The first high-quality chromosomal genome assembly of a medicinal and edible plant *Arctium lappa*

Yanyun Yang[1] | Shengnan Li[1] | Yanping Xing[1] | Zhongren Zhang[2] | Tao Liu[3] | Wuliji Ao[4] | Guihua Bao[4] | Zhilai Zhan[5] | Rong Zhao[1] | Tingting Zhang[1] | Dachuan Zhang[1] | Yueyue Song[1] | Che Bian[1] | Liang Xu[1] | Tingguo Kang[1]

[1]School of Pharmacy, Liaoning University of Traditional Chinese Medicine, Dalian, China.

[2]Novogene Bioinformatics Institute, Beijing, China.

[3]School of Pharmacy, China Medical University, Shenyang, China.

[4]School of Mongol Medicine, Inner Mongolia University for Nationalities, Tongliao, China.

[5]Traditional Chinese Medicine Resource Center, Chinese Academy of Traditional Chinese Medicine, Beijing, China

Correspondence: Liang Xu (Email: 861364054@qq.com) or Tingguo Kang (Email: kangtingguo@163.com) School of Pharmacy, Liaoning University of Traditional Chinese Medicine, Dalian 116600, China

These authors contributed equally: Yanyun Yang, Shengnan Li, Yanping Xing and Zhongren Zhang

Funding information

National Natural Science Foundation of China, Grant/Award Number: 81874338 and 81773852; Major Expenditure Increase and Reduction Project at the Central Level “Capacity Building for Sustainable Utilization of Precious Traditional Chinese Medicine Resources”, Grant/Award Number: 2060302; Mongolian Medicine R & D National Local Joint Engineering Research Center Open Fund Project Budget Funding, Grant/Award Number: MDK2019047; 2019 Liaoning Provincial Department of Education Scientific Research

Project, Grant/Award Number: L201942; National Key Research and Development Plan, Grant/Award Number:2018YFC1708200; Major Special Fund for Science and Technology of Inner Mongolia Autonomous Region, Grant/Award Number: 2019ZD004.

Abstract

Arctium lappa has a long medicinal and edible history with great economic importance. We combined Illumina and PacBio sequences to generate the first high-quality chromosome-level draft genome of *A. lappa*. The assembled genome is approximately 1.79 Gb with a N50 contig size of 6.88 Mb. Approximately 1.70 Gb (95.4%) of the contig sequences were anchored onto 18 chromosomes using Hi-C data; the scaffold N50 was improved to be 91.64 Mb. Furthermore, we obtained 1.12 Gb (68.46%) of repetitive sequences and 32,771 protein-coding genes; 616 positively selected candidate genes were identified. Additionally, we compared the transcriptomes of *A. lappa* roots at three different developmental stages and identified 8,943 differentially expressed genes (DEGs) in these tissues. Among candidate genes related to lignan biosynthesis, the following were found to be highly correlated with the accumulation of arctiin: 4-coumarate-CoA ligase (4CL), dirigent protein (DIR), and hydroxycinnamoyl transferase (HCT). These data can be utilized to identify genes related to *A. lappa* quality or provide a basis for molecular identification and comparative genomics among related species.

KEYWORDS

Arctium lappa, genome, arctiin, lignan

1 | INTRODUCTION

Arctium lappa is a biennial Asteraceae herb that is found all over the world (Chan et al.,

2011). According to Flora of China, there are about 11 species of *Arctium*, of which *A. lappa* is the most used therapeutic species. Nearly 1500 years ago the dried and mature fruit of *A. lappa*, a traditional Chinese medicine, was recorded for the first time in Ming Yi Bie Lu. It is described in Chinese Pharmacopoeia as the main treatment for the common cold of wind-heat, cough with sputum, swelling and pain of the throat, measles, rubella, and mumps. The roots and stems of *A. lappa* are also used medicinally in the Compendium of Materia Medica. Aside from its long-standing medicinal history *A. lappa* also has nutritive value; it is known as “Oriental Ginseng” in Japan and the leaves and stems of the plant can be eaten raw or stewed. In many countries the plant is considered a healthy vegetable and thought to prevent disease (Kang et al., 2013); it contains cellulose, protein, calcium, phosphorus, iron, etc. At present, there are several types of functional foods derived from *A. lappa* on the market, such as *A. lappa* tea, *A. lappa* drinks, and *A. lappa* cans. In addition to its high medicinal and nutritive value, *A. lappa* has also become a popular plant in academic research.

Studies investigating the biological activity of *A. lappa* have provided evidence of anticancer, anti-inflammatory, antibacterial, antiviral, and antioxidant properties (Liu et al., 2014, Yang et al., 2015, Wang et al., 2019). A variety of compounds have been isolated from *A. lappa*, including lignans, fatty acids, phytosterols, polysaccharides, terpenoids, and phenolic acid (Wang et al., 2019, Xu et al., 2006). Pharmacological studies on *A. lappa* has mainly focused on two dibenzylbutyrolactone lignans arctigenin ($C_{21}H_{24}O_6$) and arctiin ($C_{27}H_{34}O_{11}$). Arctigenin was first identified in *A. lappa*, and arctiin is a chemical quality marker for the quality of Arctii Fructus (Gao et al., 2018, Kang et al., 2019). Arctigenin from the extract of the fruits of *A. lappa* and *Forsythia suspensa* can inhibit the proliferation of

HepG2 cells and inhibit autophagy (Okubo et al., 2020). According to the Chinese Pharmacopoeia, The content of arctiin in dried fruits of *A. lappa* should be equal to or greater than 5.0% (PPRC, 2020). Zhou et al. showed that arctiin isolated from *A. lappa* protects mice from acute lung injury (ALI) induced by lipopolysaccharide (LPS) (Zhou et al., 2018). However, research into the synthetic pathways of arctiin and arctigenin remains unclear.

Genomic analysis has been reported in a variety of medicinal plants such as *Platycodon grandiflorus*, *Isatis indigotica*, etc. (Kang et al., 2020, Kim et al., 2020). Genomic background is an important way to study the metabolic processes of medicinal plants. Past research into the genetics of *A. lappa* focused on the mitochondrial or chloroplast genomes (Xing et al., 2019, Zhang et al., 2020). In this study, the first draft of the *A. lappa* genome (Figure 1) was generated using Illumina and PacBio sequencing. The generated genomic sequence data provides knowledge into the genetics of *A. lappa* and provide additional resources for investigating the function of key genes in various metabolic processes.

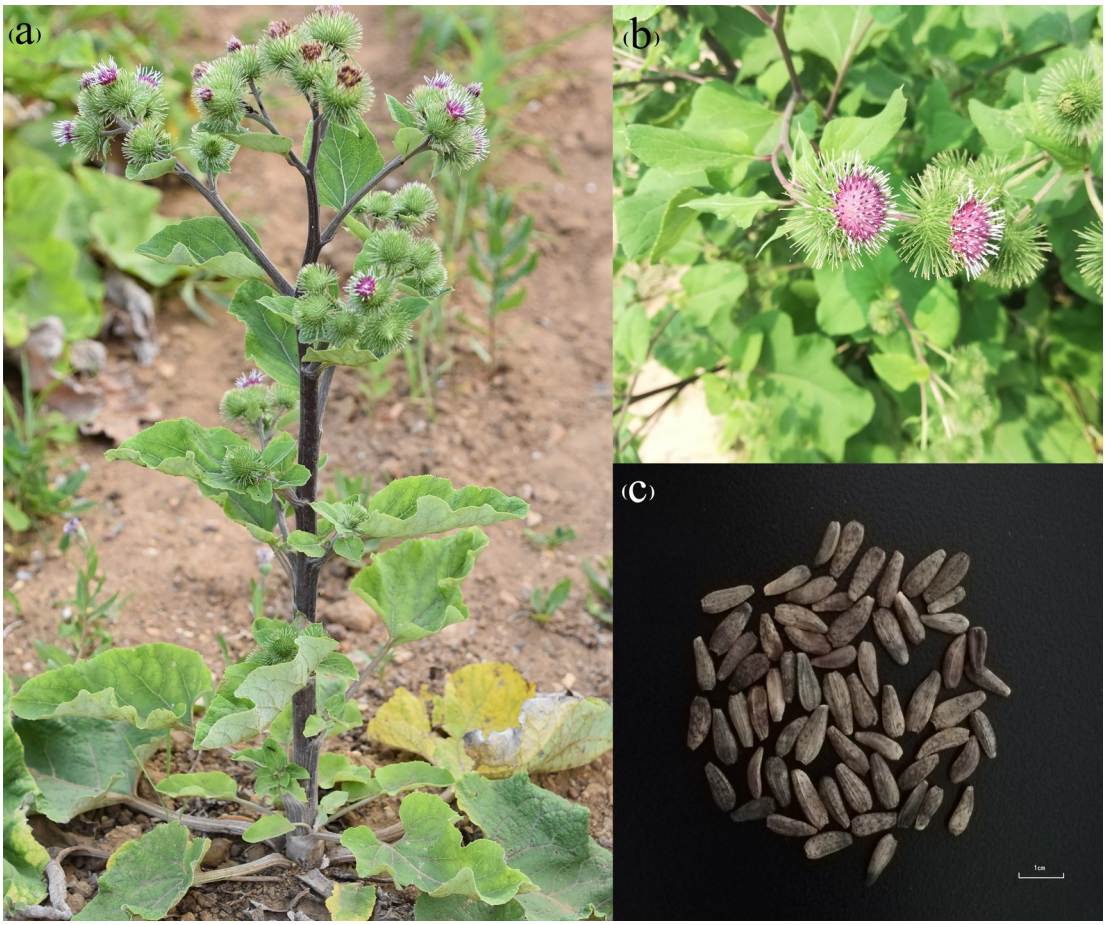


FIGURE 1 Morphological characteristics of *A. lappa.* (a) Plant. (b) Flower. (c) Fruit

2 | MATERIALS AND METHODS

2.1 | Sample collection, Illumina library preparation and sequencing

Plant material was collected from an *A. lappa* plant grown in a field at Liaoning University of Traditional Chinese Medicine (N39°03′35″ , E121°52′12″), China. Young leave tissue of *A. lappa* was collected and genomic DNA was extracted using DNAsecure Plant Kit (TIANGEN, China). Based on the manufacturer's instructions (Illumina, USA), the library construction kit was used to construct a sequence library with an insert size of 350 bp and then sequenced using the Illumina HiSeq X Ten platform, and 144.06 Gb raw reads were obtained. Finally, we constructed a 20-kb single-molecule real-time DNA sequencing library

and sequenced on the Pacbio Sequel platform (Pacific Biosciences, USA), obtaining about 189.11 Gb PacBio data (Table S1).

We extracted DNA from young leaves of the same *A. lappa* plant to construct a Hi-C library. The *A. lappa* leaf cells were lysed and the extracted chromatin was fixed with formaldehyde and digested with HindIII endonuclease. The DNA molecules were released from crosslinking by removing the proteins with protease; the purified DNA was then cut into 350-bp fragments and ligated to sequencing adaptors (Yaffe et al., 2011). The fragments labelled with biotin were collected with streptavidin beads. The PCR-enriched libraries were sequenced using an Illumina HiSeq X Ten instrument, generating approximately 244.91 Gb of raw data (Table S1).

Total RNA of fruit, perianth, stem, petiole, involucre, leaf, root, and stalk from the same *A. lappa* plant was extracted using an RNAprep Pure Plant Kit (TIANGEN, China). A cDNA library was constructed using the NEBNext Ultra RNA Library Prep Kit for Illumina (New England Biolabs, USA) and sequenced on an Illumina Hiseq X Ten platform.

2.2 | Estimation of genome size and genome assembly

K-mer (k = 17) statistics (the modified Lander-Waterman algorithm) was used to estimate the size of the *A. lappa* genome (Liu et al., 2013). Long reads obtained using PacBio SMRT sequencing were assembled *de novo* using FALCON (length_cutoff_pr = 4000, max_diff = 100, max_cov = 100), and then polished using Quiver and error-corrected with Pilon (Chin et al., 2016, Walker et al., 2014). Then, the Hi-C sequencing data were compared with the generated scaffolds using BWA-mem, and the contig sequences were anchored onto the 18 chromosomes of *A. lappa* with LACHESIS (Simão et al., 2017).

2.3 | Annotation of genome sequences

TEs of the *A. lappa* genome were identified by combining *de novo* and homology-based methods. We used RepeatModeler, LTR_FINDER, and RepeatScout to build a *de novo* repeat library, then used RepeatMasker v.4.0.5 and RepeatProteinMask against the Repbase TE library and the TE protein database, respectively (Tarailo-Graovac et al., 2009, Xu et al., 2007, Price et al., 2005). We also used the software Tandem Repeats Finder (TRF) to identify tandem repeats (Benson et al., 1999).

Protein-coding genes of the *A. lappa* genome were predicted by combining the following three methods: homology-based prediction, *de novo* prediction, and transcriptome-based prediction. TBLASTN was used to compare homologous protein sequences from 4 plant genomes (*Artemisia annua*, *Helianthus annuus*, *Chrysanthemum nankingense*, and *Lactuca sativa*) downloaded from the Ensembl Plants and NCBI to the *A. lappa* genome assembly, and a E-value cut-off of 1e-5 was used (Gertz et al., 2006). Solar software was used to conjoin the BLAST hits (Homo-set), and then GeneWise was used to predict the exact gene structure of the corresponding genomic region in each BLAST hit (Altschul et al., 1990, Birney, et al., 2020). TopHat v.2.0.8 and Cufflinks v.2.1.1 were used to map transcriptome data to assemblies (Kim et al., 2013, Ghosh, et al., 2016). We used Trinity to assemble RNA-seq sequences and then create pseudo-unigenes, which were also located on the assembly, and PASA was used to predict gene models (Campbell et al., 2006). Augustus v.2.5.5, GENSCAN v.1.0, GlimmerHMM v.3.0.2, geneid, and SNAP were used to predict coding regions in the repeat-masked genome. Gene model evidence was combined into a non-redundant set of gene structures using EVidenceModeler (EVM) (Haas et al., 2008).

BLASTP (E-value < 1e-04) was used to perform functional annotation of protein-coding genes with the SwissProt and NR databases (Altschul et al., 1997). Protein domains were annotated by searching against the InterPro and Pfam databases using the InterProScan (v.4.8) and HMMER (v. 3.1) (Finn et al., 2017, El-Gebali et al., 2019, Zdobnov et al., 2001, Finn et al., 2015). GO terms of the genes were obtained from the corresponding InterPro or Pfam entry (Ashburner et al., 2000). The KEGG database obtained through BLAST specifies the pathways that may involve genes, with an E-value cut-off of 1e-04 (Kanehisa et al., 2004).

Noncoding RNA, tRNA, miRNA, snRNA, and rRNA fragments were respectively predicted using tRNAscan-SE software, INFERNAL software, Rfam database (version 9.1), and comparing with the rRNA sequences using BLASTN with an E-value cut-off of 1e-10 (Lowe et al., 1997, Nawrocki et al., 2009, Griffithsjones et al., 2005).

2.4 | Gene family cluster, divergence time estimation and WGD

Protein-coding gene sequences from *A. lappa* and ten other plant genome sequences of *A. annua*, *Coffea canephora*, *Cynara cardunculus*, *C. nankingense*, *Daucus carota*, *H. annuu*, *O. sativa*, *Solanum lycopersicum*, *Vitis vinifera*, and *Zea mays* were used for gene family clustering. When there were multiple transcripts representing a gene, only the longest transcript in the coding region was used for further analysis; secondly, genes encoding proteins with less than 30 amino acids were removed. We obtained the corresponding protein sequences of all species through BLASP (E-value < 1E−5). Protein sequences of all species were clustered into paralogous and orthologous using OrthoMCL and the inflation parameter was set as 1.5 (Li et al., 2003).

Following the above analysis, we used MUSCLE to compare the 895 single-copy gene

protein sequences obtained by gene family clustering, and all the comparison results formed a super comparison matrix (Edgar et al., 2004). Then, we constructed the phylogenetic tree of 11 species using RAxML (Stamatakis et al., 2014); the maximum likelihood method was used with the bootstrap value = 100. Finally, the MCMCtree program of PAML was used to calculate the divergence time based on the constructed phylogenetic tree (burn-in = 35,000,000; sample number = 1,000,000; sample frequency = 50) (Yang et al., 2007).

Using BLASTP (E value < 1E-5), the protein sequences of *A. lappa*, *C. cardunculus*, *C. nankingense*, and *H. annuu* were searched against themselves for homogeneity blocks. MCScanX was then used to determine syntenic blocks, calculate the 4DTv (fourfold degenerate sites) for syntenic segments, and plot the distribution of 4DTv values (Wang et al., 2012).

2.5 | Expansion and contraction of gene families

Gene family expansion and contraction analysis was performed (p-value = 0.05) according to the results of clustering analysis of gene family, and CAFÉ program was used to filter out gene families with abnormal gene numbers in individual species (Han et al., 2013). Probabilistic graphical model (PGM) was used to calculate the probability of gene family size change.

2.6 | Positively selected genes in *A. lappa*

Multiple sequence alignments were performed on the protein sequences of single-copy genes of *C. cardunculus*, *C. nankingense*, and *H. annuus* being selected for analysis using MUSCLE; the alignment results were used as templates to generate multiple sequence alignment results corresponding to the coding sequence (CDS) (Edgar et al., 2004). For each

gene family, the branch-site model in PAML was used to test whether the gene family was positively selected in the foreground branch (*A. lappa*), and the LRT was used to determine whether there is a positive selection (Yang et al., 2007). The P-value was calculated using the χ^2 statistic, and multiple tests were corrected according to the false discovery rate (FDR) method.

2.7 | Comparative transcriptome analysis of roots at different developmental stages

DESeq2 was used for normalizing gene expression (BaseMean) in each sample and identifying DEGs for each compared group using "P-adj (adjusted p value) < 0.05" as the threshold (Love et al., 2014). GO enrichment analysis of DEGs was implemented using the GOseq R package, in which gene length bias was corrected. GO terms with corrected P-value less than 0.05 were considered significantly enriched by DEGs. We used KOBAS software to test the statistical enrichment of DEGs in KEGG pathways. Pathways with q-value < 0.05 were considered as significantly enriched.

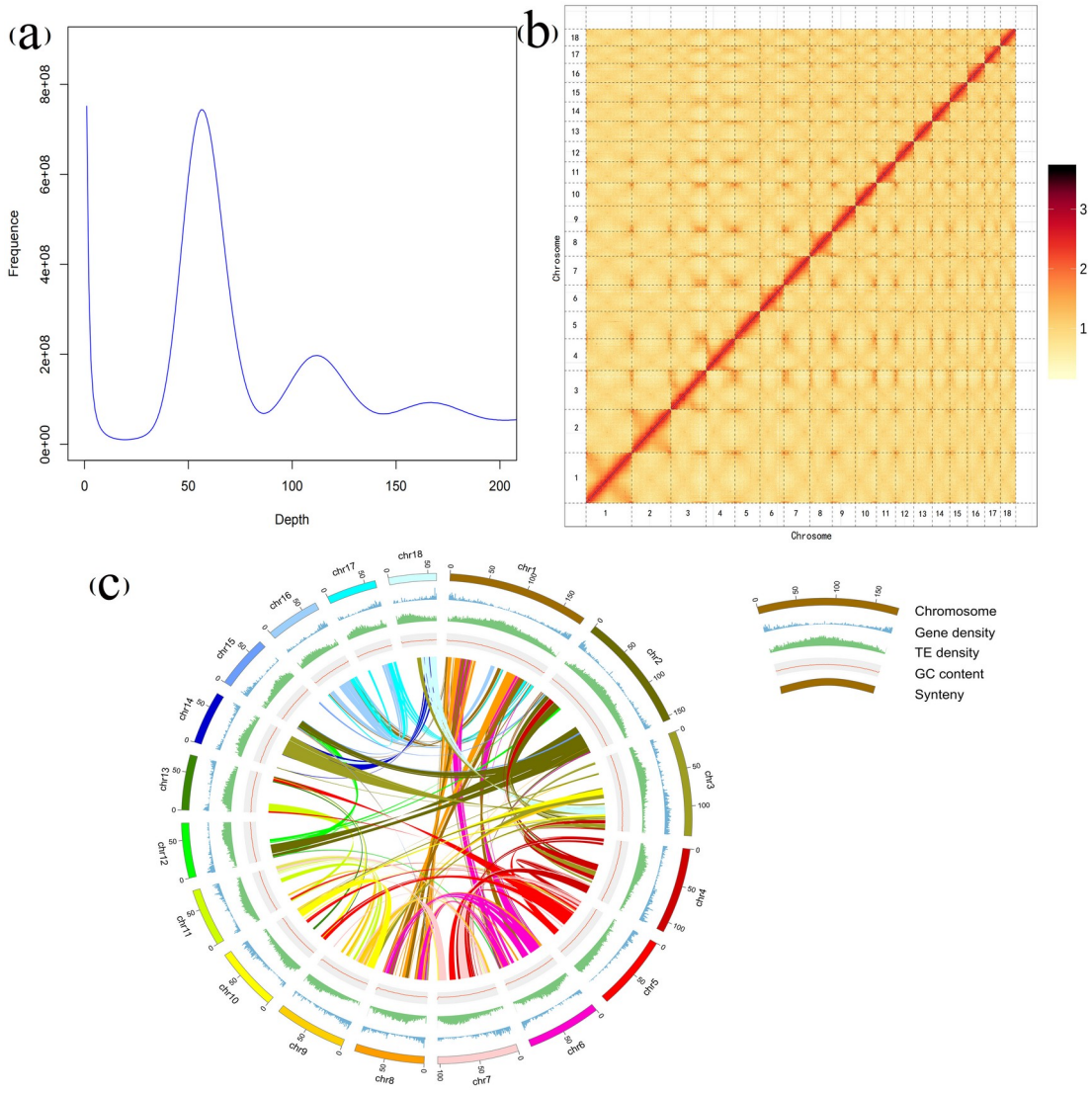
2.8 | Metabolism-related genes of *A. lappa*

Raw RNA reads were filtered and trimmed to yield clean reads and these high-quality reads were mapped to the draft reference genome using Hisat2 with default parameters. FeatureCounts was used to count the number of reads mapped to each gene (Mortazavi et al., 2008, Liao et al., 2014). Then the FPKM of each gene was calculated based on the length of the gene and reads count mapped to this gene. Alignment of genomic protein files of *A. lappa* and other Asteraceae plants using Arabidopsis protein sequences with BLAST (E-value < 1E−5) and protein (identity ≥ 50%, align_ratio ≥ 50%) sequences were extracted. Pfam_scan was used to search for proteins that contain the corresponding domains.

3 | RESULTS

3.1 | Genome size estimation and assembly

Genomic DNA was extracted from *A. lappa* and sequenced using an Illumina HiSeq platform. We obtained 144.06 Gb of 350-bp paired-end (PE) reads that were used for k-mer analysis. Analysis was based on the peak value (depth=56) from the depth and k-mer number frequency distribution curve (Figure 2a), as well as the total number of k-mers (102,732,089,104) (Table S2). The heterozygosity rate of the *A. lappa* genome was estimated to be 0.14% and the genome size was about 1,821.08 Mb. We used PacBio SMRT sequencing and Hi-C sequencing to assemble the genome of *A. lappa* (Chin et al., 2013). The assembled genome size was 1.79 Gb (Table S3) with a contig N50 = 6.88 Mb and a scaffold N50 = 91.64 Mb, respectively (Table 1). There was 1.70 Gb (95.4%) of contig sequences that were anchored onto 18 chromosomes of *A. lappa* using LACHESIS (Figure 2b, 2c, Table S4).



214

215**FIGURE 2** *A. lappa* genome assembly. (a) Sequencing depth and k-mer number frequency

216distribution curve of *A. lappa* (horizontal axis is the depth of k-mer, and vertical axis is the

217number of k-mers corresponding to the depth). (b) Hi-C contact map data analysis (each

218group represents an individual chromosome). (c) Distribution of *A. lappa* genomic features.

219 **Table 1 Summary of the final genome assembly of *A. lappa***

Sample ID	Length		Number	
	Contig (bp)	Scaffold (bp)	Contig	Scaffold
Total	1,786,740,615	1,786,776,515	2,293	1,934
Max	22,631,658	180,240,226	-	-

Number (≥2 kb)	-	-	2,293	1,934
N50	6,883,471	91,637,793	82	8
N60	5,381,283	82,935,164	112	10
N70	4,043,964	73,314,808	150	12
N80	3,129,478	70,146,235	201	14
N90	1,675,446	63,474,660	277	17

3.2 | Quality evaluation for genome assembly

Evaluation using CEGMA (Core Eukaryotic Genes Mapping Approach) showed that 95.56% of conserved genes were assembled (Table S5) (Parra et al., 2007). Furthermore, assessment using BUSCO (Benchmarking Universal Single-Copy Orthologs) found that among 1,440 orthologous single-copy genes, 89.7% of them were assembled from the *A. lappa* genome (Table S6) (Simão et al., 2017). BWA-MEN was used to map high-quality reads from short-insert-size PE libraries to the genome assembly (Li et al., 2014). SAMtools was used to calculate sequencing depth distribution of each position to evaluate the integrity of genome assembly (Li et al., 2009). The rate that could be mapped to the assembly and the coverage rate of all short reads were about 99.66% and 99.13% respectively (Table S7). Overall, we are confident our *A. lappa* genome assembly is of high quality and coverage.

3.3 | Genome annotation

Repetitive sequence content accounted for 68.46% of the *A. lappa* genome, the largest amount of which was long terminal repeat retrotransposons (62.24%). DNA, LINE, and SINE classes repeat elements respectively accounted for 3.2%, 0.85%, and 0.03% of the genome (Table S8). From the assembled *A. lappa* genome there were 32,771 predicted genes. The average transcript size was 1,158.31 bp, the average length of exons was 229.64 bp, the average length of introns was 898.2 bp, and the average number of exons per gene was 5.04

238(Table S9).

239 We also annotated noncoding RNA genes in the assembled genome. Ultimately, we 240predicted 1,648 transfer RNA (tRNA) genes, 1,740 microRNA (miRNA) genes, 2,751 small 241nuclear RNA (snRNA) genes, and 796 ribosomal RNA (rRNA) genes in the *A. lappa* genome 242(Table S10). Accounting for 98.90% of all genes in the *A. lappa* genome, 32,425 of the 243protein-coding genes were predicted to be functional. These protein-coding genes were 244further analyzed using NR (29,436, 89.8%), Swiss-Prot (24,062, 73.4%), Kyoto Encyclopedia 245of Genes and Genomes (KEGG; 22,525, 68.7%), and gene ontology (GO; 29,646, 90.5%) 246(Table S11).

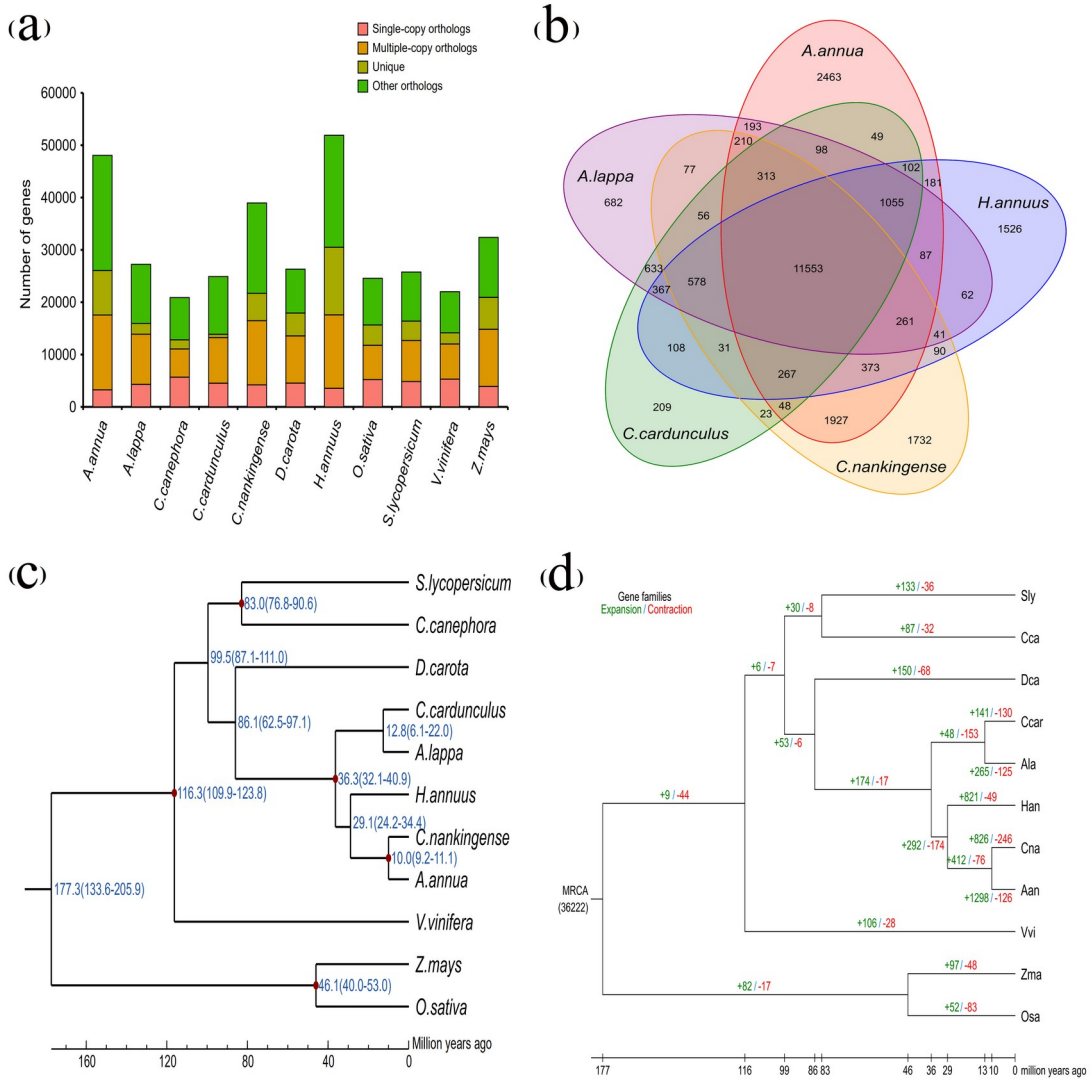
247**3.4 | Gene family cluster, divergence time estimation and whole-genome duplication 248(WGD)**

249All protein-coding genes from 11 sequenced genomes (see Section “Materials and methods”) 250were clustered into 36,240 gene families (two or more members), including 895 single-copy 251orthologs (Figure 3a). Among gene families of *A. lappa*, *A. annua*, *C. cardunculus*, *C. 252nankingense*, and *H. annuus*, 682 were unique to *A. lappa* (Figure 3b). These *A. lappa*-253specific gene families were annotated in the KEGG database, and their functional terms 254mainly include vitamin B6 metabolism, sesquiterpenoid, and triterpenoid biosynthesis (Table 255S12). These *A. lappa*-specific gene families were enriched in GO terms of terpene synthase 256activity, and cellulose synthase activity, etc. (Table S13). Moreover, many of these genes may 257play roles in the formation of the cell wall in *A. lappa* as they are involved in cellulose 258synthase (UDP-forming) activities or cellulose metabolic processes.

259 According to the phylogenetic tree, the result of inferring the divergence time was as

follows: calibration times of divergence between *Oryza sativa* and *Z. mays* (40-53 million years ago, Mya), *C. canephora* and *S. lycopersicum* (77-91 Mya), *S. lycopersicum* and *V. vinifera* (110-124 Mya), *A. annua* and *C. nankingense* (6-10 Mya), *A. annua* and *C. cardunculus* (32-41 Mya) were obtained from the TimeTree database (Hedges et al., 2006). The divergence time between *A. lappa* and *C. cardunculus* was estimated to be around 83 Mya (Figure 3c).

In *A. lappa*, 265 gene families (2,212 genes) were substantially expanded, and 125 gene families (340 genes) were contracted (Figure 3d). Among the expanded gene families, there are 51 gene families significantly enriched for GO terms (Table S14). In the KEGG pathway, functional categories of expanded gene families mainly included photosynthesis, fatty acid metabolism, and biosynthesis of unsaturated fatty acids, etc. (Table S15). Contraction genes were annotated in ABC transporters, alpha-linolenic acid metabolism, etc. in the KEGG pathway.



274

275**FIGURE 3** The genome evolution of *A. lappa*. (a) The distribution of genes in different 276species. (b) Common and unique gene families among the five species. (c) Estimation of 277divergence time. (d) Expansion and contraction of gene families.

278 Earlier studies have shown that there is whole-genome triplication (WGT-1) in Asterids II 279such as *C. cardunculus* and *H. annuus* (Badouin et al., 2017). We performed WGD analysis in 280*A. lappa*. One peak around 0.27 was observed in the 4DTv values in the *A. lappa* genome 281(Figure 4a). According to the results of inter-specific synteny analysis, it can be seen that the 282*A. lappa* genome duplication event should be WGT-1 that occurred before the diversification

of Asterids II (Figure 4b).

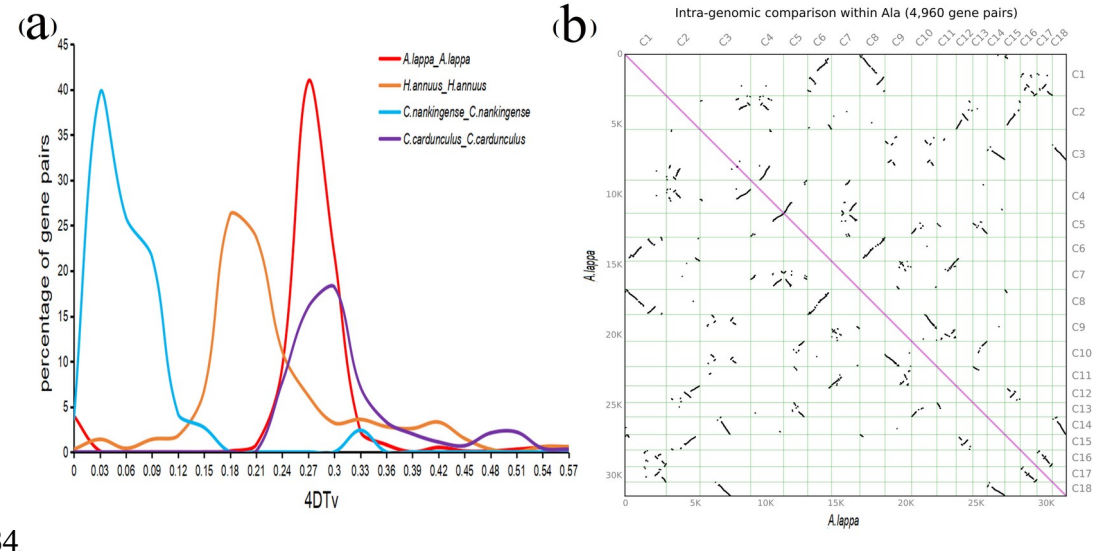


FIGURE 4 Whole-genome duplication analysis of the *A. lappa* genome. (a) 4DTv distribution in *A. lappa* and other representative plant species. (b) Inter-specific synteny analysis of *A. lappa* genome.

3.5 | Positively selected genes in *A. lappa*

A. lappa was used as the foreground branch and *C. nankingense*, *A. annua*, *H. annuus*, and *C. carduncuulus* were used as background branches; 616 candidates (P <0.01, false discovery rate <0.05) of positively selected genes were identified by likelihood ratio test (LRT) in *A. lappa* (Table S16). The GO terms of positively selected genes showed that within the biological process, assignments were mostly enriched in the organic cyclic compound metabolic process, and molecular function was mainly grouped into signal transducer activity and molecular transducer activity. The KEGG pathways of positively selected genes include nicotinate and nicotinamide metabolism and vitamin B6 metabolism, etc.

3.6 | Comparative transcriptome analysis of roots at different developmental

stages

In addition to being used for medicine, the root of *A. lappa* is a raw material for food such as *A. lappa* tea. Here we compared the transcriptome analysis of the roots of seedling (Root0), annual root (Root1), and two-year root (Root3), and determined candidate differentially expressed genes (DEGs) for each tissue. In all pairwise comparisons (Root1 vs Root0, Root2 vs Root0, and Root2 vs Root1), a total of 30,714 DEGs were identified, of which 270 DEGs overlapped (Figure 5a). The result showed that in the three-group difference analysis, Root2 vs Root0 had the most DEGs (5,624), of which 2,754 were up-regulated genes, 2,878 were down-regulated genes. In addition, Root2 vs Root1 had the least DEGs (3,082), of which 1,013 were up-regulated genes, 2,069 were down-regulated genes. There were 5,202 DEGs between Root1 and Root0, of which 2,694 genes were up-regulated and 2,508 genes were down-regulated (Figure 5b). We used cluster Profile software to perform GO function enrichment analysis on DEGs. It was classified into three parts of “cell component”, “molecular function”, and “biological process” using GO to obtain their functional definition (padj less than 0.05 was used as the threshold of significant enrichment). DEGs were mainly enriched in GO terms such as thylakoid (GO:0009579), thylakoid part (GO:0044436), photosystem (GO:0009521), etc. (Figure S1). In order to further study the biological explanation, all DEGs were mapped to the KEGG database. According to KEGG pathway enrichment analysis, the DEGs between Root1 vs Root0, Root2 vs Root0, and Root2 vs Root1 were annotated to 107, 105, and 91 pathways, respectively (Figure S2). There were 13 up-regulated genes in Phenylpropanoid biosynthesis in Root1 vs Root0, 13 up-regulated genes in Phenylpropanoid biosynthesis in Root2 vs Root0 and 5 up-regulated genes in

Phenylpropanoid biosynthesis in Root2 vs Root1.

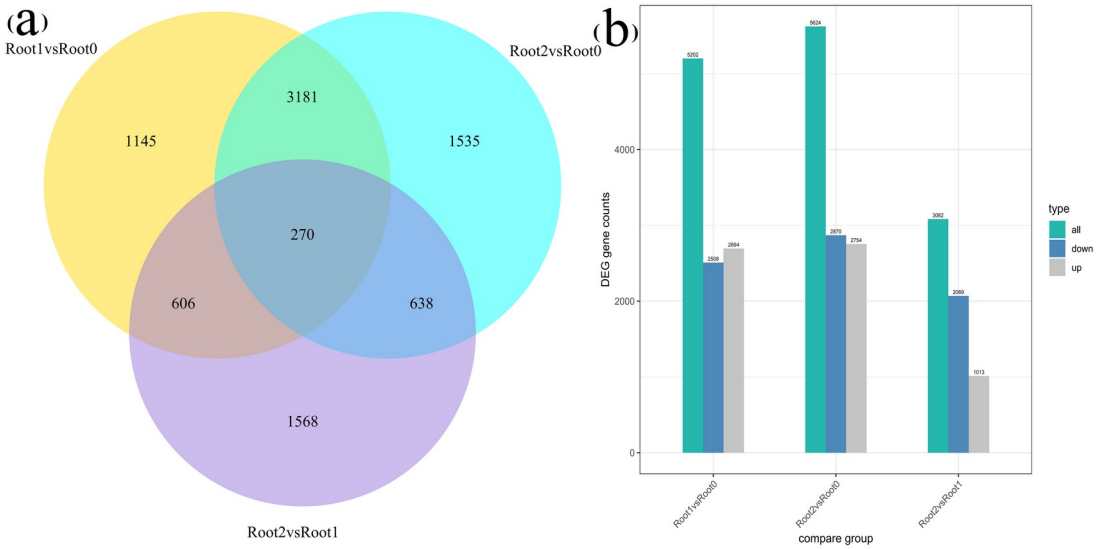


FIGURE 5 Differentially expressed genes (DEGs) at the three stages of root development. (a) Venn diagram of the number of DEGs in stage comparisons: Root1 vs Root0, Root2 vs Root0, and Root2 vs Root1. (b) The number of up- and down-regulated DEGs in three comparisons.

3.7 | Metabolism-related genes of *A. lappa*

Lignans are phenylpropanoid dimers, which can be divided into eight subgroups (Umezawa et al., 2003). Arctiin and arctigenin are dibenzylbutyrolactone lignans with numerous biological effects. The biosynthetic pathway of dibenzylbutyrolactone lignans is well established (Figure 6). The phenylpropanoid pathway is an important secondary metabolic synthesis pathway and the common starting pathway to lignans, lignins, and flavonoids in plants (Suzuki et al., 2007). Phenylpropanoid biosynthesis starts with formation of the phenylalanine. Coniferyl alcohol is a precursor of synthetic lignans derived from phenylalanine (Ralph et al., 2019, Ferrer et al., 2008). Lignans can be composed of only one enantiomer, or both enantiomers. However, lignins are composed of many substructures (Suzuki et al., 2007). Moreover,

matairesinol synthesizes arctigenin by an unidentified enzyme, which is glycosylated to arctiin by an unknown glucosyltransferase (Morimoto et al., 2013). Metabolism-related genes related to lignan biosynthesis pathway were found in *A. lappa* (Table 2). The growth of *A. lappa* fruit can be divided into the following five stages: pre-flower stage (Fruit1), early flowering stage (Fruit2), flowering stage (Fruit3), late flowering stage (Fruit4), and mature stage (Fruit5). We identified the main chemical components of the fruits of *A. lappa* at five different growth stages and identified 31 compounds, including 21 lignans. In addition, the embryonic parenchyma cells or endocarp stone cells of *A. lappa* at five different developmental stages were quantitatively analyzed. There was little accumulation of arctiin or arctigenin in embryonic parenchyma cells or endocarp stone cells of *A. lappa* at the first three stages. Embryonic parenchyma cells and endocarp stone cells produced and accumulated great quantity of arctiin when *A. lappa* fruit was in the late flowering stage and mature stage. The content of arctigenin was the most in endocarp stone cells in the late flowering stage and decreased in the mature stage (Li et al., 2019) (Figure S3). The late flowering stage and the mature stage were key stages for the massive accumulation of arctiin. The up-regulated genes indicated that these genes might be related to the synthesis of arctiin in *A. lappa* (Figure 7). Pearson correlation coefficients ($|cor|>0.7$, $P<0.05$) were calculated based on the arctiin content and related gene expression. The results showed that expression of a 4-coumarate-CoA ligase (4CL) gene (ID: evm.model.000137F.68), a dirigent protein (DIR) gene (evm.model.000171F.169), and a hydroxycinnamoyl transferase (HCT) gene (evm.model.000054F.303) were highly correlated with embryonic parenchyma cells and endocarp stone cells of *A. lappa* (Table S17). Additionally, related metabolic candidate genes

were not annotated in the expansion and contraction of gene families or the positively selected genes.

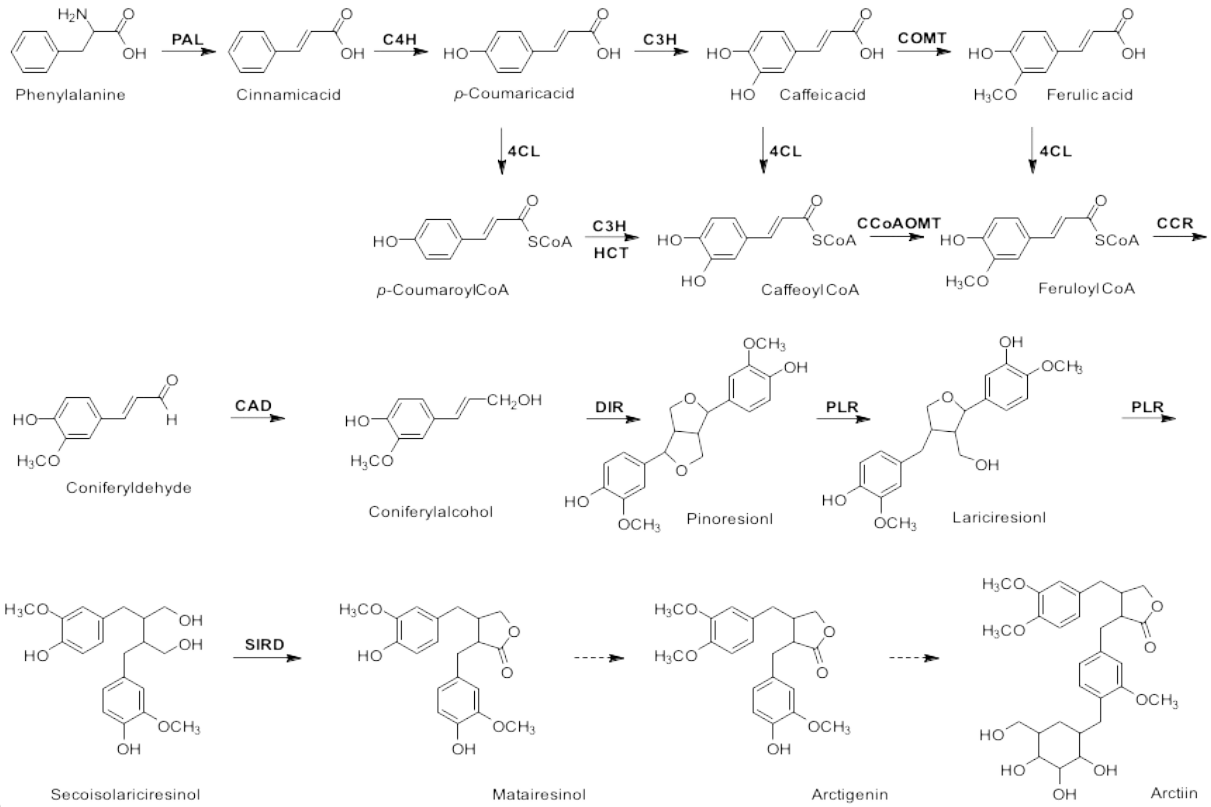
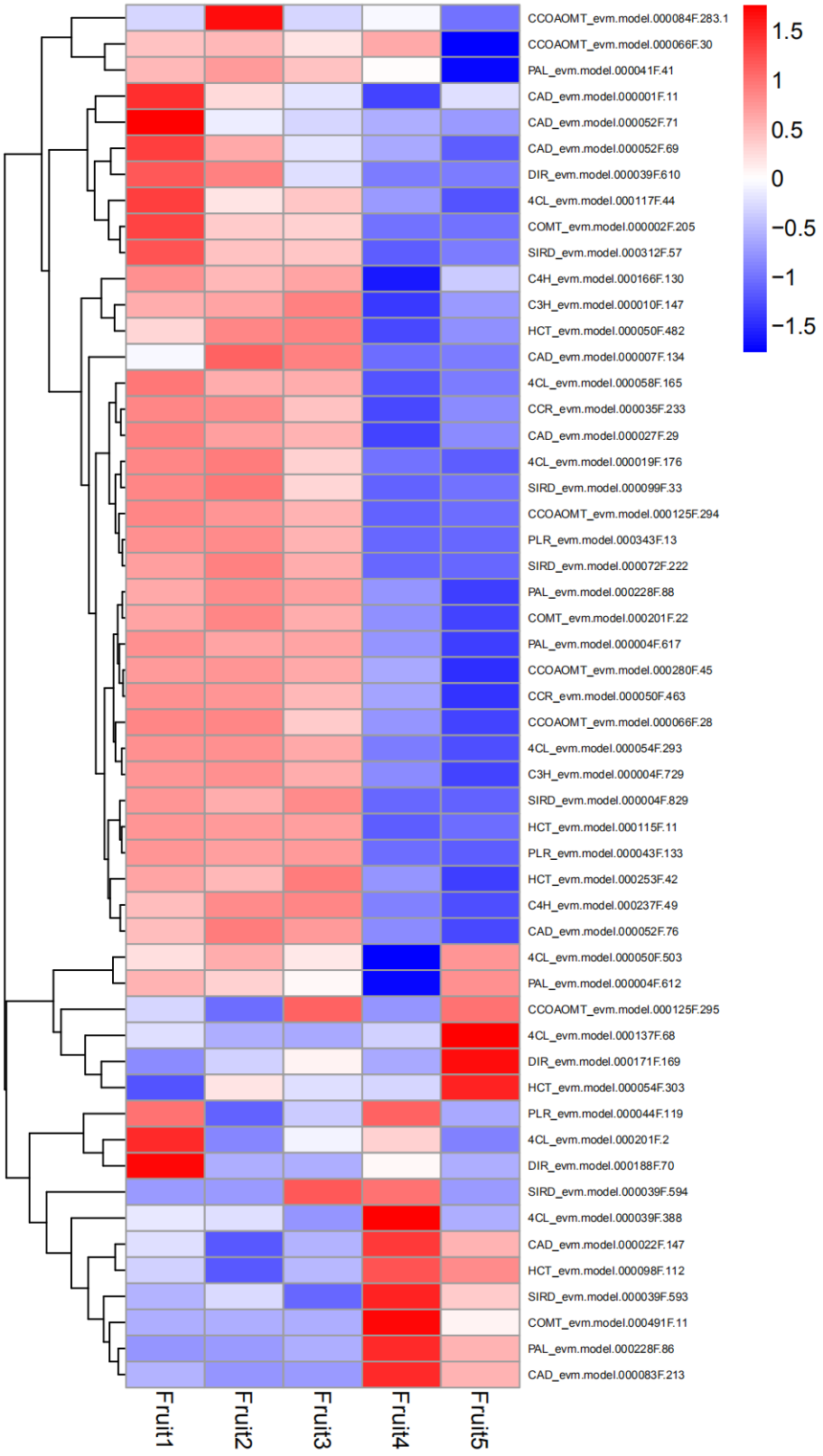


FIGURE 6 Lignan biosynthesis pathways in *A. lappa*. PAL: phenylalanine ammonia lyase; C4H: cinnamate 4-hydroxylase; C3H: coumarate 3-hydroxylase; COMT: catechol-O-methyltransferase; 4CL: 4-coumarate-CoA ligase; CCoAOMT: caffeoyl-CoA O-methyltransferase; CCR: cinnamoyl CoA reductase; CAD: cinnamyl alcohol dehydrogenase; HCT: hydroxycinnamoyl transferase; DIR: dirigent protein; PLR: pinoresinol/ lariciresinol reductase; SIRD: ecoisolariciresinol dehydrogenase.

TABLE 2 Statistics of metabolism-related genes of lignan

Gene name	***A. lappa***	***C. nankingense***	***A. annua***	***H. annuus***	***C. cardunculus***
4-coumarate-CoA ligase (4CL)	8	12	12	15	10
cinnamate-4-hydroxylase (C4H)	2	2	5	3	2
caffeoyl-CoA O-methyltransferase	6	32	14	21	6

(CCOAOMT)					
cinnamoyl-CoA reductase (CCR)	2	5	7	4	2
phenylalanine ammonia-lyase (PAL)	5	5	6	9	5
coumarate3-hydroxylase (C3H)	2	2	4	2	2
cinnamyl alcohol dehydrogenase (CAD)	8	13	7	14	10
caffeic acid/5-hydroxy-ferulic acid O-methytransferase (COMT)	5	13	20	13	2
dirigent protein (DIR)	3	7	7	6	2
hydroxycinnamoyl transferase (HCT)	5	21	13	22	5
pinoresinol/lariciresinol reductase (PLR)	3	9	13	16	7
ecoisolariciresinol dehydrogenase (SIRD)	6	15	12	15	8



368

369**FIGURE 7** Heat map of lignan biosynthesis-related genes in five different stages of *A. lappa*.

In this paper, *A. lappa* was sequenced using an Illumina HiSeq platform, and the genome sequence was assembled using PacBio SMRT sequencing and Hi-C sequencing. The assembled genome had a contig N50 = 6.88 Mb and scaffold N50 = 91.64 Mb. The 1.70 Gb contig sequence was anchored on the 18 chromosomes of *A. lappa*. The assembled *A. lappa* genome size was about 1.79 Gb, larger than those of *P. grandiflorus* (680 Mb) and *I. indigotica* (293.88 Mb), and smaller than that of *H. annuus* (2 Gb) belonging to the same family (Kim et al., 2020, Kang et al., 2020, Badouin et al., 2017). CEGMA, BUSCO, and SAMtools were used to evaluate the genome assembly quality and coverage. These analyses showed the genome assembly was of high quality and coverage.

Repetitive sequences make up a large portion of the *A. lappa* genome, accounting for 68.46% of the genome; long terminal repeat retrotransposons accounted for a large portion of repetitive sequences (62.24%). Transposable elements (TEs) can cause gene recombination or mutation, which are a great value to molecular breeding. There were 32,771 genes annotated in the *A. lappa* genome, more than that of *S. miltiorrhiza* (30,478) but less than that of *H. annuus* (52,232).

Through the analyses of gene family cluster and divergence time estimation, we found the evolutionary state of *A. lappa* and *C. cardunculus* is close. WGD events are common in plants, and can almost be detected in a large number of sequenced plant genomes. In the history of angiosperms, many polyploidization events have been discovered, including WGTs in the common ancestor of core dicots (Ren et al., 2018). The *A. lappa* WGD event identified in this study is WGT-1.

Subsequently, 682 unique genes and 265 gene families (2,212 genes) were obtained by

comparative analysis of *A. lappa* genome. The results are conducive to the location and screening of genes related to specific traits of *A. lappa* in the future. At the same time, a total of 616 positive selection candidate genes were compared, which were the result of the long-term evolution of *A. lappa*.

Comparative transcriptome analysis is a useful and routine method for studying the spatiotemporal patterns of gene expression (Higuchi et al., 1990). DEGs are one of the fundamental reasons for the diversity of cell morphology and function, and they are also the mechanistic basis for plant growth and development along with other physiological and pathological processes. In this study, we first identified three groups of DEGs in roots at different developmental stages, and then identified genes that were up-regulated or down-regulated in each comparison. Some of the up-regulated genes are involved in Phenylpropanoid biosynthesis, which is the upstream pathway of lignin and lignans biosynthesis. With the growth of plants, the roots of *A. lappa* gradually fibrosis. These data will help us understand the development process of the *A. lappa* root and genes that play a key role in *A. lappa* development.

In the analysis of lignan metabolism-related genes in *A. lappa*, in addition to *A. lappa*, other plants of the Asteraceae are also enriched in these genes. However, arctiin components were not reported in these plants, indicating that the same gene played different roles in different plants, and that there might be other factors affecting the synthesis of arctiin. These data provide a reference for the study of the synthetic pathways of arctinin and arctigenin and serve as a valuable resource for *A. lappa* biology. In the future, we will conduct further gene function verification studies on candidate genes.

CODE AVAILABILITY

The execution of this work involved many software tools, whose versions, settings and parameters are described below.

(1) FALCON: version 3.1; **(2) LACHESIS:** version 201701; **(3) BWA:** version 0.7.8, default parameters; **(4) Tandem Repeat Finder:** version 409, default parameters; **(5) RepeatMasker:** version 4.0.5, default parameters; **(6) Repbase:** version 15.02; **(7) RepeatModeler:** version 1.0.11, default parameters; **(8) RepeatScout:** version 1.0.5, default parameters; **(9) LTR_FINDER:** version 1.0.7, default parameters; **(10) Augustus:** version 2.5.5, default parameters; **(11) GENSCAN:** version 1.0, default parameters; **(12) geneid:** version 1.4, default parameters; **(13) GlimmerHMM:** version 3.0.2, default parameters; **(14) SNAP:** version 11-29-2013; **(15) BLAST:** version 2.2.26, default parameters; **(16) GeneWise:** version 2.2.0, default parameters; **(17) TopHat:** version 2.0.8, default parameters; **(18) CEGMA:** version 2.5; **(19) Trinity:** version 2.4.0, default parameters; **(20) PASA:** version 2.3.3, default parameters; **(21) EVidenceModeler:** version 1.1.1, default parameters; **(22) InterPro:** version 5.16, default parameters; **(23) Pfam database:** version 03-30-2016; (24) **InterProScan:** version 4.8, default parameters; **(25) NR database:** version 08-10-2015; (26) **KEGG database:** version 08-31-2015; **(27) SwissProt database:** version 05-24-2016; (28) **HMMER:** version 3.1b1, default parameters; **(29) tRNAscan-SE:** version 1.3.1, default parameters; **(30) BUSCO:** version 3.0.2, Embryophyta Version odb9.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (81874338 and 81773852), Major Expenditure Increase and Reduction Project at the Central Level

"Capacity Building for Sustainable Utilization of Precious Traditional Chinese Medicine Resources" (2060302), Mongolian Medicine R & D National Local Joint Engineering Research Center Open Fund Project Budget Funding (MDK2019047), 2019 Liaoning Provincial Department of Education Scientific Research Project (L201942), National Key Research and Development Plan (2018YFC1708200) and Major Special Fund for Science and Technology of Inner Mongolia Autonomous Region (2019ZD004).

AUTHOR CONTRIBUTIONS

L.X., T.G.K, Y.P.X and R.Z. conceived, initialized and guided the entire project; T.L., W.L.J.A., G.H.B., Y. Y. S, C. B. and Z.L.Z suggested experimental approaches; D.C.Z and T.T.Z collected and prepared samples; Z.R.Z performed genome assembly and data analysis; S.N.L wrote the manuscript with other authors' help; Y.Y.Y revised the manuscript; All authors read and approved the final manuscript.

DATA AVAILABILITY STATEMENT

Raw data of genome sequencing and transcriptome sequencing of *A. lappa* are deposited in the NCBI SRA database under BioProject ID PRJNA598011. Other data, such as the assembled genome sequence, gene structure annotation, predicted CDS and protein sequences, annotation of TEs, tandem repeat sequences, tRNA genes, miRNA genes, snRNA genes, and rRNA genes, are available at FigShare (10.6084/m9.figshare.11590983).

ORCID

Liang Xu http://orcid.org/0000-0002-8080-4927

Competing interests

The authors declare that they have no competing interests.

REFERENCES

460 Altschul, S. F., Gish, W., Miller, W. & Myers, E. W. (1990). Basic local alignment search tool.
461 *Journal of Molecular Biology*, *215*, 403-410. https://doi.org/10.1016/S0022-2836(05)80360-2
462 Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, J., Miller, W. & Lipman, D.
463 J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search
464 programs. *Nucleic Acids Research*, *25*, 3389-3402. https://doi.org/10.1093/nar/25.17.3389
465 Ashburner, M., Ball, C. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K.,
466 Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis,
467 S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., Sherlock, G. (2000). Gene
468 ontology: tool for the unification of biology. *Nature Genetics*, *25*, 25-29.
469 https://doi.org/10.1038/75556
470 Badouin, H., Gouzy, J., Grassa, C. J., Murat, F., Staton, S. E., Cottret, L., Lelandais-Brière,
471 C., Owens, G. L., Carrère, S., Mayjonade, B., Legrand, L., Gill, N., Kane, N. C., Bowers, J.
472 E., Hubner, S., Bellec, A., Bérard, A., Bergès, H., Blanchet, N., ... Langlade, N. B. (2017).
473 The sunflower genome provides insights into oil metabolism, flowering and Asterid
474 evolution. Nature, 546, 148–152. https://doi.org/10.1038/nature22380
475 Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic*
476 *Acids Research*, *27*, 573-580. https://doi.org/10.1093/nar/27.2.573
477 Birney, E. & Durbin, R. (2000). Using GeneWise in the Drosophila annotation experiment.
478 *Genome Research*, *10,* 547-548. https://doi.org/10.1101/gr.10.4.547
479 Campbell, M. A., Haas, B. J., Hamilton, J. P., Mount, S. M. & Buell, C. R. (2006).
480 Comprehensive analysis of alternative splicing in rice and comparative analyses with
481 Arabidopsis. *BMC Genomics*, *7*, 327. https://doi.org/10.1186/1471-2164-7-327
482 Chan, Y. S., Cheng, L. N., Wu, J. H., Chan, E., Kwan, Y. W., Lee, S. M-Y., Leung, G. P-H., Yu,
483 P. H-F. & Chan, S. W. (2011). A review of the pharmacological effects of *Arctium lappa*
484 (burdock). *Inflammopharmacology*, *19*, 245-254. https://doi.org/10.1007/s10787-010-0062-4
485 Chin, C. S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., Clum, A.,
486 Copeland, A., Huddleston, J., Eichler, E. E., Turner, S. W. & Korlach, J. (2013).
487 Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data.
488 *Nature Methods*, *10*, 563-569. https://doi.org/10.1038/nmeth.2474
489 Chin, C. S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., Dunn, C.,

O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., Cramer, G. R., Delledonne, M., Luo, C., Ecker, J. R., Cantu, D., Rank, D. R., Schatz, M. C. (2016). Phased Diploid Genome Assembly with Single Molecule Real-Time Sequencing. *Nature Methods*, *13*, 1050-1054. https://doi.org/10.1038/nmeth.4035

Chinese Pharmacopoeia Commission (2020). Pharmacopoeia of the People's Republic of China. China Medical Science and Technology Press part 1, 72.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*, 1792-1797. https://doi.org/10.1093/nar/gkh340

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J, Salazar, G. A., Smart, A., Sonnhammer, E. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. & Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Research*, *47*, D427-D432. https://doi.org/10.1093/nar/gky995

Ferrer, J. L., Austin, M. B., Stewart, C. & Noel, J. P. (2008). Structure and function of enzymes involved in the biosynthesis of phenylpropanoids. *Plant Physiology and Biochemistry*, *46*, 356-370. https://doi.org/10.1016/j.plaphy.2007.12.009

Finn, R. D., Attwood, T. K., Babbitt, P. C., Bateman, A., Bork, P., Bridge, A. J., Chang, H. Y., Dosztányi, Z., El-Gebali, S., Fraser, M., Gough, J., Haft, D., Holliday, G. L., Huang, H., Huang, X., Letunic, I., Lopez, R., Lu, S., Marchler-Bauer, A. ... Mitchell, A. (2017). InterPro in 2017—beyond protein family and domain annotations. *Nucleic Acids Research*, *45*, D190-D199. https://doi.org/10.1093/nar/gkw1107

Finn, R. D., Clements, J., Arndt, W., Miller, B. L., Wheeler, T. J., Schreiber, F. S., Bateman, A. & Eddy, S. R. (2015). HMMER web server: 2015 update. *Nucleic Acids Research*, *43*, 30-38. https://doi.org/10.1093/nar/gkv397

Gao, Q., Yang, M. & Zuo, Z. (2018). Overview of the anti-inflammatory effects, pharmacokinetic properties and clinical efficacies of arctigenin and arctiin from *Arctium lappa* L. *Acta Pharmacologica Sinica*, *39*, 787-801. https://doi.org/10.1038/aps.2018.32

Gertz, E., Yu, Y. K., Agarwala, R., Schäffer, A. A. & Altschul, S. F. (2006). Composition-based statistics and translated nucleotide searches: Improving the TBLASTN module of BLAST. *BMC Biology, 4*, 41. https://doi.org/10.1186/1741-7007-4-41

Ghosh, S. & Chan, C. K. (2016). Analysis of RNA-Seq Data Using TopHat and Cufflinks. *Plant*

Bioinformatics. New York: Academic, p. 339-361. https://doi.org/10.1007/978-1-4939-3167-5215_18

Griffithsjones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S. R. & Bateman, A. (2005). Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Research*, *33*, D121-D124. https://doi.org/10.1093/nar/gki081

Haas, B. J., Salzberg, S, L., Zhu, W., Pertea, M., Allen, J., Orvis, J., White, O., Buell, C. R. & Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology*, *9*, R7. https://doi.org/10.1186/gb-2008-9-1-r7

Han, M. V., Thomas, G. W., Jose, L. M. & Hahn, M. W. (2013). Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Molecular Biology and Evolution*, *30*, 1987-1997. https://doi.org/10.1093/molbev/mst100

Hedges, S. B., Dudley, J. & Kumar, S. (2006). TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics*, *22*, 2971-2972. https://doi.org/10.1093/bioinformatics/btl505

Higuchi, T. (1990). Lignin biochemistry: Biosynthesis and biodegradation. *Wood Science and Technology*, *24*, 23-63. https://doi.org/10.1007/BF00225306

Kanehisa, M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Research*, *32*, D277-D280. https://doi.org/10.1093/nar/gkh063

Kang, T. G. & Dou, D. Q. (2013). Research on Chinese Burdock. Liaoning Science and Technology Publishing House, 332-339.

Kang, T. G., Dou, D. Q. & Xu, L. (2019). Establishment of a Quality Marker (Q-marker) System for Chinese Herbal Medicines using Burdock as an Example. *Phytomedicine*, *54*, 339-346. https://doi.org/10.1016/j.phymed.2018.04.005

Kang, M. H., Wu, H., Yang, Q., Huang, L., Hu, Q. J., Ma, M., Li, Z. Y. & Liu, J. Q. (2020). A chromosome-scale genome assembly of *Isatis indigotica*, an important medicinal plant used in traditional Chinese medicine. *Horticulture Research*, *7*, 18. https://doi.org/10.1038/s41438-020-0240-5

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. & Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene

fusions. *Genome Biology*, *14*, R36. https://doi.org/10.1186/gb-2013-14-4-r36

Kim, J., Kang, S. H., Park, S. G., Yang, T. J., Lee, Y., Kim, O. T., Chung, O., Lee, J., Choi, J. P., Kwon, S. J., Lee, K., Ahn, B. O., Lee, D. J., Yoo, S. I., Shin, I. G., Um, Y., Lee, D. Y., Kim, G. S., Hong, C. P. ... Kim, C. K. (2020). Whole-genome, transcriptome, and methylome analyses provide insights into the evolution of platycoside biosynthesis in *Platycodon grandiflorus*, a medicinal plant. *Horticulture Research*, *7*, 112. https://doi.org/10.1038/s41438-020-0329-x

Li, H. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, *30*, 2843-2851. https://doi.org/10.1093/bioinformatics/btu356

Li, H. Handsaker, H., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*, 2078-2079. https://doi.org/10.1093/bioinformatics/btp352

Li, L., Stoeckert, C. J. & Roos D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research*, *13*, 2178-2189. https://doi.org/10.1101/gr.1224503

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomicfeatures. *Bioinformatics*, *30*, 923-930 . https://doi.org/10.1093/bioinformatics/btt656

Liu, B. H., Shi, Y., Yuan, Y. & Galaxy, Y. (2013). Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *Quantitative Biology*, *35*, 62-67. https://doi.org/10.1016/S0925-4005(96)02015-1

Liu, W., Wang, J., Zhang, Z., Xu, J., Xie, Z., Slavin, M. & Gao, X. (2014). In vitro and in vivo antioxidant activity of a fructan from the roots of *Arctium lappa* L. *International J ournal of Biological Macromolecules*, *65*, 446-453. https://doi.org/10.1016/j.ijbiomac.2014.01.062

Li, S. N., Yang, Y. Y., Xu, L., Chen, S. Y. & Kang, T. G. (2019). Simultaneous determination of three chemical compounds in embryonic parenchyma cells and endocarp stone cells of Fructus Arctii at five different growth stages by UFLC-MS/MS. *Acta Pharm Sin*, *54*, 1265-1270.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*, 550.

https://doi.org/10.1186/s13059-014-0550-8

Lowe, T. M. & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, *25*, 955-964. https://doi.org/10.1093/nar/25.5.955

Morimoto, K. & Satake, H. (2013). Seasonal alteration in amounts of lignans and their glucosides and gene expression of the relevant biosynthetic enzymes in the Forsythia suspense leaf. *Biological & Pharmaceutical Bulletin*, *36*, 1519-1523. https://doi.org/10.1248/bpb.b13-00437

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, *5*, 621-628. https://doi.org/10.1038/nmeth.1226

Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. (2009). Infernal 1.0: inference of RNA alignments. *Bioinformatics*, *25*, 1335-1337 . https://doi.org/10.1093/bioinformatics/btp157

Okubo, S., Ohta, T., Shoyama, Y. & Uto, T. (2020). Arctigenin suppresses cell proliferation via autophagy inhibition in hepatocellular carcinoma cells. *Journal of Natural Medicines*, *74*, 525-532. https://doi.org/10.1007/s11418-020-01396-8

Parra, G., Bradnam, K. & Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, *23*, 1061-1067. https://doi.org/10.1093/bioinformatics/btm071

Price, A. L., Jones, N. C. & Pevzner, P. A. (2005). De novo identification of repeat families in large genomes. *Bioinformatics*, *21*, i351-i358. https://doi.org/10.1093/bioinformatics/bti1018

Ralph, J., Lapierre, C. & Boerjan, Wout. (2019). Lignin structure and its engineering. *Current Protocols in Bioinformatics*, *56*, 240-249. https://doi.org/10.1016/j.copbio.2019.02.019

Ren, R., Wang, H., Guo, C., Zhang, Z., Zeng, L., Chen, Y., Ma, H. & Qi, J. (2018). Wide-Spread Whole Genome Duplications Contribute to Genome Complexity and Species Diversity in Angiosperms. *Molecular Plant*, *11*, 414-428. https://doi.org/10.1016/j.molp.2018.01.002

Simão, F. A., Waterhouse, R. M., Loannidis, P., Kriventseva, E. V. & Zdobnov, E. M. (2017). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*, 3210-3212. https://doi.org/10.1093/bioinformatics/btv351

610Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis
611 of large phylogenies. *Bioinformatics*, *30*, 1312-1313.
612 https://doi.org/10.1093/bioinformatics/btu033
613Suzuki, S. & Umezawa, T. (2007). Biosynthesis of lignans and norlignans. *Journal of Wood*
614 *Science*, *53*, 273-284 . https://doi.org/10.1007/s10086-007-0892-x
615Tarailo-Graovac, M. & Chen, N. (2009). Using RepeatMasker to identify repetitive elements
616 in genomic sequences. *Current Protocols in Bioinformatics*, *25*, 4.10.1-4.10.14.
617 https://doi.org/10.1002/0471250953.bi0410s05
618Umezawa, T. (2003). Diversity in lignan biosynthesis. *Phytochemistry Reviews*, *2*, 371-390.
619 https://doi.org/10.1023/B:PHYT.0000045487.02836.32
620Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A.,
621 Zeng, Q., Wortman, J., Young, S. K. & Earl, A. M. (2014). Pilon: An Integrated Tool for
622 Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *Plos*
623 *One*, *9*, e112963. https://doi.org/10.1371/journal.pone.0112963
624Wang, D. D., Bădărau, A. S., Swamy, M. K., Shaw, S., Maggi, F., Silva, L. E., López, V.,
625 Yeung, A. W. K., Mocan, A. & Atanasov, A. G. (2019). *Arctium* Species Secondary
626 Metabolites Chemodiversity and Bioactivities. *Frontiers in Plant Science*,
627 https://doi.org/10.3389/fpls.2019.00834
628Wang, Y. P., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., Lee, T., Jin, H., Marler, B.,
629 Guo, H., Kissinger, J. C. & Paterson, A. H. (2012). MCScanX: a toolkit for detection and
630 evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research*, *40*, e49 .
631 https://doi.org/10.1093/nar/gkr1293
632Zhang, Y. P., Xu, L., Chen, S. Y., Liang, Y. M., Wang, J. H., Liu, C. S., Liu, T. & Kang, T. G.
633(2019). Comparative analysis of complete chloroplast genomes sequencesof *Arctium lappa*
634and *A. tomentosum*. *Biologia Plantarum*, *63*, 565-574. https://doi.org/10.32615/bp.2019.101
635Xu, Z. H., Zhao, A. H. & Gao, X. F. (2006). Chemical constituents of antihyperglycemic
636 active fraction from *Arctium lappa*. *Chinese Journal of Natural Medicines*, *4*, 444-447.
637Xu, Z. & Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length
638 LTR retrotransposons. *Nucleic Acids Research*, *35*, W265-W268.
639 https://doi.org/10.1093/nar/gkm286

Yaffe, E. & Tanay, A. (2011). Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature Genetics*, *43*, 1059-1065. https://doi.org/10.1038/ng.947

Yang, Y. N., Huang, X. Y., Feng, Z., Jiang, J. S. & Zhang, P. C. (2015). New Butyrolactone Type Lignans from Arctii Fructus and Their Anti-inflammatory Activities. *Journal of Agricultural and Food Chemistry*, *63*, 7958-7966. https://doi.org/10.1021/acs.jafc.5b02838

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, *24*, 1586-1591. https://doi.org/10.1093/molbev/msm088

Zdobnov, E. M. & Apweiler, R. (2001). InterProScan – an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, *17*, 847-848. https://doi.org/10.1093/bioinformatics/17.9.847

Zhang, D. C., Xing, Y. P., Xu, L., Zhao, R., Yang, Y. Y., Zhang, T. T., Li, S. N. & Kang, T. G. (2020). The complete mitochondrial genome of *Arctium lappa* (Campanulales, Asteraceae). *Mitochondrial DNA Part B*, *5*, 1722-1723. https://doi.org/10.1080/23802359.2020.1749153

Zhou, B., Weng, G. H., Huang, Z. X., Liu, T. & Dai, F. Y. (2018). Arctiin Prevents LPS-Induced Acute Lung Injury via Inhibition of PI3K/AKT Signaling Pathway in Mice. *Inflammation*, *41*, 2129-2135. https://doi.org/10.1007/s10753-018-0856-x