BMC Genomic Data

DATA NOTE

Open Access

Draft genome of *Angelica Biserrata*, a traditional Chinese medicinal herb of the *Angelica* genus (Apiaceae)

Yuan-jiang Xu[1], Li Li[1], Xue Liu[1], Fang-yu Zhao[2], Yi-Quan Zhou[1*] and Xian-you Qu[1*]

Abstract

Objectives *Angelica biserrata* (commonly known as "Duhuo"), a traditional Chinese medicinal herb of the genus *Angelica* within the Apiaceae family, is clinically valued for its therapeutic effects in dispelling wind-dampness (a TCM syndrome manifesting as inflammatory arthritis, chronic headaches, and migratory pain) and alleviating arthralgia. Its pharmacological properties are primarily attributed to coumarins, to elucidate the molecular mechanisms underlying coumarin biosynthesis and facilitate the breeding of high-coumarin cultivars, we present the first draft genome assembly and annotation of *A. biserrata*. The first genome assembly of *A. biserrata* will provide novel insights into elucidating coumarin biosynthesis and advancing evolutionary biological studies.

Data description The genome of A. biserrata was sequenced using PacBio HiFi technology, generating 8.83 million high-fidelity reads with an average length of 14.2 kb (125.34 Gb, sequencing coverage 41 ×). The reads were assembled to give a draft genome of 4.52 Gb with an N50 contig length of 35.72 Mb. Chromosome-scale scaffolding was then performed using 300.87 Gb Hi-C data, resulting in a final genome assembly of 3.89 Gb with improved continuity (contig N50 = 34.42 Mb, scaffold N50 = 325.77 Mb). The genomic integrity was 96.59% (based on the embryophyta database of OrthoDB 10) through the evaluation of universal single copy direct homologous gene (BUSCO). At the same time, 3811.62 Mb long sequences were attached to 11 chromosomes, accounting for 97.86%.

Keywords Angelica Biserrata, Hi-Fi Genome, Hi-C, Genome annotation

Objective

Angelica biserrata, with the original name *Angelica pubescens* f. *biserrata*, belongs to the Apiaceae family. Its dry root is a traditional Chinese medicine [1]. Flora of China (FOC) is distributed mainly in Chongqing, Anhui, Hubei, Jiangxi, and Sichuan Provinces, etc., at altitudes of 1000–1700 m. FOC is known as "Duhuo" or "Rouduhuo" in China, "Dokwhal" in South Korea, and "Dokkatsu" in Japan; *A. biserrata* has different common names across East Asia [2]. The medicinal properties of *A. biserrata* were first recorded in "Shen Nong Ben Cao Jing" (Eastern Han Dynasty). It is used mainly to treat rheumatism, headache and other diseases [3]. Coumarins are considered the active ingredients

*Correspondence:
Yi-Quan Zhou
zhouyq1982@163.com
Xian-you Qu
quxy023@126.com
[1]Chongqing Academy of Chinese Materia Medical, Nanshan Road, Huangjueya, Nanan District, Chongqing 400065, China
[2]Food Science College, Xizang Agricultural and Animal Husbandry University, Xizang 860000, China

of *A. biserrata* and feature a wide range of pharmacological activities, including anti-inflammatory, antitumor and antimicrobial properties, and take part in plant growth and development, such as seed germination, defense against plant pathogens and stress reactions [4, 5]. However, their complex synthetic pathways are not fully known. The chromosome-level genome assembly of *Angelica sinensis* (2.3 Gb) has been successfully deciphered, providing novel genomic insights into the biosynthetic machinery of simple coumarins [6]. The coumarins in *A. biserrata* are more complex and diverse than those of *A. sinensis*. High-quality reference genomes are important for research on the resource protection of *A. biserrata* and the synthetic pathways of coumarin and other active ingredients. This study completed the first assembly and annotation of the *A. biserrata* genome. This work provides critical data for further analysis of the biosynthetic pathways of the active ingredients of *A. biserrata*, the development of molecular marker-assisted breeding systems, and the formulation of scientific conservation strategies. Moreover, sequencing the genome of *A. biserrata* will be helpful for analyzing the evolutionary history of Umbelliferae medicinal plants.

Data

Several leaves were collected from a plant of *A. biserrata* cultivated in Lanying Township, Wuxi County, Chongqing City, China (N 29.531106°, E 106.598273°), snap-frozen in liquid nitrogen, and stored in a −80 ℃ freezer in the laboratory. Prior to sequencing, high-quality DNA and total RNA were extracted from the samples. First, a DNBSEQ-T7 genomic survey and sequencing were performed on the Illumina sequencing platform, and 161.67 Gb of data (data file 1) were obtained after quality control. Subsequently, approximately 125.34 Gb of data were generated through Hi-Fi sequencing via the PacBio Revio sequencer (data file 2, data file 3, data file 4, data file 5, and data file 6). A Hi-C library was constructed via methods such as cell cross-linking and Hind III restriction enzyme digestion and then sequenced via the DNBSEQ-T7 sequencer in 150-bp paired-end mode. Hi-C sequencing yielded a total of approximately 300.87 Gb of data (data file 7). Finally, RNA-seq sequencing was performed via the Illumina X-plus sequencer, generating approximately 6 Gb of data (data file 8), which was used to assist in annotation.

The findings suggested that the genome size of *A. biserrata* is approximately 3.03 Gb, with a heterozygosity rate of approximately 0.27% and a repetitive sequence content of approximately 83.59%. It is a complex and large genome. HiFi sequencing yielded 8.83 million high-fidelity reads with an average length of 14.2 kb. Hifiasm (V 0.19) [7] was used to splice the genome, and the total length of the contig sequence was 4.55 Gb (N50 = 35.4 Mb). After nucleoside alignment and elimination of the Hi-C signal redundancy, the total length of the genome contig sequence was 3.89 Gb, and the contig N50 was 34.42 Mb. In accordance with OrthoDB 10 (http://cegg.unige.ch/orthodb) and BUSCO (V 5.2.1) [8], the integrity of the *A. biserrata* genome was determined to be 96.59%. HiC-Pro [9] (v2.10.0) was adopted to filter and evaluate the Hi-C data to obtain high-quality clean data. Through BWA [10] (version: 0.7.17-r1188; comparison mode: aln; other parameters are default), the double-end sequence data were aligned with the sequence of the assembled genome. The genome sequence was grouped, sequenced and oriented with LACHESIS [11] software, and then the final version of the genome was obtained after Hi-C

Table 1 Overview of data files

Label	Name of data file	File types (file extension)	Data repository and identifier (DOI or accession number)
Data file 1	Raw short whole genome Illumina sequencing reads	Fasta file (.fastq)	CNCB Big sub-Genome Sequence Archive (GSA) Accession number CRA024320. https://ngdc.cncb.ac.cn/gsa/browse/CRA024320/CRR1712864 [19]
Data file 2	Raw long whole genome Hi-Fi sequencing reads 1	Fasta file (.fastq)	CNCB Big sub-Genome Sequence Archive (GSA) Accession number CRA024335. https://ngdc.cncb.ac.cn/gsa/browse/CRA024335/CRR1713706 [20]
Data file 3	Raw long whole genome Hi-Fi sequencing reads 2	Fasta file (.fastq)	CNCB Big sub-Genome Sequence Archive (GSA) Accession number CRA024335. https://ngdc.cncb.ac.cn/gsa/browse/CRA024335/CRR1713707 [21]
Data file 4	Raw long whole genome Hi-Fi sequencing reads 3	Fasta file (.fastq)	CNCB Big sub-Genome Sequence Archive (GSA) Accession number CRA024335. https://ngdc.cncb.ac.cn/gsa/browse/CRA024335/CRR1713708 [22]
Data file 5	Raw long whole genome Hi-Fi sequencing reads 4	Fasta file (.fastq)	CNCB Big sub-Genome Sequence Archive (GSA) Accession number CRA024335. https://ngdc.cncb.ac.cn/gsa/browse/CRA024335/CRR1713709 [23]
Data file 6	Raw long whole genome Hi-Fi sequencing reads 5	Fasta file (.fastq)	CNCB Big sub-Genome Sequence Archive (GSA) Accession number CRA024335. https://ngdc.cncb.ac.cn/gsa/browse/CRA024335/CRR1713710 [24]
Data file 7	Raw Hi-C reads	Fasta file (.fastq)	CNCB Big sub-Genome Sequence Archive (GSA) Accession number CRA024325. https://ngdc.cncb.ac.cn/gsa/browse/CRA024325/CRR1713007 [25]
Data file 8	Raw RNA-seq reads	Fasta file (.fastq)	CNCB Big sub-Genome Sequence Archive (GSA) Accession number CRA024296. https://ngdc.cncb.ac.cn/gsa/browse/CRA024296/CRR1711832 [26]
Data file 9	Assembled genome	Fasta file (.fastq)	CNCB Big sub-Genome Warehouse (GWH) Accession number GWH-FQYN00000000.1 https://ngdc.cncb.ac.cn/gwh/Assembly/92525/show [27]
Data file 10	Genome annotation	Gff3 file (.gff3)	Figshare, https://doi.org/10.6084/m9.figshare.28740092.v1 [28]

error correction, auxiliary chromosome mounting and redundancy removal. The assembly result was 3894.84 Mb, the contig N50 was 34.42 Mb, and the scaffold N50 was 325.77 Mb (data file 9). Among them, the 3811.62 Mb genome sequence was located on 11 chromosomes, accounting for 97.86% of the genome, indicating high integrity of the assembly. Then, repeatModeler2 [12] (v2.0.1), RepeatMasker [13] (V4.1.2) and other software were used for ab initio prediction of repeat sequences; Augustus [14] (v3.1.0) and SNAP [15] (2006-07-28) were utilized for ab initio prediction of structural genes; GeMoMa [16] (V1.7) was used to predict homologous species; and PASA [17] (v2.4.1) was used for transcriptome-assisted gene prediction. Finally, EVM [18] (v1.1.1) was adopted to combine the prediction results from the above three methods, and 98.66% of the genes (a total of 44603 genes) were annotated (data file 10), indicating high annotation quality.

Limitations

Although we have successfully assembled a chromosome-level genome of *A. biserrata* (Duhuo), the biosynthetic pathways of its active compounds (e.g., coumarins) and the genetic mechanisms governing its medicinal properties require further investigation. Future studies will employ integrated multiomics approaches involving comparative genomics, metabolomics, transcriptomics, and in vitro enzymatic functional validation to systematically elucidate the complex coumarin biosynthesis network and its regulation in this medicinal species.

Abbreviations

BUSCO	Benchmarking Universal Single-Copy Orthologs
CNCB	China National Center for Bioinformation
PacBio	Pacifc Biosciences
Hic	High-throughput chromosome conformation capture
TCM	Traditional Chinese Medicine

Acknowledgements

The authors appreciate the use of biomarkers for providing high-quality sequencing.

Authors' contributions

Y. J X and L. L. are jointly responsible for sample organization and sequencing data analysis. Y. J X ultimately wrote the manuscript, and Y. Q Z and X. Y Q revised the manuscript. L. L. and X. L. assisted in genome assembly and annotation. Y.Q Z and X. Y Q conceived and designed this project. All the authors have read and approved the final version of this manuscript.

Funding

The study was funded by the Analysis of Synthetic Pathway Study of Key Enzyme Genes of Chongqing Post-Bo Foundation (grant number cstc2021jcyj-bshX0015).

Data availability

The raw sequence data reported in this paper have been deposited in the Genome Sequence Archive (GSA) of the National Genomics Data Center. The accession numbers for the Illumina, Hi-Fi, Hi-C, and RNA-seq data are CRA024320, CRA024335, CRA024325 and CRA024296, respectively. The assembled genome reported in this paper has been deposited in the Genome Warehouse (GWH) at the National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences/China National Center for Bioinformatics, under accession number GWHFQYN00000000.1. The results of the genome annotations have been uploaded to Figshare. Please see Table 1 for details and links to the data.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 11 April 2025 / Accepted: 23 September 2025
Published online: 29 October 2025

References

1. Chinese Pharmacopoeia Commission. Pharmacopoeia of the people's Republic of China, vol. Vol. Beijing: China Medical Science; 2020. p. 274.
2. Ma J, Huang J, Hua S, Zhang Y, Zhang Y, Li T, Dong L, Gao Q, Fu X. The ethnopharmacology, phytochemistry and Pharmacology of *Angelica biserrata* - A review. J Ethnopharmacol. 2019;231:152–69. https://doi.org/10.1016/j.jep.2018.10.040.
3. Lu Y, Wu H, Yu X, Zhang X, Hanyan L, Liying T. Traditional Chinese medicine of *Angelicae pubescentis* radix: A review of Phytochemistry, Pharmacology and Pharmacokinetics[J]. Front Pharmacol. 2020;11. https://doi.org/10.3389/fphar.2020.00335.
4. Venugopala KN, Rashmi V, Odhav B. Review on natural coumarin lead compounds for their pharmacological activity. BioMed Res Int. 2013;963248. https://doi.org/10.1155/2013/963248.
5. Zhao Y, He Y, Han L, Zhang L, Xia Y, Yin F, Wang X, Zhao D, Xu S, Qiao F, Xiao Y, Kong L. Two types of coumarins-specific enzymes complete the last missing steps in pyran- and furanocoumarins biosynthesis. Acta Pharm Sin B. 2024;14(2):869–80. https://doi.org/10.1016/j.apsb.2023.10.016.
6. Han X, Li C, Sun S, Ji J, Nie B, Maker G, et al. The chromosome-level genome of female ginseng (*Angelica sinensis*) provides insights into molecular mechanisms and evolution of coumarin biosynthesis. Plant J. 2022;112(5):1224–37. https://doi.org/10.1111/tpj.16007.
7. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021;18(2):170–5. https://doi.org/10.1038/s41592-020-01056-5.
8. Simão Felipe A, Waterhouse RM, Panagiotis I, et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs[J]. Bioinformatics. 2015;19. https://doi.org/10.1093/bioinformatics/btv351.
9. Servant N, et al. Genome Biol. 2015;16(1):1–11. https://doi.org/10.1186/s13059-015-0831-x. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing.
10. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25:1754–60. https://doi.org/10.1093/bioinformatics/btp324.
11. Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat Biotechnol. 2013;31(12):1119–25. https://doi.org/10.1038/nbt.2727.
12. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. Repeatmodeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A. 2020;117(17):9451–7. https://doi.org/10.1073/pnas.1921046117.
13. Tarailo-Graovac M, Chen N. Using repeatmasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinf. 2009;Chap44101–41014. https://doi.org/10.1002/0471250953.bi0410s25.
14. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 2008;24(5):637–44. https://doi.org/10.1093/bioinformatics/btn013.
15. Korf I. Gene finding in novel genomes. BMC Bioinformatics. 2004;5:59. https://doi.org/10.1186/1471-2105-5-59. Published 2004 May 14.

16. Keilwagen J, Wenk M, Erickson JL, Schattat MH, Grau J, Hartung F. Using intron position conservation for homology-based gene prediction. Nucleic Acids Res. 2016;44(9):e89. https://doi.org/10.1093/nar/gkw092.
17. Haas BJ, Delcher AL, Mount SM, et al. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 2003;31(19):5654–66. https://doi.org/10.1093/nar/gkg770.
18. Haas BJ, Salzberg SL, Zhu W, et al. Automated eukaryotic gene structure annotation using evidencemodeler and the program to assemble spliced alignments. Genome Biol. 2008;9(1):R7. https://doi.org/10.1186/gb-2008-9-1-r7. Published 2008 Jan 11.
19. Xu YJ. Illumina sequencing reads of *Angelica Biserrata.* China National center for bioinformation (CNCB)-Genome sequence archive (GSA). 2025. https://ngdc.cncb.ac.cn/gsa/browse/CRA024320/CRR1712864
20. Xu YJ. Hi-Fi sequencing reads 1 of *Angelica Biserrata*. China National Center for Bioinformation (CNCB)-Genome Sequence Archive (GSA). 2025. https://ngdc.cncb.ac.cn/gsa/browse/CRA024335/CRR1713706
21. Xu YJ. Hi-Fi sequencing reads 2 of *Angelica Biserrata*. China National Center for Bioinformation (CNCB)-Genome Sequence Archive (GSA). 2025. https://ngdc.cncb.ac.cn/gsa/browse/CRA024335/CRR1713707
22. Xu YJ. Hi-Fi sequencing reads 3 of *Angelica Biserrata*. China National Center for Bioinformation (CNCB)-Genome Sequence Archive (GSA). 2025. https://ngdc.cncb.ac.cn/gsa/browse/CRA024335/CRR1713708
23. Xu YJ. Hi-Fi sequencing reads 4 of *Angelica Biserrata*. China National Center for Bioinformation (CNCB)-Genome Sequence Archive (GSA). 2025. https://ngdc.cncb.ac.cn/gsa/browse/CRA024335/CRR1713709
24. Xu YJ. Hi-Fi sequencing reads of 5 *Angelica Biserrata*. China National Center for Bioinformation (CNCB)-Genome Sequence Archive (GSA). 2025. https://ngdc.cncb.ac.cn/gsa/browse/CRA024335/CRR1713710
25. Xu YJ. Hi-C reads of *Angelica Biserrata*. China National Center for Bioinformation (CNCB)-Genome Sequence Archive (GSA). 2025. https://ngdc.cncb.ac.cn/gsa/browse/CRA024325/CRR1713007
26. Xu YJ. RNA-seq of *Angelica Biserrata*. China National Center for Bioinformation (CNCB)-Genome Sequence Archive (GSA). 2025. https://ngdc.cncb.ac.cn/gsa/browse/CRA024296/CRR1711832
27. Xu YJ. Draft Genome of *Angelica Biserrata*. China National Center for Bioinformation (CNCB)- Genome Warehouse (GWH). 2024. Accession number GWHFQYN00000000.1 https://ngdc.cncb.ac.cn/gwh/Assembly/92525/show
28. Xu YJ. Genome annotation of *Angelica biserrata*. Figshare. 2025. Figshare, https://doi.org/10.6084/m9.figshare.28740092.v1

Publisher's Note