Data in Brief 60 (2025) 111530

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Dataset of a *de novo* transcriptome assembly for the leaves and rhizomes of a five-year-old *Atractylodes chinensis*

Yelin Tian[a], Lizhi Ouyang[a,b], Xinyu Li[a,b], Li Xiao[c], Xu Qiao[b], Yixuan Chen[a,b], Tingting Fang[b], Yimian Ma[b,*]

[a] *School of landscape architecture, Beijing University of Agriculture, Beijing 102206, China*
[b] *Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100193, China*
[c] *Institute of Grain Groups, Xinjiang Academy of Agricultural Sciences, Urumqi 830091, China*

ARTICLE INFO

Article history:
Received 6 November 2024
Revised 13 March 2025
Accepted 25 March 2025
Available online 28 March 2025

Dataset link: Dataset of transcriptome assembly of Atractylodes chinensis (Original data)

Keywords:
Differentially expressed genes
Atractylodin
Sesquiterpene
Medicinal plant
Illumina

ABSTRACT

Atractylodes (A.) chinensis (DC.) Koidz. is a traditional Chinese medicinal plant. The rhizome contains its medicinal component, which consists of abundant essential oils. Sesquiterpene and atractylodin are the main active ingredients in these essential oils. On the other hand, the leaves contain less medicinal active ingredients. Thus far, studies on the formation mechanism of the active ingredients, especially atractylodin, are still limited. This study used RNA sequencing to reveal the *de novo* transcriptome of the leaves and rhizomes of a five-year old *A. chinensis* plant with divided leaves. High-throughput sequencing data was acquired using the Illumina NovaSeq X Plus system (Illumina, USA) in PE150 mode. After the data was corrected and filtered, the clean data was used for subsequent analysis. Based on the assembled sequence file, the differentially expressed unigenes between the rhizomes and leaves of *A. chinensis* were analyzed. The assembled unigene file and table including these differentially expressed unigenes was deposited in the "Mendeley Data" database. The raw SRA data was deposited in the National Center of Biotechnology Information (NCBI) Sequence Read Archive (SRA) database.

* Corresponding author.
E-mail address: ymma@implad.ac.cn (Y. Ma).

https://doi.org/10.1016/j.dib.2025.111530

Specifications Table

Subject	Agricultural and Biological Sciences.
Specific subject area	Medicinal Plant Transcriptome Analysis.
Type of data	Table, Figure. Raw, Processed.
Data collection	Leaves and rhizomes of *A. chinensis* (DC.) Koidz. were obtained from a five-year-old plant. RNA was extracted from the leaf and rhizome samples using the RNA Easy Fast Plant Tissue Kit (Tiangen, Beijing, China). The library was constructed using the Ultima Pro RNA Library Prep Kit (Yeasen, Shanghai, China). The insert size of the library was detected using an Agilent 2100 bioanalyzer (Agilent Technologies, CA, USA). Quantitative reverse transcription polymerase chain reaction (qRT-PCR) was used to accurately quantify the effective concentration of the library (above 1.5 nM) to ensure library quality. RNA sequencing data was generated by paired-end sequencing using the Illumina NovaSeq X Plus system. Poor quality raw reads were removed using Fastp to acquire the clean data.
Data source location	Leaves and rhizomes of *A. chinensis* were collected at: • Institute of Medicinal Plant Development • City/Town/Region: No. 151 Malianwa North Road, Haidian District, Beijing • Country: China.
Data accessibility	Repository name: NCBI Sequence Read Archive (SRA) Data identification number: PRJNA1177777 Direct URL to data: https://www.ncbi.nlm.nih.gov/sra/PRJNA1177777 Repository name: Mendeley Data Data identification number: DOI: 10.17632/94hhxg2vtz.1 Direct URL to data: https://data.mendeley.com/datasets/94hhxg2vtz/1
Related research article	None.

1. Value of the Data

- These data provide transcriptome data from the leaves and rhizomes of *A. chinensis*.
- These data can be reused to study differentially expressed genes of *A. chinensis* related to the biosynthesis of active ingredients, such as sesquiterpene and atractylodin.
- These data can be reused to investigate the development processes of leaves and rhizome of *A. chinensis*.

2. Background

Atractylodes (A.) chinensis (DC.) Koidz. is a traditional medicinal plant of the Asteraceae family. The rhizome of *A. chinensis* is one of the most important medicinal materials in China. It has many uses, such as air disinfection, anti-bacterial, anti-viral, and insect killing. It can also be used for epidemic prevention in China. Sesquiterpene and atractylodin are the major active ingredients of *A. chinensis* [1]. Although the cultivation of *A. chinensis* has developed over the years, the content of their active ingredients is always lower than their wildtype. To improve the content of active ingredients of cultivated *A. chinensis*, it is important to reveal the biosynthesis pathways of sesquiterpene and atractylodin. However, details on the functional genes of *A. chinensis* are still limited. In recent years, several studies have focused on the molecular mechanism of *A. chinensis* sesquiterpene biosynthesis and anti-stress properties [2,3]. However, the formation mechanism of atractylodin, a kind of polyacetylene, remains largely unknown. Recently, Liu et al. compared the contents of the active ingredients of *A. chinensis* with different leaf shapes

Table 1
Statistics of the sequencing reads.

Sample	Raw reads	Raw bases	Clean reads	Clean bases	Q20	Q30	Error%	GC%
G1_1	20623760	6.19 Gb	20076646	6.02 Gb	98.60	96.07	0.01	44.47
G1_2	23353804	7.01 Gb	22689345	6.81 Gb	98.61	96.01	0.01	44.42
G1_3	21532028	6.46 Gb	20880865	6.26 Gb	98.62	96.04	0.01	44.26
Y1_1	20052777	6.02 Gb	19306285	5.79 Gb	98.61	96.11	0.01	45.21
Y1_2	19830840	5.95 Gb	19212117	5.76 Gb	98.64	96.18	0.01	45.17
Y1_3	20819528	6.25 Gb	20151958	6.05 Gb	98.65	96.15	0.01	45.05

Table 2
Frequency distribution of splice length.

Type	200-500bp	500bp-1kb	1-2 kb	>2kb	Total number
Number of Transcript	57,282	68,482	57,633	37,533	220,930
Number of Unigene	29,367	27,638	18,239	11,972	87,216

Table 3
BUSCO assessment of the splice quality.

Parameters	Complete (S/D)	Fragmented	Missing	Total BUSCO groups
Transcript.fasta	1394 (616/778)	162	58	1614
Unigene.fasta	1097 (1062/35)	337	180	1614

Note: Complete (S/D) represents the Complete BUSCO (Single copy BUSCO/Duplicated BUSCO).

and discovered that the plant with divided leaves exhibited the highest levels of active ingredients [4]. We also detected a high level of atractylodin from a five-year-old *A. chinensis* plant with divided leaves cultivated in our institute. The leaves and rhizome of this *A. chinensis* were used to extract RNA and high-throughput sequencing on an Illumina NovaSeq X Plus system (Illumina, USA) with a paired-end 150 (PE150) mode. The dataset generated can be used to reveal functional genes involved in the biosynthesis of sesquiterpene and atractylodin, and to reveal key genes involved in the development processes of the divided leaves of *A. chinensis*.

3. Data Description

Leaves and rhizomes were collected from *A. chinensis* plants. Libraries were prepared using the Ultima Pro DNA Library Prep Kit (Yeasen, Shanghai, China), according to the manufacturer's instructions. The raw data was generated using an Illumina NovaSeq X Plus system (Illumina, USA) with a PE150 mode. The raw data was filtered using the fastp (v 0.19.4) software [5]. The sequencing error rate and the GC content distribution were checked, and the resulting clean reads were obtained for subsequent analysis. After data filtration, about 6 Gb of clean bases for each replicate sample were acquired. The data summary table is shown in Table 1. The raw data was deposited as FASTQ files in the National Center of Biotechnology Information (NCBI) Sequence Read Archive (SRA) database under BioProject accession number PRJNA1177777 [6]. The SRA accession numbers for the raw data were SRR31156762 (Y1_1), SRR31156761 (Y1_2), SRR31156760 (Y1_3), SRR31156765 (G1_1), SRR31156764 (G1_2), and SRR31156763 (G1_3). Subsequently, clean reads were assembled as transcripts and clusters using the Trinity (v2.15.1) software and the Corset (v1.09) software [7,8]. The longest sequence in each cluster was considered a unigene, and the length frequency distribution of transcripts and unigenes are shown in Table 2. The Benchmarking Universal Single-Copy Orthologs BUSCO (v5.8.2) software was used to evaluate the splicing quality of Trinity.fasta and unigene.fasta (Table 3) [9]. The results revealed that the number of complete single BUSCO in 1614 BUSCO groups in unigene.fasta was 1062. Next, a comparison of the differentially expressed genes was performed between genes expressed in rhizomes and leaves. Out of the 32,455 unigenes expressed in the rhizomes, 15,314 unigenes exhibited higher expression levels when compared to the leaves, while 17,141 unigenes showed lower expression levels in the rhizomes than in the leaves. In total, 5755 unigenes were enriched in the Kyoto Encyclopaedia of Genes and Genomes (KEGG) categories based on their KEGG Orthology (KO) accession numbers [10]. Of the 5755 unigenes enriched, 1871 unigenes exhibited higher expression levels in rhizomes and were enriched in 117 KEGG pathways, while 3884 unigenes exhibited higher expression levels in leaves and were enriched in 122 KEGG pathways. The unigene file and the differentially expressed unigenes enriched in KEGG pathways were stored in the "Mendeley Data" database as unigene.fasta and G1vsY1.all_KEGGenrich.xls.

The data can be used to identify differentially expression genes involved in the active ingredient biosynthesis pathways, as well as those involved in the development processes of the rhizome and leaves. We particularly focused on the 1871 unigenes that demonstrated increased expression levels in the rhizome, as the rhizome is recognized as the traditional medicinal part of *A. chinensis* and is known to contain a greater concentration of active ingredients when compared to the leaves. The genes enriched in the KEGG pathways related to sesquiterpene and polyacetylene biosynthesis were manually searched. As a result, a total of 116 candidate unigenes were identified, with 52 genes related to terpenoid backbone biosynthesis (ko00900), 32 genes related to sesquiterpenoid and triterpenoid biosynthesis (ko00909), and 32 genes related to biosynthesis of unsaturated fatty acids (ko01040). The functional identification of the genes involved in these pathways are currently in progress. The identified genes will expand our understanding of the biosynthesis process of the active ingredients in *A. chinensis*. These data also can be used to study the molecular mechanism of *A. chinensis* leaves or rhizome development by searching for the specific genes involved in related KEGG pathways.

4. Experimental Design, Materials, and Methods

4.1. Plant materials

Initially, several two-year-old *A. chinensis* plants were acquired as a gift from a farmer in Qinhuangdao City, Hebei Province, China. The *A. chinensis* plants were identified by Associate Professor Xu Qiao in our group. These *A. chinensis* plants were then cultivated in the field at the Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences & Peking Union Medical College, Haidian, Beijing, China. Maintenance and fertilization were carried out based on the group standards of the Chinese Association of Chinese Medicine: Code of Multiple Cultivation for Good Agricultural Practice of Atractylodis Rhizoma (*Atractylodes chinensi*s) (T/CACM 1374.168-2024). The soil type of the field was sandy loam with good drainage, a low groundwater level, a loose structure, and rich humus. During the growth period, irrigation was conducted based on the soil moisture. Plants were cultivated for an additional three years. The leaves and rhizomes were collected from a five-year-old *A. chinensis* with divided leaves on a sunny morning in August 2024 (Fig. 1). The leaf sample comprised of five pieces of leaves, while the rhizome sample comprised of five segments of rhizome. The samples were wrapped in tin foil and stored in a −80 °C refrigerator.

4.2. RNA extraction and sequencing

RNA was extracted from the leaf samples and rhizome segments of *A. chinensis* using the TIANampGenomic DNA Kit (TIANGEN, Beijing, China). The integrity of the RNA was checked using the Bioanalyzer 2100 system (Agilent Technologies, CA, USA). The quality and quantity of the RNA was analysed using a 1 % agarose gel electrophoresis and the Qubit 4.0 (Thermo Fisher Scientific, MA, USA). A cDNA library was prepared using the Ultima Pro DNA Library Prep Kit (Yeasen, Shanghai, China). After library quality control, different libraries were pooled based on the effective concentration and targeted data amount, and then sequenced on an Illumina NovaSeq X Plus Platform. Each library was sequenced three times to reduce bias.

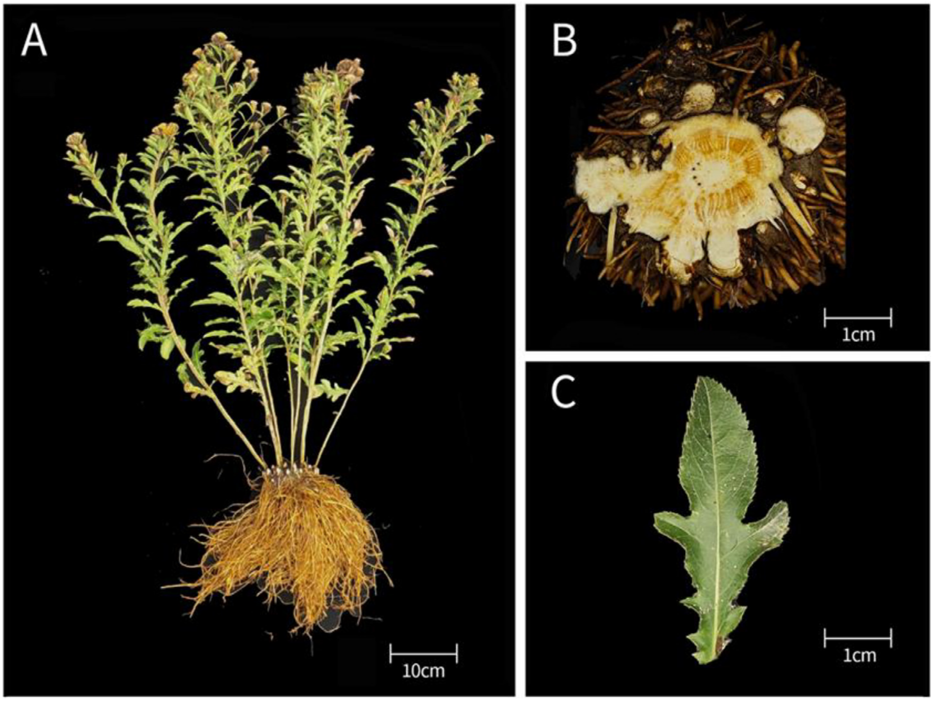

Fig. 1. Images of a five-year-old *Atractylodes chinensis*. **A.** The whole plant. **B.** A cross-section of the rhizome. **C.** The divided leaves.

4.3. Data processing and quality control

Raw data (raw reads) in a fastq format were processed using the fastp (v0.19.4) software to remove low-quality reads and those containing adapter and ploy-N. In addition, Q20, Q30, and GC content were calculated. Clean reads were assembled as transcripts using the Trinity (v2.15.1) software and resulted in a Trinity.fasta file. The transcripts were further assembled into clusters based on shared reads using Corset (v. 1.09). The longest cluster in a cluster group was considered a unigene. The splicing quality of Trinity.fasta, unigene.fa, and cluster.fasta were evaluated using the BUSCO (v5.8.2) software.

4.4. Screening differentially expressed genes

Initially, the original read count was standardised to correct the sequencing depth using the DESeq2 (v1.26.0) software [11]. Next, a statistical model was used to calculate the probability of hypothesis testing (P-value), and then multiple hypothesis testing corrections were performed to obtain the false discovery rate (FDR) value (padj is its common form, which is used below to represent FDR).

4.5. Analysis of differentially expressed genes enriched in KEGG pathways

Genes were enriched in KEGG categories using the KOBAS (v3.0) software installed in a local Linux server [12]. Genes involved in the KEGG pathways related to sesquiterpene and polyacetylene biosynthesis were identified through a search utilizing relevant keywords and existing knowledge.

Limitations

None.

Ethics Statement

The authors have read and followed the ethical requirements for publication in Data in Brief and confirmed that the current work does not involve human subjects, animal experiments, or any data collected from social media platforms.

Credit Author Statement

Yilin Tian: Conceptualization, Supervision, Writing-Original draft preparation. **Lizhi Ouyang**: Data curation, Software, Writing-Reviewing and Editing. **Xinyu Li**: Visualization, Investigation. **Li Xiao: Writing-**Reviewing and Editing**. Xu Qiao**: Investigation, Validation. **Yixuan Chen:** Data curation. **Tingting Fang**: Software, Investigation. **Yimian Ma:** Software, supervision, Data curation, Writing- Reviewing and Editing.

Data Availability

Dataset of transcriptome assembly of Atractylodes chinensis (Original data) (Mendeley Data).

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Grant No. 32160521) and the Science and Technology Project of the Inner Mongolia Autonomous Region (Grant No. 2022YFXZ0026).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] W. Wang, Y.Y. Jiang, B.H. Song, X.M. Tang, Discovery of quality markers in the rhizome of *Atractylodes chinensis* using GC–MS fingerprint and network pharmacology, Arab. J. Chem. 16 (10) (2023) 105114, doi:10.1016/j.arabjc.2023.105114.

[2] J. Zhao, C. Sun, F. Shi, S. Ma, J. Zheng, X. Du, L. Zhang, Comparative transcriptome analysis reveals sesquiterpenoid biosynthesis among 1-, 2- and 3-year old *Atractylodes chinensis*, BMC Plant Biol 21 (1) (2021) 354, doi:10.1186/s12870-021-03131-1.

[3] S. Ma, C. Sun, W. Su, W. Zhao, S. Zhang, S. Su, B. Xie, L. Kong, J. Zheng, Transcriptomic and physiological analysis of atractylodes chinensis in response to drought stress reveals the putative genes related to sesquiterpenoid biosynthesis, BMC Plant Biol 24 (1) (2024) 91, doi:10.1186/s12870-024-04780-8.

[4] X.W. Liu, L.L. Weng, C.P. Xiao, Difference of effective component content and key enzyme gene expression of biosynthesis in *Atractylodes chinensis* with different leaf shapes, Zhongguo Zhong Yao Za Zhi 49 (8) (2024) 2138–2146, doi:10.19540/j.cnki.cjcmm.20240119.103.

[5] S.F. Chen, Y.Q. Zhou, Y.R. Chen, J. Gu, fastp: an ultra-fast all-in-one FASTQ preprocessor, Bioinformatics 34 (17) (2018) i884–i890, doi:10.1093/bioinformatics/bty560.

[6] K. Katz, O. Shutov, R. Lapoint, M. Kimelman, J.R. Brister, C. O'Sullivan, The Sequence Read Archive: a decade more of explosive growth, Nucleic Acids Res 50 (D1) (2022) D387–D390, doi:10.1093/nar/gkab1053.

[7] M.G. Grabherr, B.J. Haas, B.J. Yassour, et al., Full-length transcriptome assembly from RNA-Seq data without a reference genome, Nat. Biotechnol. 29 (7) (2011) 644–652, doi:10.1038/nbt.1883.

[8] N.M. Davidson, A. Oshlack, Corset: enabling differential gene expression analysis for de novo assembled transcriptomes, Genome Biol 15 (7) (2014) 410, doi:10.1186/s13059-014-0410-6.

[9] M. Manni, M.R. Berkeley, M. Seppey, E.M. Zdobnov, BUSCO: assessing genomic data quality and beyond, Curr. Protoc. 1 (12) (2021) e323, doi:10.1002/cpz1.323.

[10] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, Y. Yamanishi, KEGG for linking genomes to life and the environment, Nucleic Acids Res 36 (Database issue) (2008) D480–D484, doi:10.1093/nar/gkm882.

[11] M.I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, Genome Biol 15 (12) (2014) 550, doi:10.1186/s13059-014-0550-8.

[12] D. Bu, H. Luo, P. Huo, Z. Wang, S. Zhang, Z. He, Y. Wu, L. Zhao, J. Liu, J. Guo, S. Fang, W. Cao, L. Yi, Y. Zhao, L. Kong, KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis, Nucleic Acids Res 49 (W1) (2021) W317–W325, doi:10.1093/nar/gkab447.