS, TPS-c (CPS), TPS-e/f (CKS and KSL), and CYP701. Therefore, specific evolutionary events at the gene level, such as gene loss or duplication, might be associated with species-specific secondary metabolite compositions.

Chromosomal evolution of Aralia-Panax

Inferring the ancestral karyotype could offer valuable insights about the structural evolution of the modern species. In order to reconstruct ancestral karyotypes and infer the evolutionary scenario, the modern chromosome structures of five species in the Araliaceae family were compared. The conserved chromosome structures of *A. cordata* were demonstrated through structure comparisons with chromosome pairs from five species by dotplots (Supplementary Figs. 30–32 and Supplementary Table 13). The species exhibiting the most conserved karyotypes were identified by comparing with the ancestral core eudicot karyotype (ACEK) (Supplementary Figs. 33–35). Consequently, twelve

chromosomes of *A. cordata* were designated as ancestral karyotypes following Araliaceae-β (Supplementary Table 14). Since Araliaceae-β was a shared duplication event within the Araliaceae family, twelve Araliaceae post-β duplication ancestral chromosomes exhibited 2:1 relationships with ACEK (Supplementary Fig. 36). Based on the self-comparisons, the majority of chromosomes exhibited a 1:1 relationship with their counterpart, with the exception of AP03, 05, 06, and 10 (Supplementary Fig. 37).

A syntenic analysis revealed paired relationships, as the basic chromosome number of Apiales was previously suggested to be six[12]. In both comparisons with ACEK and their counterparts, the structures of six ancestral chromosomes could be reorganized (Supplementary Figs. 38–41). Ancestral chromosomes, including AP02:AP04, AP07:AP12, and AP08:AP11, still exhibited relatively intact forms when compared to their pre-β duplication structures. In contrast, AP01:AP09 and AP03:AP05:AP06:AP10 were relatively fragmented. Based on the inferred ancestral chromosome relationships, it was possible to reconstruct chromosome evolution scenarios of modern Araliaceae (Fig. 4). In summary, ACEK was constructed through γ-triplication from seven ancestral eudicot karyotype (AEK)[29]. ACEK underwent reconstruction to six Araliaceae pre-β-duplication ancestral chromosomes. After the β-duplication, twelve post-β-duplication chromosomes were reconstructed. After the ancestral chromosomes 2A:6B and 5A:6A underwent reciprocal translocation (RT) events, the twelve Araliaceae basic chromosomes were reconstructed. By genus- or species-specific RT, end-to-end joining (EEJ), and large inversion events, the ancestral karyotypes of each species were reconstructed. Finally, modern species karyotypes were constructed through speciation and species-specific WGD.

Genome size variations in Aralia-Panax

The genome sizes of five species in the Araliaceae family ranged from 1.04 to 3.46 Gbp (Supplementary Table 15). Despite the recent WGD in *A. cordata*, its genome size was comparable to that of its diploid relative, *A. elata*. Additionally, the closely related genus *Panax* was larger than that of the *Aralia* genus. TE expansions frequently affect the genome size variations in angiosperms[20, 30, 31]. In previous studies, it was reported that the *Gypsy del*-type LTR retrotransposon was the primary factor contributing to genome expansion in *P. ginseng*[19, 20]. To confirm the variations in the genomic landscape, TEs of five species of the Araliaceae family were analyzed (Supplementary Table 15). The correlation between genome size and total TE was observed, with the exception of two species of the *Aralia* genus, which were a result of WGD in *A. cordata*. Whole TE exhibited correlations with LTRs, suggesting that it works as a driving force for variations in genome size. Therefore, representative LTR elements were extracted and classified at the element level to estimate and elucidate genome size variations (Supplementary Table 16).

The *Gypsy del*-type LTR element was the most prevalent in all the species that were investigated. Additionally, the composition of the *Gypsy del* in *Panax* was much larger than that of other species, indicating significant forces that influenced genome sizes. In contrast to *Gypsy del* in the *Panax* genus, no element in *A. cordata* and *A. elata* differed by more than a hundred megabases. However, the compositions of *Gypsy athila*, *Gypsy tat*, and *Copia tork* exhibited differences more than 45 Mbp in

classified LTRs, suggesting that species-specific expansions were driven by diverse forces. Therefore, a phylogenetic analysis was conducted to further classify those four types of LTR elements (Supplementary Figs. 42–45). The histogram for each LTR group was calculated to estimate insertion times and assess their contributions to the genome size of each cluster. The frequency and proportions of each cluster exhibited variations among the species. In *A. cordata*, the proportions of *del*-Cluster1 and *del*-Cluster2 were comparable, with 95.75 and 73.06 Mbp, respectively. In other species, *del*-Cluster2 was predominant in *P. ginseng* and *P. notoginseng*, while *del*-Cluster1 was major in *A. elata* and *E. senticosus*. In the *Panax*, the *del*-Cluster2, which was inserted 3 to 5 Mya, shared over 600 Mbp, 885.44 Mbp in *P. ginseng*, and 682.88 Mbp in *P. notoginseng*. It was over 500 Mbp larger than other species. This finding suggested that the recent burst of *del*-Cluster2 in *Panax* might have been a driving force in the expansion of genome sizes in the *Panax* genus, which was in concordance with previous studies[19, 20].

The genomic landscape differences between *A. cordata* and *A. elata* were affected by three LTR elements. Those three LTR elements, *Gypsy athila*, *Gypsy tat*, and *Copia tork*, expanded to a total of 190 Mbp, and most of the expansion could be specified. The *Gypsy athila* expanded by 30.04 Mbp in *athila*-Cluster1, which was inserted around 3 to 5 Mya. The *Gypsy tat* expanded by 72.29 Mbp in *tat*-Cluster2, which was inserted around 2 to 5 Mya. The *Copia tork* expanded by 64.26 Mbp and was inserted around 2 to 3 Mya. The contributions and distributions of these LTR elements to the genome expansion history were revealed through the comparison of genomic landscapes (Fig. 5). All the expanded elements between *A. cordata* and *A. elata*, *athila*-Cluster1, *del*-Cluster1, *tat*-Cluster2, and *tork* were all inserted intensely between gene synteny blocks. In *P. ginseng*, *del*-Cluster2 was inserted with high intensity and distributed whole chromosomes when compared to *A. cordata* and *A. elata*, while *del*-Cluster1 was inserted between gene synteny blocks. Despite the massive expansions in LTR elements, gene synteny blocks remained relatively conserved. The distance between common gene synteny blocks was approximately 13.6 and 15.4 Mbp in *A. cordata* Chr02A and Chr02B, respectively. Meanwhile, it was more than 50 Mbp in *A. elata* and 90 Mbp in *P. ginseng* Chr02A and Chr02B, respectively. Overall, diverse LTR element expansion in each species of the Araliaceae family suggested species-specific genomic landscape evolution.

Centromeric tandem repeat evolutions in Aralia-Panax

In many eukaryotes, centromere sequences were frequently comprised of tandem repeat (TR) arrays with a unit length ranging from 100 to 200 bp[32]. In a previous study, it was suggested that the 161 bp tandem repeat unit (TR161) consisted of centromere repeat arrays in *A. elata*[33]. Meanwhile, the most prevalent tandem repeat unit in *P. ginseng*, designated as TR167, was determined to be a non-centromeric repeat array[15, 34]. Therefore, high-frequency tandem repeats were searched from five species, including *A. cordata*, *A. elata*, *P. ginseng*, *P. notoginseng*, and *E. senticosus*, to identify the significant tandem repeat units in *Aralia* and *Panax*. In *A. cordata* and *E. senticosus*, TR160 was the most frequent TR unit (Supplementary Fig. 46, Supplementary Tables 17–18). In *A. elata*, TR70 was the most frequent TR unit. However, TR160 was also highly frequent in its genome. Meanwhile, TR160 was not significantly detected in *Panax*. Instead, TR167 was identified as the most frequent TR unit in *P. ginseng*,

as reported in a previous study[34]. Moreover, it was also one of the prevalent TR units in *P. notoginseng*. As TR161 was reported as a centromeric repeat array, we collected TR160 units from those five species. We detected the TR160 unit with a criterion of at least five copies, and it was detected in all the species.

In *Aralia*, the TR160 unit was detected across all chromosomes, while *P. ginseng*, *P. notoginseng*, and *E. senticosus* were detected in six, two, and nineteen chromosomes, respectively (Supplementary Fig. 47). All the collected TR160 units in each chromosome were comparable in all the species. The TR160 units were highly similar despite the different copy numbers in each species. Moreover, TR160 units collected in each species were grouped together, distinguishing them from other species. Except for one TR160 unit in *A. elata*, all other units from the same species were grouped together. TR160 units of *E. senticosus* were most distinct from other species, consistent with the phylogenetic relationships of those genera. These results implied that the TR160 repeat arrays were conserved not only in *Aralia-Panax* but also in *Eleutherococcus*, whose divergence time was almost 20 Mya. The concerted evolution of the TR160 could be suggested by distinguishable repeat arrays in the phylogenetic analysis. Moreover, the TR160 was suggested as a centromeric repeat array by PLOP-FISH analysis (Supplementary Table 19). The FISH signals were observed in the centromeric regions in *A. cordata*, *A. elata*, and *E. senticosus* (Fig. 6). However, it was not observed in *P. ginseng* as consistent with the low copy numbers of TR160 in *Panax*. The TR160 repeat arrays were conserved as centromeric repeats in the Araliaceae family. However, the low copy number and non-FISH signals in *Panax* implied the loss of TR160 during its evolutionary history. Moreover, the distinct features of this centromeric repeat in each species suggested concerted evolution.

In summary, the Araliaceae family underwent a shared WGD, Araliaceae-β. Afterwards, species-specific WGD occurred in *A. cordata*, *P. ginseng*, and *E. senticosus*. The activation of the DDS gene in *Aralia-Panax* and the gene loss in *Aralia* resulted in the biosynthesis of *Panax*-specific metabolite compositions, represented by ginsenosides. The tandem gene duplication, such as TPS-e/f, might be associated with the specific diterpenoid compositions in *A. cordata*. The genome size of the *Panax* genus underwent an expansion as the *Gypsy del* burst, whereas the comparable genome size between *A. cordata* and *A. elata* was contributed by the burst of diverse LTR elements in *A. elata*. Even with species- or genus-specific LTR expansions and WGD, gene syntenic blocks were conserved until the present. In addition to the genome size expansion in the *Panax* genus, it lost the common TR160 centromere repeat arrays, which were different from the other species. Therefore, the accumulation of the genus- or species-specific events along the path of evolutionary history has shaped the genome characteristics of the modern species (Fig. 6).

Discussions

Polyploidization was a common phenomenon in land plants during their genome evolution[35]. Gene duplication, chromosome doubling, and subgenome dominance could be led by polyploidization[36, 37]. Among the investigated five species, including the newly assembled *A. cordata* genome in this study, three were tetraploid, possessing 24 pairs of chromosomes (*A. cordata*, *P. ginseng*, and *E. senticosus*),

while the remaining two were diploid, possessing 12 pairs of chromosomes (*A. elata* and *P. notoginseng*)[14, 16, 17]. *A. cordata* was predicted to be an allotetraploid by the smudgeplot, like other tetraploids, such as *P. ginseng*. The naming of subgenomes could be classified by their own evolutionary histories, as allotetraploidy was the result of the hybridization of two species. As their specific evolutionary traces, subgenomes could be classified by subgenome-specific TE expansions and distinct phylogenetic locations with extant diploid species[16, 38, 39]. In this study, TE-based methods classified two subgenomes, whereas phylogeny-based methods could not classify each subgenome due to the lack of extant species. To support the TE-based method, the *K*s-based method was utilized[40]. It was assumed that the genetic distance between each subgenome and *A. elata* might be distinguishable, as they have distinct evolutionary histories until hybridization. The classification based on *K*s values of syntenic blocks was consistent with the TE-based classification, suggesting that TE and *K*s values could be employed complementarily to divide subgenomes in the absence of extant genomes and in the cases of ambiguous phylogenetic relationships.

Karyotype evolution is essential for understanding structural evolution, speciation history, and validating genome assembly[16, 41, 42]. Synteny analysis among five species demonstrated conserved chromosome structures, despite the frequent occurrence of species-specific WGD events, such as Ac-α (*A. cordata*), Pg-α (*P. ginseng*), and Es-α (*E. senticosus*). Additionally, synteny analysis indicated that *A. cordata* possesses intact chromosomes, which makes it possible to reconstruct the ancestral karyotype using *A. cordata*, regarding it as the ancestral form. A reconstructed chromosome evolutionary scenario, which was represented by syntenic gene blocks, suggested that the chromosome structures in modern Araliaceae have been conserved.

The plant genome sizes ranged from 60 Mbp in *Genlisea margaretae* to 149 Gbp in *Paris japonica*[43]. Repeat elements notably contributed to variations in genome size, alongside WGD[31, 44]. In the five analyzed Araliaceae genomes, *Gypsy*-type LTR retrotransposons occupied the largest portion, varying from 21.18% to 52.58% of their genomes. Especially, *Gypsy del* in the two *Panax* species, *P. ginseng* and *P. notoginseng*, had a notably larger proportion than other species, which was estimated to be 1,367 Mbp and 994 Mbp, respectively. In detail, *del*-Cluster2 was notably expanded, with sizes of 885 Mbp in *P. ginseng* and 683 Mbp in *P. notoginseng*. Previous studies had demonstrated that the specific *Gypsy del*, named *PgDel*1, was a driving force in genome expansion within the *Panax* genus, which consisted of approximately 25% of the *P. ginseng* genome[19, 20]. In this study, *PgDel*1 was grouped together with *del*-Cluster2. The estimated occupation of *del*-Cluster2 in *P. ginseng* was 25.62%, in accordance with the estimated occupation of *PgDel*1, indicating that the classification and results of this study could represent the genomic compositions of LTR elements.

The genome sizes of two *Aralia* species were comparable, even though *A. cordata* was a recent allotetraploid. The *Gypsy athila*, *tat*, and *Copia tork* occupied 6.94% of the *A. cordata* genome, whereas they represented 25.39% in *A. elata*. Among them, *athila*-Cluster1, *tat*-Cluster2, and *tork* contributed to repeat expansion in *A. elata*, with insertion times varying from 2 to 5 Mya. Genome expansions by LTR

elements, which mostly existed between gene synteny blocks, indicated heterochromatin region expansions similar to other plant species.

Centromere sequences in the plant species could span megabase units, which consisted of species-unique centromere repeat arrays. The centromere repeats consisted of 100–200 bp TR units with more than thousands of copies and the sequences of TR units are similar in closely related species and show correlation with the phylogenetic relationship[32, 45]. However, some species, such as *Glycine max*, have different lengths of monomers consisting of 91, 92, and 413 bp[46]. Moreover, species like *Solanum tuberosum* have more than 1 kbp of exceptionally large repeat arrays[47]. In the Araliaceae, specific TR units, TR161 and TR167, were characterized by FISH from *A. elata* and *P. ginseng*, respectively[33, 34]. TR161 was characterized as centromere repeat units in *A. elata*, while TR167 was not centromere repeat units in *P. ginseng*. The bioinformatic pipeline utilized in this study also implied that the TR160, previously TR161 in *A. elata*, might be associated with the centromere repeat arrays in *A. cordata*, *A. elata*, and even *E. senticosus*, while it was not significantly detected in both *P. ginseng* and *P. notoginseng*. Considering the phylogenetic relationships among *Aralia*, *Panax*, and *Eleutherococcus*, TR160 might be the original centromere repeat arrays in the ancestor of the *Aralia-Panax*. However, except for TR167, any other TR units were not detected in the *Panax* genus, and also TR167 was distinguishable with TR160. It was also reported that retrotransposon elements consisted of centromere regions in some plant species, such as *Gossypium hirsutus* and *Malusa domestica*[48, 49]. Therefore, the lack of specific TR units suggested that the centromere repeat arrays in the *Panax* genus might be different and there were genus-specific evolutionary drivers distinguishable from the *Aralia* genus.

Even though species-specific LTR expansions influenced genome size variations, the reconstruction of the ancestral karyotype and genomic landscape comparison indicated that the modern chromosome structures of Araliaceae remain conserved. Moreover, the centromeric TR unit was different between *Aralia* and *Panax*. However, the lack of true T2T-level genomes, most species in the Araliaceae family, and other types of genomic data hindered the clear elucidation of the centromere repeat evolutions of *Aralia-Panax*. Therefore, gap-free genome assembly with diverse genomic data, such as methylome and ChIP-seq, in both the *Aralia* and *Panax* genera might help elucidate the chromosome evolutions with epigenetic regulations and their roles in those closely related genera.

The Araliaceae family contains almost 45 genera with 1,500 species[9, 11]. This family is divided into four major groups: the Asian Palmate, *Polyscias-Pseudopanax*, *Aralia-Panax*, and Greater *Raukaua* groups[11]. The *Aralia-Panax* group was a monophyletic group within the Araliaceae family consisting of the *Aralia* and *Panax* genera. This study utilized genomic information to elucidate the impact of the genomic variations on chromosomal evolution and metabolite biosynthesis within the *Aralia-Panax* group. In the cytogenetic aspect, the basic chromosome number for the Araliaceae family was known as n = 12[12]. Both the *Aralia* and *Panax* genera have been reported to have either 12 or 24 pairs of chromosomes. Therefore, the n = x = 12 and n = 2x = 24 chromosomes could be considered diploid and tetraploid species, respectively. It has been reported that the *P. ginseng* and *P. quinquefolius* in the *Panax* genus

shared a recent WGD event[50, 51]. Furthermore, this study confirmed that *A. cordata* also underwent its specific WGD event recently.

The *Aralia* genus could be distinguished by its phenotypes and geographic distribution. The *Aralia* genus included both herbaceous and woody species[1]. Moreover, this genus was distributed in isolation in Asia and America, similarly to the *Panax* genus[52, 53, 54]. For example, *A. cordata* and *A. elata* were found in Northeast Asia, while *A. racemosa* and *A. spinosa* were found in North America. Additionally, *A. cordata* and *A. racemosa* were herbaceous plants, whereas *A. elata* and *A. spinosa* were woody plants. In addition to these distinctions, the metabolite compositions were also different, with diterpenoids in herbal *Aralia* and triterpenoids in woody *Aralia*[1]. Previous studies suggested that the distinct metabolite compositions in subtropical and temperate *Panax* species were the result of WGD and adaptation[15, 51]. Both *A. cordata* and *A. elata*, which were diploid and tetraploid *Aralia* species, had comparable genome sizes of 1.1 Gbp, even though *A. cordata* underwent WGD recently. Nevertheless, they were found in temperate regions and could overwinter, implying that WGD affected other adaptation factors in these diploid and tetraploid *Aralia* species. This study suggested that chromosomal structures have been conserved in both genomes with regard to gene blocks. Additionally, the recent burst of LTR retrotransposons, including *Gypsy athila*, *Gypsy tat*, and *Copia tork* element in *A. elata*, in the heterochromatin regions indicated recent expansions in the genome size of these species. Therefore, the recent WGD in *A. cordata* and transposon insertions in *A. elata* caused their genomes to become similar in size.

In this study, we assembled a new high-quality genome of *A. cordata* and conducted comprehensive analyses of its genomic landscape evolution. Therefore, this study provided information on how the genomic landscape changed following speciation. Still, many distinct features remained to be untangled in these species and related genera. The newly assembled genome and genomic analyses in this study could be utilized to understand how plants underwent speciation, including polyploidy, geographic isolation, and secondary metabolite biosynthesis.

MATERIALS AND METHODS

Plant materials and sequencing

Young leaves collected from the *A. cordata* cultivar ‘Baek mi hyang’ were used for sequencing, which was provided by GWARES (Ganwondo Agriculture Research and Extension Services, Chuncheon-si, Gangwon-do, Korea). Library preparation and sequencing services for long reads, short reads, and Omni-C were provided by the National Instrumentation Center for Environmental Management (Seoul National University, Seoul, Korea), Phyzen (Sungnam, Korea), and DNALINK (Seoul, Korea), respectively. Five organs, including leaves, stems, roots, flowers, and fruits, were collected for RNA sequencing. Each organ was sequenced by Illumina NovaSeq with three biological replicates. Total RNA of all organs was collected for PacBio Iso-Seq. RNA sequencing and PacBio Iso-Seq sequencing services were provided by Phyzen (Seongnam, Korea) and DNALINK (Seoul, Korea), respectively.

Genome assembly

K-mer-based methods utilizing short reads were conducted to estimate genomic characteristics. Trimmomatic (v0.39) was used to trim raw short read data (ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36) to collect cleaned short reads[55]. K-mer distribution was calculated by Jellyfish (v2.2.10) with 21-mers using cleaned short reads[56]. A calculated k-mer histogram was utilized for estimating genome size and ploidy level by using findGSE (version 0.1.0) and smudgeplot (v0.2.2), respectively[57, 58]. Raw long reads were assembled into draft contigs by the Flye assembler (2.9.3-b1797) with default parameters[59]. Hapo-G was used to polish draft contigs in two rounds using cleaned short reads[60]. Raw Omni-C reads were trimmed by Trimmomatic with the same parameters described above, and then cleaned reads were mapped to the contigs by bwa-mem2 for scaffolding[61]. Contigs were ordered and scaffolded by HapHiC (version 1.0.6) utilizing Omni-C read mapping data[62]. Mis-joined and mis-ordered scaffolds were manually curated by Juicebox (2.17.00)[63]. Suspected haplotigs were excluded by Purge Haplotigs with the "-a 85" parameter[64]. The genome completeness was evaluated with BUSCO (v5.6.1), the Omni-C contact matrix, the LTR Assembly Index (LAI), and the mapping rate of long reads and short reads[62, 65, 66]. The read mapping for long reads and short reads was conducted by minimap2 and bwa-mem2, respectively[61, 67].

Genome annotation

Transposable elements were annotated using the EDTA pipeline (v.2.1.3) with the '--sensitive 1 --anno 1' parameters[68]. To avoid repeat regions for protein-coding gene annotation, genome sequences were masked based on repeat annotation. Transcript evidence, protein homology, and *ab initio* prediction were utilized for the gene model prediction. To collect transcript evidence, short RNA reads and Iso-Seq data were mapped to the genome by HISAT2 (version 2.2.1) and minimap2 (2.26-r1175), respectively[67, 69]. Afterwards, transcripts were assembled by StringTie (v2.1.7) and Trinity (v2.15.1)[70, 71]. Additionally, *de novo* transcript assembly was conducted by Trinity. Peptide sequences of related species with other publicly available data were collected for homology-based evidence (*A. elata*, *P. ginseng*, *P. notoginseng*, *E. senticosus*, *D. carota*, *Glycine max*, *Helianthus annuus*, *Solanum lycopersicum*, *Arabidopsis thaliana*, and Swiss-Prot)[7, 14, 15, 16, 17, 72, 73, 74, 75, 76, 77]. Collected peptide sequences were mapped to the *A. cordata* using miniprot (version 0.13-r262) with default parameters[78]. Mapped information was retrieved by two criteria: more than positive 0.9 or having a stop codon and not less than positive 0.5. Filtered homology evidence was further utilized for gene predictions. The *ab initio* prediction was conducted by AUGUSTUS (3.5.0), BUSCO (v5.6.1), SNAP (version 2006-07-28), and Genemark-ES (version 4.72)[65, 79, 80, 81]. Afterwards, the consensus gene model was predicted by EVidenceModeler (v2.0.0) with the '--segmentSize 500000 and --overlapSize 100000' parameters[82]. The weight for the gene model was given priority for transcripts, peptide homology, and *ab initio* predictions. Transcripts evidence was utilized for the curation of the draft gene model by the PASA pipeline (v2.5.3) with two rounds[83]. The curated gene

model was further filtered based on the peptide length, transposable elements (TE), and long non-coding RNA (lncRNA). Peptides shorter than 50 amino acids (aa), TE domains stretched more than 50% of the genes, and putative lncRNAs were filtered[84, 85]. TE and lncRNA were searched by InterProScan (version 5.66-98.0) with '-appl pfam' and cpc2, respectively. Evidence from Iso-Seq and StringTie were assembled into transcripts by PASA and TransDecoder (5.5.0). Finally, if these transcripts were missed, they were retrieved to the gene model. The predicted gene model was functionally annotated with InterProScan, eggNOG, Swiss-Prot, the nr database (viridiplantae), the TrEMBL database (plants), and Gene Ontology (GO) retrieved from InterProScan[77, 84, 86, 87].

Genome evolution and whole-genome duplication

TE and gene density were calculated by using repeat and gene annotation with sliding window methods (window size: 3 Mb, sliding size: 1 Mb). Synteny blocks among chromosomes of *A. cordata* were inferred by MCScanX utilizing BLASTP (2.10.0+) results with '-max_hsps 20 -max_target_seqs 20 -evalue 1e-10'[88, 89]. Afterwards, those genomic characteristics were drawn by Circos[90]. To infer the phylogenetic relationships, eight other species were used to construct a phylogenetic tree. Single-copy ortholog genes were identified by OrthoFinder[91]. Each single-copy ortholog gene was aligned by MUSCLE (v3.8.1551), and then peptide sequences were converted into nucleotide sequences by PAL2NAL[92, 93]. Aligned coding sequences were concatenated into a single matrix and then utilized to construct a phylogenetic tree using IQ-TREE (version 1.6.12) with '-bb 1000 -alrt 1000'[94]. This tree topology was utilized for the divergence time estimation. Divergence time among nine species was estimated by MCMCtree, embedded in PAML (version 4.9j)[95]. The root length (124.2 Mya) and minimum age of divergence between Araliaceae and Apiaceae (37.2 Mya) were utilized as calibration points for divergence time estimation[96, 97]. Orthogroups with at least one gene included in all the species were utilized to calculate gene family expansion and contraction by CAFE5[98]. To analyze WGD in *A. cordata* and its relatives, gene pairs were inferred by BLASTP with the '-max_hsps 20 -max_target_seqs 20 -evalue 1e-10' parameters[89]. Those BLAST results were utilized by the WGDI pipeline for downstream analysis[99]. The *Ks* distributions were curated, as the substitution rates depend on the evolutionary history of each species. Two common WGD peaks, the WGT peak of *V. vinifera* and the WGD peak of *E. senticosus*, were used for the *Ks* distributions curation following the equation of previous studies[14, 100]. SubPhaser and WGDI were utilized for TE-based and phylogenetic-based classification, respectively, to divide subgenomes following the methods described in the previous study[38, 99]. Additionally, the median *Ks* values of synteny blocks were also utilized. The *Ks* values on the syntenic blocks were used to calculate the genetic distance between the subgenomes of *A. cordata* and *A. elata*. Two subgenomes of *A. cordata*, subA and subB, were classified based on those methods. To identify gene pairs affected by WGD, duplication modes of gene pairs were classified by DupGen_finder[101]. The *Ks* distributions of WGD gene pairs were calculated by KaKs_Calculator 2.0[102]. If the *Ks* was smaller than 0.2 in a gene pair, it was classified as Ac-α. If one gene's expression level was larger than 1.5 times, it was considered expression bias, indicating expression dominance.

Candidate gene mining and analysis

Candidate genes associated with the biosynthesis of triterpenoids and diterpenoids were identified. Domain-based searches were conducted to identify the draft gene set. InterPro and HMMER3 (3.3.1) were utilized to identify candidate genes with HMMER profiles (SS: PF00494, SE: PF08491, OSC: PF13243 and PF13249, GPPS and GGPPS: PF00348, TPS: PF01397 and PF03936, and CYP: PF00067)[84, 103, 104]. Pfam profiles were downloaded from the InterProScan websites (https://www.ebi.ac.uk/interpro/). Homology-based searches were performed using reference sequences obtained from SwissProt and Araport, with BLASTP '-evalue 1e-10 and identity > = 40%' parameters[76, 77, 89]. Genes (SS, SE, OSC, TPS, and CYP) that were shorter than 300 aa and genes (GPPS and GGPPS) that were shorter than 250 aa were excluded, considering them as pseudogenes. The candidate gene set was aligned by MUSCLE (v3.8.1551)[92]. These gene sets were selected and classified according to the tree topology through phylogenetic analysis. The phylogenetic tree was constructed using the ML method with the parameters '-bb 1000 -alrt 1000' by IQ-TREE[94]. StringTie was utilized to calculate and retrieve gene expression levels (TPM)[70]. RNA-seq data of leaves, stems, and roots were obtained from GinsengDB and SRA (*P. ginseng*: GinsengDB, *A. elata*: SRR15501070-155017076 and SRR17927884-17927885, *P. notoginseng*: SRR7427748-7427755), respectively[105]. JCVI was used to visualize gene structures and synteny blocks[106]. The *K*s values were calculated by KaKs_Calculator 2.0[102]. The chromosomal location of each gene was marked by TBtools-II (v2.210)[107]. Genes expressed more than 5 at TPM values in at least one organ were utilized for co-expression analysis. Co-expressed genes were clustered by Clust[108].

Ancestral karyotypes reconstruction

The WGDI pipeline was utilized to reconstruct the ancestral karyotypes of Araliaceae[99]. In summary, dotplots were drawn for the five species in the Araliaceae family (*A. cordata*, *A. elata*, *P. ginseng*, *P. notoginseng*, and *E. senticosus*). Twelve chromosomes were selected and regarded as post-β duplication ancestral karyotypes based on dotplot comparisons. Based on the syntenic relationships of twelve post-β ancestral karyotypes, relationships of six pre-β duplication ancestral karyotypes were inferred and assigned, as a previous study suggested the basic chromosome numbers of the Apiales order as six[12]. Each pair of post-β karyotypes was compared by dotplots to reconstruct six pre-β duplication ancestral karyotypes.

Comparisons of genome size variations

To assess the contributions of repeat elements to genome size variations, the EDTA pipeline (v.2.1.3) was utilized to annotate repeat elements in five species (*A. cordata*, *A. elata*, *P. ginseng*, *P. notoginseng*, and *E. senticosus*)[68]. Among the annotated LTR elements by the EDTA pipeline, only LTR elements longer than 500 bp were utilized for LTR element-level classification. The classification of LTR elements was

performed using TEsorter (version 1.3) with the '-db gydb -cov 10 -eval 1e-2' parameters[109]. The compositions of each LTR element were estimated by those classification results. Four estimated driving forces, including *Gypsy athila*, *Gypsy del*, *Gypsy tat*, and *Copia tork*, were further classified by TEids based on the EDTA annotation. The longest ones for each TEid were selected as representative LTRs based on the cumulative lengths of five domains, including AP, GAG, INT, RNaseH, and RT. Each domain was aligned by muscle and then concatenated into a single matrix[92]. FastTree (version 2.1.11) was utilized to generate a phylogenetic tree for further classification into cluster level[110]. The distributions of each LTR element on the chromosomes were visualized by in-house Python code. Gene synteny blocks were drawn by WGDI pipelines[99]. Whole genome alignments were conducted by minimap2 and then visualized by JBrowse2[111].

Tandem repeat identification

To identify significant TR units in the genome, we utilized diverse TR search methods from previous studies[112, 113, 114]. First, we collected highly frequent TR units from the genome by using TRF[115]. Afterwards, we searched low k-mer diversity regions in the genome by jellyfish using 21-mers[56]. Among highly frequent TR units, overlapped with low k-mer diversity regions and at least five copies, were extracted for the frequency comparison. Based on the previous study and frequency in the genomes, we extracted TR160 from each chromosome in each genome. The consensus TR160 sequences for each chromosome were extracted by cons in EMBOSS packages[116]. The alignment of TR160 was conducted by MAFFT and then visualized by NCBI MSA Viewer (https://www.ncbi.nlm.nih.gov/projects/msaviewer/)[117]. The phylogenetic tree of TR160 was drawn by IQ-TREE (version 1.6.12) with '-bb 1000 -alrt 1000'[94].

Centromeric FISH analysis

Pre-labeled oligonucleotide probes (PLOPs) were designed to observe the chromosomal distribution of the TR160. The synthesis service of PLOPs was provided by Bioneer (Daejeon, South Korea). The FISH procedure was conducted with the modified method described in the previous study (Waminal et al. 2018). Preparation of the 40 µL hybridization mixture was conducted by adding 100% formamide, 50% dextran sulfate, 20× SSC, 50 ng/µL of probe, and Sigma water. Afterwards, the mixture was incubated at 37°C overnight and denatured at 80°C. Slides were washed with 2× SSC at 22–25°C and dehydrated in 70%, 90%, and 100% ethanol for 3 minutes, respectively. Finally, they were counterstained with DAPI-Vectashield (Vector Laboratories, Newark, CA, USA) and observed using a Leica DFC365 FS CCD camera (Leica Microsystems, Wetzlar, Germany) in an Olympus BX53 fluorescence microscope system (Olympus, Tokyo, Japan). The images were captured using the Cytovision software (Leica Microsystems) and subsequently enhanced using Adobe Photoshop CS6 (Adobe Inc., San Jose, California, USA).

Data availability

The raw genome data of *A. cordata* was uploaded in the NCBI under the BioProject PRJNA1522235. Genome assembly and annotation files were deposited on FigShare

(https://doi.org/10.6084/m9.figshare.33413566).

Declarations

Acknowledgement

This study was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A4A1032888). We also thank to Gangwondo ARES, Wild Vegetable Research Institute, who provided *A. cordata* variety resources.

Author contributions

YSP, PM, YHS, and TJY designed the concepts of the project. KJP, JYP, and TJY collected and conserved the materials. Genome assembly, annotation, and evolutionary analyses were conducted by YSP. Experiments were designed by HJL, NTH, YJL, JYK, and HHK. The manuscript was written by YSP and TJY. All authors read and approved the manuscript.

Conflict of interest

The authors declare that there are no competing interests.

References

1. Xu Y, *et al.* Traditional uses, phytochemistry, pharmacology, toxicity and quality control of medicinal genus Aralia: A review. *Journal of Ethnopharmacology* **284**, 114671 (2022).
2. Kim H, Song M-J. Analysis and recordings of orally transmitted knowledge about medicinal plants in the southern mountainous region of Korea. *Journal of Ethnopharmacology* **134**, 676-696 (2011).
3. Matsuo H, *et al.* Utilization of inedible parts of Aralia cordata with high-content chlorogenic acid cultivated in Tokyo. *Sustainability* **16**, 2582 (2024).
4. Puzerytė V, Viškelis P, Balčiūnaitienė A, Štreimikytė P, Viškelis J, Urbonavičienė D. Aralia cordata Thunb. as a source of bioactive compounds: Phytochemical composition and antioxidant activity. *Plants* **11**, 1704 (2022).
5. Kim JG, *et al.* Diterpenoids and diacetylenes from the roots of Aralia cordata with inhibitory effects on nitric oxide production. *Journal of natural products* **84**, 230-238 (2021).
6. Kim J, *et al.* A comprehensive analysis integrating phenotypic assessment uncovering thornless cultivar lineages in Aralia elata. *Genomics* **116**, 110824 (2024).
7. Wang Y, *et al.* Deletion and tandem duplications of biosynthetic genes drive the diversity of triterpenoids in Aralia elata. *Nature communications* **13**, 2224 (2022).
8. Liu W, *et al.* A high-quality reference genome sequence and genetic transformation system of Aralia elata. *Frontiers in Plant Science* **13**, 822942 (2022).

9. Kang J-S*, et al.* Evolution of the Araliaceae family involved rapid diversification of the Asian Palmate group and Hydrocotyle specific mutational pressure. *Scientific Reports* **13**, 22325 (2023).
10. Kim K, Dong J, Wang Y, Park JY, Lee S-C, Yang T-J. Evolution of the Araliaceae family inferred from complete chloroplast genomes and 45S nrDNAs of 10 Panax-related species. *Scientific reports* **7**, 1-9 (2017).
11. Li R, Wen J. Phylogeny and diversification of Chinese Araliaceae based on nuclear and plastid DNA sequence data. *Journal of Systematics and Evolution* **54**, 453-467 (2016).
12. Yi T, Lowry PP, Plunkett GM, Wen J. Chromosomal evolution in Araliaceae and close relatives. *Taxon* **53**, 987-1005 (2004).
13. Sun BY, Kim CH, Soh WY. Chromosome numbers of Araliaceae in Korea. *Korean Journal of Plant Taxonomy* **18**, 291-296 (1988).
14. Yang Z*, et al.* Chromosomal-scale genome assembly of Eleutherococcus senticosus provides insights into chromosome evolution in Araliaceae. *Molecular Ecology Resources* **21**, 2204-2220 (2021).
15. Kim NH*, et al.* Genome and evolution of the shade-requiring medicinal herb Panax ginseng. *Plant Biotechnology Journal* **16**, 1904-1917 (2018).
16. Wang Z-H*, et al.* Reshuffling of the ancestral core-eudicot genome shaped chromatin topology and epigenetic modification in Panax. *Nature Communications* **13**, 1902 (2022).
17. Yang Z*, et al.* The chromosome-scale high-quality genome assembly of Panax notoginseng provides insight into dencichine biosynthesis. *Plant Biotechnology Journal* **19**, 869-871 (2021).
18. Song Y*, et al.* Telomere-to-telomere reference genome for Panax ginseng highlights the evolution of saponin biosynthesis. *Horticulture research* **11**, uhae107 (2024).
19. Choi HI*, et al.* Major repeat components covering one-third of the ginseng (P anax ginseng CA Meyer) genome and evidence for allotetraploidy. *The Plant Journal* **77**, 906-916 (2014).
20. Lee J*, et al.* Rapid amplification of four retrotransposon families promoted speciation and genome size expansion in the genus Panax. *Scientific reports* **7**, 9045 (2017).
21. Kushiro T, Shibuya M, Ebizuka Y. β-Amyrin synthase: cloning of oxidosqualene cyclase that catalyzes the formation of the most popular triterpene among higher plants. *European Journal of Biochemistry* **256**, 238-244 (1998).
22. Han J-Y, Kim M-J, Ban Y-W, Hwang H-S, Choi Y-E. The involvement of β-amyrin 28-oxidase (CYP716A52v2) in oleanane-type ginsenoside biosynthesis in Panax ginseng. *Plant and Cell Physiology* **54**, 2034-2046 (2013).
23. Tansakul P, Shibuya M, Kushiro T, Ebizuka Y. Dammarenediol-II synthase, the first dedicated enzyme for ginsenoside biosynthesis, in Panax ginseng. *FEBS letters* **580**, 5143-5149 (2006).
24. Han J-Y, Hwang H-S, Choi S-W, Kim H-J, Choi Y-E. Cytochrome P450 CYP716A53v2 catalyzes the formation of protopanaxatriol from protopanaxadiol during ginsenoside biosynthesis in Panax ginseng. *Plant and cell physiology* **53**, 1535-1545 (2012).

25. Han J-Y, Kim H-J, Kwon Y-S, Choi Y-E. The Cyt P450 enzyme CYP716A47 catalyzes the formation of protopanaxadiol from dammarenediol-II during ginsenoside biosynthesis in Panax ginseng. *Plant and cell physiology* **52**, 2062-2073 (2011).
26. Chen F, Tholl D, Bohlmann J, Pichersky E. The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. *The Plant Journal* **66**, 212-229 (2011).
27. Xu M*, et al.* Functional characterization of the rice kaurene synthase-like gene family. *Phytochemistry* **68**, 312-326 (2007).
28. Kanno Y*, et al.* Characterization of a rice gene family encoding type-A diterpene cyclases. *Bioscience, biotechnology, and biochemistry* **70**, 1702-1710 (2006).
29. Murat F, Armero A, Pont C, Klopp C, Salse J. Reconstructing the genome of the most recent common ancestor of flowering plants. *Nature genetics* **49**, 490-496 (2017).
30. Pulido M, Casacuberta JM. Transposable element evolution in plant genome ecosystems. *Current Opinion in Plant Biology* **75**, 102418 (2023).
31. Piegu B*, et al.* Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice. *Genome research* **16**, 1262-1269 (2006).
32. Naish M, Henderson IR. The structure, function, and evolution of plant centromeres. *Genome Research* **34**, 161-178 (2024).
33. Peniton Jr EA, Nguyen HT, Waminal NE, Yang T-J, Kim HH. Cytogenomic evaluation of regenerated Aralia elata using PLOP-FISH and flow cytometry. *Scientific Reports* **14**, 30289 (2024).
34. Waminal NE*, et al.* A refined Panax ginseng karyotype based on an ultra-high copy 167-bp tandem repeat and ribosomal DNAs. *Journal of Ginseng Research* **41**, 469-476 (2017).
35. Adams KL, Wendel JF. Polyploidy and genome evolution in plants. *Current opinion in plant biology* **8**, 135-141 (2005).
36. Alger EI, Edger PP. One subgenome to rule them all: underlying mechanisms of subgenome dominance. *Current opinion in plant biology* **54**, 108-113 (2020).
37. Bird KA, VanBuren R, Puzey JR, Edger PP. The causes and consequences of subgenome dominance in hybrids and recent polyploids. *New Phytologist* **220**, 87-93 (2018).
38. Jia KH*, et al.* SubPhaser: a robust allopolyploid subgenome phasing method based on subgenome-specific k-mers. *New Phytologist* **235**, 801-809 (2022).
39. Peng R*, et al.* Evolutionary divergence of duplicated genomes in newly described allotetraploid cottons. *Proceedings of the national academy of sciences* **119**, e2208496119 (2022).
40. Zhang R-G*, et al.* Subgenome-aware analyses suggest a reticulate allopolyploidization origin in three Papaver genomes. *Nature communications* **14**, 2204 (2023).
41. Zhao Q*, et al.* Reconstruction of ancestral karyotype illuminates chromosome evolution in the genus Cucumis. *The Plant Journal* **107**, 1243-1259 (2021).

42. Murat F, *et al.* Understanding Brassicaceae evolution through ancestral genome reconstruction. *Genome biology* **16**, 1-17 (2015).
43. Pellicer J, Fay MF, Leitch IJ. The largest eukaryotic genome of them all? *Botanical Journal of the Linnean Society* **164**, 10-15 (2010).
44. Wessler SR. Transposable elements and the evolution of eukaryotic genomes. *Proceedings of the National Academy of Sciences* **103**, 17600-17601 (2006).
45. Lim KB, *et al.* Characterization of the centromere and peri-centromere retrotransposons in Brassica rapa and their distribution in related Brassica species. *The Plant Journal* **49**, 173-183 (2007).
46. Liu Y, *et al.* Pan-centromere reveals widespread centromere repositioning of soybean genomes. *Proceedings of the National Academy of Sciences* **120**, e2310177120 (2023).
47. Gong Z, *et al.* Repeatless and repeat-based centromeres in potato: implications for centromere evolution. *The Plant Cell* **24**, 3559-3574 (2012).
48. Daccord N, *et al.* High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. *Nature genetics* **49**, 1099-1106 (2017).
49. Chang X, *et al.* High-quality Gossypium hirsutum and Gossypium barbadense genome assemblies reveal the landscape and evolution of centromeres. *Plant communications* **5**, (2024).
50. Shim H, Waminal NE, Kim HH, Yang T-J. Dynamic evolution of Panax species. *Genes & Genomics* **43**, 209-215 (2021).
51. Koo H, *et al.* Comparative transcriptome and metabolome analyses of four Panax species explore the dynamics of metabolite biosynthesis. *Journal of Ginseng Research* **47**, 44-53 (2023).
52. Wen J. Evolution of the Aralia–Panax complex (Araliaceae) as inferred from nuclear ribosomal ITS sequences. *Edinburgh Journal of Botany* **58**, 243-257 (2001).
53. Zuo Y-J, Wen J, Zhou S-L. Intercontinental and intracontinental biogeography of the eastern Asian–eastern North American disjunct Panax (the ginseng genus, Araliaceae), emphasizing its diversification processes in eastern Asia. *Molecular Phylogenetics and Evolution* **117**, 60-74 (2017).
54. Wen J, Shi S, Jansen RK, Zimmer EA. Phylogeny and biogeography of Aralia sect. Aralia (Araliaceae). *American Journal of Botany* **85**, 866-875 (1998).
55. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
56. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764-770 (2011).
57. Sun H, Ding J, Piednoël M, Schneeberger K. findGSE: estimating genome size variation within human and Arabidopsis using k-mer frequencies. *Bioinformatics* **34**, 550-557 (2018).
58. Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature communications* **11**, 1432 (2020).
59. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. *Nature biotechnology* **37**, 540-546 (2019).

60. Aury J-M, Istace B. Hapo-G, haplotype-aware polishing of genome assemblies with accurate reads. *NAR genomics and bioinformatics* **3**, lqab034 (2021).
61. Vasimuddin M, Misra S, Li H, Aluru S. Efficient architecture-aware acceleration of BWA-MEM for multicore systems. In: *2019 IEEE international parallel and distributed processing symposium (IPDPS)*). IEEE (2019).
62. Zeng X*, et al.* Chromosome-level scaffolding of haplotype-resolved assemblies using Hi-C data without reference genomes. *Nature Plants* **10**, 1184-1200 (2024).
63. Durand NC*, et al.* Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell systems* **3**, 99-101 (2016).
64. Roach MJ, Schmidt SA, Borneman AR. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC bioinformatics* **19**, 1-10 (2018).
65. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular biology and evolution* **38**, 4647-4654 (2021).
66. Ou S, Chen J, Jiang N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic acids research* **46**, e126-e126 (2018).
67. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018).
68. Ou S*, et al.* Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome biology* **20**, 1-18 (2019).
69. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature biotechnology* **37**, 907-915 (2019).
70. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature biotechnology* **33**, 290-295 (2015).
71. Grabherr MG*, et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* **29**, 644-652 (2011).
72. Iorizzo M*, et al.* A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. *Nature genetics* **48**, 657-666 (2016).
73. Schmutz J*, et al.* Genome sequence of the palaeopolyploid soybean. *nature* **463**, 178-183 (2010).
74. Badouin H*, et al.* The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. *Nature* **546**, 148-152 (2017).
75. Tomato Genome Consortium x. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**, 635 (2012).
76. Cheng CY, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD. Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. *The Plant Journal* **89**, 789-804 (2017).
77. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. *Nucleic acids research* **51**, D523-D531 (2023).

78. Li H. Protein-to-genome alignment with miniprot. *Bioinformatics* **39**, btad014 (2023).
79. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic acids research* **34**, W435-W439 (2006).
80. Korf I. Gene finding in novel genomes. *BMC bioinformatics* **5**, 1-9 (2004).
81. Gabriel L*, et al.* BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Research* **34**, 769-777 (2024).
82. Haas BJ*, et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome biology* **9**, 1-22 (2008).
83. Haas BJ*, et al.* Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic acids research* **31**, 5654-5666 (2003).
84. Jones P*, et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-1240 (2014).
85. Kang Y-J*, et al.* CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic acids research* **45**, W12-W16 (2017).
86. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molecular biology and evolution* **38**, 5825-5829 (2021).
87. Huerta-Cepas J*, et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic acids research* **47**, D309-D314 (2019).
88. Wang Y*, et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic acids research* **40**, e49-e49 (2012).
89. Camacho C*, et al.* BLAST+: architecture and applications. *BMC bioinformatics* **10**, 1-9 (2009).
90. Krzywinski M*, et al.* Circos: an information aesthetic for comparative genomics. *Genome research* **19**, 1639-1645 (2009).
91. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome biology* **20**, 1-14 (2019).
92. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**, 1792-1797 (2004).
93. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic acids research* **34**, W609-W612 (2006).
94. Nguyen L-T, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution* **32**, 268-274 (2015).
95. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution* **24**, 1586-1591 (2007).

96. Sun G, Dilcher DL, Wang H, Chen Z. A eudicot from the Early Cretaceous of China. *Nature* **471**, 625-628 (2011).
97. Magallón S, Gómez-Acevedo S, Sánchez-Reyes LL, Hernández-Hernández T. A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *New Phytologist* **207**, 437-453 (2015).
98. Mendes FK, Vanderpool D, Fulton B, Hahn MW. CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics* **36**, 5516-5518 (2020).
99. Sun P, *et al.* WGDI: a user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. *Molecular plant* **15**, 1841-1851 (2022).
100. Song X, *et al.* The celery genome sequence reveals sequential paleo-polyploidizations, karyotype evolution and resistance gene reduction in apiales. *Plant Biotechnology Journal* **19**, 731-744 (2021).
101. Qiao X, *et al.* Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. *Genome biology* **20**, 1-23 (2019).
102. Wang D, Zhang Y, Zhang Z, Zhu J, Yu J. KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics, proteomics & bioinformatics* **8**, 77-80 (2010).
103. Eddy SR. Accelerated profile HMM searches. *PLoS computational biology* **7**, e1002195 (2011).
104. Finn RD, *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic acids research* **44**, D279-D285 (2016).
105. Jayakodi M, *et al.* Ginseng Genome Database: an open-access platform for genomics of Panax ginseng. *BMC plant biology* **18**, 62 (2018).
106. Tang H, *et al.* JCVI: A versatile toolkit for comparative genomics analysis. *Imeta* **3**, e211 (2024).
107. Chen C, *et al.* TBtools-II: A “one for all, all for one” bioinformatics platform for biological big-data mining. *Molecular plant* **16**, 1733-1742 (2023).
108. Abu-Jamous B, Kelly S. Clust: automatic extraction of optimal co-expressed gene clusters from gene expression data. *Genome biology* **19**, 172 (2018).
109. Zhang R-G, *et al.* TEsorter: an accurate and fast method to classify LTR-retrotransposons in plant genomes. *Horticulture Research* **9**, uhac017 (2022).
110. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular biology and evolution* **26**, 1641-1650 (2009).
111. Diesh C, *et al.* JBrowse 2: a modular genome browser with views of synteny and structural variation. *Genome biology* **24**, 74 (2023).
112. Xu D, *et al.* CentIER: Accurate centromere identification for plant genomes. *Plant communications* **5**, (2024).
113. Shi X, *et al.* The complete reference genome for grapevine (Vitis vinifera L.) genetics and breeding. *Horticulture Research* **10**, uhad061 (2023).
114. Melters DP, *et al.* Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome biology* **14**, R10 (2013).

115. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**, 573-580 (1999).

116. Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. *Trends in genetics* **16**, 276-277 (2000).

117. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution* **30**, 772-780 (2013).

Figures

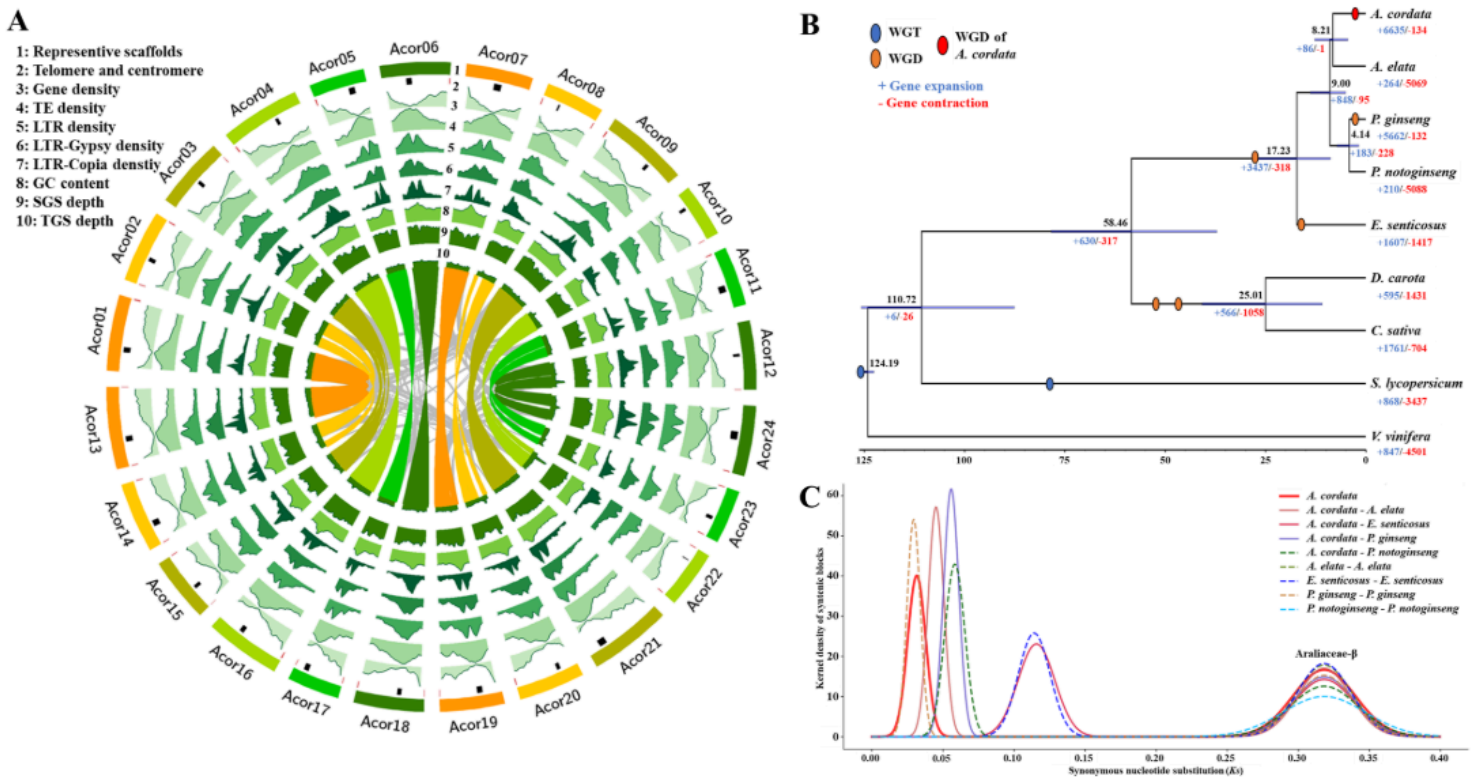


Figure 1

Structure and evolution of the *Aralia cordata* genome. A) Genomic landscape of *A. cordata*. Colored lines represent recent genome duplication (Ac-α) blocks and gray lines represent paleo genome duplication blocks (Araliaceae-β trace). (1) Representative scaffolds. (2) Telomere (red) and centromere (black) peaks. (3) Gene density. (4) TE density. (5) LTR density. (6) Gypsy-type LTR density. (7) Copia-type LTR density. (8) GC content. (9) Short reads depth. (10) Long reads depth. B) Phylogenetic tree of *A. cordata* and others. The blue dot on the branch indicated a reported WGT, the orange dot indicated a reported WGD, and the red dot indicated a WGD of *A. cordata*. The black letters on the nodes indicate the estimated divergence time. Blue letters and red letters indicate the expansion and contraction of gene families, respectively. C) Distribution of synonymous nucleotide substitution rates (*Ks*) of syntenic blocks. The X axis is synonymous nucleotide substitution rates, and the Y axis is kernel density of syntenic blocks.

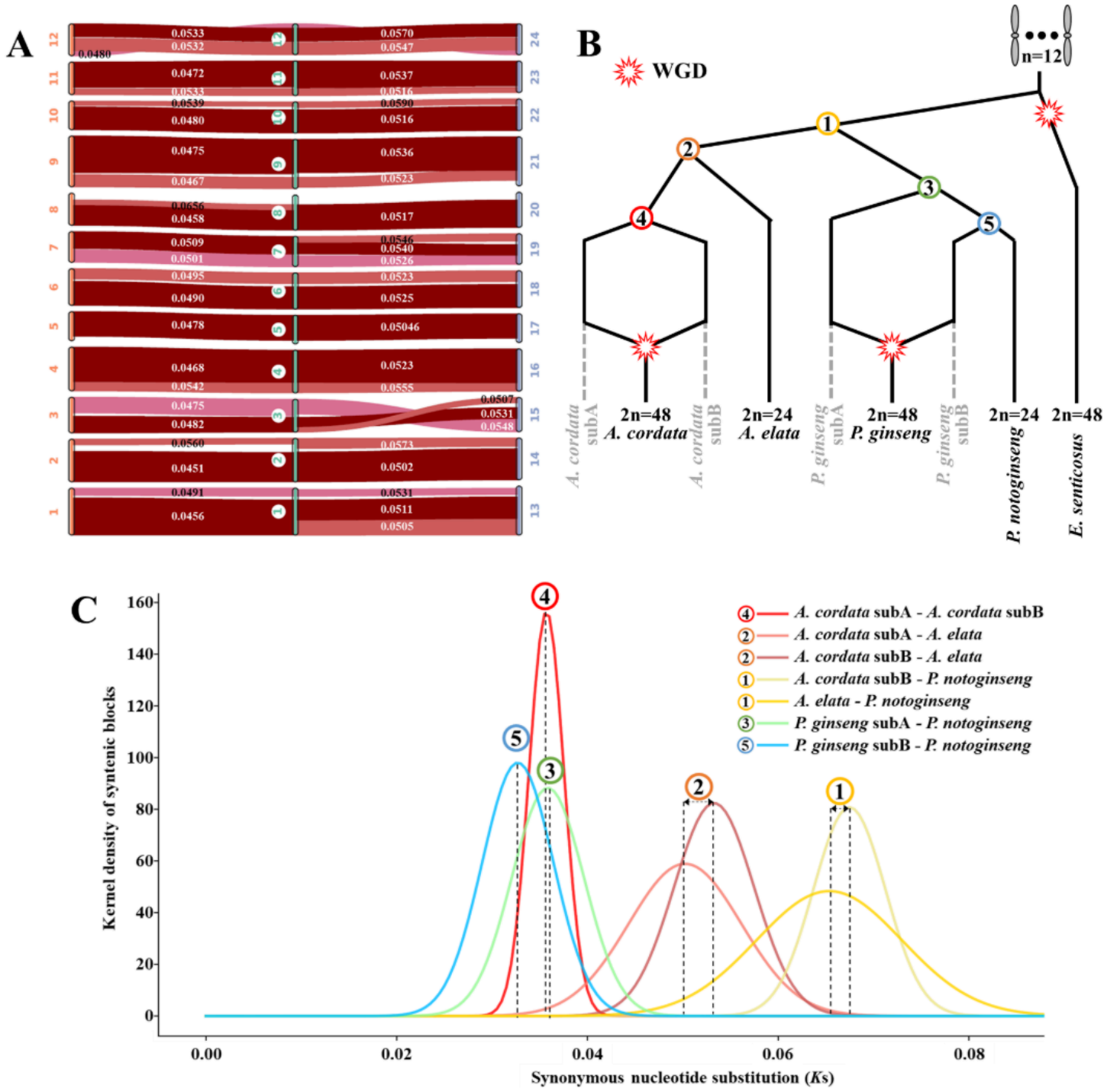


Figure 2

Subgenome evolution of *A. cordata*. A) Subgenome division based on median *Ks* of synteny blocks. B) Model of genome duplication in the *Aralia* and *Panax* genus. C) *Ks* distribution between subgenomes in the *Aralia-Panax* genus. Numbers marked in the circles of B) and C) indicate big events for genome evolution in the *Aralia-Panax*genus. □: Two peaks between *Aralia* and *Panax*. □: Two peaks between two subgenomes of *A. cordata* and *A. elata*. □: Peaks between *P. ginseng* subA and *P. notoginseng*. □ Peaks between two subgenomes of *A. cordata*. □: Peaks between *P. ginseng* subB and *P. notoginseng*.

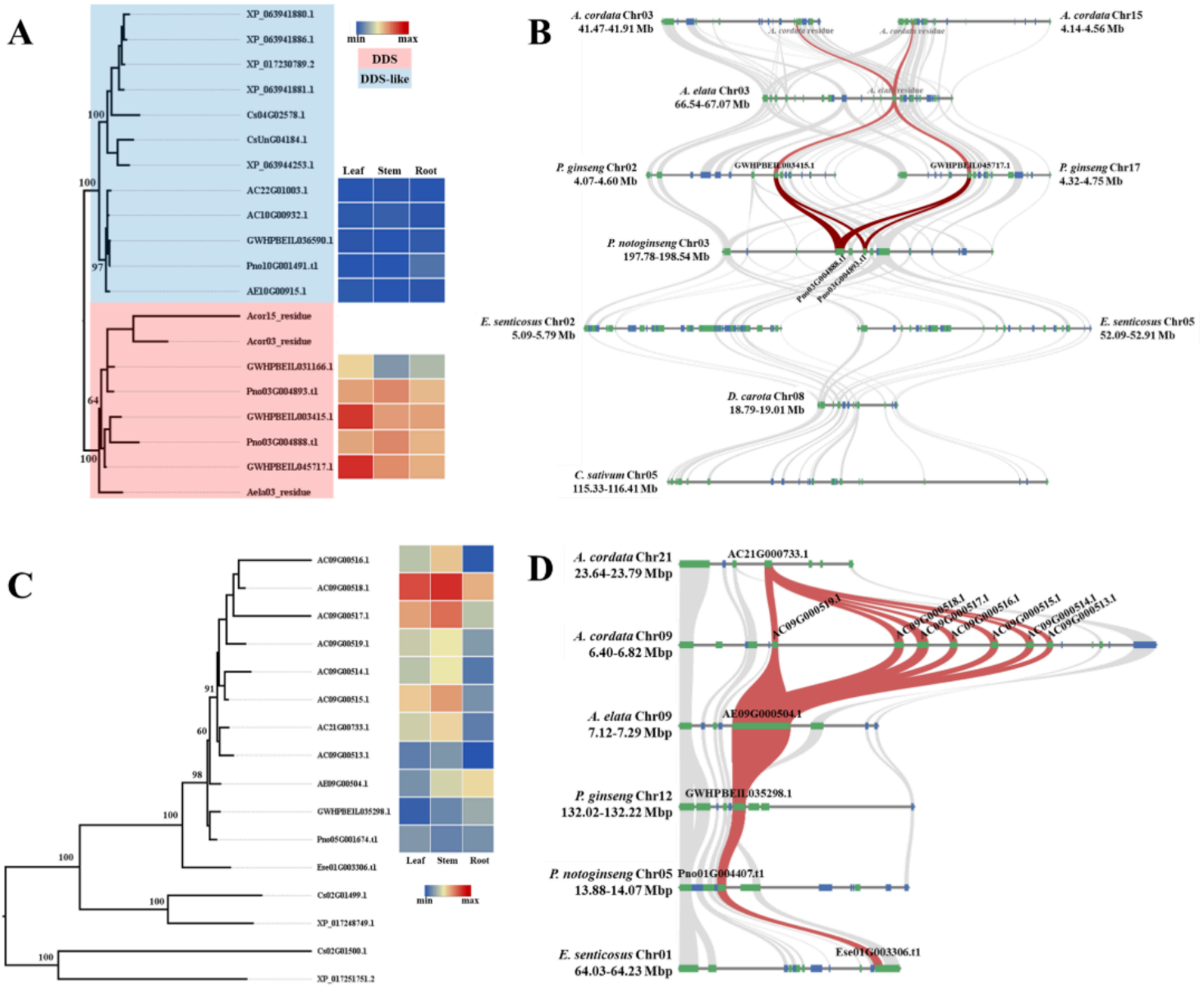


Figure 3

Genes related to secondary metabolite biosynthesis distinctions. A) Phylogenetic tree of DDS and DDS-like genes with their expression levels. The red shadow indicated DDS, while the blue shadow indicated DDS-like. B) Micro-synteny blocks of DDS and DDS residues of *A. cordata* and *A. elata*. C) Phylogenetic tree of TPS-e/f genes with their expression levels. D) Micro-synteny blocks of tandemly duplicated TPS-e/f genes of *A. cordata*.

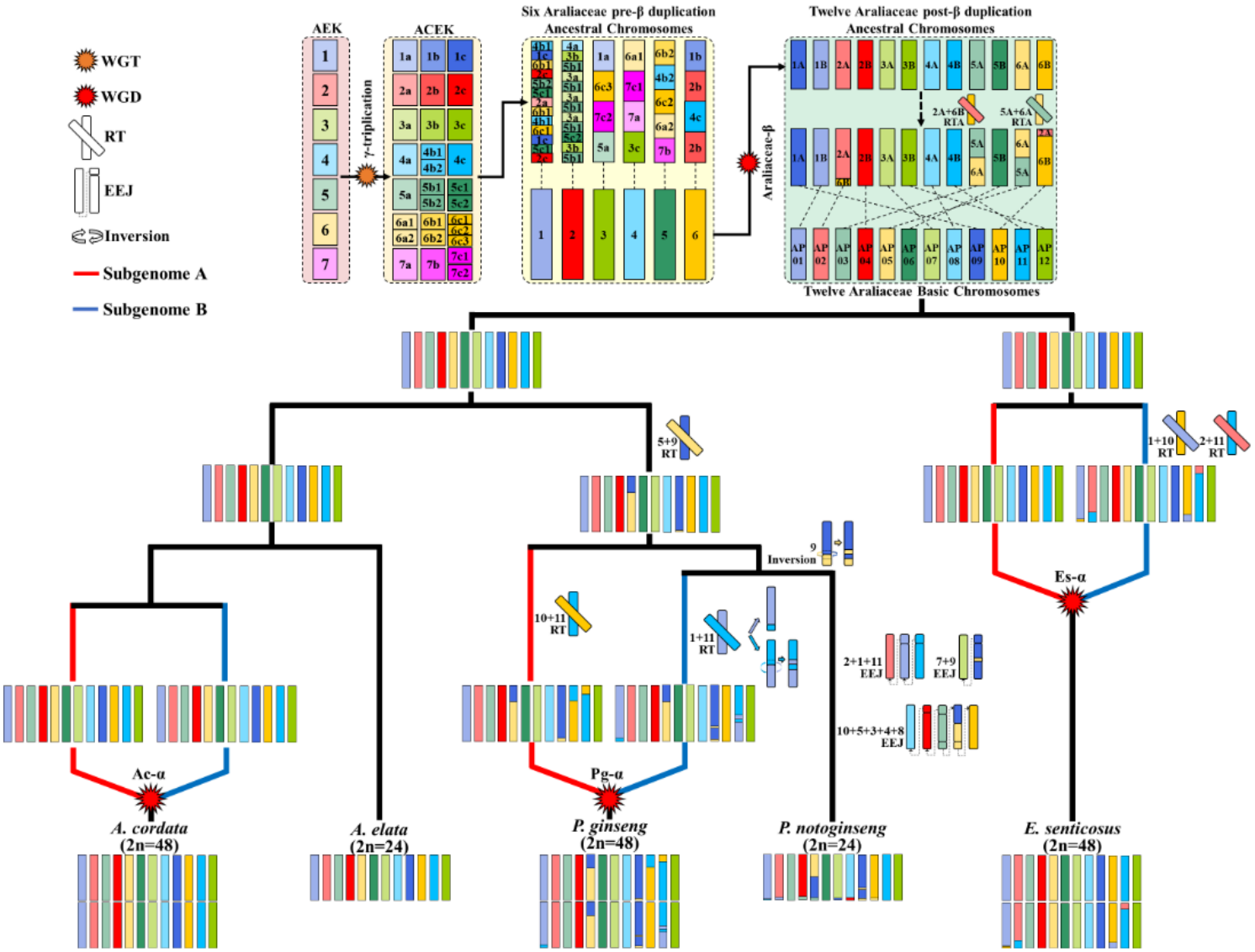


Figure 4

The chromosome evolution scenario of the Araliaceae family. Red and blue lines indicated the divergence between subgenome A and subgenome B. RT: Reciprocal translocation, EEJ: End-to-end joining.

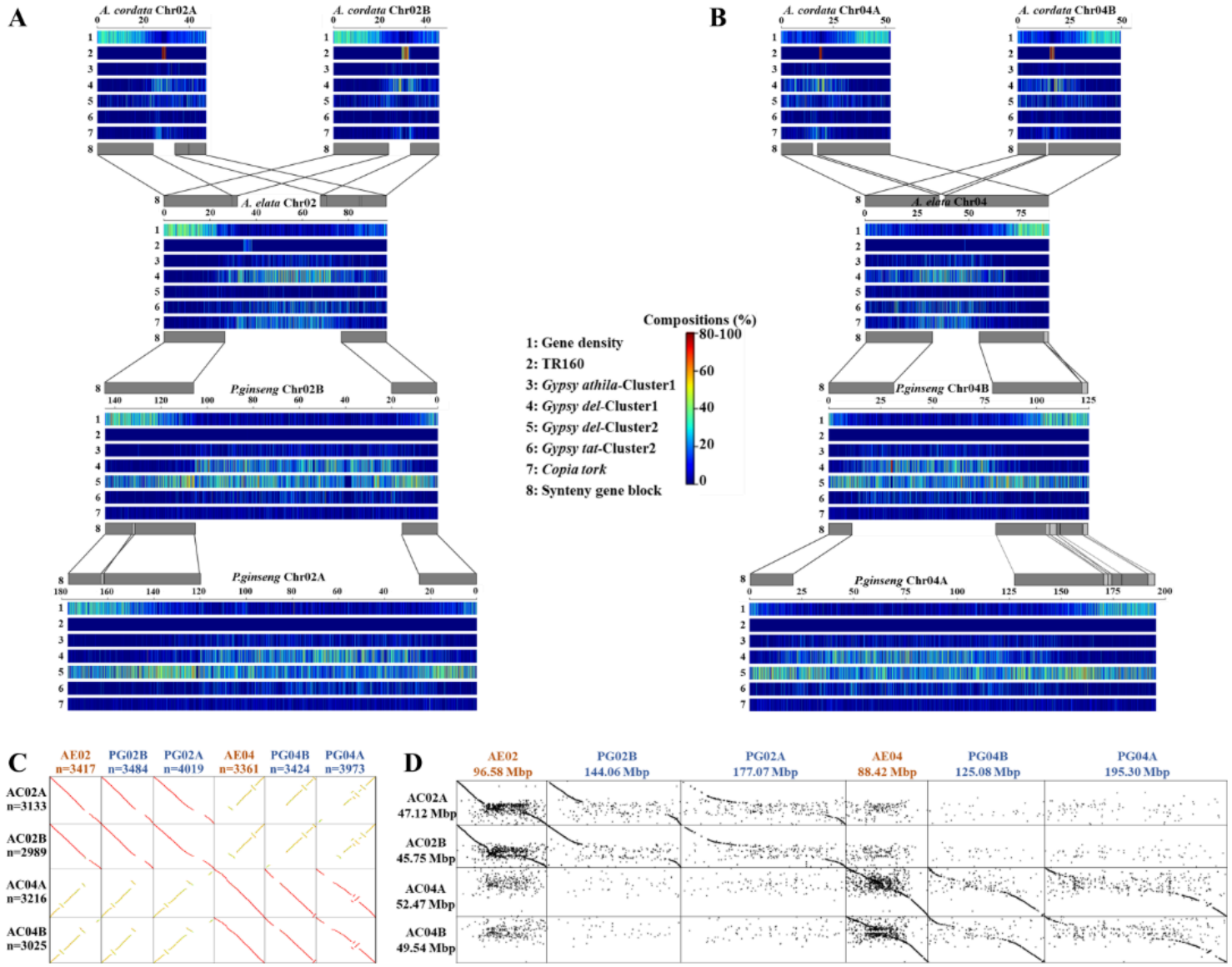


Figure 5

Chromosome-level landscape differences. A) and B) Whole chromosome relationships and genomic compositions of Chr02 and Chr04. C) and D) Alignment relationships of Chr02 and Chr04 among *A. cordata* (AC), *A. elata* (AE), and *P. ginseng* (PG). A and B represented each subgenome A and B. Black letters marked on the left indicate *A. cordata*. Dark orange and dark blue letters marked on the top indicate *A. elata* and *P. ginseng*, respectively. C) Synteny gene block alignment relationships. The number marked on the bottom of the species indicated gene numbers. D) Global genome alignment relationships. The number marked on the bottom of the species indicated the lengths of the chromosomes.

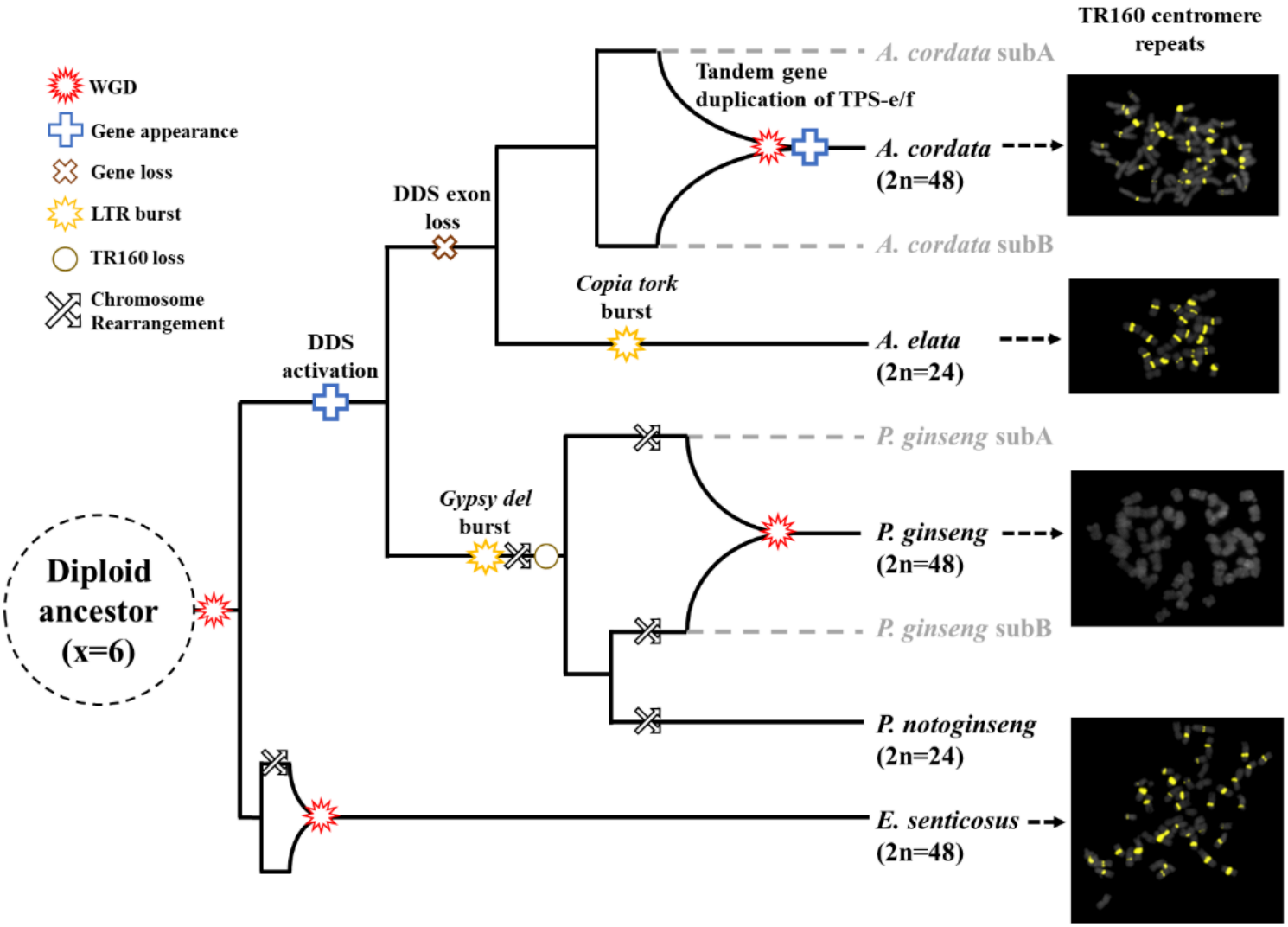


Figure 6

Evolutionary scenario of the Araliaceae family. Different speciation drivers on the evolutionary paths were marked on the branches.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AcordatagenomeSupplementary.pdf
- SupplementaryData1.xlsx