




A chromosome-level genome assembly for *Onobrychis viciifolia* reveals gene copy number gain underlying enhanced proanthocyanidin biosynthesis

Junyi He^{1,4}, Danyang Tian^{1,4}, Xue Li^{1,4}, Xuemeng Wang¹, Tingting Wang¹, Ziyao Wang¹, Hui Zang¹, Xiaofan He², Tiejun Zhang², Quanzheng Yun³, Rengang Zhang³ , Jishan Jiang¹, Shangang Jia^{1,5}  & Yunwei Zhang^{1,5} 

Sainfoin (*Onobrychis viciifolia*), which belongs to subfamily Papilionoideae of Leguminosae, is a vital perennial forage known as “holy hay” due to its high contents of crude proteins and proanthocyanidins (PAs, also called condensed tannins) that have various pharmacological properties in animal feed, such as alleviating rumen tympanic disease in ruminants. In this study, we select an autotetraploid common sainfoin ($2n = 4x = 28$) and report its high-quality chromosome-level genome assembly with 28 pseudochromosomes and four haplotypes (~1950.14 Mb, contig N50 = 10.91 Mb). The copy numbers of genes involved in PA biosynthesis in sainfoin are significantly greater than those in four selected Fabales species, namely, autotetraploid *Medicago sativa* and three other diploid species, *Lotus japonicus*, *Medicago truncatula*, and *Glycine max*. Furthermore, gene expansion is confirmed to be the key contributor to the increased expression of these genes and subsequent PA enhancement in sainfoin. Transcriptomic analyses reveal that the expression of genes involved in the PA biosynthesis pathway is significantly increased in the lines with high PA content compared to the lines with medium and low PA content. The sainfoin genome assembly will improve our understanding of leguminous genome evolution and biosynthesis of secondary metabolites in sainfoin.

¹College of Grassland Science and Technology, China Agricultural University, 100193 Beijing, China. ²School of Grassland Science, Beijing Forestry University, 100083 Beijing, China. ³Department of Bioinformatics, Ori (Shandong) Gene Science and Technology Co., Ltd., Weifang 261322, China. ⁴These authors contributed equally: Junyi He, Danyang Tian, Xue Li. ⁵These authors jointly supervised this work: Shangang Jia, Yunwei Zhang. ✉email: shangang.jia@cau.edu.cn; zywei@126.com

Sainfoin (*Onobrychis viciifolia*) is a significant perennial forage that belongs to subfamily Papilionoideae of Leguminosae and is known as “holy hay” due to its >15% crude protein content¹. It is mostly tetraploid ($2n = 4x = 28$) and cross-pollinated, which leads to high levels of genetic diversity and phenotypic variations^{2–5}. For example, there were significant differences in winter survival rate, dry matter yield, seed yield, and stem number among sainfoin germplasms^{6,7}. In addition, higher levels of crude protein, moderate levels of soluble sugar, and lower levels of acid and neutral detergent fiber contributed to the palatability of livestock feed made from *O. viciifolia*. It was reported that the palatability of *O. viciifolia* to ruminants and the animal productivity level per unit feed consumption are equivalent to or higher than those of alfalfa (*Medicago sativa*)⁸. Meanwhile, proanthocyanidins (PAs), also called condensed tannins (CTs), are highly abundant in *O. viciifolia* and distributed in almost all organs except for roots and cotyledons^{9–11}. Studies have shown that PAs play multiple key roles in ruminant feeding¹², including alleviating rumen tympanic disease in ruminants, improving protein utilization, and increasing anti-parasitic activity in the rumen^{12–18}. Owing to the high PA content and its benefits, *O. viciifolia* has the potential to be used for forage feeding and industrial production of PAs.

PAs can be found in a variety of forage species of Fabaceae, including big trefoil and bird's-foot trefoil in *Lotus*¹⁵, white clover and red clover in *Trifolium*^{19,20}, alfalfa in *Medicago*^{21,22}, and sainfoin²³. In previous studies, researchers discovered that the PA content of *O. viciifolia* was ~80 g/kg, while that of *M. sativa* was ~0.5 g/kg¹⁵. Meanwhile, PA was found in all parts of *O. viciifolia*, while in *M. sativa*, it was enriched mainly in the seed coat¹⁵. Despite the variations in PA contents, these species might share similar PA biosynthetic pathways, according to studies in *Arabidopsis thaliana*, *Medicago truncatula* and other model plants^{24,25}. PAs are produced through phenylpropanoid and flavonoid pathways and share most of the biosynthetic pathways with anthocyanins, only with a split downstream²⁶. Meanwhile, recent studies have shown that a large number of transcription factors (TFs), including R2-R3 MYB, bHLH, and WD40 TFs, are involved in facilitating multiple enzymatic steps in PA biosynthesis^{27,28}. However, the orthologous genes for PA biosynthesis in the *O. viciifolia* genome are unknown.

It is widely believed that sainfoin originated in Asia, was introduced to Europe via Arabs in the fourteenth century, and was used for hay and seed production on a large scale in the twentieth century²⁹. *O. viciifolia* phylogenetically falls into the clade of common leguminous species (Hologalegina, Papilionoideae, Fabaceae), including *Pisum sativum*, *M. sativa*, and *Cicer arietinum*. To date, multititudinous genome assemblies of leguminous species, including chickpea (*C. arietinum*), cultivated soybean (*Glycine max*), alfalfa (*M. sativa*) and Chinese milk vetch (*Astragalus sinicus*)^{30–33}, have greatly promoted our understanding of the evolutionary history of Fabaceae. Alfalfa has four genome assemblies available, including those for “Xinjiang Daye”, *M. sativa* ssp. *caerulea*, “Zhongmu No. 1”, and “Zhongmu No. 4”. The latest genome version of Zhongmu No. 4 included four haplotypes for 32 chromosomes ($2n = 4x = 32$), with a genome size of 2.74 Gb and contig N50 of 2.06 Mb, showing multiple genome evolution events in alfalfa compared to *M. truncatula*^{32,34–36}. Chromosome-level assembly of autopolyploid genomes is a major challenge, especially for species with a large genome size. Similar to alfalfa, *O. viciifolia* is also an autotetraploid legume forage, but its genome is not available for exploring genome evolutionary history. Available genomic resources for *O. viciifolia* are rare, and scientists have revealed the genetic diversity and phylogenetic relationships of sainfoin by using amplified fragment length polymorphisms (AFLPs),

inter-simple sequence repeats (ISSRs), simple sequence repeats (SSRs), expressed sequence tag-derived simple sequence repeats (EST-SSRs), and single nucleotide polymorphisms (SNPs)^{37–41}. The lack of a high-quality, chromosome-level genome assembly for sainfoin has hampered its genetic research and improvement, although the complete chloroplast genome sequence of sainfoin and some transcriptomic resources are available⁴².

In this study, we present a high-quality, chromosome-level genome assembly of common sainfoin (Fig. 1a). The assembly contains 28 pseudochromosomes with four haplotypes. We confirmed that gene expansion, which was driven by a combination of autotetraploidization, whole-genome duplication, and fragmental duplication events, increased the expression of genes associated with the production of PAs, which contributed to the enhanced PA content in *O. viciifolia* leaves. Our results provide a basis for understanding the genomic evolution of the genus *Onobrychis* and legume species and for accelerating genetic breeding in *Onobrychis* species and functional genomic studies in *O. viciifolia*.

Results

Assembly and annotation of the sainfoin genome. It was noted previously that a single haplotype of the autotetraploid genome lost many genes⁴³. Therefore, we sought to assemble the four haplotypes of autotetraploid *O. viciifolia*. We sequenced the genomic DNA of one *O. viciifolia* plant by combining Oxford Nanopore technology (ONT), Illumina NovaSeq and Hi-C and generated ~200 Gb (coverage: ~109.89 X), 175 Gb (coverage: ~96.15 X), and 300 Gb (coverage: ~164.84 X) of data, respectively (Supplementary Table 1). Using *K*-mer analysis ($k = 17$), the genome size of sainfoin was estimated to be ~1821.52 Mb, which is close to the result of ~1851.75 Mb obtained by flow cytometry (Supplementary Fig. 1a, b). To accurately assemble the sainfoin genome, we first performed the preliminary assembly of four haplotypes based on ONT clean reads and used short reads to polish the assembled contigs. First, we preliminarily assembled the contigs based on the Hi-C data and 3D-DNA, and the process effectively corrected some errors; however, there were also many errors in the output result. Then, we corrected the errors manually by Juicebox, and the process did not have a quantitative standard but followed expectations based on the principle that an interaction signal with a shorter distance is stronger than one with a longer distance in the Hi-C auxiliary assembly process. Based on our optimized pipeline, the nuclear assembly (~1950.14 Mb) and phasing placed 1044 contigs (N50 = 10.92 Mb) onto 28 pseudochromosomes (Supplementary Table 2), which represented four haplotypes with seven chromosomes (4×7 chromosomes, ~1892.58 Mb) and unplaced contigs (57.56 Mb) (Fig. 1b, c, Supplementary Fig. 2 and Supplementary Table 3). However, we found a low level of switch error and collapsed contig (Supplementary Fig. 2) due to the homologous regions among the four haplotypes when the assembled contigs were placed in 28 pseudochromosomes. Based on the ONT long reads, we found 770 collapsed regions and 40 switch errors by using Inspector program, whose lengths reached 0.17 Mb and 0.018 Mb respectively (Supplementary Fig. 3). The GC content of the entire genome was 34.62% (Supplementary Table 4), and the average content of the four haplotypes was 34.61%, which was close to the 34.97% of *Glycine max* in Fabaceae and 34.86% of *Vitis vinifera* in Pentapetalae (Fig. 1d, Supplementary Table 5 and Supplementary Data 1). Despite collapsed contigs and switch errors, the four haplotypes (labeled Haplotypes A~D) showed similar total lengths ranging from 458.4 Mb to 482.9 Mb (Fig. 1c and Supplementary Table 5).

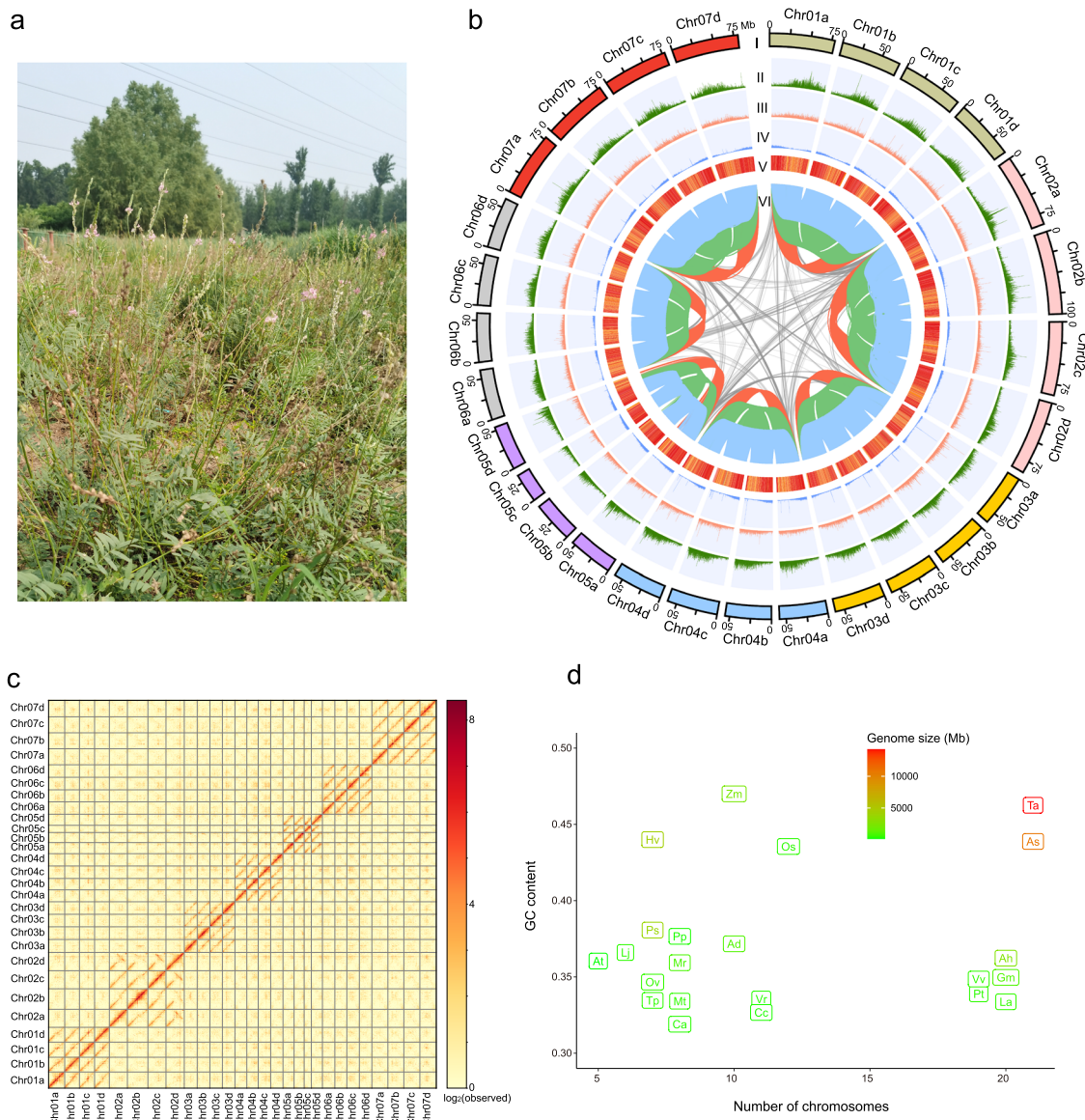


Fig. 1 Distribution of genomic features within the genome assembly of *O. viciifolia*. **a** *O. viciifolia* grown in the field, **b** genomic features of *O. viciifolia*. **I**, Twenty-eight chromosomes of *O. viciifolia*; **II**, Density of Gypsy elements; **III**, Density of Copia elements; **IV**, Density of genes; **V**, GC content; **VI**, Links of collinear gene blocks of *O. viciifolia* among four haplotypes. Blue ribbons indicate synteny blocks between chra and chrb, chrb and chrc, and chrc and chrd; green ribbons indicate synteny blocks between chra and chrc and chrb and chrd; and red ribbons indicate synteny blocks between chra and chrd. Each sliding window was 100 kb. **c** Hi-C interaction heatmap of 28 chromosomes in *O. viciifolia*. **d** GC content vs. chromosome number of Fabaceae and non-Fabaceae species, At-*A. thaliana*, Pp-*P. persica*, Pt-*P. trichocarpa*, Vv-*V. vinifera*, Ad-*A. duranensis*, Ah-*A. hypogaea*, La-*L. angustifolius*, Cc-*C. cajan*, Vr-*V. radiata*, Gm-*G. max*, Lj-*L. japonicus*, Ov-*O. viciifolia*, Ca-*C. arietinum*, Ps-*P. sativum*, Tp-*T. pratense*, Mr-*M. ruthenica*, Mt-*M. truncatula*, Zm-*Z. mays*, Os-*O. sativa*, Hv-*H. vulgare*, Ta-*T. aestivum*, As-*A. sativa*.

We identified 3,147,173 repetitive sequences with a total length of 1239.64 Mb (63.55% of the whole genome assembly) (Supplementary Table 6). We also annotated 1,033,972 long terminal repeat (LTR) retrotransposons (~36.14%), which were the most abundant transposable elements (TEs) in the genome. There were 474,059 simple repeat sequences, 381,092 Gypsy elements, and 218,580 Copia retrotransposons in the 28 pseudochromosomes. Gypsy and Copia LTR retrotransposons accounted for 17.4% and 7.7%, respectively (Supplementary Table 6). Combining ab initio prediction and evidence-based methods, 109,998 high-confidence genes were identified, which contained 719,988 exon domains (Supplementary Table 7). The average number of genes of the four haplotypes was 27,284,

similar to the 27,571 of *Cicer arietinum* and 28,251 of *Lotus japonicus* (Supplementary Table 5)^{30,44}.

To evaluate genome assembly quality, we downloaded RNA-Seq (Supplementary Table 8) data from the NCBI SRA and mapped the clean reads to the genome, along with short and long DNA reads. The mapping rates were 99.0%, 97.1%, and 91.7% for short DNA reads, long DNA reads, and RNA-Seq short reads, respectively (Supplementary Table 9). We further identified 1315 (91.3%) of the 1440 total core genes in the BUSCO (Benchmarking Universal Single-Copy Orthologs) analysis based on embryophyta_odb9, including 108 single-copy BUSCOs (7.5%), 1207 duplicated BUSCOs (83.8%), 23 fragmented BUSCOs (1.6%), and 102 missing BUSCOs (7.1%), while the integrated proteins

covered 97.2% of the complete core genes (Supplementary Table 10). We also evaluated four haplotypes of the genome assembly by BUSCOs, and the results showed that 91.3% ~ 93.0% of BUSCOs were matched for the annotated proteins of four haplotypes (Supplementary Table 5). We annotated 99,518 coding genes against the databases (90.47% of the total genes), and the majority of genes received a functional assignment, for example, 98.51% for the nonredundant Nr protein sequence database, 97.86% for Clusters of Orthologous Genes (COG), and 98.67% for the protein database translated from EMBL (TrEMBL) (Supplementary Table 11). These results indicated high integrity and quality of the genome assembly. Meanwhile, by assembling the four haplotypes, we obtained 99,145 annotated genes that could be anchored to the 28 chromosomes and accounted for 90.84% of the total number of genes anchored to the chromosomes (Supplementary Fig. 4 and Supplementary Tables 12–15). We also assessed the assembly based on the long terminal repeat assembly index (LAI). The LAIs were 6.49, 5.79, 7.05, and 4.78 for the four haplotypes (Supplementary Table 16). This result suggested that the four haplotypic assemblies reached the medium level, while the LAI was 3.67 for *C. arietinum*, 3.79 for *C. cajan*, and 10.72 for *M. truncatula*. And the average quality value of the whole genome reached up to 33.4, with an error rate of 0.00046 and the completeness of 93.98% (Supplementary Table 17).

By considering the genome information of the four haplotypes of autotetraploid *O. viciifolia*, we obtained a relatively complete genome for *O. viciifolia* (Fig. 1b and Supplementary Tables 12 and 13). Although there were some errors in phasing the highly similar contigs of the four haplotypes (Supplementary Figs. 2 and 3), the assembled genome of the four haplotypes of *O. viciifolia* is expected to have a positive impact on genetic and molecular biology research.

Gene family and evolutionary analysis. Based on 448 single-copy genes and their protein sequences obtained from OrthoFinder2, we built a phylogenetic tree of 19 representative species primarily from Fabales and showed that *O. viciifolia* phylogenetically neighbors *L. japonicus* and *C. arietinum* (Fig. 2a). We estimated the divergence time for this phylogenetic tree and found that in the Hologalegina clade, the divergence time of *O. viciifolia* from *L. japonicus* was ~44.3 million years ago (Mya), and *O. viciifolia* speciation started ~37.6 Mya, which was much earlier than that of the other Hologalegina species in the tree, for example, ~6.5 Mya for the split of *M. sativa* (Zhongmu No. 1) and *M. truncatula*. In contrast, the Fabales clade was estimated to have diverged from non-Fabales species ~99.1 Mya. We discovered 27,749 homologous gene families by comparing the common sainfoin (*O. viciifolia*) genome with the genomes of other species in the phylogenetic tree and 1205 expanded and 4371 contracted gene families in *O. viciifolia*, which were similar to those of other Fabales species. KEGG enrichment analysis revealed that the expanded gene families were mainly enriched in the pathways of biosynthesis of cofactors, amino sugar and nucleotide sugar metabolism, and flavonoid biosynthesis, while the contracted gene families were enriched in the pathways of plant hormone signal transduction, biosynthesis of various plant secondary metabolites, and cyanoamino acid metabolism (Supplementary Fig. 5). Furthermore, we identified the lowest number of multiple-copy gene families in *O. viciifolia* compared to *L. japonicus*, *C. arietinum*, and *M. sativa* (Supplementary Fig. 6), which was contrary to the higher number of copies of genes involved in PA biosynthesis (see the following results). Meanwhile, we compared gene families among the four selected species and found 11,257 common conserved gene clusters and

333 unique gene families in *O. viciifolia* (Fig. 2b). In the KEGG enrichment analysis, the unique gene families in *O. viciifolia* were enriched in the pathways of “zeatin biosynthesis”, “taurine and hypotaurine metabolism”, and “alanine, aspartate and glutamate metabolism” (Supplementary Table 18).

LTR insertion and genome evolution. We discovered numerous long terminal repeats (LTRs) in the *O. viciifolia* genome, and we calculated the LTR insertion time for six Fabales species: *O. viciifolia*, *M. sativa*, *M. truncatula*, *C. arietinum*, *C. cajan*, and *V. radiata*. The results showed that the peak insertion time of *O. viciifolia* was ~0.033 million years ago (Mya), which was similar to that of *M. sativa* (~0.038 Mya) and *M. truncatula* (~0.077 Mya) and more recent than those of *C. arietinum* (~0.529 Mya), *C. cajan* (~2.188 Mya) and *V. radiata* (~1.653 Mya) (Fig. 2c and Supplementary Data 2).

We investigated the history of genome evolution in *O. viciifolia* based on the synonymous substitution rate (K_s), with special interest in whole-genome duplication (WGD) and whole-genome triplication (WGT) events, and found that the distribution of K_s values confirmed the Papilionoideae whole-genome duplication (PWGD) for legume species (*A. hypogaea*, *C. arietinum*, *G. max*, *L. japonicus*, and *M. sativa*) and the γ polyploidization event (γ -WGT) for nonlegume species (*P. trichocarpa* and *V. vinifera*) (Fig. 3a, Supplementary Fig. 7 and Supplementary Data 3). The K_s peak at ~0.5–1.0 based on paralogous genes in five species indicated the presence of the shared PWGD, and another peak at ~0–0.5 in *G. max* and *A. hypogaea* (Supplementary Fig. 7) corresponded to one recent additional WGD event, which was consistent with previous reports^{45,46}. We found that the orthologous genes that experienced the PWGD event were mainly enriched in KEGG pathways related to signal transduction (Supplementary Table 19) and enriched in biological processes related to signal transduction, response to environment, response to hormones, and development of different organs in GO analysis (Supplementary Table 20).

Comparative genomics revealed a significant percentage of orthologous genes with a 2:2 ratio between *O. viciifolia* and the other four species, namely, *C. arietinum*, *M. sativa*, *M. truncatula*, and *L. japonicus* (Fig. 3b, Supplementary Fig. 8 and Supplementary Data 4), which suggested that these Fabales species all experienced a single WGD event, i.e., the PWGD event (Fig. 3c). Furthermore, a ratio of 4:2 was detected between *A. hypogaea* and *O. viciifolia*, which further confirmed the PWGD event in both species and one more WGD event in *A. hypogaea* (Supplementary Fig. 8d).

Genome synteny and chromosomal arrangements. Genome collinearity of *O. viciifolia* was explored among the selected species *C. arietinum*, *M. sativa*, *M. truncatula*, *P. sativum*, *L. japonicus*, *G. max*, *V. radiata*, *C. cajan*, *A. hypogaea*, *A. duranensis*, *P. persica*, *P. trichocarpa*, and *V. vinifera* (Supplementary Fig. 9), and a high level of collinearity was discovered, especially between *O. viciifolia* and some Fabales species, such as *L. japonicus*, *C. arietinum*, *P. sativum*, *M. sativa*, and *M. truncatula*, which are all closer than *P. persica*, *P. trichocarpa*, and *V. vinifera* to *O. viciifolia* in the phylogenetic tree. Several chromosomes in *C. arietinum* showed a high level of collinearity with *O. viciifolia*, such as OvChr1 vs. CaChr7, OvChr4 vs. CaChr3, OvChr5 vs. CaChr6, OvChr6 vs. CaChr5, and OvChr7 vs. CaChr4 (Supplementary Fig. 9a). Meanwhile, chromosomes with high similarity and collinear relationships were found between *O. viciifolia* and *M. sativa*, including OvChr1 vs. MsChr8, OvChr4 vs. MsChr7, OvChr5 vs. MsChr4, OvChr6 vs. MsChr3, and OvChr7 vs. MsChr1 (Supplementary Fig. 9b). The *O. viciifolia* chromosomes

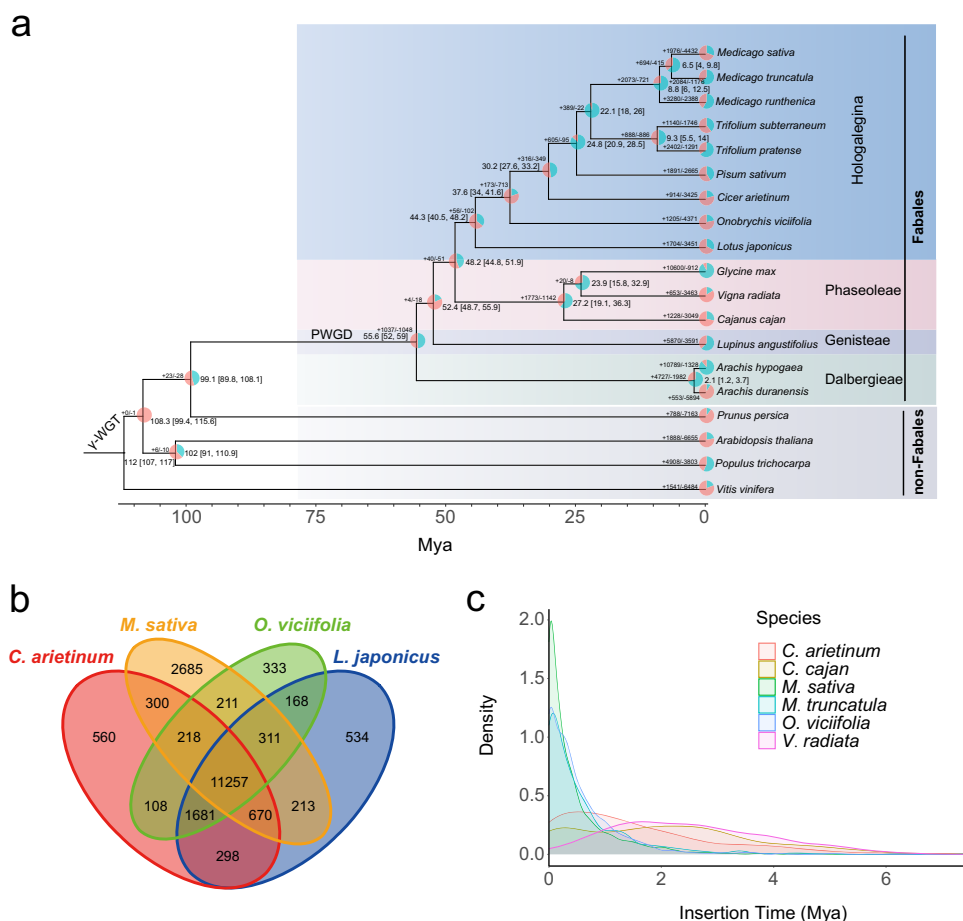


Fig. 2 Gene family and phylogenetic tree of *O. viciifolia* and selected species. **a** Phylogenetic tree of *O. viciifolia* and 18 other Fabales and non-Fabales species. The numbers at each node represent the divergence time, and the time range in brackets is based on the 95% confidence interval. The pie chart represents the expansion and contraction of gene families, with blue for expansion and red for contraction, and the numbers next to the pie chart represent the number of gene families with expansion and contraction. **b** Venn diagram for gene families in *C. arietinum*, *M. sativa*, *O. viciifolia*, and *L. japonicus*. **c** LTR insertion time and density in six Fabales plants.

OvChr1, OvChr4, OvChr5, OvChr6, and OvChr7 were relatively conserved and showed the same ancestral chromosomal structure, as observed in comparison with *C. arietinum*, *P. sativum*, *M. truncatula*, and *M. sativa* (Supplementary Fig. 9a–d). However, they still carried significant variations and chromosomal rearrangements compared to the ancestral eudicot karyotype (AEK), as the seven ancestral chromosomes were still retained in *V. vinifera* with only the γ -WGT event, and collinearity was poor between *V. vinifera* and *O. viciifolia* (Supplementary Fig. 9m).

Further karyotype evolutionary history analysis was performed to determine the ancestral Hologalegina karyotype (AHK) for *O. viciifolia* and six other Hologalegina plant species, *L. japonicus*, *O. viciifolia*, *C. arietinum*, *P. sativum*, *M. truncatula*, *M. sativa*, and *G. max*. The results showed that there were seven ancestral chromosomes in Ancestor1 (Fig. 4). The AHK for Ancestor1 experienced an enormous number of chromosomal rearrangement events and evolved into the modern Hologalegina karyotypes. It was inferred that Ancestor1 experienced 16 fission and 16 fusion events and evolved into the modern *O. viciifolia* karyotype (Supplementary Table 21). Chromosomal conservation was observed and confirmed between *O. viciifolia* and Ancestor1. Conservation could also be found in the complete synteny block between *O. viciifolia* and *C. arietinum* (Supplementary Fig. 9a).

We rebuilt the karyotype evolutionary events for the inferred intermediate ancestral nodes (Supplementary Fig. 10). The results showed that chromosomal breakage of Ancestor1 with 7

chromosomes generated up to 8 chromosomes for Ancestor2, Ancestor3, Ancestor4, and Ancestor5. Ancient *O. viciifolia* lost one chromosome, and modern *O. viciifolia* retained 7 chromosomes.

Expression of genes in the PA biosynthesis pathway. We collected 41 germplasm lines of tetraploid *O. viciifolia* from 13 countries and regions (Supplementary Table 22) and measured PA contents for 59 individuals (Supplementary Table 23). According to the PA content, these lines were divided into three groups, including high (H), medium (M) and low (L), and nine lines were selected (3 typical independent lines with similar PA contents for each group) for transcriptome sequencing to explore the expression levels of genes involved in PA biosynthesis and transport (Supplementary Fig. 11a and Supplementary Tables 23 and 24). Our RNA-seq analysis identified 567 differentially expressed genes (DEGs) for H samples compared to L samples, 631 DEGs for H vs. M samples, and 830 DEGs for M vs. L samples. We conducted KEGG enrichment analysis to identify the differentially expressed genes in the biosynthesis pathway of PAs in sainfoin leaves with high PA content. The up- and downregulated DEGs in the pairwise comparisons of the three groups were mainly enriched in pathways of “carbon metabolism”, “biosynthesis of amino acids”, “phagosome”, etc. (Supplementary Fig. 11b–d).

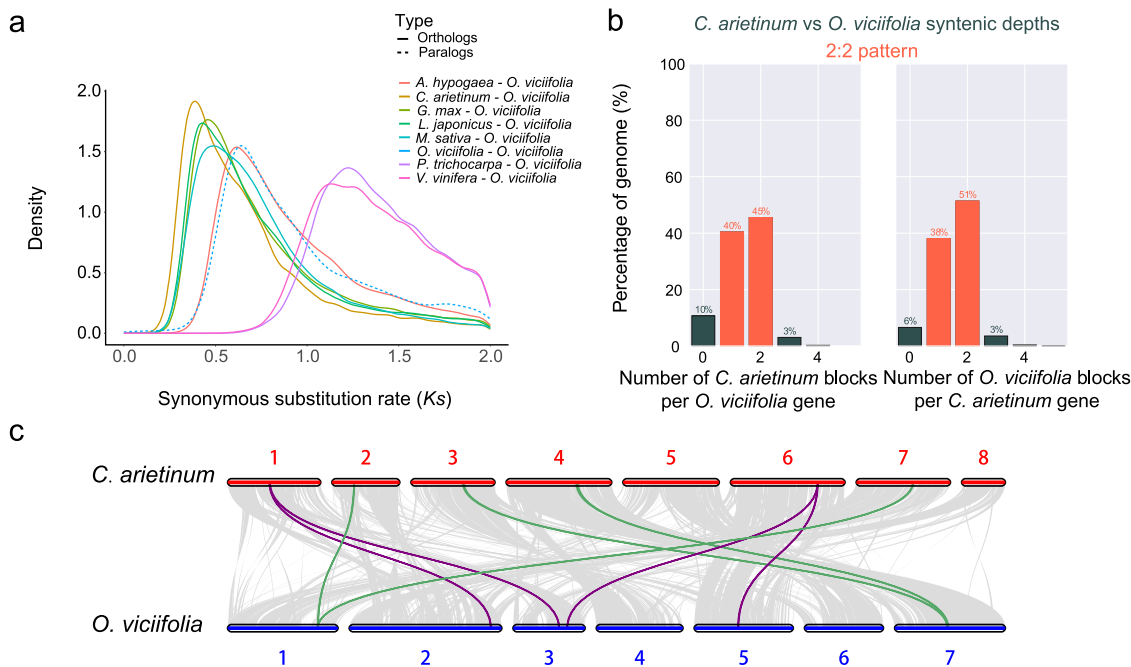


Fig. 3 WGD events and karyotype evolution of *O. viciifolia* and other species. **a** Distribution of synonymous substitution rates (Ks) of homologous gene pairs between *O. viciifolia* and other species. **b** Syntenic depths between *O. viciifolia* and *C. arietinum*. **c** Chromosome-scale synteny analysis between *O. viciifolia* and *C. arietinum*.

Weighted gene coexpression network analysis (WGCNA) was used to generate 9 gene modules after merging similar gene expression modules (Supplementary Fig. 12a). The correlation relationships between traits and gene modules showed that the MEblack module exhibited the highest correlation with PA content. In this module, LeOno02aG0012600, which was annotated as an anthocyanidin synthase (ANS) gene, was the hub gene with 37 other genes in the coexpression network (Supplementary Figs. 12b, 13 and Supplementary Table 25), suggesting a key role of ANS in PA biosynthesis.

Expansion of genes involved in PA biosynthesis. Based on the published literature, we summarized the whole pathway of PA biosynthesis, which started from phenylalanine to PAs (Fig. 5a), and multiple regulators, including synthesis enzymes and TFs, functioning in key steps of the PA biosynthesis pathway. We collected the sequences of genes involved in the pathway of PA biosynthesis in *A. thaliana*, *V. vinifera* and *G. max* (Supplementary Table 26) and determined the orthologous genes and their copy numbers in *O. viciifolia* in comparison with the other four Fabales species, *M. sativa* cultivar XinJiangDaYe”, *Lotus japonicus* “MG20”, *Medicago truncatula* “A17”, and *Glycine max* “Wm82”, based on all the annotated proteins downloaded from the NCBI database. We found that the autotetraploid *O. viciifolia* and *M. sativa* gained more copy numbers than the other three diploid Fabales species (Fig. 5a and Supplementary Fig. 14). Meanwhile, compared to autotetraploid *M. sativa*, we also revealed gene expansion and copy number gain in *O. viciifolia*, as more gene copy numbers were found, for a total of 13 related regulators in the PA biosynthesis pathway, namely, C4H, CHI, F3’/3’5’H, DFR, ANR, MATE, AHA10, LAC, MYB12, TT2, TT8, and TTG1, in *O. viciifolia* (highlighted in red in Fig. 5a). It is worth noting that the copy number of genes encoding O-methyltransferase (OMT), a key enzyme for anthocyanin biosynthesis, was significantly higher in *O. viciifolia* than in the other diploid Fabales species.

Furthermore, we plotted the homologous genes of the PA biosynthesis pathway onto the chromosomes (Fig. 5b and Supplementary Fig. 15) and found many tandemly repeated genes on the chromosomes, such as CHS genes (LeOno02aG0003100, LeOno02aG0003200, LeOno02aG0003300) on chr02a and LAC genes (LeOno04aG0201900, LeOno04aG0202000, LeOno04aG0202100) on chr04a (Fig. 5b). The intensive tandem repeat genes belonged to the paralogous genes, and similar results were found in the other three haplotypes (Supplementary Fig. 15). Based on the results, we inferred that one of the reasons for the enrichment of PAs in sainfoin may be that the generation of tandem repeats contributed greatly to the enrichment of PAs in terms of gene dosage effects.

In fact, it is known that PA content is significantly higher in *O. viciifolia* than in *M. sativa*, although both of these two species are important leguminous forage crops and autotetraploids^{47,48}. To check the effects of gene expansion, RNA-seq and qRT-PCR were performed to uncover the expression levels of genes involved in the PA biosynthesis pathway in the leaves of *O. viciifolia* and *M. sativa*. The RNA-seq results showed that genes were mostly highly expressed in the three *O. viciifolia* groups with high, medium, and low PA contents than in alfalfa cultivar “Zhongmu No. 1” (Supplementary Table 27), with GADPH, CYP, and EF1a as internal reference genes. qRT-PCR further confirmed the significantly higher expression levels of PAL, ANR, ANS, F3’5’H, and F3H in the leaves of *O. viciifolia* compared to *M. sativa* (Supplementary Fig. 16). These results suggest the contribution of gene expansion to the expression of genes involved in PA biosynthesis in *O. viciifolia*.

Furthermore, we found increased expression of genes involved in PA biosynthesis in the *O. viciifolia* lines with high PA content by using RNA-seq analysis of three *O. viciifolia* groups (H, M, and L). The expression levels of genes involved in the PA biosynthesis pathway were mostly all significantly increased in Group H (left in the heatmap of fold changes of gene expression) compared to M (middle) and L (right) (Fig. 5a). In addition, the genes *OMT* and *UGT*, which are responsible for anthocyanin

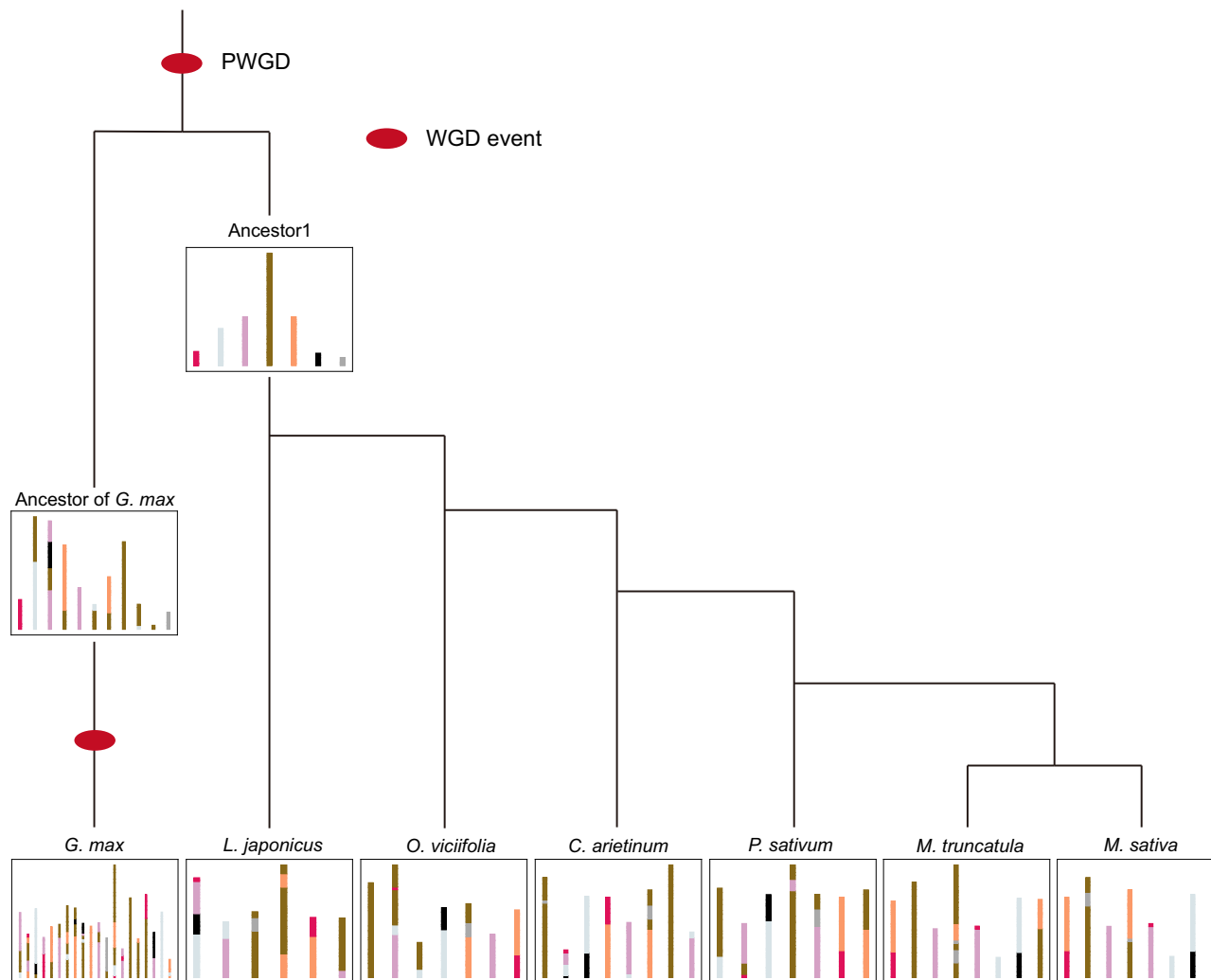


Fig. 4 Karyotype evolutionary history of selected leguminous species. Ancestor1 was the common ancestor of selected Hologalegina plants, and chromosomes are highlighted in different colors.

biosynthesis, were downregulated in Group H, which implied that the biosynthesis pipeline predominantly involved the production of PA, rather than anthocyanin, in the leaves of *O. viciifolia* lines with high PA content. Taken together, these results indicate that gene expansion is the key driving force for the enhancement of PA production in *O. viciifolia*.

Discussion

O. viciifolia is a widely cultivated leguminous forage with high contents of both protein and PA, and the *O. viciifolia* genome assembly, as the one for Hedysareae plants, is valuable for understanding the genome evolution and phylogenetic relationships of Fabales species, for which >10 genome assemblies are available⁴⁹. However, it is a major challenge to assemble the genome of *O. viciifolia* since it is autotetraploid and cross-pollinated. In our study, we applied next-generation technology and Hi-C technology to obtain a relatively complete tetraploid genome for sainfoin.

It is valuable but difficult to assemble the whole genome, including four haplotypes, of autotetraploids to obtain whole-genome information, which mainly includes unique genes existing in different haplotypes. The reason for the difficulty is the lack of ability to identify and separate the extremely similar haplotypes⁵⁰. Currently, with the help of new sequencing

technology, especially third-generation sequencing technology for long reads and high-throughput chromatin conformation capture technology, a highly prominent number of complete genomes of autotetraploid plants have been assembled and reported^{34,36,51–54}. In summary, there are three main strategies used to assemble the four haplotypes of autotetraploid genomes. One of them is to assemble one haplotype of the autotetraploid, such as *M. sativa* “Zhongmu No. 1”³², but the disadvantage is that many genes might be lost because of the mixed nature of the four haplotypes under cross-pollination. The second strategy is to assemble the four haplotypes of autotetraploids based on third-generation sequencing and Hi-C data, such as the recent allele-aware genomes of autotetraploid “Zhongmu No. 4”³⁶, autotetraploid potato “Q9”⁵³, tetraploid highbush blueberry cultivar “Draper”⁵², and autopolyploid sugarcane “Np-X”⁵⁴. The third strategy of autotetraploid assembly is to use additional methods to improve autotetraploid genome phasing, such as linkage grouping in a population of hybrid progenies⁵¹ or single-gamete sequencing⁴³. Based on the four haplotypes of autotetraploids, researchers discovered many more genes and structural rearrangements, and the results markedly promoted the development of genomics and functional genomics.

In this study, we used the second strategy and successfully assembled four haplotypes of the *O. viciifolia* chromosome-level genome. The contig N50 of 10.41 Mb was much longer than those

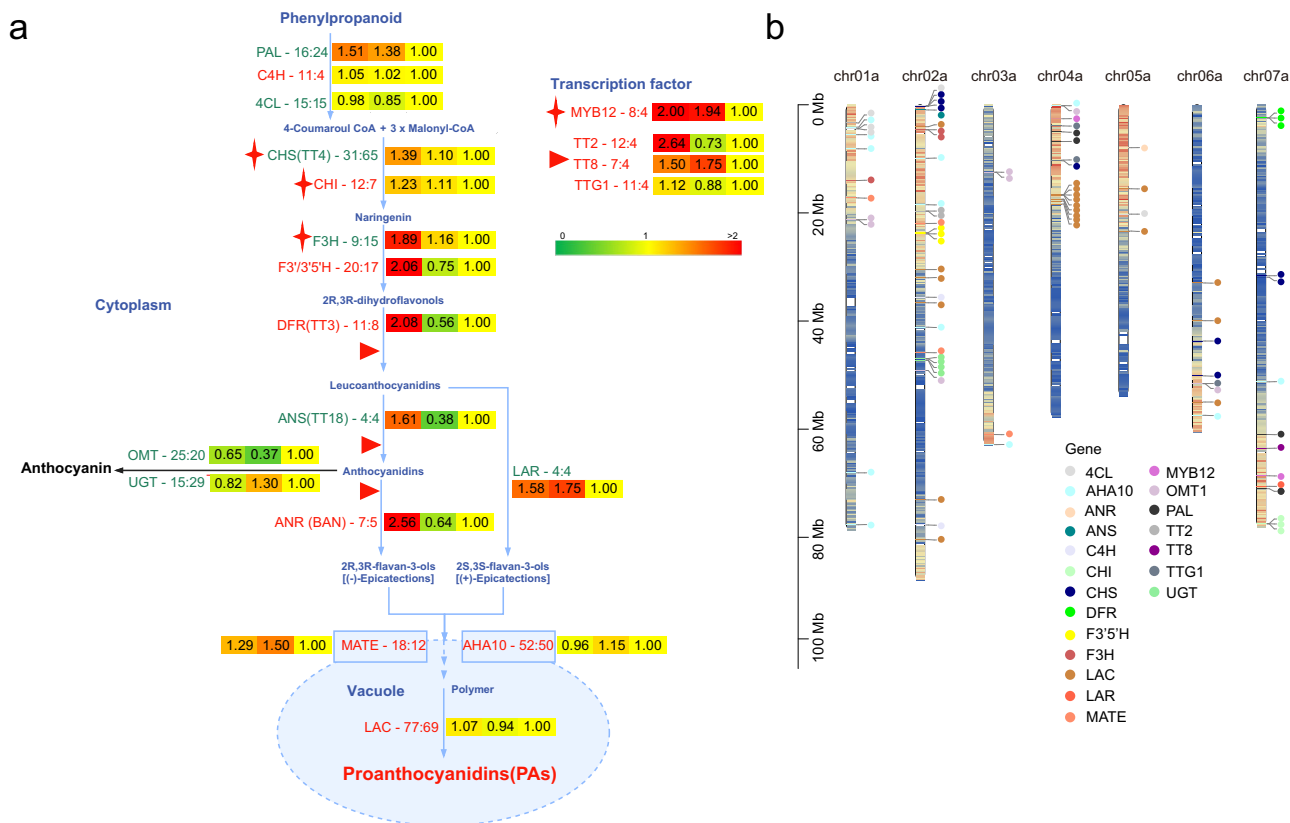


Fig. 5 Expansion and expression of genes involved in PA biosynthesis. **a** Heatmaps showing the fold changes in the expression levels of genes in the PA biosynthesis pathway among the three groups of H, M, and L (the left is for H, the center is for M, and the right is for L). The numbers in the heatmap represent the fold change values, which were calculated with the gene expression in Groups H and M compared with their counterparts in Group L. The gene copy numbers are shown following the enzymes and TFs in the whole biosynthesis pathway in a comparison of *O. viciifolia* and *M. sativa* “XinjiangDaye” (the left is for *O. viciifolia*, and the right is for *M. sativa*). The genes are highlighted in red if the copy number is higher in *O. viciifolia* than in *M. sativa*, and otherwise in green. L-phenylalanine ammonia-lyase, PAL; cinnamate 4-hydroxylase, C4H; 4-coumarate coenzyme A ligase, 4CL; chalcone isomerase, CHI; chalcone synthase, CHS; flavanone-3-hydroxylase, F3H; “flavonoid 3’ hydroxylase”/“flavonoid 3’5’ hydroxylase”, F3’/3’5’H; dihydroflavonol 4-reductase, DFR; anthocyanidin reductase, ANR; leucoanthocyanidin reductase, LAR; multidrug and toxic compound extrusion-type transporter, MATE; plasma membrane H⁺-ATPase, AHA10; laccase-like flavonoid oxidase, LAC. **b** Homologous genes involved in the PA biosynthesis pathway plotted on the seven chromosomes of haplotype A of *O. viciifolia*.

of recently reported leguminous species genome assemblies, such as those of *A. sinicus* and *M. sativa*^{33,36}. BUSCO analysis revealed 89.1%–90.4% complete genes in the four haplotypes and 91.3% in the whole genome upon genome assembly (Supplementary Tables 5 and 10). In addition, the alignment rates of RNA-seq data were 80.86%–86.22% in the four haplotypes and 90.13%–93.30% for the whole genome (Supplementary Table 24). This study provides a complete genome resource for *O. viciifolia* to search for copies of genes involved in PA biosynthesis and will promote more functional genomic studies in the future.

PAs have prominent effects on fruit flavor, forage quality, and plant defence^{55–57} and could reduce rumen bloat disease and methane emission in ruminants⁵⁸. PAs are synthesized by a long pipeline that involves several key enzymes and TFs (Fig. 5a) in plants^{27,28,59}. Gene expansion is a key strategy employed by plants for environmental adaptation under the enhancement of secondary metabolite production due to the quantitative dose effects and gene expression regulation network. Copy numbers of PAL, 4CL, and CHS genes that were key in the flavonoid pathway were higher in the genome of *Cenchrus purpureus* with more anthocyanidin than in the genomes of *C. americanus* and *Setaria italica* with less anthocyanidin⁶⁰. Tandem repeats of CHS genes, which are related to the first rate-limiting enzyme in flavonoid biosynthesis, were found in *M. truncatula*³³. In our study,

we confirmed that the enrichment of PAs in *O. viciifolia* was significantly contributed by gene expansion and expression upregulation. We revealed significant expansion of genes involved in PA biosynthesis in *O. viciifolia* compared with the closely related autotetraploid species *M. sativa*, although *M. sativa* has a large genome size of 3157 Mb and 164,632 protein-coding genes³⁴. The upregulated expression of these genes further resulted in the enhancement of PA content in *O. viciifolia*. More than half of all plant species experienced polyploidy, including autopolyploidy and allopolyploidy⁶¹. It was shown that in allo-tetraploid elephant grass (*Cenchrus purpureus* Schumach.) with high anthocyanidin content, a gene expansion strategy was also adopted to enhance the biosynthesis of anthocyanidins and flavonoids⁶⁰. Consequently, the induced autopolyploids exhibited enhanced resistance to biotic and abiotic stresses, and secondary metabolite production might be considerably increased⁶². Cases were reported that confirmed this link between the enhancement of secondary metabolite production and gene copy number increase. For example, cichoric acid⁶³, caffeic acid⁶⁴, and flavonoids⁶⁵ were found at higher levels in autotetraploid plants. In addition, in our case, the downregulated expression of both OMT and UGT in the leaves of *O. viciifolia* was one key reason that anthocyanidins were converted into PAs rather than anthocyanins.

A recent study revealed that *MtGSTF7*, a homologous gene of *AtTT19*, in *Medicago truncatula* played an important role in the accumulation and translocation of both anthocyanin and PAs⁶⁶. However, the homologous genes were not expressed in *O. viciifolia* (i.e., LeOno02aG0476700, LeOno02bG0495900, LeOno02cG0475100, and LeOno02dG0404000) and *M. sativa* leaves. Therefore, we suspected that there would be other genes involved in the accumulation and translocation of PAs, which could be discovered in the future. Meanwhile, PAs are enriched in different tissues in various plant species. For example, *O. viciifolia* and *L. japonicus* can enrich PAs in all organs, but PAs are mainly enriched in the seed coat of *M. sativa* and flowers of *T. pratense*¹⁵. Phagosomes, involved in a cellular homeostatic process, were significantly enriched in the lines of *O. viciifolia* with a high PA content (Supplementary Fig. 11b). Plants enriched in PAs synthesize monomeric flavonoids, such as flavan-3-ol and epi-flavan-3-ol, in the cytoplasm and transport them into the vacuole to synthesize PAs in the final polymerization step¹⁵. Some researchers found that autophagy involved trafficking of anthocyanin from the cytoplasm to vacuoles, as anthocyanin content was reduced in autophagy-deficient plants^{67–69}. Overexpression of MdATG18a enhanced autophagy activity and improved anthocyanin content in apple⁷⁰. The phagosome-mediated autophagosome transport mechanism of monomeric flavonoids might play a key role in PA biosynthesis in *O. viciifolia*. Autophagy and autophagosomes control peroxisome quality and are involved in the degradation of peroxisomes⁷¹, which is consistent with the downregulation of genes related to peroxisomes (Supplementary Fig. 11b). To date, several genes have been reported to play a role in PA biosynthesis^{15,72,73}, among which ANS is vital in the production pathway of flavonoids. ANS is a key enzyme that catalyses leucoanthocyanidins into anthocyanidins, which serve as substrates for the production of anthocyanins and PAs¹⁵. Our coexpression network results showed a strong correlation between PA content and the MEblack module and further identified ANS (LeOno02aG0012600) as the hub gene in the gene regulatory network for the PA biosynthesis pathway of *O. viciifolia* leaves, which was in accordance with the results found in *Tetrastigma hemsleyanum*⁷⁴. The genes in the MEblack module are valuable for further investigation of their roles in PA biosynthesis.

Materials and methods

Plant materials and preparation for sequencing. We chose one robust plant of common sainfoin that was grown in the greenhouse of CAU (China Agricultural University) and cut its leaves before immediately processing them in liquid nitrogen. Genomic DNA was extracted from the young leaves using a modified CTAB method previously described⁷⁵. An Agilent 2100 instrument and electrophoresis were used to assess DNA quality. Whole-genome sequencing was performed by combining the Oxford Nanopore technology (ONT, ~109.89×) and Illumina (~96.15×) approaches (Supplementary Table 1). Young leaf tissue was used for Hi-C library construction based on HindIII digestion. The Hi-C library was sequenced on an Illumina HiSeq 2500 platform in paired-end 150 bp mode (~164.84×). All sequencing was conducted by BioMarker Technologies Company (Beijing, China).

A total of 59 individuals of 41 germplasm lines of *O. viciifolia* were used to determine the PA content (Supplementary Tables 22 and 23). They were grouped into three groups, which corresponded to high, medium, and low levels of PA. Three independent representative individuals from each group were selected for transcriptome sequencing of leaf tissue. RNA extraction and quality assessment were performed according to

a previous method⁷⁶. We also downloaded the transcriptome data of sainfoin from the National Center for Biotechnology Information (<https://www.ncbi.nlm.nih.gov>) to annotate our genome assembly (Supplementary Table 8).

Genome size estimation. We estimated genome size roughly through the two strategies of *K*-mer analysis based on Illumina short reads of genomic DNA and flow cytometry, for which *Medicago truncatula* (cultivar “A17”), *Panicum virgatum* (cultivar “Alamo”), and *Zea mays* (B73) were used as internal references. Clean short reads were subjected to counting of each 17-mer using Jellyfish 2.0 (www.genome.umd.edu/jellyfish.html). The results of Jellyfish were input into GenomeScope2 (<http://qb.cshl.edu/genomescope/genomescope2.0>) to estimate genome size and heterozygosity (Supplementary Fig. 1).

Genome assembly. ONT raw reads were corrected by using NextDenovo v2.5.0 (parameter: `correction_options = -b`) and initially assembled into contigs using Flye v2.9 (parameter: `--keep-haplotypes --iterations 2`)⁷⁷, assembling contigs of four haplotypes as completely as possible. The contigs were polished with Illumina short reads to remove nucleotide errors using Pilon47 v1.24⁷⁸. There were small-scale collapses and redundancy in the four haplotypes, and some contigs from different haplotypes were incorrectly linked together. Haplotypic chromosome-level assembly was accomplished using the Hi-C technique. First, following a previously described pipeline⁵¹, we aligned Hi-C reads to the initial genome assembly using Juicer v1.6⁷⁹, and Hi-C-assisted chromosomal assembly was conducted by using 3D-DNA v180922⁸⁰ to correct the majority of the assembly errors. Then, based on the interaction of Hi-C data, manual inspection and adjustment, including adjusting the boundary of chromosomal segmentation and correcting visible errors as much as possible, were performed to generate the final allele-aware chromosomal assembly in Juicebox v2.14.00⁸¹. It is widely accepted that the interaction among reads with shorter distances is stronger than that among reads with long distances in Hi-C analysis. The assembly was finally generated and grouped into 28 pseudochromosomes, without significant misassembly. However, we still obtained some collapsed contigs. After manual examination, gaps were filled by using LR_Gapcloser⁸² based on ONT long reads. The final genome assembly was generated after three rounds of polishing by using Pilon47 v1.24.

We mapped the Illumina short reads, ONT long reads, and RNA-Seq reads to the genome assembly with BWA-MEM v0.7.15⁸³, Minimap2 v2.24⁸⁴, and HISAT2 v2.2.1^{85–87}, respectively. The read mapping rate was calculated to assess assembly integrity. BUSCO v5.2.2⁸⁸ was used to evaluate the quality and completeness of the assembly (parameters: `-m genome --augustus`) based on `embryophyte_odb9`. Based on the Illumina short reads, we further assessed the whole genome and four haplotypes of *O. viciifolia* by using Merqury v1.3 and Meryl v1.4⁸⁹. And based on the ONT long reads, we used Inspector-v1.2 to identify the regions with collapse and switch error in the whole genome⁹⁰.

Repeat identification and gene annotation. We identified transposable elements with the *ab initio* method by EDTA v2.0.0 (parameter: `--sensitive 1 --anno 1`)⁹¹ and used RepeatMasker (<http://www.repeatmasker.org/RepeatMasker/>) to predict repeat sequences. We identified tRNAs by using tRNAScan-SE v2.0.9⁹², rRNA by using Barrnap (<https://github.com/tseemann/barrnap>), and noncoding RNA with RfamScan⁹³.

The combined strategies of *ab initio* prediction and evidence- and homolog-based searching methods were used for gene modeling. The transcript evidence came from the two

transcriptome assemblies. The first de novo transcriptome assembly was conducted based on the high-quality RNA-seq reads using Trinity v2.13.2⁹⁴, and the second transcriptome assembly was accomplished by using HISAT2 alignment and genome-guided Trinity assembly. In total, 326,841 transcript sequences were obtained after merging these two transcriptome assemblies and removing redundant sequences by using CD-HIT v4.6 (95% identity and 95% coverage)⁹⁵. Based on the above transcript sequences, ab initio gene predictions were produced by PASA v2.4.1⁹⁶ and AUGUSTUS v3.4.0⁹⁷ from the genome assembly and optimized for five rounds by using AUGUSTUS. In addition, we also ran the MAKER2 pipeline⁹⁸ to build gene models. The alignments from 326,841 transcripts obtained by using BLASTN and 200,995 homologous proteins from *Glycine max*, *Medicago truncatula*, *Cicer arietinum*, *Vitis vinifera*, *Arabidopsis thaliana*, and *Oryza sativa* obtained by using BLASTP were run against the repeat-masked genome assembly and manipulated with Exonerate v2.4.0⁹⁹. The ab initio gene predictions from AUGUSTUS and evidence- and homolog-based gene annotations were merged using the maker2eval script packaged with MAKER2. To increase accuracy and completeness¹⁰⁰, the gene models from MAKER2 and PASA were integrated and combined into the final gene models using EvidenceModeler (EVM) v1.1.1¹⁰¹.

For gene functional annotation, the predicted gene models were subjected to homology searches against the following databases: Kyoto Encyclopedia of Genes and Genomes (KEGG), Gene Ontology (GO), and Clusters of Orthologous Groups of proteins (KOG/COG/eggNOG) by using eggNOG-mapper v2.1.6¹⁰²; NCBI nonredundant protein sequences (NR) and Swiss-Prot by using DIAMOND v2.0.14¹⁰³; and Pfam by using InterProScan v5.54-87.0¹⁰⁴.

Gene family and phylogenetic analysis. We selected 14 legume species, i.e., *M. sativa*, *M. truncatula*, *M. runthetica*, *T. subterraneanum*, *T. pratense*, *P. sativum*, *C. arietinum*, *L. japonicus*, *G. max*, *V. radiata*, *C. cajan*, *L. angustifolius*, *A. hypogaea*, and *A. duranensis*, and four other outgroup species (*P. persica*, *A. thaliana*, *P. trichocarpa*, and *V. vinifera*), and downloaded their protein sequences from the NCBI. Protein sequences of the *O. viciifolia* genome assembly and these downloaded genome assemblies were used to identify orthologues by using OrthoFinder2 v2.5.4 with a parameter of “-M msa”¹⁰⁵. In total, 448 single-copy proteins were obtained, and their sequences were concatenated to build a phylogenetic tree with the maximum likelihood method by using IQTREE v2.1.6¹⁰⁶ based on the JTT + F + R5 model and 1000 bootstraps. This ML tree and the 448 single-copy orthogroups were subjected to divergence time estimation by the MCMCTree program in the PAML package v4.10.3¹⁰⁷. Three time-calibration points were selected, including *M. sativa* vs. *O. viciifolia* 15–91 million years ago (Mya); *A. hypogaea* vs. *O. viciifolia* 25.9–120 million years ago (Mya); and *V. vinifera* vs. *O. viciifolia* 107–135 Mya, which are available on the Timetree website (<http://timetree.org/>). By determining the numbers of gene families and genes with OrthoFinder2, we obtained and compared statistical information on the gene families of *O. viciifolia*, *M. sativa*, *L. japonicus*, and *C. arietinum*.

Genome evolutionary analysis. Based on orthologous genes, genome collinearity analysis among species and between haplotypes was performed by using MCScanX and viewed in JCVI¹⁰⁸. Based on syntenic or homologous gene pairs obtained from MCScanX, Ka, Ks, and Ka/Ks values were calculated by using TBtools¹⁰⁹, and the Ks distribution was used for inferring genome

palaeopolyploidization events and occurrence time estimation. LTRs were identified by LTR_harvest V1.6.2¹¹⁰ (parameters: -similar 90 -vic 10 -seed 20 -seqids yes -minlenltr 100 -maxlenltr 7000 -mintsd 4 -maxtsd 6 -motif TGCA -motifmis 1), and LTR insertion time was calculated by LTR_retriever V2.9.0¹¹¹ (parameters: -u 7e-9).

We used CAFE 5^{112,113} to identify the expanded, contracted, and rapidly evolved gene families based on the phylogenetic tree and 27,749 homologous gene families from OrthoFinder2.

Ancestral karyotype evolution. We selected seven leguminous species, namely, *O. viciifolia*, *M. sativa*, *M. truncatula*, *P. sativum*, *C. arietinum*, and *L. japonicus*, with *G. max* as an outgroup species. Based on the published pipeline (<https://github.com/xjtu-omics/processDrimm>) (Gao et al. 2022)¹¹⁴, orthogroups of complete homologous gene sequences were identified by using Orthofinder v2.5.5 (<https://github.com/davidemms/OrthoFinder>), and the non-overlapping syntenic blocks were built by using Drimm-Synteny (<https://github.com/xjtu-omics/processDrimm/tree/master/drimm>). The longest common subsequence (LCS) algorithm was used to retrieve the gene sequences for each block and their copy number in each species. The ancestors' genomes were rebuilt based on the gene sequences by using IAGS scripts (<https://github.com/xjtu-omics/IAGS>).

Expansion of genes involved in PA biosynthesis between *O. viciifolia* and *M. sativa*. We downloaded the reference protein sequences of *A. thaliana*, *V. vinifera*, and *G. max* (Supplementary Table 25) and identified the copy numbers of genes involved in the PA biosynthesis pathway based on the protein sequences of *O. viciifolia*, *M. sativa* cultivar “XinjiangDaye” (https://figshare.com/projects/whole_genome_sequencing_and_assembly_of_Medicago_sativa/66380), *Lotus japonicus* (GenBank accession GCA_000181115.2), *Medicago truncatula* (GenBank accession GCF_003473485.1), and *Glycine max* (GenBank accession GCF_000004515.6). The cut-offs of BLASTP were set with an E-value of less than $1e^{-150}$ and identity of >50%. The homologous genes of *O. viciifolia* were plotted on the chromosomes by using TBtools v1.113.

Transcriptome analysis for gene expression in leaves. RNA-seq data from the leaves of *M. sativa* (NCBI accession PRJNA795295) and three groups of *O. viciifolia* lines that showed high (Group H), medium (Group M), and low (Group L) PA contents were analyzed based on a trimming pipeline in Trimmomatic v0.39¹¹⁵, mapping by HISAT2 v2.2.1, and determining gene expression FPKM values by StringTie v2.2.0¹¹⁶. Significantly differentially expressed genes were determined by the R package DESeq2¹¹⁷. Principal component analysis (PCA) and visualization were conducted by using the R functions prcomp() and ggbiplot(). KEGG enrichment was performed by using the R package “clusterProfiler”.

In coexpression network analysis, gene modules were generated, and their correlation with PA content and other phenotypic traits was analyzed by using the R package “WGCNA”. The expression levels of all genes related to one specific regulator (including enzymes and TFs) in the PA biosynthesis pathway were summed and compared among the three groups based on the fold change values in the formula of Group H/L or M/L. The expression of genes in leaves involved in PA biosynthesis was compared between *O. viciifolia* and *M. sativa* based on the three internal reference genes GADPH, CYP, and EF1a.

For qRT-PCR verification, total RNA was extracted from *O. viciifolia* and *M. sativa* leaves by using a Plant RNAout kit (Beijing Huayueyang Biotechnology, China), and 1 µg total RNA

was used for inverse transcription and cDNA generation with a PrimeScript™ RT reagent kit with gDNA Eraser (Takara, RR047A). The qRT-PCR experiments were performed by using 2 × ChamQ Universal SYBR qPCR Master Mix (Q711-02) and qTOWER³ G (Analytik Jena, Germany), with three biological replicates and GAPDH/EF1a as internal control genes. The relative expression levels were determined using the 2^{-ΔCt} method. All primers were designed by Primer3 and are listed in Supplementary Table 28.

Statistics and reproducibility. In our research, all resource data are available from the corresponding authors to ensure the reproducibility of the analysis. To advance the reproducibility, we have defined processing or sampling with a frequency more than two as replicate. There are three samples replicated at different PA content levels in the transcriptome sequencing, and two samples replicated for every species in the qRT-PCR verification. *T*-test was used to identify the significant difference between the two samples, and linear regression analysis was preformed using the *lm* function of the R (v4.2.3). And Benjamini-Hochberg (BH) method was applied for *p* value correction¹¹⁸.

Data availability

All the raw data, including Nanopore long reads and Illumina short reads were uploaded to the China National Center for Bioinformation GSA (Genome Sequence Archive) database under BioProject PRJCA009631. Genome assembly and gene annotation files were available in the figshare website (<https://doi.org/10.6084/m9.figshare.24155073>). The source data for Supplementary Data 1–4 is available for Figs. 1d, 2c and 3a, b respectively, while all other source data are available from S.J. and Y.Z. on reasonable request.

Received: 11 March 2023; Accepted: 28 December 2023;

Published online: 05 January 2024

References

- Norman, H. C. et al. Productivity and nutritional value of 20 species of perennial legumes in a low-rainfall Mediterranean-type environment in southern Australia. *Grass Forage Sci.* **76**, 134–158 (2021).
- Akçelik, S., Avci, S., Uzun, S. & Sancak, C. Karyotype analysis of some *Onobrychis* (sainfoin) species in Turkey. *Arch. Biol. Sci.* **64**, 567–571 (2012).
- Thomson, J. R. Cross- and self-fertility in sainfoin. *Ann. Appl. Biol.* **25**, 695–704 (1938).
- Yucel, G. et al. The chromosome number and rDNA loci evolution in *Onobrychis* (Fabaceae). *Int. J. Mol. Sci.* **23**, 11033 (2022).
- Ranjbar, M., Hajmoradi, F. & Karamian, R. An overview on cytogenetics of the genus *Onobrychis* (Fabaceae) with special reference to *O. sect. Hymenobrychis* from Iran. *Caryologia* **65**, 187–198 (2012).
- Bhattarai, S., Coulman, B., Biliget, B. & Navabi, A. Sainfoin (*Onobrychis viciifolia* Scop.): renewed interest as a forage legume for western Canada. *Can. J. Plant Sci.* **96**, 748–756 (2016).
- Carlton, A. E., Cooper, C. S., Delaney, R. H., Dubbs, A. L. & Eslick, R. F. Growth and forage quality comparisons of sainfoin (*Onobrychis viciaefolia* Scop.) and alfalfa (*Medicago sativa* L.). *Agron. J.* **60**, 630–632 (1968).
- Sheppard, S. C. et al. Sainfoin production in western Canada: a review of agronomic potential and environmental benefits. *Grass Forage Sci.* **74**, 6–18 (2019).
- Lees, G. L., Gruber, M. Y. & Suttill, N. H. Condensed tannins in sainfoin. II. Occurrence and changes during leaf development. *Can. J. Bot.* **73**, 1540–1547 (1995).
- Lees, G. L., Suttill, N. H. & Gruber, M. Y. Condensed tannins in sainfoin. I. A histological and cytological survey of plant tissues. *Can. J. Bot.* **71**, 1147–1152 (1993).
- Marais, J. P. J., Mueller-Harvey, I., Brandt, E. V. & Ferreira, D. Polyphenols, condensed tannins, and other natural products in *Onobrychis viciifolia* (Sainfoin). *J. Agric. Food Chem.* **48**, 3440–3447 (2000).
- Mueller-Harvey, I. et al. Benefits of condensed tannins in forage legumes fed to ruminants: importance of structure, concentration, and diet composition. *Crop Sci.* **59**, 861–885 (2019).
- Hatew, B. et al. Impact of variation in structure of condensed tannins from sainfoin (*Onobrychis viciifolia*) on in vitro ruminal methane production and fermentation characteristics. *J. Anim. Physiol. Anim. Nutr.* **100**, 348–360 (2016).
- Jones, G. A., McAllister, T. A., Muir, A. D. & Cheng, K. J. Effects of sainfoin (*Onobrychis viciifolia* Scop.) condensed tannins on growth and proteolysis by four strains of ruminal bacteria. *Appl. Environ. Microbiol.* **60**, 1374–1378 (1994).
- Jonker, A. & Yu, P. The occurrence, biosynthesis, and molecular structure of proanthocyanidins and their effects on legume forage protein precipitation, digestion and absorption in the ruminant digestive tract. *Int. J. Mol. Sci.* **18**, 1105 (2017).
- Lorenz, M. M. et al. Relationship between condensed tannin structures and their ability to precipitate feed proteins in the rumen. *J. Sci. Food Agric.* **94**, 963–968 (2014).
- McMahon, L. R. et al. Effect of sainfoin on in vitro digestion of fresh alfalfa and bloat in steers. *Can. J. Anim. Sci.* **79**, 203–212 (1999).
- Patra, A. K. & Saxena, J. Exploitation of dietary tannins to improve rumen metabolism and ruminant nutrition. *J. Sci. Food Agric.* **91**, 24–37 (2011).
- Sivakumaran, S. et al. Floral procyanidins of the forage legume red clover (*Trifolium pratense* L.). *J. Agric. Food Chem.* **52**, 1581–1585 (2004).
- Zeller, W. E. et al. Protein precipitation behavior of condensed tannins from *Lotus pedunculatus* and *Trifolium repens* with different mean degrees of polymerization. *J. Agric. Food Chem.* **63**, 1160–1168 (2015).
- Koupai-Abyazani, M. R. et al. Purification and characterization of a proanthocyanidin polymer from seed of alfalfa (*Medicago sativa* Cv. Beaver). *J. Agric. Food Chem.* **41**, 565–569 (1993).
- Skadhauge, B., Gruber, M. Y., Thomsen, K. K. & von Wettstein, D. Leucocyanidin reductase activity and accumulation of proanthocyanidins in developing legume tissues. *Am. J. Bot.* **84**, 494–503 (1997).
- McNabb, W. C., Peters, J. S., Foo, L. Y., Waghorn, G. C. & Jackson, F. S. Effect of condensed tannins prepared from several forages on the in vitro precipitation of ribulose-1,5-bisphosphate carboxylase (rubisco) protein and its digestion by Trypsin (EC 2.4.21.4) and Chymotrypsin (EC 2.4.21.1). *J. Sci. Food Agric.* **77**, 201–212 (1998).
- Abrahams, S., Tanner, G. J., Larkin, P. J. & Ashton, A. R. Identification and biochemical characterization of mutants in the proanthocyanidin pathway in Arabidopsis. *Plant Physiol.* **130**, 561–576 (2002).
- Xie, D. Y., Sharma, S. B. & Dixon, R. A. Anthocyanidin reductases from *Medicago truncatula* and *Arabidopsis thaliana*. *Arch. Biochem. Biophys.* **422**, 91–102 (2004).
- James, A. M. et al. Poplar MYB115 and MYB134 transcription factors regulate proanthocyanidin synthesis and structure. *Plant Physiol.* **174**, 154–171 (2017).
- Chen, W. et al. DkMYB14 is a bifunctional transcription factor that regulates the accumulation of proanthocyanidin in persimmon fruit. *Plant J.* **106**, 1708–1727 (2021).
- Lu, N., Rao, X., Li, Y., Jun, J. H. & Dixon, R. A. Dissecting the transcriptional regulation of proanthocyanidin and anthocyanin biosynthesis in soybean (*Glycine max*). *Plant Biotechnol. J.* **19**, 1429–1442 (2021).
- Mora-Ortiz, M. & Smith, L. M. *Onobrychis viciifolia*; a comprehensive literature review of its history, etymology, taxonomy, genetics, agronomy and botany. *Plant Genet. Resour.* **16**, 403–418 (2018).
- Jain, M. et al. A draft genome sequence of the pulse crop chickpea (*Cicer arietinum* L.). *Plant J.* **74**, 715–729 (2013).
- Liu, Y. et al. Pan-genome of wild and cultivated soybeans. *Cell* **182**, 162–176 (2020).
- Shen, C. et al. The chromosome-level genome sequence of the autotetraploid alfalfa and resequencing of core germplasms provide genomic resources for alfalfa research. *Mol. Plant* **13**, 1250–1261 (2020).
- Chang, D. et al. The chromosome-level genome assembly of *Astragalus sinicus* and comparative genomic analyses provide new resources and insights for understanding legume-rhizobial interactions. *Plant Commun.* **3**, 100263 (2022).
- Chen, H. et al. Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa. *Nat. Commun.* **11**, 2494 (2020).
- Li, A. et al. A chromosome-scale genome assembly of a diploid alfalfa, the progenitor of autotetraploid alfalfa. *Hortic. Res.* **7**, 194 (2020).
- Long, R. et al. Genome assembly of alfalfa cultivar Zhongmu-4 and identification of SNPs associated with agronomic traits. *Genomics Proteom. Bioinforma.* **20**, 14–28 (2022).
- Bhattarai, S. et al. Genetic diversity and relationship of sainfoin (*Onobrychis viciifolia* Scop.) germplasm as revealed by amplified fragment length polymorphism markers. *Can. J. Plant Sci.* **98**, 543–551 (2018).
- Kempf, K., Mora-Ortiz, M., Smith, L. M., Kolliker, R. & Skot, L. Characterization of novel SSR markers in diverse sainfoin (*Onobrychis viciifolia*) germplasm. *BMC Genet.* **17**, 124 (2016).
- Mora-Ortiz, M. et al. De-novo transcriptome assembly for gene identification, analysis, annotation, and molecular marker discovery in *Onobrychis viciifolia*. *BMC Genomics* **17**, 756 (2016).

40. Shen, S. et al. Development of polymorphic EST-SSR markers and characterization of the autotetraploid genome of sainfoin (*Onobrychis viciifolia*). *PeerJ* **7**, e6542 (2019).
41. Zarrabian, M., Majidi, M. M. & Ehtemam, M. H. Genetic diversity in a worldwide collection of sainfoin using morphological, anatomical, and molecular markers. *Crop Sci.* **53**, 2483–2496 (2013).
42. Jin, Z., Jiang, W., Yi, D. & Pang, Y. The complete chloroplast genome sequence of Sainfoin (*Onobrychis viciifolia*). *Mitochondrial DNA Part B Resour.* **6**, 496–498 (2021).
43. Sun, H. et al. Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. *Nat. Genet.* **54**, 342–348 (2022).
44. Li, H., Jiang, F., Wu, P., Wang, K. & Cao, Y. A high-quality genome sequence of model legume *Lotus japonicus* (MG-20) provides insights into the evolution of root nodule symbiosis. *Genes* **11**, 483 (2020).
45. Schmutz, J. et al. Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178–183 (2010).
46. Zhuang, W. et al. The genome of cultivated peanut provides insight into legume karyotypes, polyploid evolution and crop domestication. *Nat. Genet.* **51**, 865–876 (2019).
47. Lagrange, S., Lobón, S. & Villalba, J. J. Gas production kinetics and in vitro degradability of tannin-containing legumes, alfalfa and their mixtures. *Anim. Feed Sci. Technol.* **253**, 56–64 (2019).
48. Verma, S., Salminen, J. P., Taube, F. & Malisch, C. S. Large inter- and intraspecies variability of polyphenols and proanthocyanidins in eight temperate forage species indicates potential for their exploitation as nutraceuticals. *J. Agric. Food Chem.* **69**, 12445–12455 (2021).
49. Kreplak, J. et al. A reference genome for pea provides insight into legume genome evolution. *Nat. Genet.* **51**, 1411–1422 (2019).
50. Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat. Plants* **5**, 833–845 (2019).
51. Bao, Z. et al. Genome architecture and tetrasomic inheritance of autotetraploid potato. *Mol. Plant* **15**, 1211–1226 (2022).
52. Colle, M. et al. Haplotype-phased genome and evolution of phytonutrient pathways of tetraploid blueberry. *Gigascience* **8**, giz012 (2019).
53. Wang, F. et al. The autotetraploid potato genome provides insights into highly heterozygous species. *Plant Biotechnol. J.* **20**, 1996–2005 (2022).
54. Zhang, Q. et al. Genomic insights into the recent chromosome reduction of autopolyploid sugarcane *Saccharum spontaneum*. *Nat. Genet.* **54**, 885–896 (2022).
55. Aron, P. M. & Kennedy, J. A. Flavan-3-ols: nature, occurrence and biological activity. *Mol. Nutr. Food Res.* **52**, 79–104 (2008).
56. Dixon, R. A., Xie, D. Y. & Sharma, S. B. Proanthocyanidins—a final frontier in flavonoid research? *N. Phytol.* **165**, 9–28 (2005).
57. Zheng, Q. et al. Comparative transcriptome analysis reveals regulatory network and regulators associated with proanthocyanidin accumulation in persimmon. *BMC Plant Biol.* **21**, 356 (2021).
58. Jonker, A. & Yu, P. The role of proanthocyanidins complex in structure and nutrition interaction in alfalfa forage. *Int. J. Mol. Sci.* **17**, 793 (2016).
59. Gou, L. et al. Multigene synergism increases the isoflavone and proanthocyanidin contents of *Medicago truncatula*. *Plant Biotechnol. J.* **14**, 915–925 (2016).
60. Yan, Q. et al. The elephant grass (*Cenchrus purpureus*) genome provides insights into anthocyanidin accumulation and fast growth. *Mol. Ecol. Resour.* **21**, 526–542 (2021).
61. Muntzing, A. The evolutionary significance of autopolyploidy. *Hereditas* **21**, 363–378 (1936).
62. Gantait, S. & Mukherjee, E. Induced autopolyploidy—a promising approach for enhanced biosynthesis of plant secondary metabolites: an insight. *J. Genet. Eng. Biotechnol.* **19**, 4 (2021).
63. Abdoli, M., Moieni, A. & Badi, H. N. Morphological, physiological, cytological and phytochemical studies in diploid and colchicine-induced tetraploid plants of *Echinacea purpurea* (L.). *Acta Physiol. Plant* **35**, 2075–2083 (2013).
64. Xu, C. et al. A comparative study of bioactive secondary metabolite production in diploid and tetraploid *Echinacea purpurea* (L.) Moench. *Plant Cell Tissue Organ Cult.* **116**, 323–332 (2013).
65. Chung, H. H., Shi, S. K., Huang, B. & Chen, J. T. Enhanced agronomic traits and medicinal constituents of autotetraploids in *Anoectochilus formosanus* Hayata, a top-grade medicinal orchid. *Molecules* **22**, 1907 (2017).
66. Wang, R. et al. *MTGSTF7*, a TT19-like GST gene, is essential for accumulation of anthocyanins, but not proanthocyanins in *Medicago truncatula*. *J. Exp. Bot.* **73**, 4129–4146 (2022).
67. Chanoca, A. et al. Anthocyanin vacuolar inclusions form by a microautophagy mechanism. *Plant Cell* **27**, 2545–2559 (2015).
68. Masclaux-Daubresse, C. Autophagy controls carbon, nitrogen, and redox homeostasis in plants. *Autophagy* **12**, 896–897 (2016).
69. Pourcel, L. et al. The formation of anthocyanin vacuolar inclusions in *Arabidopsis thaliana* and implications for the sequestration of anthocyanin pigments. *Mol. Plant* **3**, 78–90 (2010).
70. Sun, X. et al. *MdATG18a* overexpression improves tolerance to nitrogen deficiency and regulates anthocyanin accumulation through increased autophagy in transgenic apple. *Plant Cell Environ.* **41**, 469–480 (2018).
71. Shibata, M. et al. Highly oxidized peroxisomes are selectively degraded via autophagy in *Arabidopsis*. *Plant Cell* **25**, 4967–4983 (2013).
72. He, F., Pan, Q. H., Shi, Y. & Duan, C. Q. Biosynthesis and genetic regulation of proanthocyanidins in plants. *Molecules* **13**, 2674–2703 (2008).
73. Winkel-Shirley, B. Flavonoid biosynthesis. A colorful model for genetics, biochemistry, cell biology, and biotechnology. *Plant Physiol.* **126**, 485–493 (2001).
74. Yue, E. et al. Comparative analysis of proanthocyanidin metabolism and genes regulatory network in fresh leaves of two different ecotypes of *Tetrastigma hemsleyanum*. *Plants* **11**, 211 (2022).
75. Jia, S., Morton, K., Zhang, C. & Holding, D. An exome-seq based tool for mapping and selection of candidate genes in maize deletion mutants. *Genomics Proteom. Bioinforma.* **16**, 439–450 (2018).
76. Jia, S. et al. Deletion of maize RDM4 suggests a role in endosperm maturation as well as vegetative and stress-responsive growth. *J. Exp. Bot.* **71**, 5880–5895 (2020).
77. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
78. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
79. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
80. Dudchenko, O. et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
81. Durand, N. C. et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016).
82. Xu, G. C. et al. LR_GapCloser: a tiling path-based gap closer that uses long reads to complete genome assembly. *Gigascience* **8**, giv157 (2019).
83. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* <https://doi.org/10.48550/arXiv.1303.3997> (2013).
84. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
85. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
86. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
87. Siren, J., Valimaki, N. & Makinen, V. Indexing graphs for path queries with applications in genome research. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **11**, 375–388 (2014).
88. Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
89. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merquy: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
90. Chen, Y., Zhang, Y., Wang, A. Y., Gao, M. & Chong, Z. Accurate long-read de novo assembly evaluation with Inspector. *Genome Biol.* **22**, 312 (2021).
91. Ou, S. et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).
92. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
93. Kalvari, I. et al. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res.* **46**, 335–342 (2018).
94. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
95. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
96. Haas, B. J. et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
97. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntetically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644 (2008).
98. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinforma.* **12**, 491 (2011).
99. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinforma.* **6**, 31 (2005).
100. Cook, D. E. et al. Long-read annotation: automated eukaryotic genome annotation based on long-read cDNA sequencing. *Plant Physiol.* **179**, 38–54 (2019).
101. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVIDENCEModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, R7 (2008).

102. Huerta-Cepas, J. et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
103. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–63 (2015).
104. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
105. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
106. Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
107. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
108. Tang, H. B. et al. Synteny and collinearity in plant genomes. *Science* **320**, 486–488 (2008).
109. Chen, C. et al. TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant* **13**, 1194–1202 (2020).
110. Ellinghaus, D., Kurtz, S. & Willhoeft, U. *LTRharvest*, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinforma.* **9**, 18 (2008).
111. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).
112. Hahn, M. W., De Bie, T., Stajich, J. E., Nguyen, C. & Cristianini, N. Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Res.* **15**, 1153–1160 (2005).
113. Mendes, F. K., Vanderpool, D., Fulton, B. & Hahn, M. W. CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics* **36**, 5516–5518 (2020).
114. Gao, S. et al. IAGS: inferring ancestor genome structure under a wide range of evolutionary scenarios. *Mol Biol Evol.* **39**, msac041 (2022).
115. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
116. Kovaka, S. et al. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278 (2019).
117. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
118. Wu, T. et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation* **2**, 100141 (2021).

Acknowledgements

This work was supported by the Beijing Natural Science Foundation (6212019), National Natural Science Foundation of China (32071870), China Agricultural University Project (2022TC020), the Key R&D project of Sichuan Science and Technology Program

(2023YFSY0012), and Elite alfalfa cultivar selection and its industrialization demonstration.

Author contributions

J.H. carried out field and lab work, conducted the data analysis and wrote the manuscript. D.T. and X.L. conducted data analysis of comparative genomics and expansion of genes related to PA biosynthesis, and involved in draft. X.W. conducted the RNA-seq analysis. T.W., Z.W. and H.Z. measured the phenotypic traits of sainfoin germplasm lines. X.H., T.Z., Q.Y. and R.Z. participated in data analysis. S.J. and J.J. revised the manuscript. S.J. and Y.Z. designed the study, supervised the work, and contributed equally as corresponding authors. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42003-023-05754-6>.

Correspondence and requests for materials should be addressed to Shangang Jia or Yunwei Zhang.

Peer review information *Communications Biology* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editor: David Favero.

Reprints and permission information is available at <http://www.nature.com/reprints>

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024