Chromosome-level genome assembly of *Aristolochia contorta* provides insights into the biosynthesis of benzylisoquinoline alkaloids and aristolochic acids

Xinyun Cui†, Fanqi Meng†, Xian Pan, Xiaoxiao Qiu, Sixuan Zhang, Caili Li* and Shanfa Lu*

Medicinal Plant Cultivation Research Center, Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences & Peking Union Medical College, Haidian District, Beijing 100193, China

*Corresponding authors. E-mail: clli@implad.ac.cn, sflu@implad.ac.cn, sflu2004@aliyun.com

†These authors contributed equally.

Abstract

Aristolochic acids (AAs) and their derivatives are present in multiple Aristolochiaceae species that have been or are being used as medicinal materials. During the past decades, AAs have received increasing attention because of their nephrotoxicity and carcinogenicity. Elimination of AAs from medicinal materials using biotechnological approaches is important for improving medication safety. However, it has not been achieved because of the limited information available on AA biosynthesis. Here, we report a high-quality, reference-grade genome assembly of the AA-containing vine *Aristolochia contorta*. The total size of the assembly is 209.27 Mb, and it is assembled into 7 pseudochromosomes. Synteny analysis, \( K_s \) distribution, and 4DTv suggest an absence of whole-genome duplication (WGD) events in *Aristolochia contorta* after the angiosperm-wide WGD. Based on genomic, transcriptomic, and metabolic data, pathways and candidate genes were proposed for benzylisoquinoline alkaloid (BIA) and AA biosynthesis in *A. contorta*. Five O-methyltransferase genes, including *AcOMT1–3*, *AcOMT5*, and *AcOMT7*, were cloned and functionally characterized. The results provide a high-quality reference genome for AA-containing species of Aristolochiaceae. They lay a solid foundation for further elucidation of AA biosynthesis and regulation and for the molecular breeding of Aristolochiaceae medicinal materials.

Introduction

The safety of aristolochic acids (AAs) and their derivatives has been a matter of concern for researchers worldwide. In the 1990s, the occurrence of “Belgian weight-loss drug (which contained *Aristolochia fangchi*)”, “Gentian purging liver pills (which once contained *A. manshuriensis*)” and other events, as well as the increasing number of cases of renal function damage caused by AA-containing Chinese patent medicines, prompted researchers to find that AAs had obvious nephrotoxicity and carcinogenicity [1, 2]. AAs induce gene mutations to directly damage renal tubular epithelial cells [1], and their nitro and methoxy groups are the key determinants of nephrotoxicity [3]. In 2008, the International Agency for Research on Cancer (IARC) of the World Health Organization (WHO) listed AAs and AA-containing plants as category 1 carcinogens (renal cancer) [4]. AAs exist mainly in Aristolochiaceae genera such as *Aristolochia*, *Asarum* [5], *Saruma* [6], and *Thottea* [7], many of which have been used as medicinal materials. For instance, in the Chinese market, 22 Aristolochiaceae herbs, more than 300 Chinese patent medicines, and 1000 prescriptions that contain Aristolochiaceae herbs have been used [8]. After the discovery of AA nephrotoxicity and carcinogenicity, the use of Aristolochiaceae herbs met with great challenges. AA-containing herbs began to be deleted from the Chinese Pharmacopoeia in 2005. In the most recent Chinese Pharmacopoeia, the vast majority of AA-containing herbs have been deleted, and only one remains (*Asarum sieboldii*) [9].

Over the years, many methods have been used for attenuation of AAs in medicinal herbs, such as processing, concerted application, and traditional breeding. However, these methods have a number of disadvantages, including being unsuitable for quantitative detection, difficult to supervise, or time-consuming and costly. More importantly, these methods cannot eliminate AAs from the herbs. Modern biotechnological tools, such as the CRISPR/Cas9 technique, show promise for the elimination of AAs from sources of medicinal materials. These tools require an understanding of the AA biosynthetic pathway and key enzyme genes. Unfortunately, there is little information available on this pathway, which was analyzed only in a few studies in the 1960s [10–13]. In addition, only one gene (TyrDC) that is probably involved
upstream of the pathway has been cloned [14]. In the AA biosynthetic pathway proposed in the 1960s, norlaudanosoline, orientaline, orientalinone, and orientalinol were considered to be intermediates of AA biosynthesis. However, norlaudanosoline was detected only in Macleaya cordata [15] and has been proposed to be virtually nonexistent in plants in recent studies. In addition, orientaline, orientalinone, and orientalinol have never been found in Aristolochiaceae plants. Therefore, it is important to revisit the proposed AA biosynthetic pathway. Recently, the biosynthesis of benzylisoquinoline alkaloids (BIAs) has been intensively studied [16]. Because the AAs are BIA derivatives [17], elucidation of BIA biosynthetic pathways provides a foundation for speculation about the AA biosynthetic pathway.

A. contorta is a vine from the family Aristolochiaceae in Pipersales with significant medicinal and economic value. It is a garden plant that is widely distributed in northern China, Korea, Japan, and Russia. A. contorta has been an original plant in traditional Chinese medicinal materials and was included in the official Chinese Pharmacopoeia. All of its organs, including roots, stems, leaves, flowers, and fruits, have been used medicinally for centuries, mainly to treat cough, asthma, and excessive phlegm [18]. Modern pharmacological studies have shown that A. contorta has anti-inflammatory, antihypertensive, antibacterial, anti-tumor, antifertility, analgesic, sedative, and other activities [19–21]. Phytochemical analysis showed that it contains abundant bioactive compounds, including alkaloids (such as AAs [22], aristolactams [23], magnoflorine and allantoin [24]), monoterpenes and sesquiterpenes (such as caryophyllene [25] and aristolactone [26]), β-sitosterol [24], pinitol, daucosterol [27], and other components.

With the aim of elucidating the biosynthetic pathways of AAs, we sequenced and assembled the whole genome of A. contorta using Illumina paired-end reads, Nanopore sequencing, and Hi-C technology. Genomic studies of A. contorta provide insight into the evolutionary position of the magnoliids and the absence of whole-genome duplication (WGD) events in Aristolochia. By combining genome sequencing with analysis of the transcriptome and metabolome, we identified candidate genes involved in the biosynthetic pathways of sanguinarine, berberine, and magnoflorine in A. contorta. In addition, possible biosynthetic pathways of aristolochic acids were proposed for further validation. The A. contorta genome assembly provides insights into the phylogenomics of magnoliids and AA biosynthesis. It also lays a solid foundation for the molecular breeding of Aristolochiaceae medicinal materials.

### Results and discussion

#### Genome sequencing, assembly and annotation

The sample used for whole genome sequencing was from an accession of A. contorta growing in the Medicinal Botanical Garden of the Institute of Medicinal Plant Development in Beijing, China. The size of the A. contorta genome was estimated to be 290.64 Mb, with repeat and heterozygosity percentages of 47.14% and 0.83%, respectively, as estimated by k-mer analysis based on next-generation sequencing data (Fig. S1). Illumina HiSeq technology, Oxford Nanopore Technologies (ONT), and Hi-C technology were integrated to sequence and assemble a chromosome-level genome assembly. A total of 26.86 Gb of reads were obtained from Illumina HiSeq, and the sequencing depth was 83.04×. A total of 38.59 Gb of clean data with an N50 of 42.26 kb were obtained after filtering 46.75 Gb of raw ONT reads generated using the PromethION platform. The longest ONT read was 798.932 kb, and the average depth of sequencing was approximately 183×. The primary assembly of ONT reads was performed with Canu, wtdbg, and SMARTdenovo, and the assembly was then corrected using the Illumina data. These steps resulted in a primary assembled genome of 210.53 Mb, which was composed of 173 contigs, with a contig N50 of 2.63 Mb (Table 1).

During construction of the Hi-C library, a total of 149,355,995 raw paired-end reads were generated, from which 12,799,624 valid interaction reads were obtained using HiC-Pro. Then, 99.4% of the genome sequences were anchored into 7 (2n = 14) pseudochromosomes using LACHESIS software. The intense intrachromosomal interaction signals indicated a high-quality Hi-C assembly (Fig. S2). The final genome assembly of A. contorta was 209.27 Mb (99.4% coverage of the genome). It was composed of 214 contigs with a contig N50 of 2.32 Mb and a scaffold N50 of 30.38 Mb. Among the contigs, 147 (198.56 Mb) were anchored with an exact order and direction (Tables 1, S1 and S2).

BUSCO and CEGMA analyses showed that 90.28% and 97.38% of the conserved core genes of eukaryotes, respectively, were identified in the A. contorta genome (Table S3). Moreover, mapping of the high-quality 350-bp Illumina reads to the assembled genome showed that the mapping rate was 96.85%. These results suggest the high completeness and contiguity of the final genome assembly (Fig. 1).

Based on the construction of a species-specific repetitive sequence database, 80.56 Mb of repetitive sequences

### Table 1. Statistics and characteristics of the A. contorta genome

| Assembly feature                  | Statistics     |
|-----------------------------------|----------------|
| Estimated genome size (Mb)        | 290.64         |
| Nanopore sequencing               |                |
| Contig number                      | 173            |
| Contig N50 (Mb)                    | 2.63           |
| Genome size (Mb)                   | 210.53         |
| Hi-C assembly                      |                |
| Scaffold number                     | 84             |
| Scaffold N50 (Mb)                  | 30.38          |
| Longest scaffold (Mb)              | 38.74          |
| Contig number                      | 224            |
| Contig N50 (Mb)                    | 2.32           |
| Assembled genome size (Mb)         | 209.27         |
| Assembly % of genome               | 99.4           |
were identified, constituting 38.26% of the whole genome. Among these sequences, approximately 33.05 Mb were long terminal repeats (LTRs), accounting for 41.04% of all repetitive sequences. This result suggests that LTRs are the dominant repetitive sequences in the A. contorta genome. Further analysis of LTRs showed that Copia (21.99 Mb) and Gypsy (10.87 Mb) were the most common subtypes, accounting for 27.31% and 13.49%, respectively (Table S4). Ab initio prediction, homology-based prediction, and unigenes from RNA-seq were used to predict protein-coding genes (Fig. S3). After integration and modification, 18,311 protein-coding genes were obtained (Table S5). They were composed of 27.95 Mb of exons with an average gene length of 1.53 kb (Table S6). Noncoding RNA prediction identified 46 microRNAs, 475 tRNAs, and 404 rRNAs. GeneWise and genBlastA were used to find 770 pseudogenes. Functional analysis of the assembled genome using the Kyoto Encyclopedia of Genes and Genomes (KEGG), eukaryotic orthologous groups (KOG), gene ontology (GO), TrEMBL, and nonredundant protein sequence (Nr) databases annotated 97.02% of the predicted genes (Table S7).

### Phylogenetic analysis

A. contorta belongs to the magnoliids. Currently, there are controversial views on the taxonomic status of magnoliids. To address this problem, the genome assembly of A. contorta was compared with 17 other sequenced genomes from two ANA-grade angiosperm species, Amborella trichopoda and Nymphaea colorata; eight eudicot species, Papaver somniferum, M. cordata, Nelumbo nucifera, Vitis vinifera, Glycine max, Arabidopsis thaliana, Coffea canephora and Helianthus annuus; three magnoliid species, Piper nigrum, Cinnamomum kanehirae, and Liriodendron chinensis; three monocots, Lemna minor from the order Alismatales, Oryza sativa from the order Poales, and Dendrobium officinale from the order Asparagales; and the gymnosperm species Ginkgo biloba, which served as an outgroup.

A phylogenomic analysis of 641 single-copy orthologous gene sequences placed monocots as a sister clade to the magnoliids + eudicots clade with 100% bootstrap support (Figs. 2a and S4). This result is in agreement with results from a phylotranscriptomic analysis of 92 streptophytes and land plants [28], an angiosperm phylogeny of 26 species [29], a phylogenomic analysis of the stout camphor tree [30], and a phylogenomic analysis of one thousand plant transcriptomes [31]. However, it contrasts with the classification of the sister relationship between magnoliids and a clade including monocots and eudicots in APG IV (Angiosperm Phylogeny Group, http://www.mobot.org/MOBOT/research/APweb/) and with results from phylogenomic analyses of angiosperms, such as Magnolia biondii [32], L. chinensis [33], and Phoebe bournei [34]. In addition, Qin et al. performed an analysis of genome structural rearrangements in Aristolochia fimbriata and various other plant species [35]. Their results showed that eudicots were sister to the magnoliid + monocot clade [35].

To further investigate the taxonomic status of magnoliids, we constructed a phylogenetic tree using the chloroplast genomes of A. contorta and 35 other plant species. The results showed that magnoliids were sister to the monocot + eudicot clade (Fig. S5). This result is inconsistent with the classification in the phylogenomic analysis of 18 representative plant genomes, above. Phylogenetic discordance for the position of magnoliids is probably due to incomplete lineage sorting [36] that resulted from the rapid radiation of early-diverging lineages within angiosperms (such as the magnoliids, monocots, and eudicots) that occurred within a very short period of time of <5 million years [32]. Ancient hybridization and parallel evolution that probably occurred during plant evolution may also have interfered with the results of phylogenetic analysis [35]. In addition, deviation may also be caused by aspects of lineage sampling, such as incomplete taxon sampling and the selection of Gramineae that could increase the likelihood of long branch attraction [37]. Moreover, different analytical methods (maximum likelihood, Bayesian [34], concatenation, coalescent, supermatrix, and ASTRAL methods [36]) and different sequence types (amino acid, nucleotide, and partitioned codons) could cause phylogenetic discrepancies. In fact, phylogenomic analyses of A. fimbriata and different taxon samples using the ASTRAL and supermatrix methods showed obscure topologies of monocots, eudicots, and magnoliids [35]. Therefore, pure phylogenomic analysis is not the best way to resolve the evolutionary position of magnoliids. With the increasing number of
sequenced genomes and the availability of more rigorous analytical methods and balanced taxon sampling, more powerful phylogenomic evidence will be obtained for the placement of magnoliids. We estimated the time of divergence between magnoliids and eudicots to be 147–165 million years ago (MYA) using the MCMC tree with fossil calibration. This estimate is supported by 95% highest posterior density. In addition, *A. contorta* and *Piper nigrum* (black pepper, also in the order Piperales), which diverged 89–135 MYA, had the closest evolutionary relationship. Piperales was classified as a sister group to a clade comprising Magnoliales and Laurales among the magnoliids (Fig. 2a). Piperales first diverged from Magnoliidae–Laurales approximately 137–154 MYA (Fig. 2a).

It is worth noting that the phylogenetic relationships within the order Piperales are still uncertain. APG IV accepted three families (Saururaceae, Piperaceae, and Aristolochiaceae) in Piperales. However, recent studies have shown that it is more reasonable to consider Piperales as an order with six families: Saururaceae, Piperaceae, Aristolochiaceae, Asaraceae, Hydnoraceae, and Lactoridaceae [38]. The reference-grade genome of *Aristolochia contorta* can lay the foundation for further resolving the phylogenetic positions.
of magnoliids in angiosperms and Aristolochiaceae in Piperales.

Analysis of whole-genome duplication

Many plants have experienced shared ancestral whole-genome duplication (WGD) events or lineage-specific WGD events, which make important contributions to species diversification [39]. For instance, an ancient WGD occurred just before the divergence of Laurales and Magnoliales, and a more recent WGD was shared by all lineages of Lauraceae [34, 36]. To assess the WGD events in A. contorta, we performed a comparative analysis of a range of species, including P. somniferum (in which a WGD event occurred ~7.8 MYA) [40], L. chinense (in which a WGD event occurred ~116 MYA) [33], P. nigrum (in which a recent WGD event occurred ~17 MYA) [41] and V. vinifera (a representative of the closest modern species to the ancestral eudicot karyotype [AEK]) [42, 43]. Synonymous substitution rate (Ks) distributions and fourfold synonymous (degenerative) third-codon transversion (4DTv) (Fig. 2b, Fig. 2c) confirmed the WGT-y in V. vinifera [42] and recent species-specific WGD events in P. somniferum, L. chinense, Nymphaea colorata, P. nigrum and N. nucifera [33, 40, 41, 44, 45]. Intragenomic synteny analysis of A. contorta using MCScanX showed 50 syntenic regions throughout the whole genome; these included 1202 genes that represented 6.56% of all genes. The dot and line plot of the synteny analysis showed that 99% of the replications were intra-chromosomal duplications (Fig. 1 and Fig. S6). Very sparse syntenic blocks existed in the A. contorta genome, consistent with a synteny analysis of the A. fimbriata genome [35]. Based on the Ks distributions and 4DTv rates, a major peak (Ks value of 0.2; 4DTv value of 0.01) was observed in the A. contorta Ks and 4DTv curve (Fig. 2b, Fig. 2c). This kind of peak could be caused by tandem repeats [46] but not indicate the occurrence of a recent WGD event in A. contorta. In addition, Amborella trichopoda and A. fimbriata are known not to have undergone any lineage-specific WGDs after their divergence from the last common ancestor of extant flowering plants [35, 47]. Comparative genomic synteny analysis of A. contorta with A. trichopoda and A. fimbriata revealed 1:1 and 1:1 syntenic depth ratios in the comparisons of A. contorta with A. trichopoda and A. fimbriata, respectively (Fig. 2d). These results strongly suggest the absence of a lineage-specific whole-genome duplication event in A. contorta. In addition, transcriptome phylogenetic analysis of 1000 plants showed that Aristolochia elegans had also not undergone any WGD after the angiosperm-wide WGD [31]. Because no WGD was identified in any of the three Aristolochia species (A. contorta, A. fimbriata and A. elegans) analyzed to date, it is very likely that no Aristolochia lineage-specific WGD event has occurred in the evolutionary history of the genus.

Because of a lack of WGD events and subsequent subgenome rearrangements, the genomes of Aristolochia species are relatively concise, and this can help us to identify WGDs in magnoliids and other early angiosperms through genome comparison. By comparison of intergenicom syntenic blocks of A. fimbriata and P. nigrum, Qin et al. [35] identified another two ancient WGDs in P. nigrum that were not reported in the genome article [41]. Here we performed comparative genomic synteny analysis for A. contorta and P. nigrum. The dot plot clearly showed a syntenic depth ratio of 1:8 in the comparison of A. contorta and P. nigrum (Fig. S7). The results are consistent with those obtained by Qin et al. [35]. A. contorta has a smaller genome than A. trichopoda and A. fimbriata. It can serve as another useful reference for phylogenomic studies of angiosperms. Moreover, the highly conserved synteny between Amborella and Aristolochia provides insights into the reconstruction of the ancestral angiosperm genome [35]. We expect that the release of more Aristolochia genomes will enable the construction of a highly credible ancestral angiosperm genome. This will promote the analysis of quantitative changes and evolutionary schemes of A. contorta genes retained from the ancestral genome after the angiosperm-wide WGD.

Comparative genomic analysis

In order to identify potential genes encoding enzymes in the biosynthetic pathways of AAs or other important bioactive compounds, the A. contorta genome assembly was compared with 11 other sequenced genomes from two basal angiosperms, Amborella trichopoda and Nymphaea colorata; four eudicot species, P. somniferum, M. cordata, N. nucifera, and V. vinifera; three magnoliid species, P. nigrum, C. kanehira, and L. chinense; and two monocots, O. sativa and Zea mays. A total of 47786 gene families were identified, including 3495 shared gene families and 201 gene families that were unique to the A. contorta genome (Fig. S8). Gene copy number analysis showed that A. contorta contained 63.7% single-copy genes and 16.9% two-copy genes (Fig. S9). Patterns of gene-family sharing were studied among A. contorta, P. somniferum, L. chinense, N. nucifera, and V. vinifera; three magnoliid species, P. nigrum, C. kanehira, and L. chinense; and two monocots, O. sativa and Zea mays. A total of 11 contracted and 25 expanded gene families were predicted (Fig. S10). According to GO enrichment annotations and KEGG pathway classification, the expanded families were enriched in sesquiterpene biosynthesis. This coincides with the fact that A. contorta plants have pungent odors and contain large amounts of volatile oil, the largest proportion of which is sesquiterpenoids. Among these compounds, caryophyllene is the main component, accounting for one third of the total volatile oil [25]. It has anti-asthmatic effects and is one of the active ingredients for the treatment of aged chronic bronchitis, as well as a local anesthetic. Borneol has diaphoretic, excitement, and antispasmodic effects, whereas pinene has antitussive and expectorant properties and is antifungal [48]. The efficacy of A. contorta in traditional Chinese medicine may be closely related to the pharmacological activities of these volatile components that it contains.
In order to study the unique functions of genes in A. contorta, positive selection analysis was performed among L. chinense, C. micranthum, A. contorta, M. cordata, and N. nucifera. It resulted in the identification of 94 A. contorta genes that were subjected to significant positive selection. GO enrichment analysis of these genes indicated that most of them were related to cellular nitrogen compound metabolic process, metal ion binding, methyltransferase activity, nucleolus, and rRNA processing. Classification of positively selected genes by KEGG pathways identified those for basal transcription factors, RNA degradation, ether lipid metabolism, and so forth. In particular, the term cellular nitrogen compound metabolic process may represent the metabolism of A. contorta–specific compounds, AAs, in vivo. Methyltransferase activity was one of the most enriched molecular function GO terms. It may be related to new functions of O-methyltransferases (OMTs), N-methyltransferases (NMTs), and other methyltransferases in A. contorta.

In addition, specific gene families in A. contorta were analyzed using GO and KEGG enrichment analysis (Fig. S11). There were several terms related to secondary metabolism in GO enrichment analysis of unique genes, including cellular aromatic compound metabolic process, sesquiterpene biosynthetic process, methylation, oxidoreductase activity, sesquiterpene synthase activity, S-adenosylmethionine-dependent methyltransferase activity, and others. Various OMT, NMT, and norcocludarine synthase (NCS) genes were identified as unique genes related to the GO terms. KEGG analysis showed that the specific gene families were enriched in limonene and pinene degradation, glucosinolate biosynthesis, and the biosynthesis of stilbenoids, diarylheptanoids, and gingerols. In addition, a variety of cytochrome P450 gene families, such as CYP706, CYP86B, CYP82C, and CYP71 involved in secondary metabolism, were annotated as specific gene families of A. contorta. These results indicate the significance of A. contorta OMTs and CYPs in the biosynthesis and metabolism of unique compounds such as AAs and BIAs.

**Identification and characterization of the AcCYP gene family**

O-methylation in the biosynthesis of multiple secondary metabolites is primarily performed by S-adenosyl-l-methionine (SAM)-dependent O-methyltransferases (OMTs) encoded by the OMT gene family. Plant OMTs deliver the methyl group of SAM to the hydroxyl group of alkaloids, flavonoids, phytalexins, and lignin, resulting in the production of their methyl ether derivatives [49]. The addition of a methyl group to an alkaldoid structure has a significant effect on its chemical properties and biological activity [50]. To study the functions of the OMT gene family in the biosynthesis of BIAs and their derivatives in A. contorta, genome-wide identification of AcOMTs was performed. The results indicated that there were 25 AcOMT genes in A. contorta. Neighbor-joining phylogenetic analysis of 25 AcOMT proteins and six BIA-related OMTs from P. somniferum (Ps6OMT, Ps4’OMT1, and Ps4’OMT2), Coptis japonica (Cj6OMT and Cj4’OMT) and Thalictrum flavum (Tf6OMT) [51–55] using MEGA 7.0 with default parameters and 1000 bootstraps identified seven full-length AcOMT homologs: AcOMT1 (EVM0011328), AcOMT2 (EVM0008621), AcOMT3 (EVM0008782), AcOMT4 (EVM0006633), AcOMT5 (EVM0014567), AcOMT6 (EVM0012543), and AcOMT7 (EVM0014836) (Fig. 3a). MEME analysis of the seven identified AcOMTs and six BIA-related OMTs from other plant species showed ten conserved motifs (Fig. 3b). The results suggested that these OMTs could have similar functions in BIA biosynthesis, although their actual catalytic functions remain to be investigated.
Metabolome profiling and analysis of AA contents in A. contorta

To date, phytochemical studies on the alkaloids of A. contorta have focused mainly on AAs and aristolactams (ALs). The AAs identified include aristolochic acids I–VII, E, F, G, and their derivatives, such as aristolochic acid Ia, aristolochic acid IVb, and 7-hydroxyaristolochic acid I. The ALs identified include aristolactams I, II, III, IV and their derivatives, such as aristolactam IIIa and aristololactam Ia-N-β-D-glucopyranoside. In addition, caryophyllene, pinene, borneol, aristolactone, aristolone, β-sitosterol, allantoin, magnoflorine, pinitol, and daucosterol have also been identified in A. contorta. However, there are few reports on BIAs and intermediate compounds of the AA biosynthetic pathway. To comprehensively identify the metabolites produced in A. contorta, the metabolome in four tissues of A. contorta was analyzed. This resulted in the identification of 1147 compounds, including protopine, norsanguinarine, morphine, noscapine, anonaine,
and other BIAs (Table S9). In addition, the contents of AAs in different parts of A. contorta plants were determined using UPLC. The results showed that AAs accumulated primarily in roots, followed by flowers. The accumulation of AAs, particularly aristolochic acid A and aristolochic acid B, was much lower in stems and leaves (Fig. 4).

Identification of candidate genes involved in BIA biosynthetic pathways in A. contorta

BIAs are a group of plant-specific metabolites with multiple structures that have a long history of investigation [62]. The pharmacological activities and biosynthetic pathways of various BIA compounds, such as antimicrobial sanguinarine and berberine, anti-diabetic magnoflorine, and analgesic morphine, have been comprehensively studied in multiple plant species, including P. somniferum, M. cordata, and Coptis japonica [15, 63–66]. BIA compounds have a collective upstream biosynthetic pathway, which includes the condensation of tyrosine derivatives dopamine and 4-hydroxyphenylacetaldehyde and four other enzymatic steps to yield the critical branch node intermediate (S)-reticuline [52]. The biosynthesis of BIAs, including morphinan (morphine and codeine), benzophenanthridine (sanguinarine), and protoberberine (berberine) [62], proceeds from the intermediate (S)-reticuline. To date, little is known about the biosynthetic pathways of BIAs in A. contorta.

Based on previous studies of magnoflorine, sanguinarine, and berberine [61, 64, 67–69], we proposed the biosynthetic pathways of BIAs in A. contorta. As shown in Fig. 5a, the common biosynthetic pathway of BIAs (orange) begins with the condensation of L-tyrosine derivatives, dopamine and 4-hydroxyphenylacetaldehyde, to yield (S)-norcoclaurine through the catalysis of NCS [70, 71]. The important intermediate (S)-reticuline that acts as the central branch point of several BIAs is then synthesized from (S)-norcoclaurine through the catalysis of four enzymes: 6OMT, CNMT, NMCH, and 4′OMT [67, 72–76]. Finally, magnoflorine (yellow), sanguinarine (blue), and berberine (green) are produced from (S)-reticuline through the catalysis of multiple enzymes [15, 64, 77]. Genome-wide analysis predicted 91 nonredundant genes that were probably involved in the biosynthesis of BIAs in A. contorta (Table S10). Transcriptome analysis of A. contorta roots, stems, leaves, and flowers showed that the identified candidate genes were differentially expressed (Fig. 5b). The expression of fifteen candidate genes was validated by qRT-PCR (Fig. S14). These data provide the basis for functional analysis of genes involved in BIA biosynthetic pathways.

Enzyme assay of norcoclaurine 6-O-methyltransferase

(S)-norcoclaurine-6-O-methyltransferase (6OMT) is a critical rate-limiting enzyme that participates in the biosynthesis of reticuline, the versatile intermediate shared by various end-products of BIA biosynthesis [52]. To identify the 6OMT genes in A. contorta, seven of the nine 6OMT candidate genes (Fig. 5b), which were highly expressed in at least one analyzed tissue, were cloned and introduced into Escherichia coli cells to obtain recombinant protein. In vitro assays showed that E. coli cells produced a main band at the theoretical molecular mass of His-tagged AcOMTs (~47 kDa) after IPTG induction (Figs. S15 and S16). With the exception of AcOMT3-transformed E. coli, this band was clearly observed in the supernatant of the cell lysate, indicating the production of abundant soluble AcOMT protein. SDS-PAGE analysis of purified soluble AcOMT1, AcOMT2, AcOMT5, and AcOMT7 proteins and renatured inclusion bodies of AcOMT3 are shown in Figs. S15, S17, and S18. Norcoclaurine was tested as a potential substrate. The enzyme activity of AcOMTs was analyzed in Tris–HCl buffer and detected using UPLC. As shown in Fig. 3d, a coclaurine peak was detected in the reaction of AcOMT1 and AcOMT2 enzymes. UPLC analyses demonstrated that AcOMT1 and AcOMT2 could convert norcoclaurine into coclaurine (Fig. 3d). An unknown peak with a retention time of 3.08 min was observed in the assay of AcOMT1.
Twenty-nine nonredundant genes that are probably involved in AA biosynthesis are shown in red. Heat map values represent relative gene expression (orange), magnoflorine (yellow), norsanguinarine (blue), and berberine (green) in A. contorta. Compounds detected in the metabolic data are shown in red. 

Expression profile of 91 candidate genes potentially involved in BIA biosynthesis. Gene expression values (FPKM) were scaled by column normalization. TYDC, tyrosine decarboxylase; 3OMT, tyrosine 3-hydroxylase; TyrAT, tyrosine aminotransferase; 4HPPDC, 4-hydroxyphenylacetate dehydrogenase; NCS, norcoclaurine synthase; 6OMT, norcoclaurine 6′-methyltransferase, this peak may have resulted from the conversion of the hydroxyl group at the 7 or 4′ position to methoxy. Thus, AcOMT1 and AcOMT2 probably encode 6OMT in A. contorta, whereas the functions of the other AcOMT genes remain to be elucidated.
Possible biosynthetic pathways of AAs in *A. contorta*

Based on genomic and transcriptomic data from *A. contorta* and the results obtained from previous BIA biosynthetic pathway studies, we deduced and predicted possible pathways and candidate genes for AA biosynthesis in *A. contorta*. AAs are biogenetically derived from BIA precursors, and magnoflorine is always identified together with AAs in various Aristolochia species [11]. In addition, norlaudanosoline has been considered to be an intermediate of AA biosynthetic pathways for several decades [10–14]. However, the distribution of norlaudanosoline in plants is narrow: it has only been found in *M. cordata* [15]. Thus, the important intermediate compound in the AA biosynthetic pathway could be (S)-norcoclaurine but not norlaudanosoline. Based on information from BIA biosynthetic pathway analysis, we propose that the upstream portion of AA biosynthetic pathways is essentially the same as that of the BIA biosynthetic pathways shown in Fig. 5a (orange). It involves enzymes from the norcoclaurine synthase (NCS) family, O-methyltransferase (OMT) family, N-methyltransferase (NMT) family and cytochrome P450 (CYP) superfamily and leads to the formation of the intermediate (S)-reticuline (Fig. 5a). Then, a novel C–C bond is formed between carbon 8 and carbon 2′ on the benzyl of (S)-reticuline through a C–C phenol-coupling reaction, resulting in the production of the basic skeleton of aporphine alkaloids. Subsequently, the methoxy and hydroxyl groups on carbon 6 and carbon 7 are condensed to generate the methylenedioxy bridge, and the oxidative ring-opening of the nitrogen heterocycle occurs to form the final backbone of AAs. Substitutions of different functional groups at different positions form different types of AAs, such as aristolochic acids A, B, C, and D. Because laurelliptine, nandigerine, anonaine, and dehydroaporheine were identified in the metabolome (Table S9), we speculate that (S)-coclaurine and (S)-N-methylcoclaurine [78] may also serve as intermediates for the C–C phenol-coupling reaction, methylenedioxy bridge generation, and other subsequent reactions (Fig. 5a). Based on gene expression patterns and AA contents in different tissues, 29 nonredundant genes probably involved in AA biosynthesis were identified for further validation (Fig. 5b, shown in red).

Conclusions

A chromosome-level reference genome assembly of *A. contorta* was constructed using the Nanopore and Hi-C technologies. The assembled genome is 209.27 Mb in size, with contig and scaffold N50 sizes of 2.63 Mb and 30.38 Mb, respectively. The sequence was anchored into 7 pseudochromosomes. Phylogenetic analysis showed conflicting results for the taxonomic status of magnoliids. Synteny analysis, Ks distribution, and 4DTv rates suggested the absence of lineage-specific WGD events in *A. contorta*. Comparative genomic analysis identified 11 contracted and 25 expanded gene families and indicated the importance of *A. contorta* OMTs and CYPs in the biosynthesis of AAs and BIAs. Genome-wide analysis identified 25 AcOMT and 241 AcCYP genes in *A. contorta*. In addition, 91 genes encoding enzymes likely to be involved in BIA biosynthetic pathways were predicted. Among them, AcOMT1 and AcOMT2 were experimentally confirmed to encode 6OMT in *A. contorta*. AA biosynthetic pathways were also proposed for further validation. Aristolochia species include a variety of medicinal plants and have been used in Asia and other regions for a long time. AAs and their derivatives are one of the most important groups of compounds produced in Aristolochia species. To date, significant attention has been paid to Aristolochia-derived herbal drugs because of aristolochic acid nephropathy (AAN) [79,80]. This chromosome-level genome assembly of *A. contorta* provides a high-quality reference genome for AA-containing plant species and provides insights into the phylogenomics of magnoliids and the molecular breeding of Aristolochiaceae.

Materials and methods

Sample collection and genome sequencing

DNA was extracted from young leaves of *A. contorta* collected from the Medicinal Botanical Garden of the Institute of Medicinal Plant Development. Two 350-bp Illumina DNA libraries were constructed and sequenced using an Illumina sequencer with a read length of 150 bp. The nanopore library was constructed with high-molecular-weight DNA according to the ONT protocol using the Ligation Sequencing kit (SQK-LSK109) and sequenced using the PromethION Flow Cell Priming Kit (EXP-FLP001.PRO.6) (Oxford Nanopore Technologies, UK) according to the manufacturer’s instructions. The ONT reads were self-corrected using Canu [81], and the corrected reads were assembled into contigs using SMARTdenovo (https://github.com/ruanjue/smartdenovo). The assembled contigs were polished using Racon [82] and Pilon [83]. The quality of the genome assembly was assessed by BUSCO (https://busco.ezlab.org/) and CEGMA [84] analyses.

Hi-C fragment libraries were constructed and sequenced using the Illumina platform with insert sizes of 300–700 bp. The raw reads were trimmed to obtain clean Hi-C reads, which were then mapped to the genome assembly using BWA [85] (version 0.7.10-r789). The uniquely alignable paired reads (mapping quality > 20) were retained for scaffold assembly using LACHESIS [86] with the parameters “CLUSTER_MIN_RE_SITES, 60; CLUSTER_MAX_LINK_DENSITY, 2; ORDER_MIN_N_RES_IN_TRUN, 58; ORDER_MIN_N_RES_IN_SHREDS, 55” after filtering invalid reads with HiC-Pro [87] v2.8.1.

Gene prediction and functional annotation

To predict repeated sequences in the *A. contorta* genome, a de novo repeat library was constructed using LTR_FINDER
A. trichopoda and other 11 species, described above. Protein sequences of copy genes in at least 10 species using the same method was constructed for 12 species based on 1199 single-phylogenomic tree. 1000 bootstraps. The ML tree was constructed with MAFFT v7.487 [111], and gaps were removed using Gblocks (117–130 MYA) were used to calibrate the divergence times. Published data from A. contorta was used as the initial query sequences to search the NCBI nucleotide database (117–130 MYA) were used to construct the phylogenetic tree branch were calculated based on differentiation times and gene family clustering using CAFE v4.2 [119]. Positive selection analysis was performed using the CodeML program in PAML. Single-copy genes from A. contorta, L. chinense, C. micranthum, M. cordata, and N. nucifera were analyzed using MAFFT, PAL2NAL, model A and the null model of PAML, and the branch-site model of CodeML to perform likelihood ratio tests (P < 0.01). The Bayes method was used to identify genes that were subjected to significant positive selection. Single-copy genes in A. contorta with a Ka/Ks ratio > 1 and posterior probability >0.95 were considered to be under significant positive selection.

Synteny and WGD event analysis
Gene sequences from two species were compared, and similar gene pairs were determined using Diamond v0.9.29.130 with E-value <1 × 10^{-5} and C score > 0.5. The C score value was filtered by JCVI software [120]. MCScanX [121] and the gff3 file for the genome assembly were used to determine whether similar gene pairs were adjacent on the chromosome. The maximum gap between the anchor genes was set to 30 to obtain all syntenic gene blocks. Whole genome duplication (WGD) events were identified by a Ks distribution analysis using wgd v1.1.1 [122], and fourfold synonymous third-codon transversion (4DTv) rate was determined using a Perl script (https://github.com/infengChen/Scripts).

Gene identification and functional characterization of AcOMTs
The hidden Markov model (HMM) corresponding to the O-methyltransferase (OMT) domain, including methyltransf_2 (PF00891) and methyltransf_3 (PF01596), was acquired from the Pfam database (http://pfam.sanger.ac.uk/). OMT genes were identified using HMMER 3.0 in A. contorta genome. Default parameters were used, and the threshold was set to 1e-5. Six OMTs, Ps6OMT, Tf6OMT, Cj6OMT, Ps4OMT1, Ps4OMT2, and Cj4OMT, were used as the initial query sequences to search the A. contorta protein database using the BLASTP program. The results of the HMM and BLASTP programs were merged. After deleting all redundant sequences, the hypothetical OMT protein sequences were submitted to CDD (https://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi) and Pfam to confirm the presence of the conserved OMT domain. MEME (https://meme-suite.org/meme/tools/meme) was used to analyze conserved motif structures.

Full-length cDNAs of the AcOMTs were acquired by PCR amplification with the primers listed in Table S12. The cDNAs were cloned into the pET30a vector and...
introduced into BL21(DE3). Transformed E. coli cells were cultured in LB liquid medium supplemented with 50 μg mL⁻¹ kanamycin and induced with 0.5 mM IPTG at 16°C for 5 hours at a low speed. The cells were collected by centrifugation and vortexed in binding buffer. After that, the suspension was sonicated at 200 W. Cell debris was eliminated by centrifugation at 13000 rpm for 8 min. The soluble protein was purified using Ni-NTA Resin (Transgen Biotech, Beijing) according to the manufacturer’s instructions. To isolate AcOMT3 protein, the purified inclusion body was denatured with 6 M guanidine hydrochloride, renatured by dialysis, and then dissolved in Tris–HCl. SDS-PAGE and Coomassie blue staining were used to confirm the purity of the His-tagged proteins. The purified proteins were analyzed with the BCA Protein Assay kit (Takara Biomedical Technology, Beijing) to determine the protein concentration.

The enzyme assay for OMT activity was performed in a 50-μl reaction system that comprised 25 mM Tris–HCl (pH 8.0), 25 mM sodium ascorbate (Sigma), 0.1 mM S-adenosyl-L-methionine (Sigma), 100 mM substrate, and 10 μg purified enzyme. The reactions were incubated at 37°C for 1 h and terminated by adding 100 μL of methanol. Controls were performed with total protein from E. coli transformed with the empty pET-30a vector [52, 56]. After centrifugation, 10-μL supernatants were injected into the ACQUITY UPLC system (Waters, Milford, MA, USA). The samples were separated on an ACQUITY UPLC BEH C18 column (1.7 μm, 100 × 2.1 mm) at 35°C. The mobile phases were methanol (A) and water containing 0.1% formic acid (B). The mobile phases changed according to the following gradient with a flow rate of 0.3 mL min⁻¹: 0–0.5 min, holding at 5% A; 0.5–4 min, increasing from 5% A to 95% A; 4–6 min, holding at 95% A; 6–6.1 min, falling to 5% A; 6.1–8 min, holding at 5% A. Coelcuraine was detected using a photodiode array detector at an absorbance of 280 nm.

Cytochrome P450 gene superfamily identification and analysis

The HMM file for cytochrome P450 (PF00067) was downloaded from the Pfam database (http://pfam.xfam.org/) and used to search for CYP genes in the A. contorta genome with hmmsearch (HMMER 3.2.1). The A. thaliana CYP genes were downloaded from the Arabidopsis Information Resource (https://www.arabidopsis.org/) and used to search for CYP genes in the A. contorta genome using BLASTP with E-value < 1e⁻⁴⁰. After removing redundant sequences from the HMM and BLASTP search results, the presence of the conserved CYP domain in the putative CYP protein sequences was confirmed using CDD (https://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi) and SMART (http://smart.embl-heidelberg.de/#). A manual analysis was performed to identify full-length sequences. Based on the criteria of the CYP nomenclature system [123], in which amino acid sequence similarity of P450 families is >40% and that of subfamilies is >55% (https://www.a

RNA sequencing and gene expression analysis

Total RNA was isolated from four tissues of A. contorta: roots, stems, leaves, and flowers. Sequencing libraries were generated by mRNA isolation, fragmentation, first-strand cDNA synthesis, second-strand cDNA synthesis, and purification using the NEBNext Ultra RNA Library Prep kit according to the manufacturer’s instructions. The library preparations were sequenced on an Illumina platform (NEB, USA). Clean reads were obtained after removing adaptor sequences and low-quality sequence reads from the raw data and then mapped to the draft genome of A. contorta using HISAT2. Gene expression levels were estimated by FPKM values (fragments per kilobase of transcript per million fragments mapped) using StringTie [124].

Metabolite analyses

Roots, stems, leaves, and fruits of A. contorta were used for metabolite analysis on an LC/MS system. The UHPLC separation was performed using a 1290 Infinity series UHPLC System (Agilent Technologies). MS/MS spectra were obtained using the TripleTOF 6600 mass spectrometry system (AB Sciex) on an information-dependent basis (IDA). Analyst TF 1.7 (AB Sciex), ProteoWizard, and XCMS [125] software were used for data analysis.

Validation of gene expression levels by quantitative real-time PCR (qRT-PCR)

To validate the digital expression data, the transcript levels of 15 selected genes in different tissues were measured by qRT-PCR. Total RNA was extracted using the EASYspin Plus Complex Plant RNA kit (Aidlab, Beijing, China). First-strand cDNA was synthesized using the PrimeScript RT reagent kit with gDNA Eraser (Takara). The qRT-PCR primers designed using Oligo 7 are shown in Table S13. AcActin was used as the reference gene. qRT-PCR reactions were performed using TB Green Premix Ex Taq II (Takara) on a Bio-Rad CFX96 Real-Time system. Data for gene expression levels in different tissues were analyzed using the 2⁻ΔΔCt method.

Identification of candidates in the BIAs pathway of A. contorta

Local BLASTP was used to identify candidate genes involved in the BIA pathways. Reference protein sequences from P. somniferum, C. japonica, M. cordata, and Eschscholzia californica were obtained from the NCBI and UniProt databases and aligned against the A. contorta genome
using BLASTP with a maximum E-value of 1e⁻¹⁰. The results were filtered by identity >40% and coverage >50% to allow for distant species relationships. Based on the gene family characteristics of NCS sequences, the identity value was set to 30% [71]. A heatmap was generated using TBtools with column scaling [126].

**Aristolochic acid content analysis**

*A. contorta* tissues were ground in liquid nitrogen. Samples of powder (0.1 g) were dissolved in 2 mL of 70% methanol. The mixture was allowed to stand at room temperature for 15 min and then sonicated for 30 min. After centrifugation, the supernatant was collected. The sediment was dissolved in 70% methanol and then sonicated for 30 min. This supernatant was collected after centrifugation and combined with the previous supernatant. All supernatants were filtered through a 0.22-μm membrane and injected into the UPLC. Aristolochic acids A, B, C, and D were quantitatively detected via an ACQUITY UPLC system (Waters, Milford, MA, USA) with an ACQUITY UPLC BEH C18 column (1.7 μm, 100 x 2.1 mm) at 30°C. The mobile phases were acetonitrile and water containing 0.2% acetic acid. The elution was carried out at a flow rate of 0.4 mL min⁻¹ with 40% acetonitrile. AAs were detected using a photodiode array detector at 254 nm absorbance.

**Acknowledgements**

This work was supported by the CAMS Innovation Fund for Medical Sciences (CIFMS) (2021-I2M-1-029).

**Author contributions**

S.L., C.L., X.C., and F.M. conceived the project. X.C. and F.M. analyzed the data and wrote the manuscript. S.L. revised the manuscript. X.C. cloned the genes and performed the enzymatic reactions. X.P. contributed to the data analysis. X.Q. and S.Z. helped with the preparation of samples. All authors read and agreed to the final article.

**Data availability**

The assembled *A. contorta* genome sequence and raw transcriptome sequencing data are available from the NCBI database under project ID PRJNA738351 and PRJNA742069. The annotation information has been uploaded to the CoGe website (https://genomevolution.org/coge/SearchResults.pl?s=Aristolochia%20contorta&p=genome).

**Conflict of interest statement**

The authors declare no conflicts of interest.

**Supplementary data**

Supplementary data is available at Horticulture Research online.

**References**

1. Stiborová M, Arlt VM, Schmeiser HH. DNA adducts formed by aristolochic acid are unique biomarkers of exposure and explain the initiation phase of upper urothelial cancer. *Int J Mol Sci*. 2017;18:2144.
2. Zhou Q, Pei J, Poon J et al. Worldwide research trends on aristolochic acids (1957–2017): suggestions for researchers. *PLoS One*. 2019;14:1–23.
3. Balachandran P, Wei F, Lin R et al. Structure activity relationships of aristolochic acid analogues: toxicity in cultured renal epithelial cells. *Kidney Int*. 2005;67:1797–805.
4. Grosse Y, Baan R, Straif K et al. A review of human carcinogens—part a: pharmaceuticals. *Lancet Oncol*. 2009;10:13–4.
5. Feng T, Lee MR, Hsiao SS et al. Analysis of aristolochic acid in nine sources of Xixin, a traditional Chinese medicine, by liquid chromatography/atmospheric pressure chemical ionization/tandem mass spectrometry. *J Pharm Biomed Anal*. 2003;33:831–7.
6. Zhao H, Jiang H. Determination of aristolochic acid a in Saruma henryi by HPLC. *Guiguia*. 2009;29:548–51.
7. Nishida R, Weintraub JD, Feeny P et al. Aristolochic acids from Thottea spp. (Aristolochiaceae) and the osmeterial secretions of *Thottea*-feeding troidine swallowtail larvae (Papilionidae). *J Chem Ecol*. 1993;19:1587–94.
8. Luan Y. Thinking and discussion on the carcinogenicity of aristolochic acids, a class of compounds from traditional Chinese medicine: a summary of the presentations in the international symposium on the safety of traditional Chinese medicine. *Mod Tradit Chinese Med Mater Medica-World Sci Technol*. 2019;1–7.
9. Chinese Pharmacopoeia Commission. Pharmacopoeia of the People Republic of China 2020 Volume 1. (China Medical Science Press, 2020).
10. Spenser DI, Tiwar1 HP. Biosynthesis of aristolochic acid Chem Commun. 1966;55–6.
11. Conner F, Tiwari HP, Spenser ID. Biosynthesis of aristolochic acid. *Can J Chem*. 1969;47:481–7.
12. Schutte HR, Orban U, Mothes K. Biosynthesis of aristolochic acid. *Eur J Biochem*. 1967;1:70–2.
13. Sharma V, Bhakuni DS, Kapil RS. Biosynthesis of aristolochic acid. *JCS Perkin 1*. 1982;1153–5.
14. Wang X, Hui F, Yang Y et al. Deep sequencing and transcriptome analysis to identify genes related to biosynthesis of aristolochic acid in *Asarum heterotropoides*. *Sci Rep*. 2018;8:1–14.
15. Liu X, Liu Y, Huang P et al. The genome of medicinal plant *Macleaya cordata* provides new insights into benzylisoquinoline alkaloids metabolism. *Mol Plant*. 2017;10:975–89.
16. Singh A, Menéndez-Perdomo IM, Facchini PJ. Benzylisoquinoline alkaloid biosynthesis in opium poppy: an update. *Phytotax. Rel*. 2019;18:1457–82.
17. Dembitsky VM, Gloriozova TA, Porokov VV. Naturally occurring plant isoquinoline N-oxide alkaloids: their pharmacological and SAR activities. *Phytomedicine*. 2015;22:183–202.
18. Comission CP. Pharmacopoeia of the People Republic of China Vol. 1. (China Medical Science Press, 2015).
19. Chen M, Zhu Z. Progress on study of pharmacological functions of *Aristolochia L. Plant*. 2007;3:59–62.
20. Li K. The pharmacology of *Aristolochia* species. *Acta Acad Med CPAFF*. 2000;9:230–2.
21. Zhao H, Liu X. Review on the studies of Chinese *Aristolochia L*. herbs. *J Henan Univ Naturai Sci*. 2003;33:73–7.
22. Tian H, Cheng X, Wang C et al. Simultaneous determination of four constituents in *Aristolochia debilis* by HPLC. *Chinese Tradit Pat Med*. 2016;38:560–5.
Liu J, Dai Z, Cheng X et al. Research progress on detection and analysis of aristolochic acids. Drug Eval Res. 2019;1644–9.

Lou F, Linsheng D, Li L. Chemical studies on Aristolochia contorta Bunge. Chinese Tradit Herb Drugs. 1986;17:390–1.

Zhang C, Yu J, Wang B et al. Analysis of volatile oil constituents in the fruits of Aristolochia contorta by GC-MS. Chin J Nat Med. 2004;2:126–8.

Sun R, Zhang L, Luo Z et al. Studies on the chemical constituents of Aristolochia contorta Bunge. J Anhui Agri Sci. 2011;39:14620–1.

Chen Y, Yu L. Studies on the chemical constituents of Aristolochia Contorta. Yunnan Norm Univ. 2005;25:41–4.

Wickett NJ, Mirarab S, Nguyen N et al. Phylogenomic analysis of the origin and early diversification of land plants. Proc Natl Acad Sci U S A. 2014;111:E4859–68.

Chen L, Zhang Q, Sun R et al. Resolution of deep angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. Nat Commun. 2014;5:4956.

Chen J, Hao Z, Guang X et al. Lirioidendron genome sheds light on angiosperm phylogeny and species–pair differentiation. Nat Plants. 2019;5:63–73.

One Thousand Plant Transcriptomes Initiative. One thousand plant transcriptomes and the phylogenomics of green plants. Nature. 2019;574:679–85.

Dong S, Liu M, Liu Y et al. The genome of Magnolia biondii Pamp. Provides insights into the evolution of Magnoliales and biosynthesis of terpenoids. Hortic Res. 2021;8.

Chen J, Hu Z, Guang X et al. Lirioidendron genome sheds light on angiosperm phylogeny and species–pair differentiation. Nat Plants. 2019;5:18–25.

Chen SP, Sun WH, Xiong YF et al. The Phoebe genome sheds light on the evolution of magnolids. Hortic Res. 2020;7:146.

Qin L, Hu Y, Wang J et al. Insights into angiosperm evolution, floral development and chemical biosynthesis from the Aristolochia fimbriata genome. Nat Plants. 2021;7:1239–53.

Chen YC, Li Z, Zhao YX et al. The Litsea genome and the evolution of the laurel family. Nat Commun. 2020;11:1675.

Susko E, Roger AJ. Long Branch attraction biases in Phylogenetics. Syst Biol. 2021;70:838–43.

Jost M, Samain MS, Marques I et al. Discordant Phylogenomic placement of Hydnoraceae and Lactoridaceae within Piperales using data from all three genomes. Front Plant Sci. 2021;12.

Ren R, Wang H, Guo C et al. Widespread whole genome duplications contribute to genome complexity and species diversity in angiosperms. Mol Plant. 2018;11:414–28.

Guo L, Winzer T, Yang X et al. The opium poppy genome and morphinan production. Science. 2018;362:343–7.

Hu L, Xu Z, Wang M et al. The chromosome-scale reference genome of black pepper provides insight into piperine biosynthesis. Nat Commun. 2019;10:4702.

Jalil I, Aury J, Noel B et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature. 2007;449:463–7.

Salse J. Ancestors of modern plant crops. Curr Opin Plant Biol. 2016;30:134–42.

Zhang L, Chen F, Zhang X et al. The water lily genome and the early evolution of flowering plants. Nature. 2020;577:79–84.

Shi T, Rahmani RS, Gugger PF et al. Distinct expression and methylation patterns for genes with different fates following a single whole-genome duplication in flowering plants. Mol Biol Evol. 2020;37:2394–413.

Badouin H, Gouzy J, Grassa CJ et al. The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. Nature. 2017;546:148–52.

Amborella Genome Project. The Amborella genome and the evolution of flowering plants. Science. 2013;342:1241089–9.

Ghelardini C, Galeotti N, Di Cesare Mannelli L et al. Local anaesthetic activity of β-caryophyllene. ChemMedChem. 2001;5:387–9.

Liu X, Luo Y, Wu H et al. Systematic analysis of O-methyltransferase gene family and identification of potential members involved in the formation of O-methylated flavonoids in citrus. Gene. 2016;575:458–72.

Morris JS, Facchini PJ. Molecular origins of functional diversity in benzylisoquinoline alkaloid methyltransferases. Front Plant Sci. 2019;10.

Sato F, Tsujita T, Katagiri Y et al. Purification and characterization of S-adenosyl-L-methionine:norcocaurine 6-O-methyltransferase from cultured Coptis japonica cells. Eur J Biochem. 1994;225:125–31.

Robin AY, Giustini C, Grindorge M et al. Crystal structure of norcoclaurine-6-O-methyltransferase, a key rate-limiting step in the synthesis of benzylisoquinoline alkaloids. Plant J. 2016;87:641–53.

Ounaroon A, Decker G, Schmidt J et al. (R,S)-Reticuline 7-O-methyltransferase and (R,S)-norcoclaurine 6-O-methyltransferase of Papaver somniferum - cDNA cloning and characterization of methyl transfer enzymes of alkaloid biosynthesis in opium poppy. Plant J. 2003;36:808–19.

Ziegler J, Diaz-Chávez ML, Kramell R et al. Comparative macroarray analysis of morphine containing Papaver somniferum and eight morphine free Papaver species identifies an O-methyltransferase involved in benzylisoquinoline biosynthesis. Planta. 2005;222:458–71.

Morishige T, Tsujita T, Yamada Y et al. Molecular characterization of the S- Adenosyl-l-methionine:3′-Hydroxy-N-methylcoclaurine 4′-O- Methyltransferase involved in Isoquinoline alkaloid biosynthesis in Coptis japonica. J Biol Chem. 2000;275:23398–405.

Mizutani M, Ohita D. Diversification of P450 genes during land plant evolution. Annu Rev Plant Biol. 2010;61:291–315.

Weitzel C, Simonsen HT. Cytochrome P450 enzymes involved in the biosynthesis of mono- and sesquiterpenes. Phytochem Rev. 2015;14:7–24.

Gou M, Ren X, Martin DW et al. The scaffold proteins of lignin biosynthetic cytochrome P450 enzymes. Nat Plants. 2018;4:299–310.

Schoenbohm C, Martens S, Eder C et al. Identification of the Arabidopsis thaliana flavonoid 3′-hydroxylase gene and functional expression of the encoded P450 enzyme. Biol Chem. 2000;381:749–53.

Mizutani M, Sato F. Unusual P450 reactions in plant secondary metabolism. Arch Biochem Biophys. 2011;507:194–203.

Ikezawa N, Iwasa K, Sato F. Molecular cloning and characterization of CYP80G2, a cytochrome P450 that catalyzes an intramolecular C=C phenol coupling of (S)-reticuline in magnoflorine biosynthesis, from cultured Coptis japonica cells. J Biol Chem. 2008;283:8810–21.

Facchini FJ, Hagel JM, Liscoome DK et al. Opium poppy: blueprint for an alkaloid factory. Phytochem Rev. 2007;6:97–124.

Hagel JM, Facchini FJ. Benzylisoquinoline alkaloid metabolism: a century of discovery and a brave new world. Plant Cell Physiol. 2013;54:647–72.
105. Altschul SF, Gish W, Miller W et al. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
106. Marchler-Bauer A, Lu S, Anderson JB et al. CDD: a conserved domain database for the functional annotation of proteins. Nucleic Acids Res. 2011;39:D225–9.
107. Koonin EV, Fedorova ND, Jackson JD et al. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol. 2004;5:R7.
108. Kanehisa M. KEGG: Kyoto Encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28:27–30.
109. Boeckmann B. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 2003;31:365–70.
110. Nguyen LT, Schmidt HA, Haeseler A et al. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32:268–74.
111. Katoh K, Asimenos G, Toh H. Multiple alignment of DNA sequences with MAFFT. 2009;39–64.
112. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 2006;34:W609–12.
113. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol. 2007;56:564–77.
114. Kalyaanamoorthy S, Minh BQ, Wong TKF et al. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017;14:587–9.
115. Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. Bioinformatics. 1997;13:555–6.
116. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20:238.
117. Mi H, Muruganujan A, Huang X et al. Protocol update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). Nat Protoc. 2019;14:703–21.
118. Yu G, Wang L, Han Y et al. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16:284–7.
119. Han MV, Thomas GWC, Lugo-Martinez J et al. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. Mol Biol Evol. 2013;30:1987–97.
120. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12:59–60.
121. Wang Y, Tang H, Debarry JD et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 2012;40:e49–9.
122. Zwaenepoel A, Van de Peer Y. Wgd—simple command line tools for the analysis of ancient whole-genome duplications. Bioinformatics. 2019;35:2153–5.
123. Nelson DR, Koymans L, Kamataki T et al. P450 superfamily: update on new sequences, gene mapping, accession numbers and nomenclature. Pharmacogenetics. 1996;6:1–42.
124. Florea L, Song L, Salzberg SL. Thousands of exon skipping events differentiate among splicing patterns in sixteen human tissues. F1000Research. 2013;2:188.
125. Smith CA, Want EJ, O’Maille G et al. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal Chem. 2006;78:779–87.
126. Chen C, Chen H, Zhang Y et al. TBtools: an integrative toolkit developed for interactive analyses of big biological data. Mol Plant. 2020;13:1194–202.1